Learning from Positive and Unlabeled Data under the Selected At Random Assumption

Jessa Bekker JESSA.BEKKER@CS.KULEUVEN.BE
Jesse Davis JESSE.DAVIS@CS.KULEUVEN.BE
KU Leuven, Belgium

Abstract

For many interesting tasks, such as medical diagnosis and web page classification, a learner only has access to some positively labeled examples and many unlabeled examples. Learning from this type of data requires making assumptions about the true distribution of the classes and/or the mechanism that was used to select the positive examples to be labeled. The commonly made assumptions, separability of the classes and positive examples being selected completely at random, are very strong. This paper proposes a weaker assumption that assumes the positive examples to be selected at random, conditioned on some of the attributes. To learn under this assumption, an EM method is proposed. Experiments show that our method is not only very capable of learning under this assumption, but it also outperforms the state of the art for learning under the selected completely at random assumption.

1. Introduction

When learning binary classifiers from fully labeled data, algorithms have access to the class labels for all examples. However, in practice, many data sets only provide positive labels for some examples, with the remaining data being unlabeled and containing both positive and negative examples. Learning from positive and unlabeled data (PU learning) attempts to learn a classifier from this data. PU learning is closely related to semi-supervised learning and one-class classification (Khan and Madden, 2014).

The problem of positive and unlabeled data arises often in practice. Medical records, for example, list the diseases that patients have been diagnosed with. However, some diseases, like diabetes, often go undiagnosed. In this case, it is wrong to assume that an undiagnosed patient does not have the disease. Another examples are bookmarked pages, as these are but a subset of the pages of interest (Lee and Liu, 2003; Liu et al., 2003), and knowledge bases, which only contain a subset of facts (Zupanc and Davis, 2018).

There are roughly three established assumptions that enable PU learning. 1) Assuming the unlabeled data to be negative (Neelakantan et al., 2015). 2) Assuming separability, i.e., that the negative examples are very different from the positive ones. Learning then consists of two steps: finding reliable negative examples and then applying standard machine learning (Li and Liu, 2003; Yu, 2005; Nguyen et al., 2011). 3) Assuming that the labeled examples are selected completely at random from the set of positive examples, a classifier can be learned by incorporating the probability of labeling an positive example (Denis,

1998; Lee and Liu, 2003; Liu et al., 2003, 2005; Denis et al., 2005; Zhang and Lee, 2005; Elkan and Noto, 2008; Mordelet and Vert, 2014; Claesen et al., 2015).

The aforementioned assumptions are very strong. Our goal in this paper is to enable learning under weaker assumptions by introducing a new assumption to enable learning from positive and unlabeled data: the *Selected At Random (SAR)* assumption. It is related to the third category of approaches, but instead of assuming a constant probability for all positive examples to be labeled, it assumes that the probability is a function of the attributes, called the propensity score. The propensity score originates from causal inference, but has also been applied in semi-supervised learning (Imbens and Rubin, 2015; Schnabel et al., 2016). In order for it to be possible to learn in this setting, our proposed method SAR-EM assumes that the propensity score only depends on a known subset of the attributes.

Our contributions are the following: 1) formulate the SAR assumption for learning from positive and unlabeled data, 2) propose a method SAR-EM that works under this assumption, 3) show that a special case of SAR-EM which assumes the SCAR assumption outperforms the state-of-the-art methods for estimating the class prior in PU data, 4) show that incorrectly making the SCAR assumption hurts the classifier performance, and 5) show that SAR-EM can reconstruct propensity score functions and learn good classifiers.

2. Background

The goal of learning from positive and unlabeled (PU) data is to train a binary classifier while only having access to positively labeled and unlabeled examples. An example is represented by $\{x, y, s\}$, where x are its attributes, y the true class and s the label. Only positive examples are labeled: $s = 1 \Rightarrow y = 1$, and unlabeled examples $s = 0$ can be of any class y. Bold letters depict sets of variables, i.e., the dataset is represented by $\{\mathbf{x}, \mathbf{y}, \mathbf{s}\}$.

To enable learning with positive and unlabeled data, assumptions about the population of positives and negatives in the instance space are necessary. The two most popular assumptions are 1) separability and 2) selected completely at random. The separability assumption assumes that within the considered class of models a model exists that can perfectly separate the positive from the negative examples. This assumption is violated if the attributes do not contain enough information to deterministically determine the class. When separability holds, two-step approaches can be used to solve the problem (Li and Liu, 2003; Yu, 2005; Li et al., 2009; Nguyen et al., 2011).

Learning without the separability assumptions is challenging because examples with a low probability of being positive could still have a positive label. In this case, the sampling mechanism to select positive examples to be labeled needs to be considered. The most common assumption is that the labeled positives are a random subsample of the true positives. This is the selected completely at random assumption.

2.1. Selected Completely at Random Assumption

Under the *Selected Completely At Random (SCAR)* assumption, every positive example has exactly the same probability to be selected to be labeled (Elkan and Noto, 2008). Given the class, this probability (the *label frequency* c), is conditionally independent of the attributes:

$$c = \Pr(s = 1|\mathbf{x}, y = 1) = \Pr(s = 1|y = 1). \tag{1}$$

The SCAR assumption was introduced in analogy with the *Missing Completely A Random assumption (MCAR)* that is common when working with missing data (Rubin, 1976; Little and Rubin, 2014). However, there is a notable difference between the two assumptions. In MCAR data, the missingness of the variable cannot depend on the value of the variable, where in PU learning this is necessarily the case because all negative labels are missing. The class values are missing completely at random only if just the population of positive examples is considered.

