ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230457

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

Abstract. We discuss Empirical Risk Minimization approach in conjunction with one-class classification method to learn classifiers for biased Positive Unlabeled (PU) data. For such data, probability that an observation from a positive class is labeled may depend on its features. The proposed method extends Variational Autoencoder for PU data (VAE-PU) introduced in [16] by proposing another estimator of a theoretical risk of a classifier to be minimized, which has important advantages over the previous proposal. This is based on one-class classification approach using generated pseudo-observations, which turns out to be an effective method of detecting positive observations among unlabeled ones. The proposed method leads to more precise estimation of the theoretical risk than the previous proposal. Experiments performed on real data sets show that the proposed VAE-PU+OCC algorithm works very promisingly in comparison to its competitors such as the original VAE-PU, SAR-EM and LBE methods in terms of accuracy and F1 score. The advantage is especially strongly pronounced for small labeling frequencies.

1 Introduction

In the paper we consider a binary classification task, for which a usual observability scenario fails, in the sense that observations from both classes, between which we would like to distinguish, are *not* available. Instead, in a Positive Unlabeled (PU) case considered here, we have at our disposal a labeled sample from a positive class (labeled examples) and a group of unlabeled data which consists of negative and positive observations. Traditional PU methods require labeled sample to be an unbiased sample from a positive class. However, in practice it is common that probability of being labeled depends on a characteristics of an item in question and, consequently, the labeled data is a *biased* sample from the positive class. This is frequently called selection bias, in contrast to a special case of Selected Completely at Random (SCAR) scenario when probability of labeling is constant and, consequently, the selected sample is a sample of a random size from the positive class. Moreover, in a selection bias case considered here, unlabeled data is also a biased sample from a general population. We note that selection bias is an important future of many data gathering techniques (see e.g. [9]).

PU learning is used in situations when it is difficult or costly to obtain reliable negative examples, including text and image annotation when annotators label only objects of specific types, and omit objects of other kinds. Also, for medical evaluations, lack of diagnosis does not mean that a patient does not have a disease in question. The last example indicates that the selection bias occurs naturally: the people who are diagnosed with a certain disease may be better educated and thus aware that should undergo diagnosis at a certain age, for example. Other areas where such type of problem occur include e.g. biology of ecosystems [28], survey analysis [2] and recommendation systems [20]. We also note that the type of partial observability analyzed here is frequently considered as missing labels situation (see e.g. [22]).

Although majority of research focuses on inferential approaches for PU data when SCAR assumption is valid (see e.g. [2] for a review) some methods have been developed already which account for the more realistic scenario of biased selection of labeled items [7, 3, 6]. This is usually done attempting to learn a propensity score, defined as a probability of an item from a positive class being labeled given its feature vector x. We stress that such an approach is complicated one as there is an intrinsic unidentifiability issue here which makes it hard to disentangle the labeling mechanism from the mechanism of assignment to the positive class based on available data. This usually requires imposing assumptions on the interplay between the propensity score and the posterior probability of belonging to the positive class, such as their co-monotonicity [12]. Approaches which avoid assuming SCAR will be referred to as no-SCAR methods in the following. In this paper we follow alternative route proposed in [16] where labeling process is not modeled explicitly but is treated as a latent process which is recovered by a variational inference approach (see [4] for an overview of variational inference).

We adopt the following framework for PU problem which allows for the selection bias. We consider an iid sample $(X_i, Y_i, S_i), i = 1, \ldots, n$, where $X_i \in R^p$ is a feature vector for i^{th} item, $Y_i \in \{-1, 1\}$ is its class indicator and $S_i \in \{0, 1\}$ is its label ($S_i = 1$ means that the observation is labeled). It is assumed that the item can be labeled ($S_i = 1$) only when it belongs to the positive class P i.e. $Y_i = 1$. Negative class corresponding to observations with $Y_i = -1$ is denoted by N. Probability of labeling in a positive class given a feature vector $X = x$ is called propensity score $e(x) = P(S = 1 \mid Y = 1, X = x)$ and selection bias means that the propensity score depends on x. Solely the sample $(X_i, S_i), i = 1, \ldots, n$ is observed. From this data we would like to learn classification rule $d(X)$ recovering unobservable class Y indicator of a new item X. We stress that

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

we adopt the scenario defined above, called a single sample scenario and not a case-control scenario, which is also frequently considered. Examples of approaches devoted to biased PU case-control data include [12, 10].

2 VAE-PU method: description and discussion

In the following we will consider a classification function corresponding to the classification rule $d(\cdot)$, that is a function $g : R^p \to R$ such that $g(x) > 0 \equiv d(x) = 1$. We adopt an Empirical Risk Minimization (ERM) approach originating from statistical decision theory which considers loss function $\ell : R \to [0, \infty)$, usually nonincreasing and convex, such that $\ell(Yg(X))$ corresponds to a loss incurred when value of the decision function equals $g(X)$ and the class assignment is Y (see e.g. [8] for a general introduction and [5] for a recent ERM approach to biased PU data). The objective is to minimize empirical version of the expected loss $R(g) = E\,\ell(Yg(X))$ over some family of functions which is given e.g. as an output of neural network with a fixed architecture. In the paper the following equality plays a prominent role (see [16], Theorem 3.1):

Lemma 1 *We have*

$$\begin{aligned} R(g) = E_{X,Y}\,\ell(Yg(X)) &= P(S=1)E_{X|S=1}\,\ell(g(X)) \\ &+ P(Y=1,S=0)E_{X|Y=1,S=0}\,\ell(g(X)) - \ell(-g(X)) \\ &+ P(S=0)E_{X|S=0}\,\ell(-g(X)). \end{aligned} \tag{1}$$

Note that the above equality can be justified by the following reasoning: the first term in (1) is a correct part of risk corresponding to an assignment of all labeled elements to the positive class, the third term corresponds to assigning all unlabeled items to the negative class, and the second term is the correction of the latter which accounts for the error committed in the case of positive unlabeled (PU) data which belong to the positive class. Note that $R(g)$ is not directly estimable as the second term involves calculation of the expected value with respect to the distribution of the positive unlabeled cases, which is not observed. Its estimation will involve reconstruction of such elements. Namely, the empirical counterpart $R_{emp}(g)$ of $R(g)$ is

$$\begin{aligned} R_{emp}(g) = \frac{\pi_{PL}}{|\chi_{PL}|}\sum_{x^{(pl)}\in\chi_{PL}}\ell\left(g(x^{(pl)})\right) + \frac{\pi_{PU}}{|\tilde{\chi}_{PU}|}\sum_{\tilde{x}^{(pu)}\in\tilde{\chi}_{PU}}\ell\left(g(\tilde{x}^{(pu)})\right) \\ + \max\left\{0, -\frac{\pi_{PU}}{|\tilde{\chi}_{PU}|}\sum_{\tilde{x}^{(pu)}\in\tilde{\chi}_{PU}}\ell\left(-g(\tilde{x}^{(pu)})\right) \right. \\ \left. + \frac{\pi_U}{|\chi_U|}\sum_{x^{(u)}\in\chi_U}\ell\left(-g(x^{(u)})\right)\right\}, \end{aligned} \tag{2}$$

