Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 225–233
DOI: 10.15439/2023F1402
ISSN 2300-5963 ACSIS, Vol. 35

# Enhancing naive classifier for positive unlabeled data based on logistic regression approach

Mateusz Płatek
Warsaw University of Technology
Faculty of Mathematics and Information Science
Koszykowa 75, 00-662 Warsaw, Poland
Email: mateusz.platek.poczta@gmail.com

Jan Mielniczuk[0000−0003−2621−2303]
Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warsaw, Poland
and
Warsaw University of Technology
Faculty of Mathematics and Information Science
Koszykowa 75, 00-662 Warsaw, Poland
Email: jan.mielniczuk@ipipan.waw.pl

***Abstract*—It is argued that for analysis of Positive Unlabeled (PU) data under Selected Completely At Random (SCAR) assumption it is fruitful to view the problem as fitting of misspecified model to the data. Namely, it is shown that the results on misspecified fit imply that in the case when posterior probability of the response is modelled by logistic regression, fitting the logistic regression to the observable PU data which *does not* follow this model, still yields the vector of estimated parameters approximately colinear with the true vector of parameters. This observation together with choosing the intercept of the classifier based on optimisation of analogue of F1 measure yields a classifier which performs on par or better than its competitors on several real data sets considered.**

## I. Introduction

In the paper classification problem is analysed for partially observable data scenario for which in the case of some observations class indicators assigned to them (positive or negative in the case of binary classification) are unknown. More specifically, for positive and unlabeled data considered here, it is assumed that some observations from the positive class are labeled, whereas the rest of the observations (either positive or negative) are unlabeled. Such scenario is called Positive Unlabelled (PU) scenario. Thus in the PU setting the true binary class indicator $Y \in \{0, 1\}$ is not observed directly but only through binary label $S$. One knows that if $S = 1$ (labelled case), $Y$ has to be 1 (positive), but for $S = 0$ (unlabeled case) $Y$ may be either 1 or 0 (positive or negative). Besides, each object is described by the vector of features $x$. This setup encompasses a legion of practical situations, in which effective inference methods about class indicator $Y$ are sought. Examples include disease data (diagnosed patients with a specific disease detected, and patients yet to be diagnosed who may be ill or not), web pages preferences of a specific user (pages bookmarked as of interest and pages not yet viewed, thus of unknown interest) and ecological examples when environments are labeled provided a specific specimen inhabits them, and unlabeled, where this specimen has not been yet looked for). Such scenario is also relevant for survey data, when questions concerning socially reproachable behaviour may not be answered truthfully.

One of the popular approaches to learn from PU data is to impose certain parametric assumptions on distribution of $(X, Y)$ as it commonly done in classical classification task together with some assumptions on labeling mechanism $S$. This is partly necessitated by the fact that in general situation the posterior distribution of $Y$ as well as prior probability $P(Y = 1)$ is not identifiable. It is thus common to consider logistic type of dependence for the posterior distribution $P(Y = 1|X = x)$ and assume that censoring mechanism acts indiscriminately of $x$ and is described only by the label frequency $c = P(S = 1|Y = 1)$ (SCAR assumption discussed below). Majority of learning approaches has been developed under such assumptions; see [1] for an extensive review of the proposed methods. Recently the JOINT method has been proposed in [11] which consists in minimisation of empirical risk for the observed data $(X_i, S_i), i = 1, \ldots, n$ with respect to parameter of logistic distribution *and* label frequency. JOINT method can be considered as a generic method with specific algorithms depending on optimisation technique used. The issue is delicate as it turns out that the empirical risk is *not* a convex function of its parameters and thus it may posess multiple local minima. In particular [11] used BFGS algorithm, whereas approach in [6] has been based on Minorization-Maximization (MM) technique. Among other methods important group consists of approaches based on weighted empirical risk minimisation in which weights of observations depend on labeling frequency $c$ (see [1], section 5.3.2).