A very useful property follows from the SCAR assumption: An example's probability of belonging to the positive class is directly proportional to the probability of that example being labeled (Elkan and Noto, 2008):

$$\Pr(y = 1|\mathbf{x}) = \frac{1}{c} \Pr(s = 1|\mathbf{x}). \tag{2}$$

A probabilistic model to predict $\Pr(y = 1|\mathbf{x})$ can be obtained by training a model that predicts $\Pr(s = 1|\mathbf{x})$ from the data and scaling the output probabilities with $1/c$. This method only works with well-calibrated models and is therefore not always robust. Other methods have been introduced to cope with this, they integrate the label frequency in the training process (Denis, 1998; De Comité et al., 1999; Liu et al., 2003; Zhang and Lee, 2005; Elkan and Noto, 2008).

The previously mentioned methods require the label frequency, or equivalently the class prior $\alpha = \Pr(y = 1) = \Pr(s = 1)/c$, to be known. The class prior can be known from domain knowledge or could be estimated by labeling a small random sample of the dataset.

A substantial effort has been done to estimate the label frequency c directly from the data (Elkan and Noto, 2008; du Plessis et al., 2015; Jain et al., 2016a; Ramaswamy et al., 2016; Bekker and Davis, 2018a,b). The common assumption that these methods make is that in some region in the instance space, the positive class probability should be 1. We refer to this assumption as the *local certainty assumption*.

3. Selected at Random Assumption

This paper introduces the notion of positive examples being *Selected At Random (SAR)* to be labeled. This assumes that the selection probability is completely determined by its attributes. Just like SCAR has MCAR as a counterpart in the missing data literature, SAR is based on the *Missing At Random (MAR)* assumption (Rubin, 1976; Little and Rubin, 2014).

3.1. Propensity score

The probability that a positive example is selected to be labeled is the *propensity score* $e(x)$:

$$e(x) = \Pr(s = 1|y = 1, x).$$

Note that the propensity score only applies to positive examples; negative examples never get selected. This seems to imply that negative examples are never considered to be labeled. However, this is not necessarily the case, the labels could be lost afterwards. For example, doctors test both ill and healthy patients for diseases but may only store positive results in the patients' records. In this case, the propensity score is the unconditional probability of testing a patient and the conditional probability of storing the disease in the record.

3.2. Learning with a Known Propensity Score

The propensity score is known when the selection mechanism for positive examples is understood. For example, a hospital might have a protocol for testing people. An unbiased classifier can then be trained from positive and unlabeled data by taking the propensity score into account in an equivalent way as the label frequency under the SCAR assumption: by either scaling the output probabilities with $1/e(x)$ or integration in the training process.

3.3. Learning with an Unknown Propensity Score

This paper's goal is to learn from positive and unlabeled data under the SAR assumption with an unknown propensity score. It is an ill-defined problem because any unlabeled example can be explained by both a low positive class probability and a low labeling probability. This section evaluates which reasonable additional assumptions would enable learning.

Propensity Attributes The first assumption is that the propensity function only requires a subset of the attributes $x_e \in x$:

$$\Pr(s = 1|y = 1, x) = \Pr(s = 1|y = 1, x_e)$$
$$e(x) = e(x_e).$$

This is often a reasonable assumption: not all of the attributes might have been available at selection time, some of the attributes might be difficult to interpret for a labeler, or it might be known which variables have the highest correlation with the class and the labeler only used those.

SAR as Multi-SCAR Given discrete propensity attributes, the classification problem under the SAR assumption can be reduced to multiple classification problems under the SCAR assumption by partitioning the population into strata based on assignments to values of the propensity attributes. While being suboptimal in practice, this approach is insightful for theoretical analysis. Indeed, the conditions needed for training a correct classifier in each of the strata are sufficient for a correct classifier over the entire population.

Local Certainty Assumption for Any Propensity Configuration To enable learning with an unknown label frequency under the SCAR assumption, *local certainty* is commonly assumed. This means that the probability of examples belonging to the positive class needs to be 1 in some region of the attribute domain. A sufficient condition for learning under the SAR assumption to be possible is therefore that *local certainty* holds in all the propensity strata. Although seemingly very strong, this assumption is not implausible. It holds, for example, if one of the non-propensity attributes always provides certainty of the positive class. In the task of classifying web pages as commercial, any page with a "buy" button on it is commercial, regardless of the tags that the labeler had access to when choosing the pages to label.

We argue that the local certainty assumption for any propensity configuration can be relaxed if the propensity score function and the classification function have a certain smoothness over the propensity attributes. In other words, some similarity between the classification function when conditioned on different propensity configurations is expected. With this insight, we propose to simultaneously train a classifier and a propensity score function while promoting local certainty as much as possible.

Algorithm 1: SAR-EM

Input: attributes $\mathbf{x}$, labels $\mathbf{s}$, propensity attributes $\mathbf{x_e}$, local certainty parameter d
Output: classifier f, propensity score model e