where χ_{PU} denotes positive unlabeled sample, χ_U and χ_{PL} are analogously defined, and $|A|$ stands for the sample size of set A. Moreover, $\pi_{PL} = P(S=1), \pi_{PU} = P(Y=1,S=0), \pi_U = P(S=0)$. Importantly, $\tilde{\chi}_{PU}$ in (2) denotes some estimate of positive unlabeled sample which we are about to construct. Consider now the reason why max term has been introduced in (2). This is due to the fact the term under max operator is an estimator of a *nonnegative* summand of the risk equal to

$$E\ell(g(-X))I\{S=0\} - E\ell(g(-X))I\{S=0,Y=1\} \tag{3}$$

and it is desirable that its estimator is also non-negative. However, simple truncation at 0 may lead to a substantial loss of information. In the paper we account for this drawback and avoid truncation using one-class classification. It is assumed in the following, similarly to [16], that $\pi = P(Y=1)$ is known. This is often realistic assumption as, for example, π may be estimated with arbitrary accuracy from existing independent data base for general population. Note that

$$\pi_{PU} = P(Y=1,S=0) = P(Y=1) - P(S=1)$$

and may be estimated by plugging $n_{S=1}/n$ for $P(S=1)$, where $n_{S=i}$ for $i = 0, 1$ denote sample sizes of labeled and unlabeled group, respectively.

The second important building block of VAE-PU proposed in [16] is construction of Variational Auto-Encoder (VAE), which assumes existence of two latent multivariate random variables h_y corresponding to Y and h_s corresponding to S (denoted by h_o in [16]). Latent h_y is generated from mixture of two normal distributions corresponding to negative and positive examples and h_s follows the standard normal multivariate distribution. Then it is assumed that a feature vector is generated from the distribution parameterized by a function of (h_y, h_s) and observed labeling is generated from Bernoulli distribution with probability being a function of h_s. Parameters of the corresponding distribution are determined by variational inference approach [4] consisting in minimization of the Evidence Lower Bound (ELBO) (see Section 3.3 in [16]). As the result, the values of latent variables for observed data are obtained. By matching similar labeled and unlabeled instances and decoding their combined representations one can obtain a set of pseudo-observations from positive unlabeled (PU) population. Methods for matching the examples are further discussed in web appendix A[1]. Constructed pseudo-sample of PU observations, denoted by $\tilde{\chi}_{PU}$ is used instead of χ_{PU} in the formula for the empirical risk $R_{emp}(g)$ in (2). Observe that by doing this we replace part of original χ_U by constructed pseudo-observations. This sample is used to train target classifier (predicting class variable Y) via risk defined by equation (2).

VAE-PU training consists of minimization of $R_{emp}(g)$ concurrently with optimization of penalized ELBO criterion (incorporating two supplementary losses, adversarial generation loss and label loss, for details see (10) in [16]). Optimizing ELBO corresponds to fitting variational autoencoder, composed of two *encoders* (corresponding to latent representations h_y and h_s), *decoder* (attempting to reconstruct input observation x_i) and *observation classifier* (attempting to reconstruct observation status, i.e. label s_i for the observation). As a result of the procedure described above, pseudo-sample pertaining to PU is created, which is adjusted by the first term in penalization (adversarial loss which is used to train *discriminator*) to yield sample similar to U sample, and by the second term (label loss) to resemble PU observations. Then $R_{emp}(g)$ minimization aims to improve *target classifier* (performing final positive/negative classification). The training procedure alternates between autoencoder module (ELBO) training and target classifier updates. For detailed description of the baseline algorithm and technical model details, please refer to [16].

3 One-class classification enhancement of VAE-PU

3.1 Motivation

We explain now motivation for VAE-PU to be modified using one-class classification approach. Consider the reason why max term has been introduced in (2). This is mainly due to the fact that the observations in χ_{PU} are replaced by pseudo-observations in $\tilde{\chi}_{PU}$, *which*

[1] This and other appendices are available online in the GitHub repository: https://github.com/wawrzenczyka/VAE-PU-OCC-web-appendix

do not form a subset of χ_U and thus the estimator of (3) is not necessarily positive. For case-control PU truncation at 0 results in significant improvement for the corrected estimator nnPU over its uncorrected version uPU ([25], Chapter 11). Truncation leads to a substantial loss of information, however. Truncation is meant to decrease bias of $R_{emp}(g)$ and, of course, it does the trick when we know that the theoretical counterpart of the term we replace by 0 is necessarily non-negative. However, once the term is truncated by 0, we can not modify it to make it asymptotically unbiased for the respective theoretical term. We argue that the truncation is not necessary if the set of pseudo-observations $\tilde{\chi}_{PU}$ is replaced by the subset of observations from χ_U which is similar to χ_{PU} . Indeed, suppose for a moment that $\tilde{\chi}_{PU}$ in (2) is replaced back by the true χ_{PU}. The corresponding part of the empirical risk is

$$-\frac{\pi_{PU}}{|\chi_{PU}|}\sum_{x^{(pu)}\in\chi_{PU}}\ell\left(-g(x^{(pu)})\right)+\frac{\pi_U}{|\chi_U|}\sum_{x^{(u)}\in\chi_U}\ell\left(-g(x^{(u)})\right).$$

The expression above is bound to be positive as in view of Law of Large Numbers $\pi_{PU}/|\chi_{PU}| \approx \pi_U/|\chi_U| \approx n^{-1}$ and the second sum in the above expression is larger then the first sum *for original data*. Thus, were $\tilde{\chi}_{PU}$ a subset of χ_U, satisfying the approximate weight equality, introduction of max correction would not be necessary. Thus, our aim is to determine a subset of χ_U which would correspond to positive unlabeled observations. We show that instead of using the generated PU items directly, it is substantially more beneficial to take advantage of the generated dataset to find the observations which are likely to be true-PU items *in unlabeled dataset*. The remaining part of this paper will be dedicated to the discussion of the ways to achieve this goal by one-class classification and the results of such an approach.

In order to apply one-class classification we will treat $\tilde{\chi}_{PU}$ as the sample from nominal population described by P_{PU} distribution and χ_U as the sample corresponding to the mixture of P_{PU} and P_N.

3.2 One-Class Classification OCC

One-class classifiers (OCC) are a family of methods which, given a training dataset drawn from some nominal distribution P_X, test which of the new items are outliers or anomalies, in the sense that they are drawn from a different distribution than the nominal one; for a recent review see [18]. This task is also frequently known as anomaly (or novelty, outlier, out of distribution) detection, or learning from the positive class only [15, 17]. There are many practical situations where such scenario occurs e.g. medical analysis, fraud detection and forensic analysis. Note that there is a substantial difference between one-class classification and biased PU problem. In contrast to one-class classification when unbiased sample from positive class is available in the latter case, as has been said, we observe only biased observations from the positive class and the general population. Nevertheless, we are able to reduce biased PU problem to one-class classification problem, by treating $\tilde{\chi}_{PU}$ as the sample from the nominal population pertaining to P_{PU} and χ_U as the sample generated from the mixture of P_{PU} and P_U. Note that apart from the fact that $\tilde{\chi}_{PU}$ are generated from the distribution which is only close to P_{PU}, the second difference between PU and one class problems consists in that our primary objective is to detect nominal data, not anomalies in χ_U.

