Metadata of the chapter that will be visualized in SpringerLink

Book Title	Machine Learning, Optimization, and Data Science	
Series Title		
Chapter Title	Estimating the F_1 Score for Learning from Positive and Unlabeled Examples	
Copyright Year	2020	
Copyright HolderName	Springer Nature Switzerland AG	

Corresponding Author	Family Name	**Tabatabaei**
	Particle	
	Given Name	**Seyed Amin**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	Vrije Universiteit Amsterdam
	Address	Amsterdam, Netherlands
	Division	
	Organization	Elsevier B.V.
	Address	Amsterdam, Netherlands
	Email	s.tabatabaei@vu.nl
Author	Family Name	**Klein**
	Particle	
	Given Name	**Jan**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Centrum Wiskunde & Informatica
	Address	Amsterdam, Netherlands
	Email	j.g.klein@cwi.nl
Author	Family Name	**Hoogendoorn**
	Particle	
	Given Name	**Mark**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	Vrije Universiteit Amsterdam
	Address	Amsterdam, Netherlands
	Email	m.hoogendoorn@vu.nl

Abstract

Semi-supervised learning can be applied to datasets that contain both labeled and unlabeled instances and can result in more accurate predictions compared to fully supervised or unsupervised learning in case limited labeled data is available. A subclass of problems, called Positive-Unlabeled (PU) learning, focuses on cases in which the labeled instances contain only positive examples. Given the lack of negatively labeled data, estimating the general performance is difficult. In this paper, we propose a new approach to approximate the F_1 score for PU learning. It requires an estimate of what fraction of the total number of positive instances is available in the labeled set. We derive theoretical properties of the approach and apply it to several datasets to study its empirical behavior and to compare it to the most well-known score in the field, LL score. Results show that even when the estimate is quite off compared to the real fraction of positive labels the approximation of the F_1 score is significantly better compared with the LL score.

Estimating the F_1 Score for Learning from Positive and Unlabeled Examples

Seyed Amin Tabatabaei[1,2](✉), Jan Klein[3], and Mark Hoogendoorn[1]

[1] Department of Computer Science, Vrije Universiteit Amsterdam,
Amsterdam, Netherlands
{s.tabatabaei,m.hoogendoorn}@vu.nl
[2] Elsevier B.V., Amsterdam, Netherlands
[3] Centrum Wiskunde & Informatica, Amsterdam, Netherlands
j.g.klein@cwi.nl

Abstract. Semi-supervised learning can be applied to datasets that contain both labeled and unlabeled instances and can result in more accurate predictions compared to fully supervised or unsupervised learning in case limited labeled data is available. A subclass of problems, called Positive-Unlabeled (PU) learning, focuses on cases in which the labeled instances contain only positive examples. Given the lack of negatively labeled data, estimating the general performance is difficult. In this paper, we propose a new approach to approximate the F_1 score for PU learning. It requires an estimate of what fraction of the total number of positive instances is available in the labeled set. We derive theoretical properties of the approach and apply it to several datasets to study its empirical behavior and to compare it to the most well-known score in the field, LL score. Results show that even when the estimate is quite off compared to the real fraction of positive labels the approximation of the F_1 score is significantly better compared with the LL score.

1 Introduction

There has been a keen interest in algorithms that can learn a good classifier by using both labeled and unlabeled data. The field addressing such data is called semi-supervised learning (cf. [2]). Semi-supervised learning algorithms exploit the labeled data just like supervised learning algorithms do, but in addition take the structure seen in the unlabeled data into account to improve learning. Based on this combination, the algorithms are able to surpass the performance of fully supervised and unsupervised algorithms on partially labeled data (see e.g. [7]).

One category of problems in semi-supervised learning focuses on learning from datasets that only have positively labeled and unlabeled data, referred to as Positive-Unlabeled (PU) learning. PU learning is seen in multiple application domains (see e.g. [2,10,12]). The F_1 score is a prominent metric in classification problems in general, because taking both the precision and recall into account is desirable. This allows one to select the best model. However, in PU learning, since there are no negatively labeled examples available it is impossible to directly

G. Nicosia et al. (Eds.): LOD 2020, LNCS 12565, pp. 1–12, 2020.
https://doi.org/10.1007/978-3-030-64583-0_15

compute the F_1 score. Attempts to mitigate this problem have been proposed. For example, the LL score (cf. [5]) shows approximately the same behavior as the F_1 score without the need to have negatively labeled examples. However, in absolute terms, it can be quite off from the real F_1 score.