$f, e = \text{initialize\_models}(\mathbf{x}, \mathbf{s}, \mathbf{x_e})$
repeat
  // Expectation
  $\hat{\mathbf{y}}_\mathbf{f} = f(\mathbf{x}[\mathbf{s} == 0])$
  $\hat{\mathbf{s}} = d \cdot e(\mathbf{x_e}[\mathbf{s} == 0])$ // Decay propensity scores
  $\hat{\mathbf{y}} = \frac{\hat{\mathbf{y}}_\mathbf{f}(1-\hat{\mathbf{s}})}{\hat{\mathbf{y}}_\mathbf{f}(1-\hat{\mathbf{s}})+(1-\hat{\mathbf{y}}_\mathbf{f})}$
  // Maximization
  $\mathbf{y_w} = \text{ones}(\text{size}(\mathbf{s})) :: \text{zeros}(\text{size}(\mathbf{s}[\mathbf{s} == 0]))$
  $\mathbf{x_w} = \mathbf{x}[\mathbf{s} == 1] :: \mathbf{x}[\mathbf{s} == \mathbf{0}] :: \mathbf{x}[\mathbf{s} == \mathbf{0}]$
  $\mathbf{w} = \mathbf{s}[\mathbf{s} == 1] :: \hat{\mathbf{y}} :: (1 - \hat{\mathbf{y}})$
  $f = \text{fit}(\mathbf{x_w}, \mathbf{y_w}, \mathbf{w})$
  $\mathbf{s_w} = \mathbf{s}[\mathbf{s} == 1] :: \mathbf{s}[\mathbf{s} == 0]$
  $\mathbf{x_{ew}} = \mathbf{x_e}[\mathbf{s} == 1] :: \mathbf{x_e}[\mathbf{s} == \mathbf{0}]$
  $\mathbf{w} = \mathbf{s}[\mathbf{s} == 1] :: \hat{\mathbf{y}}$
  $e = \text{fit}(\mathbf{x_{ew}}, \mathbf{s_w}, \mathbf{w})$
until *Converged or maximum iterations reached*
return f, e

4. An EM Method for PU Learning under the SAR Assumption

We use an Expectation Maximization (EM) approach to solve the learning problem. The true class values y are hidden variables, while attributes x and labels s are observed and the propensity attributes are known. The data is generated by the following process:

$$\begin{aligned}(x, y, s) &\sim \Pr(x, y, s) \\ &\sim \Pr(x)\Pr(y|x)\Pr(s|x, y) \\ &\sim \Pr(x)\Pr(y|x)\Pr(s|x_e, y),\end{aligned}$$

where x_e are the propensity attributes. We assume that the process of classifying examples and labeling examples can be modeled using parameters θ and ϕ respectively:

$$(x, y, s) \sim \Pr(x)\Pr(y|x, \theta)\Pr(s|x_e, y, \phi).$$

The goal is to find the parameters θ and ϕ that explain the observed data the best.

4.1. Expectation Maximization

Optimizing parameters θ and ϕ means setting them to maximize the log likelihood of observing (x, s). Expectation maximization repeats two steps until the models e. During the *expectation* step, it finds the expected values $\hat{y}$ for y given the current models. During the *maximization* step, it retrains the models to optimize the log likelihood of observing $(x, s, \hat{y})$. The steps are derived below.

Expectation (Algorithm 1, lines 3-7)

$$\Pr(y = 1|x, s = 1, \theta, \phi) = 1$$

$$\begin{aligned}
&\Pr(y = 1|x, s = 0, \theta, \phi) \\
&\quad = \frac{\Pr(x, s = 0|y = 1, \theta, \phi) \Pr(y = 1|\theta, \phi)}{\Pr(x, s = 0|\theta, \phi)} \\
&\quad = \frac{\Pr(x|y = 1, \theta, \phi) \Pr(s = 0|x, y = 1, \theta, \phi) \Pr(y = 1|\theta, \phi)}{\Pr(x, s = 0|\theta, \phi)} \\
&\quad = \frac{\frac{\Pr(y=1|x,\theta) \Pr(x|\theta,\phi)}{\Pr(y=1|\theta,\phi)} \Pr(s = 0|x_e, y = 1, \phi) \Pr(y = 1|\theta, \phi)}{\Pr(x, s = 0|\theta, \phi)} \\
&\quad = \frac{\Pr(y = 1|x, \theta) \Pr(x|\theta, \phi) \Pr(s = 0|y = 1, x_e, \phi)}{\Pr(s = 0|x, \theta, \phi) \Pr(x|\theta, \phi)} \\
&\quad = \frac{\Pr(y = 1|x, \theta) \Pr(s = 0|y = 1, x_e, \phi)}{\Pr(s = 0|x, \theta, \phi)} \\
&\quad = \frac{\Pr(y = 1|x, \theta) \Pr(s = 0|y = 1, x_e, \phi)}{\Pr(y = 1|x, \theta) \Pr(s = 0|y = 1, x, \phi) + \Pr(y = 0|x, \theta) \Pr(s = 0|y = 0, x, \phi)} \\
&\quad = \frac{\Pr(y = 1|x, \theta) \Pr(s = 0|y = 1, x_e, \phi)}{\Pr(y = 1|x, \theta) \Pr(s = 0|y = 1, x_e, \phi) + \Pr(y = 0|x, \theta)}
\end{aligned}$$

Maximization (Algorithm 1, lines 8-16)

$$\begin{aligned}
\max_{\theta,\phi} \mathbb{E}_{y|x,s,\theta,\phi} \log \Pr(x, s, y|\theta, \phi) &= \max_{\theta,\phi} \mathbb{E}_{y|x,s,\theta,\phi} \log \left[\Pr(x) \Pr(y|x, \theta) \Pr(s|y, x_e, \phi)\right] \\
&= \max_{\theta,\phi} \mathbb{E}_{y|x,s,\theta,\phi} \log \left[\Pr(y|x, \theta) \Pr(s|y, x_e, \phi)\right] \\
&= \max_{\theta,\phi} \mathbb{E}_{y|x,s,\theta,\phi} \left[\log \Pr(y|x, \theta) + \log \Pr(s|y, x_e, \phi)\right] \\
&= \max_{\theta} \mathbb{E}_{y|x,s,\theta,\phi} \log \Pr(y|x, \theta) + \max_{\phi} \mathbb{E}_{y|x,s,\theta,\phi} \log \Pr(s|y, x_e, \phi)
\end{aligned}$$

4.2. Local Certainty

Applying pure EM ensures converging to a combination of classifier and propensity score model that explains the observed data well. However, care must be taken because a propensity score model that always returns 1 and a classifier that predicts the probability of observing a label explains the observed data perfectly but is not the desired solution. As stated in Section 3.3, classifiers with local certainty are preferred, so this needs to be taken into account during the EM optimization process.