Usually, the one-class classification methods output score value for each new sample. We mention in this context GAN-based methods which use the reconstruction loss as an anomaly score, compare e.g. ADGAN algorithm [19]. This is in contrast to classification, where often we can interpret the results in terms of posterior probability or class assignment. This score based approach makes evaluation of new data difficult to handle – for instance, defining a decision function (in practice usually via a threshold) is difficult; for some methods (e.g. One-Class SVM) such a boundary might be defined naturally, but many others fail to give any statistical guarantees on their outputs. Several approaches exist to tackle this issue [27, 26]; here we use p-values for scores based on validation sample which ensure that p-value for observation stemming from nominal population will be super-uniform and thus probability of false signal (i.e. erroneous detection of an outlier) can be easily controlled [1].

We will now discuss some specific one class methods we used in our VAE-PU+OCC algorithm. We stress that our aim is not to construct a new OCC method but to verify how the representative examples of the existing ones perform in our task. The classical one-class classification methods include One-Class SVM (OC-SVM) and Isolation Forest. In One-Class SVM [21] approach the coordinate center is treated as the only anomalous observation and hyperplane is sought with maximum margin separation from it for data from the nominal class. Isolation Forest [14] is based on an idea that anomalies can be detected in random forests (thus one uses random subsamples of the data and random sets of features) by finding the leaves corresponding to the shortest paths in trees constituting the forest. Recently, ECOD (Empirical Cumulative distribution based Outlier Detection) and A^3 method has been proposed. ECOD [13] is based on a premise that anomalies of multivariate distributions usually exhibit atypical behavior for marginal distributions of this distribution corresponding to one or several dimensions and thus the constructed anomaly measure has similar motivation to Fisher test statistic (see e.g. [1]). Activation Anomaly Analysis $\left(A^3\right)$ method [23] is a significantly more complex approach based on neural networks. It employs three components: *target network*, which performs a task unrelated to anomaly detection (for example, autoencoder); *anomaly network*, generating anomalous examples based on input (in the simplest form, it might be a random sample generator, similarly to GAN); and *alarm network*, which discerns normal and anomalous samples based on hidden activations of the target network. This method is designed to work in unsupervised setting, but including some anomalous samples was shown to improve results.

3.3 VAE-PU+OCC – method introduction

The main idea of VAE-PU+OCC method is a straightforward one. The approach treats pseudo-sample $\tilde{\chi}_{PU}$ as generated from the nominal class, and uses the fact that χ_U is the sample consisting of a mixture of positive unlabeled elements (which are similar to elements in $\tilde{\chi}_{PU}$) and negative elements which are considered as anomalies. One class classifiers are now used to screen anomalies (elements with $Y = 0$) from regular elements (elements such that $Y = 1$ and $S = 0$).

The main idea of VAE-PU+OCC is combining the VAE-PU powerful generative capabilities and one-class classification. That means:

- The task is to find Positive Unlabeled (abbreviated as PU) observations χ_{PU} in the unlabeled dataset,
- Due to the biased labeling, Positive Labeled (PL) sample χ_{PL} may not be used for this task as a benchmark representative, as it has a different distribution from that of PU observations χ_{PU} contained in the Unlabeled dataset χ_U,
- Instead of using unavailable Positive (P) sample χ_P, we will train

the classifier on PU pseudo-observations generated by VAE-PU itself,
- As we know PU sample only, and we do not have any information on distribution of Negative (N) observations χ_N, we apply one-class classifier instead of the traditional binary classification method. This classifier is be used to filter the true-PU items out of the unlabeled dataset (which is a mixture of PU and N instances).

VAE-PU+OCC method applies a learned VAE-PU model. That means training all elements of the model – both the generative part (encoder, decoder, observation classifier, discriminator) and the target classifier itself. The trained target classifier is needed to include label loss in process of VAE-PU training. Using target classifier pretrained in that way instead of reinitializing the model is also beneficial for the quality of final model.

The final part of the training starts with generation of the PU pseudo-observations. Then, an OCC classifier of choice is trained on the generated data – as noted before, the method does not require any specific model. In this step, generated data is split into two (in our case, equal) parts: training and calibration; training set is used to train OCC model, while calibration dataset is reserved for subsequent p-value calculation. Then, based on the trained OCC model, all of the observations in the unlabeled set are evaluated. Resulting scores are then converted to marginal p-values by calculating the fraction of scores from the (pseudo)-nominal observations exceeding the score of an item examined. Based on the class prior provided to VAE-PU, we can then evaluate the proportion of the PU items in the U dataset. Instead of using a predefined cutoff, we can then take observations from unlabeled dataset corresponding to the largest p-values and use that as an input to the VAE-PU risk function. This procedure is summed up in Algorithm 1. We stress again that using true-PU samples also solves the issue of loss reduction exceeding the original loss, which means that the $\max$ term in risk function $R_{emp}(g)$ is no longer necessary.

Algorithm 1 VAE-PU+OCC training

Require: π – class prior estimate, n – number of training items
1: Train VAE-PU model (encoder, decoder, target classifier, observation classifier and discriminator); this process is described in detail by Na et al. [16].
2: **while** not converged **do**
3: Generate pseudo-sample $\tilde{x}_{PU}$ using trained VAE,
4: Train OCC classifier of choice on $\tilde{x}_{PU}$,
5: Use OCC classifier to calculate *marginal p-values* for U dataset,
6: Calculate estimated proportion p of PU samples in U corresponding to $P(Y = 1|S = 0)$:

$$p = \frac{\pi - \frac{n_{S=1}}{n}}{\frac{n_{S=0}}{n}} = \frac{n\pi - n_{S=1}}{n_{S=0}}$$

7: Choose proportion p of all samples in χ_{PU} with the highest p-values in U as the candidate PU sample,
8: Update VAE-PU target classifier with the risk function $R_{emp}(g)$ and candidate PU sample.
9: **end while**

Algorithm 1 contains only a general outline of the training procedure. Two of the more specific improvements we used in our implementation are as follows:

1. In order to improve the diversity of the training set for the OCC classifier, the generation process is repeated several times. Due to inherent randomness of the decoder, generated observations will be slightly different in each generated batch. In our case, the process was repeated until generated sample size was of original dataset size,
2. We implemented early stopping to avoid overfitting the OCC procedure. Using OCC to train VAE-PU target classifier usually causes its precision to increase, but the recall often decreases slightly as a tradeoff. In order to balance both the precision and the recall values, early stopping metric was the F1-score on validation dataset. Procedure iteration limit was set to 100 iterations, and early stopping usually decreased this to 10 to 20 epochs. Some cases required only a few (3-5) iterations, but the core algorithm 1 hardly ever repeats over 50 times.

It is also important to emphasize that there are several limitations on performance of the OCC variant:

- Dependence on generated sample quality (or, in general, baseline VAE-PU performance) – when more adequate observations are generated, OCC is trained on more representative dataset and its predictions become more accurate,
- Dependence on label frequency which is due to the VAE-PU nature of generation process (i.e. matching each positive sample to its closest neighbor in the latent space). When there are few labeled samples, regardless of the process repeats, the amount of information in the generated PU pseudosample is also limited – all of the generated examples are based on this small set of PL items.

Additional changes. Besides OCC block, we have also introduced some minor changes in the original VAE-PU implementation. We have replaced the reverse sigmoid loss by logistic loss. Although sigmoid loss is a monotone function, it is not convex and the corresponding loss does not have stationary point apart from trivial cases, whereas for the logistic loss minimizer of the risk is a monotone function of the posterior probability (see Appendix G). Moreover, we have replaced h_s-matching by h_y-matching (see Appendix A).