In this paper, we present a novel approach to estimate the F_1 score for a PU learning scenario. This estimator assumes an additional piece of information on top of the performance on the positively labeled data, namely an estimation of what fraction of labeled cases is available compared to the entire number of positive samples in the dataset. This assumption is in many cases not unrealistic and we show that even when the estimation is somewhat off, the proposed estimator still performs better than the popular LL score. We mathematically specify the approach and perform a mathematical analysis whereby we determine the sensitivity of the novel approach to mistakes in the estimation of the fraction of positive labels. On top of that, we conduct a number of experiments, both using generated and real life data. We compare the estimates of both the LL score and our newly introduced approach and show that the estimates using our approach are: (1) significantly closer to the true F_1 score, and (2) better at selecting the "best" model out of a set of models.

The rest of this paper is organized as follows. The formal problem description is given in Sect. 2. Related work is presented in Sect. 3, while our proposed approach is introduced in Sect. 4 together with the mathematical analysis of the approach. The experimental setup and accompanying results are described in Sect. 5 and 6 respectively. Finally, Sect. 7 concludes the paper.

2 Problem Formulation

Let us begin with formally specifying PU learning. Assume instances $i \in \mathcal{M}$ which are specified by their feature vector $\mathbf{x}_i \in \mathbb{R}^d$, corresponding label $y_i \in \{-1, 1\}$ and by the availability of the label $s_i \in \{0, 1\}$. Here, $\mathcal{M} := \{1, \ldots, M\}$ is the set of observations and d is the number of features. If, for an instance i, the label is available ($s_i = 1$), then it is always positive ($y_i = 1$). If the label is not available ($s_i = 0$), it can be either positive or negative. More specifically, let $\mathcal{P} \subseteq \mathcal{M}$ be the set of observations with a positive label. Let $\mathcal{S} \subseteq \mathcal{P}$ be the subset of observations for which the positive label is provided. Consequently, the labels of the observations in $\mathcal{U} := \mathcal{M} \backslash \mathcal{S}$ are not known.

An important assumption in most PU learning algorithms is that positive labeled instances are *Selected Completely At Random* among positive examples (SCAR assumption). This assumption lies at the basis of most PU learning algorithms [1]. Hence, $\mathcal{S}$ is a random subset of $\mathcal{P}$ under SCAR.

Using $\mathcal{S}$ and $\mathcal{U}$ we want to build a classifier f which can predict the label of the cases in $\mathcal{U}$, i.e. ideally $f(\mathbf{x}_i) = y_i$ for $i \in \mathcal{U}$. It should be stressed that, during the training and validation process, the final target of instances outside of $\mathcal{S}$ is not available. Therefore, learning should be done based on $\mathcal{S}$ combined with properties from the unlabeled data $\mathcal{U}$.

3 Literature Review

In this section, to provide an intuition of PU learning algorithms, we briefly introduce the commonly used two-step strategy. This is followed by metrics which estimate the performance of the resulting models.

3.1 PU Learning Algorithms: Two-Step Strategy

A well-known class of PU learning algorithms is the two-step strategy (cf. [7]). In step 1, a set of reliable negative instances is chosen from the unlabeled instances $\mathcal{U}$. It divides $\mathcal{U}$ into two sets: $\mathcal{N}_R$ and $\mathcal{U}\backslash\mathcal{N}_R$. In step 2, the algorithm iteratively adds more instances to $\mathcal{N}_R$, which are used as negative examples in the next iterations. This procedure is repeated until a convergence criterion is met or when no more instances are added to $\mathcal{N}_R$. There are several techniques for each of these steps. For example, the spy technique [8] and the Ricchio technique [6] are used for the first step. The EM algorithm [8] can be a natural choice for the second step. A deeper review about two-step techniques can be found in [7].

3.2 Performance Estimation

To select the classifier with the best generalizable performance, some evaluation is needed. In normal supervised learning, the F_1 score is a common performance measurement for binary classifiers. It is expressed as follows:

$$F_1 = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}},$$

with

$$\begin{aligned}\text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\sum_{i \in \mathcal{M}} \mathbf{1}_{\{f(\mathbf{x}_i)=1, y_i=1\}}(i)}{\sum_{i \in \mathcal{M}} \mathbf{1}_{\{y_i=1\}}(i)} = \frac{P_1}{P} \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{P_1}{\sum_{i \in \mathcal{M}} \mathbf{1}_{\{f(\mathbf{x}_i)=1\}}(i)} = \frac{P_1}{M_1}.\end{aligned}$$

Here, $\mathbf{1}_{\{\cdot\}}$ represents the indicator function. Moreover, $P := |\mathcal{P}|$ is the number of positive instances and P_1 is the number of positive instances which are also predicted to be 1, i.e. the number of true positives (TP). M_1 is the total number of observations which are predicted as positive.