To guide the learning towards local certainty, lower propensity scores need to be encouraged. To this end, the predicted propensity scores can simply be decayed at every iteration during the expectation step (Algorithm 1, line 6). This enhances the expected class probabilities and makes sure that a more positively inclined classifier is trained.

Algorithm 2: initialize_models

Input: attributes $\mathbf{x}$, labels $\mathbf{s}$, propensity attributes $\mathbf{x_e}$
Output: classifier f, propensity score model e

1 // Fit f to predict s from x
2 $f = \text{fit}(\mathbf{x}, \mathbf{s})$
3 // SCAR assumption with minimum c
4 $\hat{\mathbf{s}} = f(\mathbf{x}[\mathbf{s} == 0])$
5 $c = 1/\max(\hat{\mathbf{s}})$
6 // Fit f to predict s from x with c
7 $\hat{\mathbf{y}} = \frac{1-c}{c}\frac{\hat{s}}{1-\hat{s}}$
8 $\mathbf{y_w} = \text{ones}(\text{size}(\mathbf{s})) :: \text{zeros}(\text{size}(\mathbf{s}[\mathbf{s} == 0]))$
9 $\mathbf{x_w} = \mathbf{x}[\mathbf{s} == 1] :: \mathbf{x}[\mathbf{s} == \mathbf{0}] :: \mathbf{x}[\mathbf{s} == \mathbf{0}]$
10 $\mathbf{w} = \mathbf{s}[\mathbf{s} == 1] :: \hat{\mathbf{y}} :: (1 - \hat{\mathbf{y}})$
11 $f = \text{fit}(\mathbf{x_w}, \mathbf{y_w}, \mathbf{w})$
12 // Train e to return c for any input
13 $\mathbf{s_w} = \text{ones}(\text{size}(\mathbf{s})) :: \text{zeros}(\text{size}(\mathbf{s}))$
14 $\mathbf{x_{ew}} = \mathbf{x_e} :: \mathbf{x_e}$
15 $\mathbf{w} = c \cdot \text{ones}(\text{size}(\mathbf{s})) :: (1 - c) \cdot \text{ones}(\text{size}(\mathbf{s})$
16 $e = \text{fit}(\mathbf{x_{ew}}, \mathbf{s_w}, \mathbf{w})$
17 **return** *f,e*

4.3. Initialization

To start EM with reasonable models, a classifier and propensity model are trained under the SCAR assumption. The label frequency is estimated by training a model to predict $\Pr(s = 1|x)$ and setting the label frequency so that the maximum predicted class probability becomes 1 (Algorithm 2, lines 1-5). This is estimator 3 of Elkan and Noto (2008), which gives a fairly unstable estimate, but it is fine for initialization.

Subsequently, the classifier is trained to predict y by weighting the examples using the label frequency (Elkan and Noto, 2008) (Algorithm 2, lines 6-11). The propensity score model is trained to always return c, by providing all examples as both positive and negative, but giving weight c when positive and $1 - c$ otherwise (Algorithm 2, lines 12-16).

4.4. Convergence

Convergence is assumed when the propensity score predictions do not change much over several iterations. Change is quantified as the average absolute slope of the minimum mean square error line through the predictions:

$$\text{slope}(s, t, n) = \frac{n\sum_{i=0}^{n-1}(is_{t-n+1+i}) - \sum_{i=0}^{n-1} i \sum_{i=0}^{n-1} s_{t-i}}{n\sum_{i=0}^{n-1} i^2 - (\sum_{i=0}^{n-1} i)^2}$$

$$\frac{1}{|\hat{\mathbf{s}}|}\sum_{\hat{s}\in\hat{\mathbf{s}}} |\text{slope}(s, t, n)| < \epsilon,$$

Table 1: Datasets

Dataset	# Examples	# Vars	$\Pr(y=1)$
Breast Cancer	683	9	0.350
Mushroom	8,124	21	0.482
Adult	48,842	14	0.761
IJCNN	141,691	22	0.096
Cover Type	536,301	54	0.495
20ng Comp - Rec	5,287	200	0.450
20ng Comp - Sci	5,279	200	0.450
20ng comp - Talk	4,856	200	0.401
20ng rec - Sci	4,752	200	0.499
20ng Rec - Talk	4,329	200	0.450
20ng Sci - Talk	4,321	200	0.451

Table 2: Class prior estimation: SAR-EM has the best rank and lowest absolute error.

Method	Average $\lvert\hat{\alpha}-\alpha\rvert$ rank +/- SD	Average $\lvert\hat{\alpha}-\alpha\rvert$ +/- SD
SAR-EM	2.87 +/- 1.83	0.08 +/- 0.07
KM2	3.38 +/- 2.38	0.10 +/- 0.13
TI*c*E	3.86 +/- 2.13	0.09 +/- 0.06
AlphaMax_N	4.04 +/- 2.15	0.13 +/- 0.10
AlphaMax	4.16 +/- 1.72	0.12 +/- 0.09
KM1	4.92 +/- 2.01	0.11 +/- 0.09
EN	6.75 +/- 1.78	0.27 +/- 0.16
PE	7.02 +/- 1.61	0.29 +/- 0.14
pen-L1	7.81 +/- 2.03	0.37 +/- 0.21

where t is the current iteration, n the number of iterations over which the slope is taken, s_i the propensity score prediction during iteration i, and ϵ the minimum average absolute slope for non-convergence. Additionally, a maximum number of iterations can be set.