4 Tests and results

4.1 Experimental setup

We assessed VAE-PU+OCC performance on several datasets to prove effectiveness of the OCC-based sample selection for construction of classifiers. Four benchmark datasets resulting in 6 different tasks were used:

- **MNIST**[2] – two different tasks, 3 versus 5 (images of digit *3* are positive, *5* – negative, abbreviated to 3v5) and OvE (images of *odd* digits are positive, *even* – negative),
- **CIFAR-10**[3] – two different tasks, Car versus Truck (*automobile* images are positive, *truck* – negative) and Machine versus Animal (*airplane*, *automobile*, *ship* and *truck* images are positive, *bird*, *cat*, *deer*, *dog*, *frog* and *horse* – negative),
- **STL-10**[4] – identical classes (but more complex images) as in CIFAR-10, Machine versus Animal split is only considered,
- **Gas Concentrations**[5] – *Ethanol* samples are positive, *Ammonia* – negative.

[2] http://yann.lecun.com/exdb/mnist/
[3] https://www.cs.toronto.edu/~kriz/cifar.html
[4] https://cs.stanford.edu/~acoates/stl10/
[5] https://archive.ics.uci.edu/ml/datasets/gas+sensor+array+drift+dataset

Detailed description of the data sets is given in the Appendix B.

For no-SCAR PU modeling, several labeling schemes were applied in order to construct artificially labeled datasets from those above. In contrast to analysis of VAE-PU in [16], our experiments consider multiple label frequency values (defined as overall probability of positive samples being labeled $c = P(S = 1 \mid Y = 1)$) and thus instead of constant probability of labeling, a feature-dependent labeling process is used. For MNIST dataset, samples were labeled according to digit „boldness" – values of each pixel (normalized to (0-1) range) were averaged, and the samples with highest average value were labeled. Number of examples to be labeled is calculated consistently with label frequency, and taking samples with maximum boldness value ensures that the task is substantially harder than the SCAR scenario – labeled dataset is biased sample from the positive ($Y = 1$) distribution. CIFAR-10 and STL-10 datasets used „redness" measure defined as $r(x) = (R(x) - G(x)) + (R(x) - B(x))$, where $R(\cdot), G(\cdot), B(\cdot)$ correspond to R, G and B channel pixel values of input image x. Similarly, images with the highest values of $r(x)$ were labeled. Gas Concentrations dataset used Strategy 1 described by [7] – samples were labeled according to their distance from classification boundary, obtained after fitting logistic regression model to the data. Observations located the furthest away from the boundary had the largest probability to be labeled.

Main objective of this paper is measuring the improvement provided by OCC-based sample selection on VAE-PU learning. To this end, performance of VAE-PU+OCC was compared to baseline VAE-PU. The VAE-PU model was reimplemented in order to incorporate major performance improvements, which allowed for study of increased array of experiments. Two implementations were prepared (as included in the result tables), one incorporating modifications described at the end of section 3.3, and the *orig* version, which preserves model settings from the original paper. As VAE-PU was shown to outperform multiple models (uPU, nnPU, PUbN\N, GenPU, PAN, PUSB; [16]) – those comparisons will not be repeated here, but instead two additional methods will be considered: SAR-EM [3] and LBE [7]. These two methods reflect a current state-of-the-art in biased PU classification. We also note that our proposed method and SAR-EM method are similar in that both are Empirical Risk Minimization methods and thus it is especially worthwhile to compare their performance. In case of SAR-EM, we use the implementation provided by the authors[6]. In order for the algorithm to work, it needs a list of data attributes which are potential propensity features – in our case, all attributes will be considered as such. We also prepared a custom implementation of LBE-LF architecture.

VAE-PU+OCC model allows for an arbitrary choice of embedded one-class classifier. We tested four OCC models: One-Class SVM [21], Isolation Forest [14], A^3 [23] and ECOD [13]. It is important to note that in the experiments a slightly modified version of ECOD was used. In the official implementation[7] training and test dataset are concatenated, and then used to calculate ECDF during prediction. The modified version uses only training data to calculate ECDF for future predictions. Original VAE-PU paper [16] considered only very low label frequencies (e.g. $c = 0.02$ for MNIST datasets). Although it is important to consider scenarios where training data information is severely limited, such a task is very difficult, especially considering its no-SCAR nature, and it is also rare that the training datasets exhibit that large sample imbalance. That lead us to expanding the test cases to the multitude of label frequency values, including the larger ones; for each dataset, five different label frequency values are considered: $c \in \{0.02, 0.1, 0.3, 0.5, 0.7\}$. For each label frequency, dataset and method the training and evaluation procedure is repeated 10 times, each time with different random seed equal to experiment number. Each such experiment is performed with a different training-validation-testing split (70-15-15 ratio). We evaluated classification performance in terms of widely used metrics: $\mathit{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$, $\mathit{Precision} = \frac{TP}{TP+FP}$, $\mathit{Recall} = \frac{TP}{TP+FN}$, and $\mathit{F1\ score} = \frac{2*\mathit{Precision}*\mathit{Recall}}{\mathit{Precision}+\mathit{Recall}} = \frac{2*TP}{2*TP+FP+FN}$. Code for modified VAE-PU (called baseline in the following) as well as its *orig* version, VAE-PU+OCC and performed experiments is publicly available at Github[8]. The repository also contains detailed instruction for computational experiment reproduction, including all software packages and their versions.

4.2 Motivational example

Before examining efficiency of the combined VAE-PU+OCC algorithm, we examine its one aspect only. Namely, we will illustrate one-class classifier potential in discerning positive examples in unlabeled data. Consider initial steps of Algorithm 1 as follows: using generated PU sample $\tilde{\chi}_{PU}$ obtained via trained VAE-PU model, we train OCC classifier; then, such a classifier is evaluated on each observation in unlabeled dataset. Figure 1 depicts distribution of p-values obtained by A^3 classifier in this scenario on two datasets. MNIST 3v5 (Fig. 1a) is a straightforward example, where both distributions behave as expected; negative examples tend to have very low p-values, whereas PU p-value distribution is approximately uniform. CIFAR MachineAnimal (Fig. 1b) is an example of dataset which proved hard for most of the tested OCC methods; we can see that even though most of the true negative items are concentrated around 0, there are multiple cases when their p-value is really high, whereas positive examples are even more skewed, with almost all of their p-values being close to 1. Even though the latter dataset shows that the separation of PU and N parts of unlabeled dataset can be often imperfect, *overall* ability of OCC methods to find positive examples in U set is still remarkable. This serves as a basis of the following experiments, which focus on overall classification performance of VAE-PU+OCC.

4.3 No-SCAR results

Tables 1 and 2 summarize experiments described in section 4.1. For each experimental setting, given as a combination of dataset, label frequency and a particular method, we report the mean accuracy and F1 score, as well as the standard error of the respective metric. In each table results of the benchmark methods (two VAE-PU versions, denoted as „Baseline (original)" and „Baseline (modified)", SAR-EM, and LBE) are separated from the proposed OCC variants based either on A^3 method, ECOD, Isolation Forest or One-Class SVM.