In PU learning, the target label y_i is not available for unlabeled instances (with $s_i = 0$). Therefore, calculating the precision and recall is not directly possible. However, under the SCAR assumption, we expect the fraction of predicted positives in $\mathcal{S}$ to be the same as the fraction of predicted positives in $\mathcal{P}$:

$$\mathbf{E}\left(\frac{S_1}{S}\right) \overset{SCAR}{=} \frac{P_1}{P},$$

with S_1 the number of predicted positives in $\mathcal{S}$ and $S := |\mathcal{S}|$. Because of SCAR, the behavior of the classifier on $\mathcal{S}$ represents its behavior on $\mathcal{P}$. Hence, the recall can be estimated by

$$\overline{\text{rec}} = \frac{\sum_{i \in \mathcal{S}} \mathbf{1}_{\{f(\mathbf{x}_i)=1\}}(i)}{S} = \frac{S_1}{S}.$$

However, it is difficult to approximate the value of the precision, because it is less straightforward to obtain an estimate of P_1/M_1. This also means it is hard to estimate the F_1 score in PU learning. To solve this, multiple approaches exist, of which the LL score is commonly used. This score is given by

$$\text{LL} = \frac{\overline{\text{rec}}^2}{M_1/M} = \frac{S_1^2 \cdot M}{S^2 \cdot M_1}.$$

It can be directly calculated from a validation set, which contains positive and unlabeled examples. Moreover, it is shown that

$$\frac{\text{recall}^2}{M_1/M} = \frac{P_1^2 \cdot M}{P^2 \cdot M_1} = \frac{(P_1/M_1) \cdot (P_1/P)}{P/M} = \frac{\text{precision} \cdot \text{recall}}{P/M}.$$

Therefore, the LL score also has an estimation of the precision in its definition. It is claimed that the LL score has roughly the same behavior as the F_1 score: a high value of the LL score means both precision and recall are high, while a low value means that either recall or precision is low [5].

4 Estimating the F_1 Score

In this section, we present our approach to estimate the F_1 score in a PU learning problem. It is based on the assumption that we have an approximation of the fraction of positive instances that are labeled. Moreover, we analyze our approach mathematically.

4.1 Approach to Estimate F_1-score

First, we show how the precision can be estimated with the fraction ρ, defined as $\rho := S/P$. Under SCAR, $\mathbf{E}(S_1/\rho) \overset{SCAR}{=} P_1$, which yields

$$\overline{\text{prec}} = \frac{(1/\rho) \sum_{i \in \mathcal{S}} \mathbf{1}_{\{f(\mathbf{x}_i)=1\}}(i)}{\sum_{i \in \mathcal{M}} \mathbf{1}_{\{f(\mathbf{x}_i)=1\}}(i)} = \frac{S_1}{\rho \cdot M_1}.$$

The F_1 score can now be estimated by

$$\overline{F_1} := 2 \cdot \frac{\overline{\text{rec}} \cdot \overline{\text{prec}}}{\overline{\text{rec}} + \overline{\text{prec}}} = 2 \cdot \frac{(S_1/S) \cdot (S_1/(\rho \cdot M_1))}{(S_1/S) + (S_1/(\rho \cdot M_1))} = 2 \cdot \frac{S_1}{\rho \cdot M_1 + S},$$

while the actual F_1 score is given by

$$F_1 = 2 \cdot \frac{\rho \cdot P_1}{\rho \cdot M_1 + S}.$$

We are interested in how the approximated $\overline{F_1}$ differs from the actual F_1 score given a dataset and trained classifier f. Hence, we define the variable ΔF_1 as:

$$\Delta F_1 := \overline{F_1} - F_1 = 2 \cdot \frac{S_1 - \rho \cdot P_1}{\rho \cdot M_1 + S}.$$