4.5. SCAR with Label Frequency Estimation as a Special Case of SAR-EM

SAR-EM can be used to train a classifier and estimate the label frequency under the SCAR assumption, by assigning no propensity attributes: $x_e = \emptyset$. Training the propensity score model then reduces to estimating the label frequency c given the expected class values.

4.6. Learning with a Known Propensity Score as a Special Case of SAR-EM

If the propensity score (or label frequency) is known, SAR-EM can still optimize the classifier. The algorithm remains the same, except for not training the propensity score model, i.e., removing lines 13-16 in Algorithm 1 and 12-16 in Algorithm 2.

Training the classifier using SAR-EM is more stable than predicting the labels and adjusting the output probabilities (Section 3.2). By optimizing the expected likelihood of the observed data, the influence of unexpectedly labeled low probability positives is reduced.

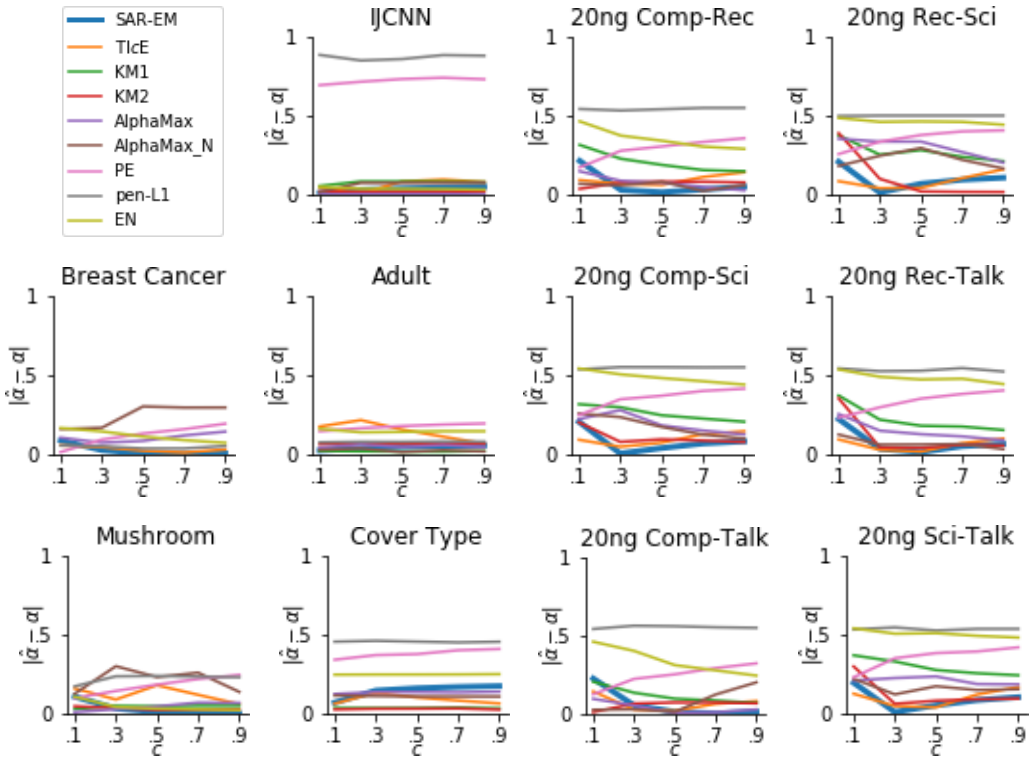

Figure 1: Class prior estimation comparison

5. Empirical Evaluation

Our experiments aim to evaluate the performance of SAR-EM. Because, to the best of our knowledge, no other methods exist to learn from positive and unlabeled data under the SAR assumption, we will first compare our method under the SCAR assumption to other methods that assume SCAR and estimate the class prior in this data. Next, the SAR assumption is considered. Here, our method is compared to SCAR methods, supervised methods and SAR-EM when the propensity score function is known. Lastly, we analyze the performance in unbalanced domains.

5.1. SAR-EM Settings

Logistic regression is used for both the classifier and the propensity score model because it is a simple classifier that is known to predict well-calibrated probabilities (Niculescu-Mizil and Caruana, 2005). All experiments have local certainty parameter d=0.9 and convergence parameters $n = 10$, $\epsilon = 0.0001$.

5.2. Performance under the SCAR Assumption

To evaluate how well SAR-EM does in the SCAR setting, the same datasets (Table 1) are used as in Bekker and Davis (2018a) to benchmark methods for label frequency estimation.