It is apparent that SAR-EM and LBE methods is significantly outperformed, both by base VAE-PU and its modifications. This is especially pronounced in low label frequency setting, but remains true even when label frequency increases. This strongly suggests that it is beneficial to construct observation which mimic those from PU class, especially when size of P sample is small. Notable exceptions are high label frequency experiments (for $c = 0.7$) on CIFAR MachineAnimal, where SAR-EM outperformed (but barely) VAE-PU+OCC variants in terms of accuracy, and STL, where even though it achieved the highest accuracy for $c = 0.7$, its F1 score is significantly smaller than several OCC-based models. LBE method, on the

[6] https://github.com/ML-KULeuven/SAR-PU

[7] Implemented in PyOD: https://github.com/yzhao062/pyod/blob/master/pyod/models/ecod.py

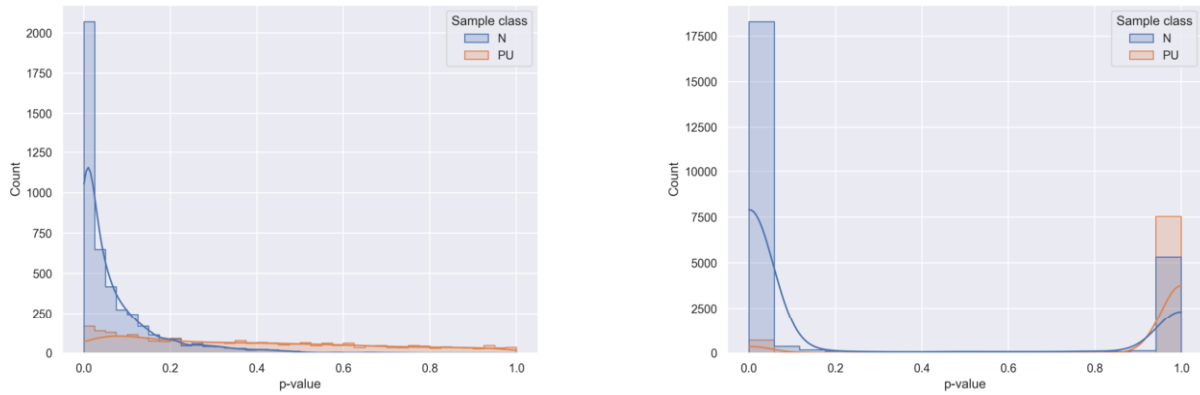
[8] https://github.com/wawrzenczyka/VAE-PU-OCC

Table 1: Accuracy values per dataset. Green ticks („✓") correspond to the cases when t-test rejected equality of accuracies of the VAE-PU-OCC method considered and that of the baseline original method in favor of the former one at $\alpha = 0.05$. Dashes („−") indicate the failure to reject (see appendix F). Bold entries correspond to maximum mean accuracy for a given dataset and label frequency combination.

c	Method	MNIST 3v5	MNIST OvE	CIFAR CarTruck	CIFAR MachineAnimal	STL MachineAnimal	Gas Concentrations
0.02	Baseline	79.99 ± 1.04	71.55 ± 1.03	78.63 ± 2.91	87.71 ± 1.03	75.82 ± 0.52	78.89 ± 2.78
	Baseline (orig)	78.18 ± 0.97	60.62 ± 0.60	79.36 ± 2.51	87.45 ± 1.59	73.62 ± 0.37	71.38 ± 3.16
	LBE	47.78 ± 0.29	49.65 ± 0.15	50.06 ± 0.39	60.20 ± 0.22	60.08 ± 0.27	39.13 ± 0.58
	SAR-EM	47.79 ± 0.31	49.54 ± 0.14	50.27 ± 0.43	60.80 ± 0.21	60.30 ± 0.32	44.90 ± 0.73
	A^3	80.21 ± 1.07 −	74.89 ± 1.62 ✓	82.33 ± 1.49 −	**90.73 ± 0.27** ✓	**79.34 ± 0.69** ✓	**89.84 ± 1.57** ✓
	IsolationForest	80.63 ± 1.16 −	75.63 ± 1.54 ✓	**87.39 ± 2.17** ✓	90.08 ± 0.47 −	75.64 ± 0.45 ✓	80.17 ± 3.24 ✓
	ECODv2	80.44 ± 1.08 −	73.83 ± 1.44 ✓	80.05 ± 1.68 −	90.39 ± 0.27 ✓	75.17 ± 0.60 ✓	80.16 ± 2.98 ✓
	OC-SVM	**80.73 ± 1.23** −	**75.70 ± 1.58** ✓	80.33 ± 1.71 −	90.39 ± 0.27 ✓	75.19 ± 0.60 ✓	81.14 ± 3.14 ✓
0.10	Baseline	85.11 ± 0.87	74.24 ± 1.59	87.70 ± 1.07	82.38 ± 2.32	81.86 ± 0.82	61.63 ± 0.75
	Baseline (orig)	81.44 ± 0.57	65.01 ± 0.84	85.67 ± 0.96	81.70 ± 2.68	81.82 ± 0.86	61.70 ± 0.76
	LBE	51.54 ± 0.35	51.48 ± 0.17	50.93 ± 0.38	62.47 ± 2.10	60.73 ± 0.29	39.30 ± 0.58
	SAR-EM	51.25 ± 0.31	49.36 ± 0.13	52.58 ± 0.40	65.27 ± 0.40	61.82 ± 0.27	47.64 ± 1.12
	A^3	90.01 ± 0.47 ✓	83.14 ± 1.41 ✓	89.37 ± 0.53 ✓	**92.35 ± 0.28** ✓	**83.31 ± 0.33** −	**88.42 ± 1.78** ✓