The actual F_1 score is fixed, but $\overline{F}_1$ depends on which subset of $\mathcal{P}$ is chosen to be labeled. The number of predicted positive observations S_1 has a hypergeometric distribution [9] with P the population size, P_1 the number of 'success states' in the population and S the number of draws. An observation is 'successful' in this context if it is a true positive. Thus, $S_1 \sim \text{Hypergeometric}(P, P_1, S)$. Then,

$$\mathbf{E}(S_1) = S \cdot \frac{P_1}{P} = \rho P_1, \mathbf{Var}(S_1) = S \cdot \frac{P_1(P-P_1)(P-S)}{P^2(P-1)} = \frac{\rho(1-\rho)P_1(S-\rho P_1)}{S-\rho}.$$

We have for the approximated recall, precision and F_1 score:

$$\begin{aligned}
\mathbf{E}(\overline{\text{rec}}) &= \frac{\rho \cdot P_1}{S} = \text{recall}, \quad &\mathbf{Var}(\overline{\text{rec}}) &= \frac{\mathbf{Var}(S_1)}{S^2},\\
\mathbf{E}(\overline{\text{prec}}) &= \frac{\rho \cdot P_1}{\rho \cdot M_1} = \text{precision}, \quad &\mathbf{Var}(\overline{\text{prec}}) &= \frac{\mathbf{Var}(S_1)}{\rho^2 M_1^2}\\
\mathbf{E}(\overline{F_1}) &= 2 \cdot \frac{\rho \cdot P_1}{\rho \cdot M_1 + S} = F_1, \quad &\mathbf{Var}(\overline{F_1}) &= \frac{4 \cdot \mathbf{Var}(S_1)}{(\rho M_1 + S)^2}.
\end{aligned}$$

Since the expected values of the estimators are equal to the actual performance metrics, the estimators are unbiased.

4.2 Estimating ρ

Since $\rho = S/P$, and the size S of $\mathcal{S}$ is given, estimating P means estimating ρ. There are different approaches to estimate this value. We will not elaborate on this for the sake of brevity, but known approaches exploit domain knowledge or prior experiences with similar datasets. Or they use a classifier to make this estimate. In the remainder of the paper we use a classifier-based approach following [4] and also evaluate how well it works for real life cases.

4.3 Behavior Under Noisy ρ

Now, we analyse what the theoretical implications of a noisy ρ are on our estimators of the recall, precision and F_1 score. We assume that we do not know the real value of $\rho \in (0, 1]$. Let the random variable $\overline{\rho}$ indicate an estimator of ρ. In this case, ρ represents the probability that a positive observation is labeled. Since our estimator of the recall does not involve ρ, the distribution of $\overline{\text{rec}}$ remains the same when ρ is replaced by $\overline{\rho}$. However, the estimator of the precision does change:

$$\overline{\text{prec}}_{\overline{\rho}} = \frac{S_1}{\overline{\rho} \cdot M_1}. \quad \text{Consequently,} \quad \mathbf{E}(\overline{\text{prec}}_{\overline{\rho}}) = \frac{\mathbf{E}(S_1)}{M_1} \cdot \mathbf{E}\left(\frac{1}{\overline{\rho}}\right) = \frac{\rho P_1}{M_1} \cdot \mathbf{E}\left(\frac{1}{\overline{\rho}}\right)$$

$$\begin{aligned}\mathbf{Var}(\overline{\mathrm{prec}}_{\overline{\rho}}) &= \frac{1}{M_1^2}\left[\mathbf{E}(S_1^2)\cdot\mathbf{E}\left(\frac{1}{\overline{\rho}^2}\right) - \mathbf{E}(S_1)^2\cdot\mathbf{E}\left(\frac{1}{\overline{\rho}}\right)^2\right] \\ &= \frac{\mathbf{Var}(S_1)\cdot\mathbf{E}\left(\frac{1}{\overline{\rho}^2}\right) + \rho^2 P_1^2\cdot\mathbf{Var}\left(\frac{1}{\overline{\rho}}\right)}{M_1^2}.\end{aligned}$$

This means $\overline{\mathrm{prec}}_{\overline{\rho}}$ is an unbiased estimator only when $\mathbf{E}(1/\overline{\rho}) = 1/\rho$, which in general is not true. More specifically, consider the convex function $\varphi : (0,1] \to [1,\infty)$ given by $\varphi(x) = 1/x$. Hence, by Jensen's inequality, $\varphi(\mathbf{E}(X)) \leq \mathbf{E}(\varphi(X))$ for random variable X and convex function φ. Thus, $1/\rho \leq \mathbf{E}(1/\overline{\rho})$, and so

$$\mathbf{E}(\overline{\mathrm{prec}}_{\overline{\rho}}) = \frac{\rho P_1}{M_1}\cdot\mathbf{E}\left(\frac{1}{\overline{\rho}}\right) \geq \frac{P_1}{M_1} = \text{precision}.$$