In the present contribution attention is called to the fact that in order to construct a reasonable classifier one can use a logistic model fitted to observable data $(X_i, S_i), i = 1, \ldots, n$ in order to recover the direction of the separating hyperplane and then shift it to the optimal position by maximising observable analogue of $F1$ score. In this approach the direction is obtained by minimising the misspecified *convex* empirical risk (equal to minus log-likelihood) for the observed data. The justification of the method is based on properties of misspecified logistic regression which are valid for PU model under SCAR condition considered here. It is argued that considering

fitting parametric models to PU data as the misspecification problem gives new insights to the established properties and leads to new solutions. In particular, results on behaviour of estimators under misspecification (see e.g. [14], [12]) can be used to assess the performance of the naive classifier and its modifications.

## II. Notions and auxiliary results

We first introduce basic notations. Let $X$ be a multivariate random variable corresponding to feature vector, $Y \in \{0,1\}$ be a true class label and $S \in \{0,1\}$ an indicator of an example being labeled ($S=1$) or not ($S=0$). We consider $X$ as a column vector and let $X = (1, \tilde{X}^T)^T \in R^{p+1}$, where the first coordinate of $X$ corresponds to an intercept and coordinates of $\tilde{X}$ relate to $p$ collected characteristics of an observation. We assume that there is some unknown distribution $P_{Y,X,S}$ such that $(Y_i, X_i, S_i)$, $i = 1, \ldots, n$ are independent observations drawn from this distribution. Observed data consists of $(X_i, S_i)$, $i = 1, \ldots, n$. This is the single sample scenario as opposed to case-control scenario when the samples from positive class and the general population are given. Only positive examples ($Y=1$) can be labeled, i.e. $P(S=1|X, Y=0) = 0$. Thus we know that $Y=1$ when $S=1$ but when $S=0$, $Y$ can be either 1 or 0. Our primary aim is to construct a classifier which predicts $Y$ class based on PU data. Note that this corresponds to a specific censored data problem as we only observe samples from distribution of $(X, S)$, where $S = Y$ with a certain probability.

To this end we define binary posterior probability of $S=1$ given $X=x$ equal $s(x) = P(S=1|x)$ and propensity score function $e(x) = P(S=1|Y=1, X=x)$. In this paper we adopt Selected Completely At Random (SCAR) assumption which stipulates that $e(x)$ does not depend on $x$, thus $e(x) = P(S=1|Y=1) := c$, where $c$ will stand for labeling frequency. This means that labeling is not influenced by feature vector $x$ and in this case labeled data is a random sample (of a random size) from a positive class. This commonly adopted assumption is restrictive but it serves as an useful approximation especially in situations when the possibility of labeling bias is recognised and one tries to avoid it. We note that as we have $P(S=1, Y=0|X=x) = 0$ it holds

$$\begin{aligned} s(x) &= P(S=1|x) = P(S=1, Y=1|x) \\ &= P(S=1|Y=1, x)P(Y=1|x) \\ &= e(x) \times y(x) = c \times y(x), \end{aligned} \tag{1}$$

where we let $y(x) = P(Y=1|X=x)$ denote posterior probability of class 1 and the last equality follows from SCAR assumption. We note that SCAR is equivalent to the property that $S$ and $X$ are conditionally independent given $Y$. We stress, however, that it is valid only when the label value is assigned with a fixed probability regardless of characteristics of an item. Under this assumption it is easy to see that $P_{X|S=1} = P_{X|Y=1}$ whereas $P_{X|S=0}$ is a mixture

$$P_{X|S=0} = \frac{\alpha - \alpha c}{1 - \alpha c} P_{X|Y=1} + \frac{1-\alpha}{1-\alpha c} P_{X|Y=0}$$

and $\alpha = P(Y=1)$ is a prior probability of $Y=1$. We also note that $c = P(S=1|Y=1) = P(S=1)/P(Y=1) = P(S=1)/\alpha$. We do not assume any previous knowledge of $c$ (although it is frequently imposed see, e.g. [1]) and thus we only know that $0 < c \leq 1$.We will adopt an parametric model for posterior probability $y(x)$ assuming that $Y$ is governed by logistic response:

$$y(x) = \frac{\exp\big(x^T\beta\big)}{1 + \exp(x^T\beta)} = \sigma(x^T\beta), \tag{2}$$

where $\sigma(s) = \exp(s)/(1+\exp(s))$ is a logistic function, $\beta^T$ stands for transposed column vector $\beta$ and $\beta = (\beta_0, \beta_{-0}^T)^T \in R \times R^p$ is an unknown but fixed vector value. Thus in view of (1) and (2) we have

$$P(S=1|x) = c \times \sigma(x^T\beta).$$

## III. Misspecified logistic modelling

Assume that (2) holds and consider naive approach when the logistic model is fitted to $(X, S)$ data using Maximum Likelihood method i.e. we maximise a log-likelihood

$$\mathcal{L}_n(b) = \sum_{i=1}^{n} S_i \log\big(\sigma(X_i^T b)\big) + (1-S_i)\log\big(1 - \sigma(X_i^T b)\big). \tag{3}$$

Maximisation of $\mathcal{L}_n(\cdot)$ is a concave optimisation problem. Note that this is equivalent to assuming (erroneously) that all unlabeled observations belong to the negative class and thus misspecified logistic model is fitted to the data for which posterior probability is governed by (1). Obviously, one can write down the complete correct log-likelihood for $(X_i, S_i)_{i=1}^n$:

$$\tilde{\mathcal{L}}_n(b, c) = \sum_{i=1}^{n} S_i \log\big(c\sigma(X_i^T b)\big) + (1-S_i)\log\big(1 - c\sigma(X_i^T b)\big) \tag{4}$$

and maximise it wrt $(b, c)$. Such method, named JOINT, has been proposed and investigated in [11]. However, finding global maximum of (4) is hindered by the fact that due to the presence of multiplicative constant $c$ in the form of posterior probability $P(S=1|x)$ given in (1) log-likelihood $\tilde{\mathcal{L}}_n(b, c)$ is no longer concave wrt $b$, in contrast to $\mathcal{L}_n(b)$. There are some attempts to account for this, either by using Minorization-Maximization algorithm or modelling $\tilde{\mathcal{L}}_n(\cdot, c)$ as the difference of two concave functions ( [13]).

Frequently, our aim is not to approximate $(\beta, c)$ but to construct a classification rule based on training data $(X_i, S_i)_{i=1}^n$. For review of such methods see e.g. [1]. In such a case one can ask whether the classifier based on maximiser of $\mathcal{L}_n(b)$ can not be modified to yield approximation of Bayes classifier of $Y$. The answer is affirmative and it relies on the crucial observation that $\mathcal{L}_n(b)$ can be viewed as log-likelihood of misspecified logistic regression fitted to data corresponding to posterior probability $q(x^T\beta) = c \times \sigma(x^T\beta)$. This was noticed already in the context of estimation of $\beta$ in [11] using Ruud's theorem [9] stated below, however its useful consequences have been never explored for PU

classification. Here we try to fill this gap by showing that the naive classifier can be improved by adjusting its intercept, the step which has significant influence on its performance. Below we state Ruud's theorem [9] for a logistic loss, for the general statement see [8].