	IsolationForest	90.52 ± 0.41 ✓	**83.60 ± 1.28** ✓	**90.80 ± 0.38** ✓	90.21 ± 0.57 ✓	83.22 ± 0.28 −	62.84 ± 0.95 −
	ECODv2	**90.82 ± 0.37** ✓	81.97 ± 1.30 ✓	89.90 ± 0.31 ✓	92.09 ± 0.30 ✓	83.10 ± 0.32 −	66.45 ± 1.89 ✓
	OC-SVM	90.75 ± 0.43 ✓	83.57 ± 1.28 ✓	89.89 ± 0.29 ✓	92.10 ± 0.30 ✓	83.17 ± 0.31 −	63.66 ± 1.04 −
0.30	Baseline	84.50 ± 0.50	76.38 ± 1.56	86.09 ± 1.47	75.17 ± 1.97	81.73 ± 0.53	60.98 ± 0.57
	Baseline (orig)	85.57 ± 0.59	72.03 ± 0.65	82.71 ± 1.29	80.76 ± 2.04	81.22 ± 0.95	61.00 ± 0.57
	LBE	62.72 ± 0.44	58.58 ± 0.84	80.62 ± 4.85	73.07 ± 2.05	69.93 ± 3.10	81.32 ± 7.27
	SAR-EM	60.85 ± 0.27	52.09 ± 0.19	64.83 ± 0.39	76.37 ± 0.49	70.43 ± 0.38	66.94 ± 1.86
	A^3	92.47 ± 0.32 ✓	90.49 ± 0.23 ✓	89.87 ± 0.65 ✓	**93.45 ± 0.12** ✓	85.16 ± 0.41 ✓	**90.30 ± 2.59** ✓
	IsolationForest	92.68 ± 0.31 ✓	90.59 ± 0.23 ✓	**92.67 ± 0.25** ✓	92.02 ± 0.44 ✓	85.13 ± 0.36 ✓	72.12 ± 3.10 ✓
	ECODv2	92.50 ± 0.32 ✓	89.55 ± 0.21 ✓	90.66 ± 0.35 ✓	93.08 ± 0.12 ✓	85.08 ± 0.35 ✓	77.92 ± 2.90 ✓
	OC-SVM	**92.71 ± 0.31** ✓	**90.67 ± 0.22** ✓	90.67 ± 0.37 ✓	93.08 ± 0.12 ✓	**85.22 ± 0.34** ✓	74.90 ± 1.97 ✓
0.50	Baseline	86.37 ± 0.59	80.51 ± 0.99	87.44 ± 0.90	80.42 ± 2.71	81.39 ± 0.36	66.74 ± 3.03
	Baseline (orig)	88.74 ± 0.58	80.40 ± 0.82	83.56 ± 1.14	77.91 ± 1.81	81.97 ± 0.71	61.32 ± 0.61
	LBE	72.72 ± 0.43	66.42 ± 1.45	90.34 ± 1.11	79.91 ± 4.99	82.60 ± 2.32	**93.84 ± 2.73**
	SAR-EM	70.51 ± 0.24	59.13 ± 0.20	81.96 ± 0.47	87.94 ± 0.40	83.37 ± 0.37	83.52 ± 1.49
	A^3	92.75 ± 1.11 ✓	92.24 ± 0.25 ✓	92.23 ± 0.32 ✓	**93.44 ± 0.12** ✓	**87.00 ± 0.35** ✓	83.15 ± 3.82 ✓
	IsolationForest	92.92 ± 1.02 ✓	92.72 ± 0.22 ✓	**93.50 ± 0.21** ✓	92.49 ± 0.30 ✓	86.92 ± 0.37 ✓	71.49 ± 3.58 ✓
	ECODv2	**92.93 ± 1.04** ✓	91.97 ± 0.23 ✓	92.31 ± 0.19 ✓	93.41 ± 0.15 ✓	86.85 ± 0.36 ✓	81.05 ± 2.42 ✓
	OC-SVM	92.80 ± 1.05 ✓	**92.79 ± 0.21** ✓	92.90 ± 0.24 ✓	93.40 ± 0.14 ✓	86.98 ± 0.36 ✓	70.27 ± 3.19 ✓
0.70	Baseline	90.20 ± 0.68	85.61 ± 1.08	87.01 ± 0.65	85.01 ± 1.44	83.73 ± 0.29	61.17 ± 0.53
	Baseline (orig)	90.55 ± 0.52	87.87 ± 0.66	87.74 ± 1.19	87.14 ± 1.42	84.87 ± 0.52	61.22 ± 0.53
	LBE	82.33 ± 0.50	66.56 ± 3.24	91.49 ± 1.27	87.97 ± 3.83	83.13 ± 2.55	**98.10 ± 0.46**
	SAR-EM	80.45 ± 0.21	79.92 ± 0.13	92.66 ± 0.19	**94.14 ± 0.13**	**88.77 ± 0.30**	92.92 ± 0.63
	A^3	93.54 ± 0.79 ✓	94.09 ± 0.31 ✓	93.21 ± 0.23 ✓	93.99 ± 0.07 ✓	88.31 ± 0.33 ✓	95.13 ± 0.70 ✓
	IsolationForest	94.02 ± 0.62 ✓	94.36 ± 0.27 ✓	**93.62 ± 0.20** ✓	93.74 ± 0.10 ✓	88.36 ± 0.36 ✓	72.28 ± 2.39 ✓
	ECODv2	93.55 ± 0.73 ✓	93.70 ± 0.29 ✓	93.44 ± 0.23 ✓	93.81 ± 0.09 ✓	88.51 ± 0.28 ✓	94.22 ± 1.03 ✓
	OC-SVM	**94.05 ± 0.66** ✓	**94.39 ± 0.28** ✓	93.47 ± 0.23 ✓	93.77 ± 0.09 ✓	88.44 ± 0.33 ✓	91.95 ± 0.98 ✓