The approximated F_1 score with noisy ρ is given by

$$\overline{F_1}_{\overline{\rho}} := 2\cdot\frac{S_1}{\overline{\rho}\cdot M_1 + S}.$$

The expected value of this estimator is at least equal to the actual F_1 score, which means it is biased. We show this again using Jensen's inequality and the convex function $\varphi : (0,1] \to (\frac{1}{M_1+S}, \frac{1}{S}]$ given by $\varphi(x) = \frac{1}{M_1\cdot x+S}$. Now,

$$\begin{aligned}\mathbf{E}(\overline{F_1}_{\overline{\rho}}) &= 2\mathbf{E}(S_1)\cdot\mathbf{E}\left(\frac{1}{\overline{\rho}\cdot M_1 + S}\right) \\ &\geq 2\mathbf{E}(S_1)\cdot\frac{1}{\mathbf{E}(\overline{\rho})\cdot M_1 + S} = 2\cdot\rho\cdot P_1\frac{1}{\rho\cdot M_1 + S} = F_1.\end{aligned}$$

Consequently, when the fraction of labeled observations among the positive instances is deemed stochastic with an arbitrary distribution, then both the estimators of the precision and F_1 score are expected to overestimate.

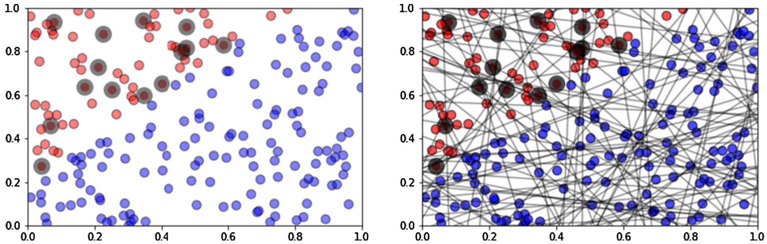
5 Experimental Setup

In order to evaluate our approach we empirically compare it to the real F_1 score and to the behavior of the LL score on four different datasets [1]. For these datasets the ground truth for all instances is available.

5.1 Datasets and Setup

Generated Dataset. The first dataset, or actually set of datasets, is generated randomly. These datasets contain two features $X_1, X_2 \in [0,1]$ and points are generated uniformly random. In order to assign the points to one of the two classes ($y = 0$ or $y = 1$), their position is compared to a randomly generated line

[1] All code is available on Github: https://github.com/SEYED7037/PU-Learning-Estimating-F1-LOD2020-.



(a) Example of generated dataset.

(b) Randomly generated linear classifiers

Fig. 1. Example of a generated dataset with accompanying classifiers. Positive labels are red and negatives are blue. Labeled samples are highlighted with a gray circle. (Color figure online)

$(X_1 - X_2 + 0.2 = 0)$. The classes are assigned to the points based on their position compared to this line. We then select a random sample (SCAR assumption) of size $\rho \cdot P$ of the positive examples ($y = 1$) to act as $\mathcal{S}$ (with value $s = 1$) and take the rest as $\mathcal{U}$ (with $s = 0$). Figure 1a shows a randomly generated dataset, and Fig. 1b shows a set of linear classifiers which are generated randomly.

IRIS Dataset is a popular dataset in pattern recognition and machine learning literature [3]. It contains 3 flower classes (*setosa*, *versicolor* and *virginica*) of 50 instances each. There are 4 features available for each instance. By taking the last class (*virginica*) as positive and the two others as negative, it transfers to a binary classification problem. These two classes are not linearly separable. We again made a random fraction ρ of all positively labeled data points available as $\mathcal{S}$, the rest being $\mathcal{U}$. 4-D hyper planes were generated randomly to act as linear classifiers.

Heart Disease Dataset is well-known in pattern recognition literature [3]. The data contains both numerical and categorical features. The goal of models applied to this dataset is to predict the presence of heart disease in a patient. In our experiments, we trained random forest models with different numbers of estimators $(randint(1, 100))$ and maximum depth $(randint(1, 10))$.