### A. Colinearity under misspecification: general case

Assume that the distribution of random vector $(X, S)$ is such that posterior probability $P(S = 1|X = x) = q(x^T\beta)$ for some unknown response function $q : R \to (0, 1)$ which is possibly different from logistic function. Let $\beta^*$ be the maximiser of expected normalised log-likelihood in (3) for such distribution:

$$\begin{aligned} n^{-1}E_{(X,S)}\mathcal{L}_n(b) &= E_X\{q(x^T\beta)\log\sigma(x^Tb) \\ &+ (1-q(x^T\beta))\log\big(1-\sigma(x^Tb)\big)\} \end{aligned} \quad (5)$$

We note that $\beta^*$ can be interpreted as the minimiser of the averaged Kullback-Leibler (KL) divergence between binary distribution $(q(X^T\beta), 1 - q(X^T\beta))$ and family of logistic models $\{\sigma(X^Tb)\}_{b\in R^{p+1}}$ (see [3] for the definition and properties of KL divergence) and thus corresponds to the Kullback-Leibler projection of the true distribution on this family. It also follows that $\beta^*$ satisfies the following vector equality

$$EXq(X^T\beta) = EX\sigma(X^T\beta^*). \quad (6)$$

The obvious consequence of (6) is that when $q(s) \equiv \sigma(s)$ and the projection is unique, then $\beta^* = \beta$.
We say that $X$ satisfies Linear Regressions Condition $(LRC(b))$ for vector $b \in R^{p+1}$ if

$$E(\tilde{X}|\tilde{b}^T\tilde{X} = w) = \gamma w + \gamma_0 \quad (7)$$

for some $\gamma = \gamma(\tilde{b}), \gamma_0 = \gamma_0(\tilde{b}) \in R^p$. We note that $LRC(b)$ condition is satisfied for the multivariate normal distribution for any $b \in R^{p+1}$ and, more generally, by the class of eliptically contoured distributions.

*Theorem 1:* [9] Assume that $X$ satisfies $LRC(\beta)$ condition and moreover covariance matrix of $\Sigma_{\tilde{X}}$ of vector $\tilde{X}$ is strictly positive definite. Additionally, $P(Y = 1|X = x) = q_0(x^T\beta)$ for some unknown function $q_0$ and for some $\beta \in R^{p+1}$. Then minimiser $\beta^*$ of (5) satisfies

$$\beta^*_{-0} = \eta\beta_{-0},$$

where $\beta = (\beta_0, \beta^T_{-0})^T$ and $\beta^* = (\beta^*_0, \beta^{*T}_{-0})^T$. Moreover, $\eta > 0$ provided that $\mathrm{Cov}(Y, X) > 0$ and $LRC(\beta^*)$ holds.

For the proof of the first part see e.g. [8]. The second part follows from normal equations (6) and the fact that vector $\gamma$ in (7) equals $(\beta^T_{-0}\Sigma_{\tilde{X}}\beta_{-0})^{-1}\Sigma_{\tilde{X}}\beta_{-0}$.

Theorem above implies that under the stated conditions despite the misspecification of the fitted model we still retain colinearity of true parameter $\beta$ and the vector $\beta^*$ of its Kullback-Leibler projection when the first coordinate in both vectors corresponding to intercept is omitted. This has an obvious relevance in classification if one recalls that Bayes classifier when logistic model is valid equals under conditions of Theorem 1:

$$\begin{aligned} &\hat{Y}(X) = I\{(\tilde{X}^T\beta_{-0} + \beta_0 > 0\} \\ = \;& I\{\eta\tilde{X}^T\beta_{-0} + \eta\beta_0 > 0\} = I\{\tilde{X}^T\beta^*_{-0} + \eta\beta_0 > 0\} \end{aligned} \quad (8)$$

Thus the direction of the optimal separating hyperplane $\tilde{X}^T\beta_{-0} + \beta_0 = 0$ is given by projection $\beta^*_{-0}$ which is easily estimable and only the intercept $\eta\beta_0$ needs to be recovered. Let $\hat{\beta}^*$ denote maximiser of (3). As Maximum Likelihood estimator $\hat{\beta}^*$ consistently estimates $\beta^*$ under mild conditions (see [14]) one can use $\hat{\beta}^*_{-0}$ as the vector defining the direction of the separating hyperplane $w^Tx + w_0$ and then adjust its intercept appropriately.