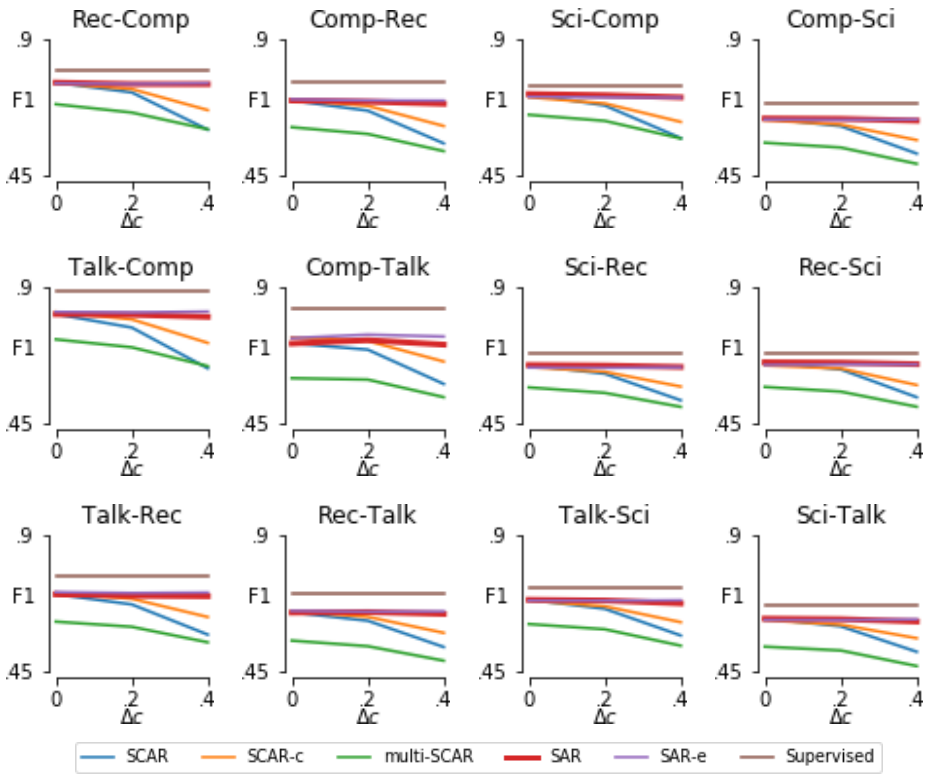

Figure 2: Propensity score with one variable, centered around $\bar{c} = 0.3$. Assuming SCAR ($\Delta c = 0$) hurts the performance when the reality deviates from it. Knowing the propensity function does not result in a big benefit over estimating it.

IJCNN originates from the IJCNN 2001 neural network competition.[1] The other dataset are available on the UCI repository.[2] The preprocessed datasets of Bekker and Davis (2018a) are used here, which means that the multivalued features are binarized and numerical features scaled between 0 and 1. To generate binary classification datasets from the twenty newsgroups (20ng), different categories are compared (computer, recreation, science and talk) and the features are generated using bag of words with the 200 most frequent words, disregarding nltk stopwords.

The following class prior estimation methods are compared: EN (Elkan and Noto, 2008), PE (du Plessis and Sugiyama, 2014), pen-L1 (du Plessis et al., 2015), KM1 and KM2 (Ramaswamy et al., 2016), AlphaMax (Jain et al., 2016b) and AlphaMax_N (Jain et al., 2016a). The datasets were turned into PU datasets in the same way as done by Bekker and Davis (2018a): the positive examples are selected to be labeled with label frequencies $c \in [0.1, 0.3, 0.5, 0.7, 0.9]$.

As usual, the algorithms based on their absolute error in class prior $|\hat{\alpha} - \alpha|$.

1. Available on: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
2. http://archive.ics.uci.edu/ml/

5.2.1. Results

SAR-EM performs very well at estimating the class prior, as can be seen in Figure 1 and Table 2. It has the best rank and lowest absolute error on average.

5.3. Performance under the SAR Assumption

The 20 newsgroups datasets are used to simulate SAR datasets. They always consider two categories to classify. With four categories there are six combinations and 12 datasets in total by switching which category is the positive class.

The attributes of 20 newsgroups are all binary: they indicate if a word appears in the article or not. To select propensity attributes that are sure to have an impact on the labeling, we only considered the attributes to have a frequency between 30% and 70% in the data, which leaves between five and eight attributes per dataset.

We consider two types of propensity scores. The first one only depends on one attribute. We analyze how deviating from the SCAR setting affects learners that assume SCAR. To this end, the difference between the two propensity scores Δc is varied with steps of 0.2, centering them around $\overline{c} = 0.3$ or $\overline{c} = 0.5$:

$$e(x_e) = x_e(\overline{c} \ +/- \ \Delta c/2) + (1 - x_e)(\overline{c} \ -/+ \ \Delta c/2)$$

The second type of propensity score is based on three attributes, where all attributes contribute independently: 0.9 for the attribute being 1 and 0.5 otherwise:

$$e(x_e) = 0.9^{\sum x_e} \cdot 0.5^{3-\sum x_e}$$

The labels were generated for each combination of attributes, at random according to the propensity scores functions.

The following methods are compared:

Supervised Logistic regression with all class labels available.
SAR SAR-EM as described in this paper.
SAR-e SAR-EM with access to the true propensity score (Section 4.6).
SCAR SAR-EM is used for the SCAR setting, as it outperformed the others (Section 5.2.1).
multi-SCAR Independent SCAR models for all propensity attributes configurations.
SCAR-c SAR-EM with access to a propensity score which is the true label frequency.

We compare the algorithms based on their classification F1 score and propensity score accuracy. The datasets were randomly divided into five folds where four were used for learning and one for evaluation. The folds are assigned in five different ways.

5.3.1. Results

As expected, the more the dataset deviates from the SCAR assumption, the worse the results get when this is assumed. This result supports the need for methods like SAR-EM that do not make this strong assumption (Figures 2 and 3).

Interestingly, learning with an unknown propensity score function performs almost equally well as with a known one, even when multiple attributes are involved (Table 3).

Knowing the label frequency for the SCAR assumption does help when the assumption does not hold. This is due to the label frequency always being an overestimate when the assumption does not hold as a result of the local certainty assumption.