Health Dataset was obtained from the VU University Medical Center and contains event logs of more than 300,000 patients. For more information about this dataset, please take a look at [10]. The goal is to identify certain types of patients based on their event log. Part of these patients are labeled as having kidney disease, others are labeled as diabetes, and the rest have another disease. For each disease, a fraction ρ of positive examples were randomly selected as labeled examples $\mathcal{S}$, while the rest were taken as unlabeled examples $\mathcal{U}$. Following [10] two features are present in the dataset to predict the label, namely $X_1, X_2 \in \mathbb{Z}$ which summarize the care paths of patients in a way that patients with that disease are optimally separable. A classifier is defined by a set of two thresholds, (θ_1, θ_2). An instance will be predicted as positive if $X_1 > \theta_1$ and $X_2 > \theta_2$.

5.2 Experimental Conditions and Performance Metrics

We compared our approach to the LL score based on two metrics: (1) distance to the real F_1 score; and (2) percentage of inversions. We compute the RMSE to measure the distance to the F_1 score. Computing the percentage of inversions, which is the key in showing that the right model was selected and thus our most important outcome, is a bit more difficult. The inversions were used to show how often the wrong model was selected based on either $\overline{F_1}$ or the LL score compared to the actual F_1 score. To this end, we took the different classifiers for each dataset and compared them pairwise. Each time a classifier that has a higher F_1 score compared to the other classifier is ranked lower we call this an *inversion*. Hence, we want to minimize the number of inversions. We compared the results using a Wilcoxon paired test to show possibly significant differences.

We conducted three types of experiments, namely: (1) empirically studying the assumptions and theoretical results of our approach; (2) evaluating the performance of the approach with the true value of ρ being available; and (3) evaluating the performance with noisy ρ. Each is explained in more detail below.

Empirical Evaluation of Assumptions. As has become clear, we make the assumption that ρ can be estimated. We have presented various approaches to estimate ρ, one of which involves a classifier g. This estimator is exactly correct if $g(x) = \Pr(s = 1|\mathbf{x})$ for all $\mathbf{x}$, but usually this condition does not hold in practice. To show the applicability of this technique (and how easily we can obtain the crucial ρ) we used the *IRIS data* and the *generated data* with different values of ρ and estimated the value of ρ using a trained classifier on the labeled data points. We conducted this experiment 100 times per value of ρ. To get more accurate results, we used the one-leave-out cross-validation technique.

Secondly, we evaluated another part of our approach, namely our result on the bounds. In this experiment, we again used the *generated* and *IRIS datasets* and took one randomly selected linear classifier with a real F_1 score of 0.44. We then took different values for ρ and for each value drew a random sample 100 times, thereby estimating the F_1 score using our approach. We used these results to compute the mean estimated value and the confidence bounds. We compared these to the bound following our mathematical result.

Performance Evaluation with Correct ρ. To evaluate the approach compared to the LL score, we first assume ρ to be known and correct. For these experiments, we selected the value of ρ ranging from 0 to 1 with increments of 0.01. For each setting of ρ for the *generated data* we generated 200 datasets and 100 random lines to act as classifiers. For the *IRIS* and the *health* dataset we generated 100 random classifiers. We measured the performance for both the deviation of the F_1 score and number of inversions.

Performance Evaluation with Noisy ρ. For the noisy ρ we varied the noise level and use the same experimental setup as presented under the *correct* ρ case. The noise level was varied from a 50% underestimation to a 100% overestimation. Due to the computational complexity, we only studied this part on the *generated data* and *IRIS* dataset and measure the percentage of inversions. Table 1 gives a brief overview of the datasets used for the various experiments.

Table 1. Datasets used for the various experiments.

Experiment	Generated data	Iris	Heart Disease	Health
Estimating ρ	X	X		
Evaluating bounds	X	X		
Perf. Eval. Correct ρ - F_1	X	X	X	X
Perf. Eval. Correct ρ - Inversions	X	X	X	X
Perf. Eval. Noisy ρ - Inversions	X	X		

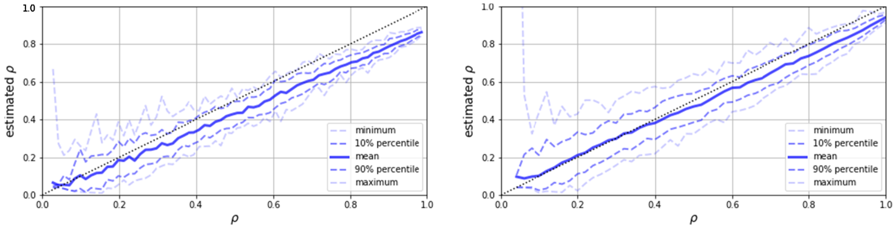
6 Results

First, we report the results of the empirical evaluation of the assumptions followed by experiments in which the correct value of ρ was known. Then, we explore the cases where the value of ρ was noisy (either under- or overestimated).