### B. Collinearity under misspecification: PU case

Consider now Positive Unlabeled data case and assume that posterior probability of $Y$ given $X$ is given by logistic model defined in (2). Then in the view of (1) when logistic model is fitted to $(S, X)$, the model is misspecified as $P(S = 1|X = x) = c \times \sigma(^T\beta)$. However, under conditions of Theorem 1 we have $\beta^*_{-0} = \eta\beta_{-0}$ and moreover (6) yields

$$cEX\sigma(X^T\beta) = EX\sigma(\beta^*_0 + \eta\tilde{X}^T\beta_{-0}).$$

This shows how parameter $\eta$ depends on labeling frequency $c$ and distribution of $X^T\beta$. When $X$ is multivariate normal this can be restated more explicitly.

*Theorem 2:* Assume that $X \sim N(0, \Sigma)$ and conditions of Theorem 1 are satisfied. (i) Then we have for any $j = 1, \ldots, p$:

$$\frac{\eta}{c} = \eta\frac{EX_j\sigma(\beta_0 + \tilde{X}^T\beta_{-0})}{EX_j\sigma(\beta^*_0 + \eta\tilde{X}^T\beta_{-0})} = \frac{E\sigma'(\beta_0 + \tilde{X}^T\beta_{-0})}{E\sigma'(\beta^*_0 + \eta\tilde{X}^T\beta_{-0})} \quad (9)$$

(ii) If $c \le 1/2$ then $\beta^*_0 < 0$ for any $\beta_0$.

Proof. The first equality in (9) is just a consequence of (6) when $j^{th}$ coordinate is considered. The second equality follows from Stein's lemma, which states that $\mathrm{Cov}(h(Z_1), Z_2) = Eh'(Z_1)\mathrm{Cov}(Z_1, Z_2)$ for bivariate normal vector $(Z_1, Z_2)$. It implies that

$$\begin{aligned} &EX_j\sigma(\beta_0 + \tilde{X}^T\beta_{-0}) = \mathrm{Cov}(X_j, \sigma(\beta_0 + \tilde{X}^T\beta_{-0})) \\ = \;& E\sigma'(\beta_0 + \tilde{X}^T\beta_{-0})\mathrm{Cov}(X_j, \beta_0 + \tilde{X}^T\beta_{-0}) \end{aligned} \quad (10)$$

and, analogously

$$\begin{aligned} &EX_j\sigma(\beta^*_0 + \eta\tilde{X}^T\beta^*_{-0}) = \mathrm{Cov}(X_j, \sigma(\beta^*_0 + \tilde{X}^T\beta^*_{-0})) \\ = \;& E\sigma'(\beta^*_0 + \eta\tilde{X}^T\beta_{-0})\mathrm{Cov}(X_j, \beta^*_0 + \eta\tilde{X}^T\beta^*_{-0}). \end{aligned} \quad (11)$$

Applying normal equations again we obtain the second equality.
In order to prove (ii) note that for any symmetric univariate random variable $Z$ we have

$$E\sigma(a + Z) < 1/2 \iff a < 0.$$

Indeed

$$E\sigma(a + Z) = 1 - E\sigma(-a - Z) = 1 - E\sigma(-a + Z),$$

where the second equation is due to symmetry of $Z$. This, and the fact that $\sigma(a+Z) < \sigma(-a+Z)$ is equivalent (due to monotonicity of $\sigma(\cdot)$) to $a < 0$ justify the claim. However, note that normal equations for the first coordinate being 1 imply that

$$E\sigma(\beta_0^* + \eta\tilde{X}^T\beta_{-0}) = cE\sigma(\beta_0 + \tilde{X}^T\beta_{-0}) < c \leq \frac{1}{2}$$

and thus $\beta_0^* < 0$.

*Remark 3.1:* Part (ii) explains why the naive classifier applied to $(S, X)$ data will work poorly, especially for small $c$: its intercept is likely to be negative regardless the sign of the intercept $\eta\beta_0$ in (8). Thus it has to be modified to enhance the performance of naive classifier.