Table 2: F1 score values per dataset. Green ticks („✓") correspond to the cases when t-test rejected equality of F1 scores of the VAE-PU-OCC method considered and that of the baseline original method in favor of the former one at $\alpha = 0.05$. Dashes („−") indicate the failure to reject (see appendix F). Bold entries correspond to maximum mean F1 score for a given dataset and label frequency combination.

c	Method	MNIST 3v5	MNIST OvE	CIFAR CarTruck	CIFAR MachineAnimal	STL MachineAnimal	Gas Concentrations
0.02	Baseline	**82.59 ± 0.85**	75.85 ± 0.77	75.01 ± 4.94	86.11 ± 0.87	67.54 ± 1.73	83.62 ± 1.60
	Baseline (orig)	80.91 ± 0.85	68.48 ± 0.50	77.23 ± 3.82	86.08 ± 1.38	68.14 ± 1.45	81.06 ± 1.79
	LBE	3.64 ± 0.37	2.14 ± 0.20	0.81 ± 0.19	0.32 ± 0.07	1.31 ± 0.34	0.21 ± 0.08
	SAR-EM	2.96 ± 0.28	1.73 ± 0.09	2.19 ± 0.50	3.51 ± 0.33	2.91 ± 0.37	14.56 ± 1.23
	A^3	82.56 ± 0.86 −	76.97 ± 1.17 ✓	82.75 ± 1.31 −	**88.96 ± 0.30** ✓	**75.24 ± 0.62** ✓	**91.73 ± 1.37** ✓
	IsolationForest	82.53 ± 0.92 −	77.26 ± 1.12 ✓	**87.72 ± 2.01** ✓	88.22 ± 0.47 −	70.92 ± 0.65 ✓	84.81 ± 2.06 −
	ECODv2	**82.59 ± 0.92** −	76.50 ± 0.98 ✓	79.62 ± 1.81 −	88.57 ± 0.29 ✓	70.49 ± 0.68 −	84.96 ± 1.69 −
	OC-SVM	82.54 ± 0.96 −	**77.27 ± 1.11** ✓	79.73 ± 1.89 −	88.57 ± 0.29 ✓	70.49 ± 0.67 −	85.96 ± 1.79 ✓
0.10	Baseline	87.50 ± 0.63	79.17 ± 1.04	88.94 ± 0.79	82.11 ± 1.92	79.89 ± 0.55	76.04 ± 0.52
	Baseline (orig)	84.26 ± 0.47	72.43 ± 0.30	87.34 ± 0.70	81.62 ± 2.22	79.61 ± 0.49	76.08 ± 0.53
	LBE	17.18 ± 0.74	12.11 ± 0.71	4.78 ± 0.70	8.97 ± 6.99	4.70 ± 0.93	0.77 ± 0.20
	SAR-EM	16.13 ± 0.43	9.60 ± 0.20	11.70 ± 0.76	24.02 ± 1.22	9.98 ± 0.56	22.71 ± 2.98
	A^3	90.65 ± 0.39 ✓	83.95 ± 1.13 ✓	90.18 ± 0.43 ✓	**90.81 ± 0.32** ✓	**80.66 ± 0.28** ✓	**91.31 ± 1.14** ✓
	IsolationForest	91.12 ± 0.36 ✓	84.32 ± 1.08 ✓	**91.31 ± 0.36** ✓	88.65 ± 0.55 ✓	80.60 ± 0.25 ✓	76.52 ± 0.60 −
	ECODv2	**91.34 ± 0.32** ✓	83.20 ± 1.06 ✓	90.59 ± 0.28 ✓	90.51 ± 0.33 ✓	80.53 ± 0.27 −	78.33 ± 1.02 ✓
	OC-SVM	91.30 ± 0.37 ✓	**84.33 ± 1.07** ✓	90.58 ± 0.25 ✓	90.51 ± 0.33 ✓	80.56 ± 0.26 −	76.93 ± 0.65 −
0.30	Baseline	87.16 ± 0.36	80.99 ± 1.05	87.66 ± 1.14	76.34 ± 1.52	80.57 ± 0.43	75.73 ± 0.45
	Baseline (orig)	87.64 ± 0.44	77.02 ± 0.30	85.15 ± 0.96	80.68 ± 1.60	80.17 ± 0.66	75.74 ± 0.45
	LBE	47.99 ± 0.75	40.57 ± 1.00	71.66 ± 9.77	56.89 ± 7.47	53.68 ± 8.28	74.83 ± 11.37
	SAR-EM	43.02 ± 0.43	30.21 ± 0.29	47.12 ± 0.85	59.32 ± 1.19	43.02 ± 0.79	61.74 ± 3.13
	A^3	92.84 ± 0.29 ✓	90.65 ± 0.23 ✓	90.51 ± 0.55 ✓	**91.97 ± 0.12** ✓	82.74 ± 0.36 ✓	**92.86 ± 1.57** ✓
	IsolationForest	93.07 ± 0.28 ✓	90.78 ± 0.22 ✓	**92.70 ± 0.25** ✓	90.46 ± 0.43 ✓	82.67 ± 0.35 ✓	81.63 ± 1.80 ✓
	ECODv2	92.91 ± 0.28 ✓	89.92 ± 0.20 ✓	91.15 ± 0.33 ✓	91.50 ± 0.13 ✓	82.69 ± 0.33 ✓	84.88 ± 1.65 ✓
	OC-SVM	**93.10 ± 0.27** ✓	**90.89 ± 0.20** ✓	91.17 ± 0.35 ✓	91.50 ± 0.13 ✓	**82.78 ± 0.31** ✓	82.95 ± 1.12 ✓
0.50	Baseline	88.47 ± 0.49	83.90 ± 0.68	88.70 ± 0.72	80.81 ± 2.18	80.61 ± 0.27	78.12 ± 1.53
	Baseline (orig)	90.08 ± 0.47	83.11 ± 0.53	85.78 ± 0.88	78.39 ± 1.39	81.05 ± 0.58	75.89 ± 0.46
	LBE	67.41 ± 0.58	64.63 ± 1.53	90.81 ± 0.93	79.75 ± 3.31	80.05 ± 1.77	**94.27 ± 2.71**
	SAR-EM	62.77 ± 0.19	51.08 ± 0.26	78.75 ± 0.71	82.96 ± 0.64	75.04 ± 0.61	84.07 ± 1.74
	A^3	93.29 ± 0.95 ✓	92.49 ± 0.24 ✓	92.47 ± 0.27 ✓	**92.06 ± 0.13** ✓	**84.42 ± 0.39** ✓	87.88 ± 2.30 ✓
	IsolationForest	**93.42 ± 0.89** ✓	92.90 ± 0.21 ✓	**93.46 ± 0.20** ✓	91.11 ± 0.29 ✓	84.37 ± 0.32 ✓	80.91 ± 1.97 ✓
	ECODv2	93.39 ± 0.94 ✓	92.20 ± 0.22 ✓	92.54 ± 0.21 ✓	91.96 ± 0.16 ✓	84.31 ± 0.32 ✓	86.18 ± 1.57 ✓
	OC-SVM	93.31 ± 0.91 ✓	**92.96 ± 0.20** ✓	93.09 ± 0.22 ✓	91.95 ± 0.15 ✓	**84.42 ± 0.32** ✓	80.21 ± 1.64 ✓
0.70	Baseline	91.40 ± 0.57	87.54 ± 0.83	88.35 ± 0.52	84.14 ± 1.27	82.47 ± 0.25	75.82 ± 0.42
	Baseline (orig)	91.41 ± 0.44	89.04 ± 0.51	89.00 ± 0.95	86.08 ± 1.27	83.41 ± 0.43	75.85 ± 0.42
	LBE	82.33 ± 0.45	73.48 ± 1.62	91.92 ± 1.07	87.50 ± 3.08	81.71 ± 2.03	**98.42 ± 0.38**
	SAR-EM	78.42 ± 0.21	77.94 ± 0.16	92.42 ± 0.23	92.53 ± 0.16	84.98 ± 0.38	93.79 ± 0.59
	A^3	93.96 ± 0.71 ✓	94.20 ± 0.29 ✓	93.31 ± 0.24 ✓	**92.58 ± 0.09** ✓	85.84 ± 0.28 ✓	96.13 ± 0.54 ✓
	IsolationForest	94.33 ± 0.58 ✓	94.46 ± 0.26 ✓	**93.62 ± 0.22** ✓	92.28 ± 0.10 ✓	**85.93 ± 0.35** ✓	81.53 ± 1.51 ✓
	ECODv2	93.95 ± 0.68 ✓	93.86 ± 0.27 ✓	93.51 ± 0.25 ✓	92.35 ± 0.11 ✓	**85.93 ± 0.26** ✓	95.50 ± 0.77 ✓
	OC-SVM	**94.38 ± 0.62** ✓	**94.49 ± 0.27** ✓	93.56 ± 0.24 ✓	92.30 ± 0.11 ✓	**85.93 ± 0.27** ✓	93.83 ± 0.69 ✓



(a) (b)

Figure 1: Distribution of p-values obtained by evaluation of trained A^3 classifier on unlabeled (U) dataset, separated by true sample class. (a) MNIST 3v5, (b) CIFAR MachineAnimal.

other hand, managed to reach significantly better accuracy and F1 score on Gas Concentrations dataset (for high enough c values) than its competitors. In other test cases, however, its performance is relatively subpar – similarly to SAR-EM, its F1 score plummets when decreasing the proportion of labeled positive examples. This establishes low label frequency performance as a substantial advantage of generative models over the algorithms which model the propensity score explicitly such as SAR-EM and LBE. We investigate the possible causes of this behavior further on in this section.

VAE-PU+OCC achieves excellent classification results on all of the benchmark datasets, as measured by the accuracy and F1 score. All of the proposed OCC variants outperform baseline VAE-PU models in all of the test cases. In some cases performance increase is very slight (e.g. MNIST 3v5, $c = 0.02$), where only a fraction of a percent increase in accuracy is observed; but there are also cases where classification performance rises dramatically (see Gas Concentration results), up to tens of percentage points (pps). Overall, applying the OCC variants results in a substantial increase in both the accuracy and F1 score, in most cases by a several pps. Difference between different VAE-PU+OCC flavors presented in the tables are usually slight (generally below 1pp for both accuracy and F1 score). Nevertheless, there are also scenarios where one of the methods performs significantly better that the other – A^3 dominates competitors on Gas Concentrations dataset, while Isolation Forest performs significantly better for CIFAR CarTruck data. Overall, A^3 variant is the most noteworthy – it outperforms competitors on multiple datasets, while remaining competitive in scenarios which proved more difficult for the method. Closer look at the results reveals that competitor methods (VAE-PU, SAR-EM and LBE) tend to introduce precision-recall imbalance (for VAE-PU, the recall is often much higher of the two, while SAR-EM and LBE are skewed towards the precision; for detailed metric values and discussion, refer to appendix C), while OCC variants of VAE-PU achieve balanced results in nearly all test cases.