6.1 Checking the Assumptions

Figure 2 shows the results on the estimation of ρ through our classifier including confidence bounds. We see that as ρ increases the variability of the estimation decreases, which makes sense as a small sample will make the estimation very sensitive to the sample drawn. However, it can be seen that estimations are very reasonable. We do not observe any obvious difference in the estimation behavior between the *generated dataset* (Fig. 2a) and the *IRIS dataset* (Fig. 2b).

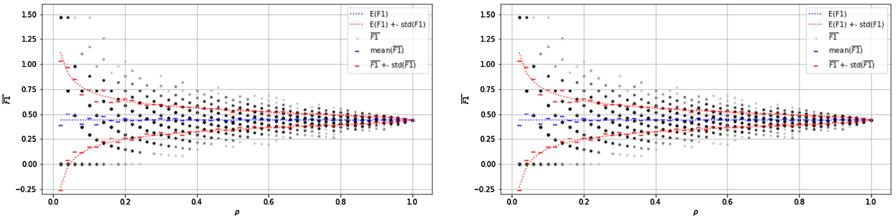
Our second study about the underlying assumptions concerns our estimation of the bounds. Figure 3 shows the empirical results for various values of ρ, the empirical mean and bounds and the computed mean and bounds based on our mathematical results for both the *generated* and *IRIS datasets*. Results show that the two align very well for both datasets.



(a) Generated Dataset

(b) IRIS Dataset

Fig. 2. Estimating ρ using a classifier (see [4]). For each value of ρ, this experiment is conducted 50 times. Mean, minimum, maximum, 10 and 90 percentiles are reported.



(a) Generated Dataset

(b) Iris Dataset

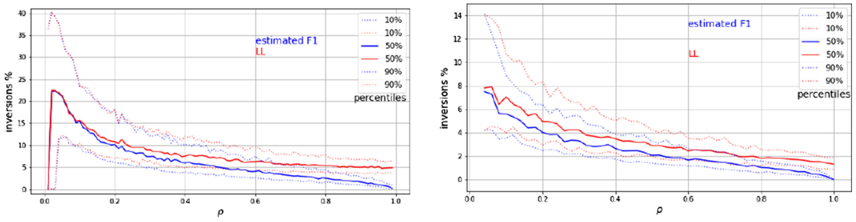
Fig. 3. Expected value and standard deviation of estimated $\overline{F_1}$ for different values of ρ. Each gray point shows $\overline{F_1}$ for one set of labeled data points. Points which overlap become darker. The empirical mean and bounds of $\overline{F_1}$ are shown by the blue and red dashed line respectively while the mean and bounds computed based on our mathematical result are shown by blue and red dashes respectively.

Table 2. RMSE of real F_1 vs. $\overline{F_1}$ and F_1 vs. LL score. For all cases, $\rho = 0.30$.

	Generated	IRIS	Heart disease	Health
F1	0.064	0.060	0.089	0.060
LL-score	0.772	0.420	0.344	0.623