*Remark 3.2:* The case when no intercept is included in both the true and the fitted model has been considered in [11]. It is shown there that then $0 < \eta \leq c < 1$. Thus in this case coefficients of logistic model corresponding to genuine predictors are shrunk towards 0.

### C. Choice of the intercept

We propose to choose the intercept $\widehat{w}_0$ of the separating hyperplane $\tilde{x}^T\hat{\beta}_{-0}^* + \widehat{w}_0 = 0$, where $\widehat{w}_0$ is an estimator of $\eta\beta_0$ (see (8)), by maximising the analogue of $F1$ measure on training data. We let, for a given classifier $\hat{Y} = \hat{Y}(X)$ learnt on the training data $\mathcal{D}^{train}$:

$$r = P(\hat{Y}(X) = 1|Y = 1) \qquad p = P(Y = 1|\hat{Y}(X) = 1)$$

be population recall and precision of $\hat{Y}$, respectively. Here, $(X, Y)$ stands for unobservable random variable having distribution $P_{X,Y}$ which is independent of $\mathcal{D}^{train}$. We define $F1$ measure as their harmonic mean

$$F1 = \frac{r \times p}{(r+p)/2}. \tag{12}$$

Thus in order to have large $F1$ value, both the precision and recall should be large. We also note that simple derivation yields $F1 = 2 \times P(Y = 1, \hat{Y} = 1)/(P(Y = 1) + P(\hat{Y} = 1))$. Moreover, note that for PU data under SCAR we have that $P(\hat{Y}(X) = 1|Y = 1, \mathcal{D}^{train}) = P(\hat{Y}(X) = 1|S = 1, \mathcal{D}^{train})$ as $\hat{Y}(X)$ given $\mathcal{D}^{train}$ depends on $X$ only and $P(X|Y = 1) = P(X|S = 1)$.

This means that the recall $r$ can be easily estimated from $(X, S)$ sample. The precision, however is unobservable, and thus we consider the following analogue of $F1$ introduced in [7], Section 4, (see also [10]) defined as

$$F1_{PU} = \frac{r \times p}{P(Y = 1)}. \tag{13}$$

$F1_{PU}$ is proportional to squared geometric mean of the precision and the recall i.e. Fowlkes-Mallows index [5]. Note that one obtains

$$\frac{P(Y = 1|\hat{Y}(X) = 1)}{P(Y = 1)} = \frac{P(\hat{Y}(X) = 1|Y = 1)}{P(\hat{Y}(X) = 1)}$$

which in terms of the precision and the recall means that $p = r \times P(Y = 1)/P(\hat{Y}(X) = 1)$ and thus

$$F1_{PU} = \frac{r^2}{P(\hat{Y}(X) = 1)}. \tag{14}$$

Let $\hat{Y}_z(x) = I\{\tilde{x}^T\hat{\beta}_{-0}^* + z > 0\}$, where $\hat{\beta}^*$ is maximiser of (3) and define $\widehat{F1}_{PU}(z)$ to be a sample analogue of $F1_{PU}$ for the classifier $\hat{Y}_z(X)$. We propose to choose $\widehat{w}_0$ as maximiser of

$$\widehat{w}_0 = \text{argmax}_z \widehat{F1}_{PU}(z) \tag{15}$$

We will call the classifier $\hat{Y}(x) = I\{\tilde{x}^T\hat{\beta}_{-0}^* + \widehat{w}_0 > 0\}$ the enhanced naive classifier. The pseudo-code for enhanced classifier is given in Algorithm 1. We show below when analysing its behaviour on real data sets that modification of the intercept of the naive classifier is crucial for its performance.

**Algorithm 1** Enhanced naive classifier

**Input:** Observed data $(x_i, s_i)$, $i = 1, \ldots, n$.
**Step 1:** Obtain estimator $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_{-0}^*)$ by fitting logistic regression to observed data $(x_i, s_i)$.
**Step 2:** Calculate intercept $\hat{w}_0$ as $\text{argmax}_z \widehat{F1}_{PU}(z)$.
**Result:** Parameters $(\hat{w}_0, \hat{\beta}_{-0}^*)$ of the separating hyperplane.