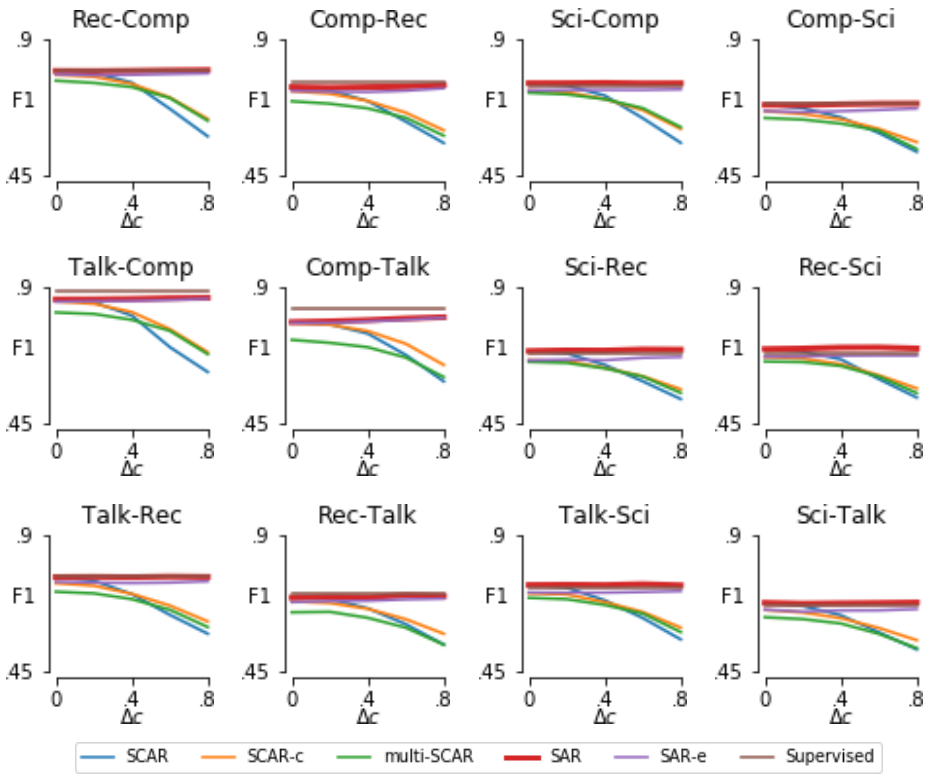

Figure 3: Propensity score with one variable, centered around $\bar{c} = 0.5$. Assuming SCAR ($\Delta c = 0$) hurts the performance when the reality deviates from it.

5.4. Effect of Unbalanced Classes

Unbalanced classes are simulated for the datasets with three propensity features. To this end, either the positive or negative class was subsampled to 30%, assuming completely balanced classes this results in class priors $\alpha = 0.23$ and $\alpha = 0.77$.

5.4.1. Results

Figure 4 compares the F1 scores of the trained classifiers for the different class priors. Handling unbalanced domains clearly becomes harder with PU data. Knowing the propensity score (SAR-e) or the class prior (SCAR-c) gives an advantage here.

6. Conclusions

This paper considers learning from positive and unlabeled data (PU learning) under the Selected At Random (SAR) assumption. That assumption states that the probability of selecting a positive example to be labeled depends on the attributes, in contrast to being constant, as commonly assumed in PU learning under the Selected Completely At Random

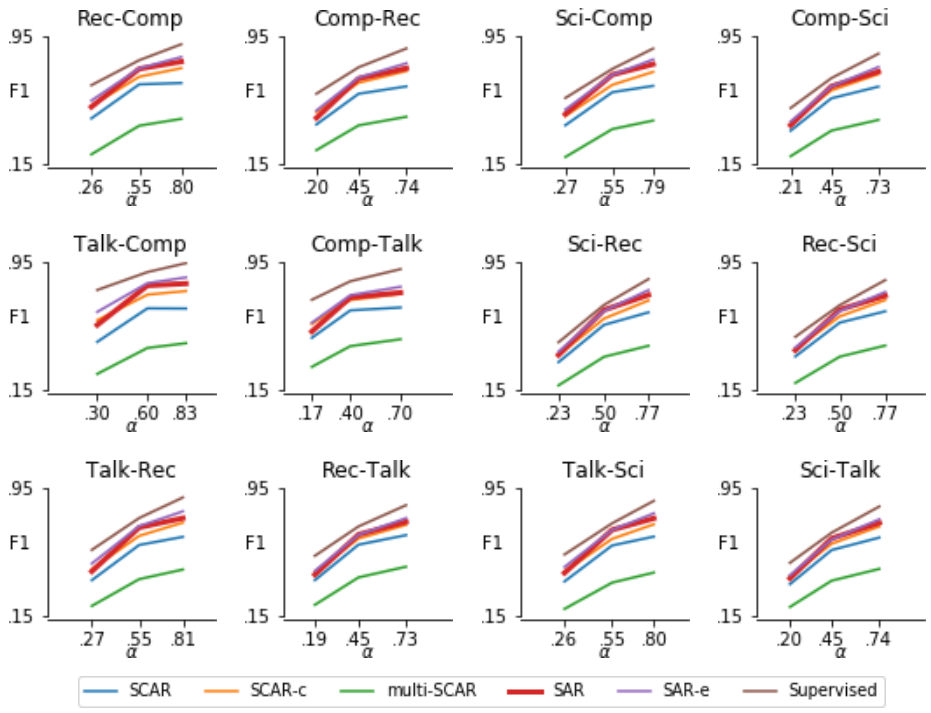

Figure 4: Unbalanced domains are relatively harder when learning from positive and unlabeled data than from fully supervised data. Knowing the true class prior (SCAR-c and SAR-e) gives an advantage.