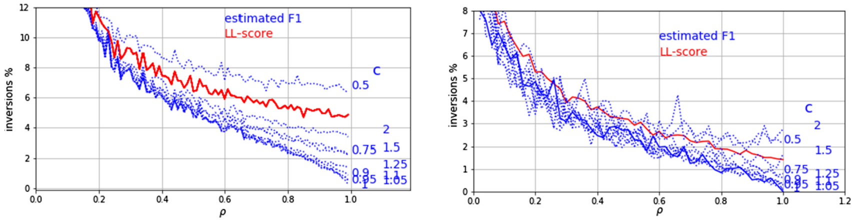
6.2 Correct ρ

Let us move on to measuring the performance of our approach. We start by considering the case in which ρ was equal to the true value. Table 2 reports the RMSE for different datasets. The RMSE for our approach is much smaller compared to the LL score. This was also to be expected as the proposed score is an estimation of F1, while the LL score aims to approximate the behavior of the F_1 score and not necessarily its actual value. Most important to observe (as our aim is model selection and hyperparameter optimization) is that our estimated values are monotonically increasing with the true F_1 score. Therefore, our central metric is the number of inversions when performing model selection. Figure 4.a shows the results for the *generated dataset* for varying values of ρ. We see that as ρ increases the difference in performance between our approach and the LL score increases in favor of the approach we put forward. Also the confidence intervals become smaller as ρ increases. We also see that our approach never performs worse. Results of a paired Wilcoxon signed-rank test [11] show that for values of $\rho > 0.05$ the number of inversions caused by sorting classifiers based on the $\overline{F_1}$ score is significantly lower than those by the LL score. Moving on to the *IRIS dataset*, Fig. 4.b shows the average number of inversions for different values of ρ. Our approach is significantly better when $\rho > 0.02$. For the *heart disease dataset*, the Wilcoxon paired test shows a significant better performance for our approach for $\rho > 0.02$. Finally, for the real-life *health datasets* our approach is significantly better when $\rho > 0.08$ for kidney disorder and $\rho > 0.02$ for diabetes.



(a) the artificially generated data (b) The Iris [3] data

Fig. 4. Number of inversions of both the LL score (red line) and the proposed $\overline{F_1}$ (blue line) including confidence bounds for different values of ρ.



(a) Generated Dataset (b) IRIS Dataset

Fig. 5. Effect of error in estimated ρ on the percentage of inversions.

6.3 Noisy ρ

In many cases, we do not know the exact value of ρ and might only be able to estimate it (see our first set of experiments). Figure 5 shows how under- and overestimations influence the number of inversions of our proposed approach for the *generated data* and the *IRIS dataset*. Here, the true value of ρ is multiplied with a value c. When considering the *generated dataset*, we see that only for a value of $c = 0.5$, i.e. an extreme underestimation of ρ, the proposed approach scores worse compared to the LL score. For the *IRIS dataset*, we see a similar pattern, except that also for a value of $c = 2$, i.e. a severe underestimation, our performance is worse. This shows that suffering from a bit of noise does not hamper our approach.

7 Conclusion

In this paper we have introduced a novel way of estimating the F_1 score to enable model selection and hyperparameter tuning in PU learning. This novel method is based on the assumption that an estimation can be made on the fraction of labeled positive cases. A mathematical analysis was performed to show the expected value of the estimation with respect to the real F_1 score. Also, we analyzed what the influence of stochasticity in ρ is on the estimations.

We showed that the estimators become biased when ρ is noisy, while they are unbiased when there is no noise.

Furthermore, we conducted experiments to evaluate our assumptions empirically, showing that the approach is practically applicable. On top, we have empirically compared our proposed approach to a well-known metric for model selection, namely the LL score. Results show that our approach (1) is closer to the true F_1 score, and (2) has fewer wrong selections of models (i.e. inversions) compared to the LL score for a variety of datasets. Both cases only hold for sufficiently large samples of training data, though the approach never performs worse. When considering wrongly estimating the fraction of positive labels we see that only severe underestimations hamper performance compared to the LL score. Our approach also brings advantages that the whole family of F scores can now be estimated.

References

1. Bekker, J., Davis, J.: Learning from positive and unlabeled data: A survey. arXiv preprint arXiv:1811.04820 (2018)
2. Denis, F., Gilleron, R., Tommasi, M.: Text classification from positive and unlabeled examples (2002)
3. Dua, D., Graff, C.: UCI machine learning repository (2017) http://archive.ics.uci.edu/ml
4. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 213–220. ACM (2008)
5. Lee, W.S., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. ICML. **3**, 448–455 (2003)
6. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. IJCAI. **3**, 587–592 (2003)
7. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, Berlin (2007)
8. Liu, B., Lee, W.S., Yu, P.S., Li, X.: Partially supervised classification of text documents. In: ICML. vol. 2, pp. 387–394. Citeseer (2002)
9. Skala, M.: Hypergeometric tail inequalities: ending the insanity. arXiv preprint arXiv:1311.5939 (2013)
10. Tabatabaei, S.A., Lu, X., Hoogendoorn, M., Reijers, H.A.: Identifying patient groups based on frequent patterns of patient samples. arXiv preprint arXiv:1904.01863 (2019)
11. Wilcoxon, F.: Some rapid approximate statistical procedures. Ann. New York Acad. Sci. **52**(1), 808–814 (1950)
12. Zhao, Y., Kong, X., Philip, S.Y.: Positive and unlabeled learning for graph classification. In: 2011 IEEE 11th International Conference on Data Mining. pp. 962–971. IEEE (2011)

Author Queries

Chapter 15

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city and country names in affiliation. Please check and confirm if the inserted city and country names is correct. If not, please provide us with the correct city and country names.	