Another significant feature of the VAE-PU+OCC variants is stabilization of the results. Note that the standard error of the mean (SEM) for all proposed methods decreases significantly (as compared to the baseline VAE-PU) in a majority of test cases. Notable exception is the Gas Concentration dataset, where even though SEM usually increases, it occurs with a simultaneous significant classification performance improvement. VAE-PU+OCC also offers drastically lowered training time (up to 10 times shorter, compared to alternatives such as SAR-EM and LBE; for detailed time values please refer to appendix D), which makes it attractive even in rare cases where its accuracy is lower. Naturally, it is also slower than baseline VAE-PU, but this loss doesn't usually exceed 20% extra training time.

Even though due to already long time required to obtain experimental results we limited the number of tested OCC methods to four, we feel that it is a representative sample of approaches, incorporating both classic and modern models and ranging from simple, statistical methods to neural-network based classifiers. As a result of the experiments we suggest that A^3-based variant of VAE-PU+OCC to be recommended in most practical scenarios, due to exceptional performance in several scenarios while maintaining strong baseline accuracy in general case.

The results in SCAR setting which follow similar patterns to no-SCAR scenario, are discussed in Appendix E.

5 Conclusions

VAE-PU+OCC builds upon an innovative VAE-PU model, which proved the strength of generative approaches in no-SCAR PU data modeling. Through the application of one-class classification methods, both modern and traditional, the extended model has shown excellent results in a diverse array of experiments. The highlight is the outstanding accuracy of the models in medium label frequency settings – for low label frequencies, the gains of OCC-based models are minor relative to VAE-PU baselines, whereas for high label frequencies classical, non-generative algorithms such as SAR-EM and LBE remain competitive. This paper proves that application of one-class classification techniques in no-SCAR PU learning provides a substantial improvement. One of the important issues is a construction of one-class classifier specially designed for PU data.

Further possible improvement of the presented approach would be to avoid assumption that $P(Y = 1)$ is known. We have experimented with using Storey's method [24], of estimating the proportion of valid null hypotheses (proportion of PU elements in U sample in our case), but this resulted in deteriorated performance, possibly due to quite complicated inter-relations between optimization processes. We believe, though, that there is still a room for improvement here.

References

[1] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia, 'Testing for outliers with conformal p-values', *The Annals of Statistics*, **51**(1), 149 – 178, (2023).

[2] Jessa Bekker and Jesse Davis, 'Learning from positive and unlabeled data: a survey', *Machine Learning*, **109**(4), 719–760, (April 2020).

[3] Jessa Bekker, Pieter Robberechts, and Jesse Davis, 'Beyond the selected completely at random assumption for learning from positive and unlabeled data', in *Proceedings of ECMLPKDD'2019*, pp. 71–85. Springer, Cham, (2019).

[4] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, 2006.

[5] Olivier Coudray, Christine Keribin, Pascal Massart, and Patrick Pamphile, 'Risk bounds for positive-unlabeled learning under the selected at random assumption', *Journal of Machine Learning Research*, **24**, 1–31, (2023).

[6] Walter Gerych, Thomas Hartvigsen, Luke Buquicchi, Emmanuel Agu, and Elke Rundensteiner, 'Recovering the propensity score from biased positive unlabeled data', in *AAAI*, (2022).

[7] Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane Jia You, Jian Yang, and Dacheng Tao, 'Instance-dependent positive and unlabeled learning with labeling bias estimation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 4163–4177, (2022).

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer, Second Edition, 2009.

[9] J. Heckman, 'Sample selection bias as a specification error', *Econometrica*, **47**, 153–161, (1979).

[10] Shantanu Jain, Justin Delano, Himanshu Sharma, and Predrag Radivojac, 'Class prior estimation with biased positives and unlabeled examples', *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 4255–4263, (04 2020).

[11] Xu Ji, João F Henriques, and Andrea Vedaldi, 'Invariant information clustering for unsupervised image classification and segmentation', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874, (2019).

[12] Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama, 'Positive-unlabeled learning with non-negative risk estimator', in *Proceedings of the NIPS'17*, NIPS'17, pp. 1674—1684, Red Hook, NY, USA, (2017). Curran Associates Inc.

[13] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen, 'ECOD: Unsupervised outlier detection using empirical cumulative distribution functions', in *IEEE Transactions on Knowledge and Data Engineering*, (2022).

[14] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, 'Isolation Forest', in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, (2008).

[15] M. M. Moya, M. W. Koch, and L. D. Hostetler, 'One-class classifier networks for target recognition applications', *NASA STI/Recon Technical Report*, **N 93**, (January 1993).

[16] Byeonghu Na, Hyemi Kim, Kyungwoo Song, Weonyoung Joo, Yoon-Yeong Kim, and Il-Chul Moon, 'Deep generative positive-unlabeled learning under selection bias', in *Proceedings of CIKM'20*, CIKM '20, pp. 1155—1164, New York, NY, USA, (2020). ACM.

[17] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko, 'A review of novelty detection', *Signal Processing*, **99**, 215–249, (2014).

[18] L. Ruff, J. Kauffmann, R. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. Dietterich, and K. Müller, 'A unifying review of deep and shallow anomaly detection', *Proceedings of the IEEE*, **PP**, 1–40, (February 2021).

[19] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs, 'Unsupervised anomaly detection with generative adversarial networks to guide marker discovery', in *Information Processing in Medical Imaging*, pp. 146–157. Springer, (2017).

[20] Thobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims, 'Recommendation as treatments: debiasing learning and evaluation', *ICML*, **48**, 1670–1679, (2016).

[21] Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson, 'Estimating support of a high-dimensional distribution', *Neural Computation*, **13**, 1443–1471, (July 2001).

[22] Erik Shultheis, Rohit Babbar, Marek Wydmuch, and Krzysztof Dembczyński, 'On missing labels, long-tails and propensities in extreme multi-label classification', in *KDD'22*, pp. 1547–1557, (2022).

[23] Philip Sperl, Jan-Philipp Schulze, and Konstantin Böttinger, 'Activation anomaly analysis', in *Machine Learning and Knowledge Discovery in Databases*, 69–84, Springer International Publishing, (2021).

[24] John Storey, 'Direct approach to false discovery rates', *Journal of the Royal Statistical Society. Series B (Methodological)*, **64**, 479–498, (2002).

[25] Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, Tomoya Sakai, and Niu ang, *Machine Learning from Weak Supervision*, MIT Press, 2022.

[26] Vladimir Vovk, Alex Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, Springer Science & Business Media, New York, NY, January 2005.

[27] Vladimir Vovk, Alexander Gammerman, and Craig Saunders, 'Machine-learning applications of algorithmic randomness', in *Proceedings of ICML'99*, ICML '99, pp. 444—453, San Francisco, CA, USA, (June 1999). Morgan Kaufmann Publishers Inc.

[28] G. Ward, T. Hastie, S. Barry, J. Elith, and J. Leathwick, 'Presence-only data and the EM algorithm', *Biometrics*, **65**, 554–563, (2009).