Table 3: Classifier F1 score (left) and propensity (right) score accuracies. 'Supervised' has no propensity model. SAR-EM with the SAR assumption performs similar to learning with a known propensity score function. The SCAR assumption is clearly insufficient.

Data	Supervised	SAR-e		SAR		multi-SCAR		SCAR		SCAR-c	
Sci-Rec	0.68	0.64	0.65	**0.65**	**0.64**	0.36	0.29	0.56	0.54	0.60	0.59
Rec-Sci	0.68	0.64	0.64	**0.65**	**0.63**	0.36	0.30	0.57	0.53	0.61	0.58
Talk-Rec	0.76	0.71	0.65	**0.70**	**0.63**	0.38	0.29	0.59	0.53	0.65	0.59
Rec-Talk	0.71	0.66	0.62	**0.66**	**0.61**	0.39	0.34	0.60	0.51	0.63	0.55
Sci-Comp	0.74	0.71	0.67	**0.71**	**0.66**	0.37	0.25	0.60	0.58	0.65	0.63
Comp-Sci	0.69	0.64	0.64	**0.64**	**0.63**	0.36	0.30	0.56	0.52	0.61	0.58
Rec-Comp	0.80	0.75	0.67	**0.75**	**0.65**	0.39	0.25	0.65	0.57	0.70	0.62
Comp-Rec	0.75	0.69	0.64	**0.68**	**0.63**	0.39	0.30	0.59	0.51	0.66	0.58
Talk-Comp	0.89	0.82	0.67	**0.80**	**0.65**	0.41	0.25	0.66	0.55	0.75	0.63
Comp-Talk	0.83	0.74	0.63	**0.73**	**0.61**	0.43	0.33	0.65	0.50	0.71	0.56
Talk-Sci	0.73	0.69	0.65	**0.69**	**0.64**	0.36	0.28	0.59	0.55	0.63	0.60
Sci-Talk	0.67	0.63	0.63	**0.64**	**0.62**	0.37	0.33	0.56	0.51	0.60	0.56

(SCAR) assumption. The SCAR assumption is clearly often violated in practice and our experiments show that using it when it does not hold severely hurts the performance.

In this work, we analyzed common assumptions in PU learning and investigated how they can be used under the SAR assumption, by formulating the problem as a multiple problems under SCAR assumption. This results in a simple yet effective EM-based method. In our experiments, we show that this method is very promising, as it does virtually equally well as assuming that the labeling mechanism is known.

Future work will investigate if the method is still effective for more complex propensity score functions. More powerful classifiers could then be preferred, but they might need calibration of their output probabilities.

Acknowledgements

JB is supported by IWT (SB/141744). JD is partially supported by KU Leuven Research Fund (C14/17/070 and C22/15/015), FWO-Vlaanderen (SBO-150033, EOS No. 30992574) and Interreg V A project NANO4Sports.

References

J. Bekker and J. Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *AAAI*, 2018a.

J. Bekker and J. Davis. Positive and unlabeled relational classification through label frequency estimation. In *ILP 2017, Revised Selected Papers*, 2018b.

M. Claesen, F. De Smet, J. Suykens, and B. De Moor. A robust ensemble approach to learn from positive and unlabeled data using svm base models. *Neurocomputing*, 2015.

F. De Comité, F. Denis, R. Gilleron, and F. Letouzey. Positive and unlabeled examples help learning. In *ALT*, 1999.

F. Denis. Pac learning from positive statistical queries. In *ALT*, 1998.

F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *TCS*, 2005.

M.C. du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE Transactions*, 2014.

M.C. du Plessis, G. Niu, and M. Sugiyama. Class-prior estimation for learning from positive and unlabeled data. *ML*, 2015.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *SIGKDD*, 2008.

G. Imbens and D.B. Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

S. Jain, M. White, and P. Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *NIPS*, 2016a.

S. Jain, M. White, M. W. Trosset, and P. Radivojac. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016b.

S.S. Khan and M.G. Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 2014.

W. Sun Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, 2003.

X-L Li and B. Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, 2003.

X-L Li, P.S. Yu, B. Liu, and S-K Ng. Positive unlabeled learning for data stream classification. In *SDM*, 2009.

R. Little and D.B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

B. Liu, Y. Dai, X-L Li, W.S. Lee, and P.S. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, 2003.

Z. Liu, W. Shi, D. Li, and Q. Qin. Partially supervised classification–based on weighted unlabeled samples support vector machine. In *ADMA*, 2005.

F. Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 2014.

A. Neelakantan, B. Roth, and A. McCallum. Compositional vector space models for knowledge base completion. *ACL*, 2015.

M. Nhut Nguyen, X-L Li, and S-K Ng. Positive unlabeled leaning for time series classification. In *IJCAI*, 2011.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *ICML*, 2005.

H.G. Ramaswamy, C. Scott, and A. Tewari. Mixture proportion estimation via kernel embedding of distributions. In *ICML*, 2016.

D.B. Rubin. Inference and missing data. *Biometrika*, 1976.

T. Schnabel, A.Swaminathan, A.Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *ICML*, 2016.

H. Yu. Single-class classification with mapping convergence. *ML*, 2005.

D. Zhang and W.S. Lee. A simple probabilistic approach to learning from positive and unlabeled examples. In *UKCI*, pages 83–87, 2005.

K. Zupanc and J. Davis. Estimating rule quality for knowledge base completion with the relationship between coverage assumption. In *Proceedings of the Web Conference*, 2018.