## IV. Numerical experiments

In the numerical experiments we have considered the following classifiers:

- Naive classifier based on fitting logistic regression model to $(X, S)$ data called Naive and the classifier Enhanced proposed here;
- Classifiers based on JOINT and MM estimators discussed above;
- Weighted classifiers introduced in [1], Section 5.3.1 using two alternative estimators of $c$: proposed in [4] (denoted by $e_1$, p.214) and TIcE estimator introduced in [2]. For the discussion of both estimators of $c$ see e.g. [6]. They will be called EN and TIcE classifiers, respectively.

The implementation of Enhanced estimator is given in github directory[1]. Maximisation of $\widehat{F1}_{PU}(z)$ in (15) is achieved by looking for maximal value among the values of this quantity, noting that numerators of numerator and denominator of the ratio defining it may change by $\pm 1$ when moving along ordered values of intercept for which predictions of considered classifiers change, i.e. values $z_i = \tilde{x}_i^T\hat{\beta}_{-0}^*$.

### A. Synthetic data

In order to check how Ruud's theorem works in practice and the performance of the proposed classifier, we considered a simple synthetic example where vector of predictors $\tilde{X}$ has three-dimensional normal distribution with mean $m = (1, 1, -1)^T$, variances equal to 1 and covariances $Cov(X_1, X_2) = 0.2$, $Cov(X_1, X_3) = -0.2$ and

[1]https://github.com/MateuszPlatek/PU_Enhanced_Naive_Classifier

$Cov(X_2, X_3) = 0$. Thus $X_1$ is positively correlated with $X_2$ and negatively correlated with $X_3$. Moreover posterior probability of $Y = 1$ given $X = x$ is logistic with $\beta = (-1, -1, 1, 1)^T$. We investigated the angle between $\hat{\beta}_{-0}$ and $\beta_{-0}$ for all considered estimators, the performance of corresponding classifiers for $c = 0.3, 0.6$ and several values of $n$ ranging from 500 to 5000. The results are shown in Figure 1 situated at the end of the paper. The first row of the panel exhibits goodness of fit of the considered estimators measured by the mean differences of their angles and the angle of $\beta_{-0}$. It indicates that in concordance with Ruud's theorem the direction of $\beta_{-0}$ is approximately recovered by direction of naive estimator $\hat{\beta}_{-0}$ for sample sizes larger than 1000 and the accuracy increases with increasing sample size. Moreover the accuracy of $\hat{\beta}_{-0}$ measured by mean difference of angles for naive, MM and JOINT estimators approximately coincides and is consistently better than that of EN and TIcE estimators. In terms of F1 measure shown in the second row the introduced enhanced naive classifier works consistently better than its competitors and in terms of Balanced Accuracy (defined as the average of the recall and the specificity; the third row) it is only outperformed by EN classifier for $c = 0.3$.

### *B. Real datasets*

We have analysed performance of the estimators on six data sets from UCI directory with sample sizes ranging from around 300 to 30 000 and number of features from 3 to 166 (the main characteristics of the data sets are given in Table I). The figures show mean performance with the regard of F1 measure (Figure 2) and Balanced Accuracy (Figure 3), for values of $c$ ranging from $0.1$ to $0.9$, based on 200 random splits of the data into training and testing subsamples. Standard errors for the mean are smaller than 0.01 in most cases for both F1 and BA measure with the only exception of F1 measure on `credit-a` and `diabetes` data set and the maximal value of SE is 0.026 for JOINT estimator on `credit-a`. Note that the results for the naive classifier are truncated from below in Figure 2: F1 measure for naive classifier is very low for $c \leq 0.5$ and approach 0 for $c$ close to 0. The first immediate observation is that the change of the intercept estimator, which is the only difference between the naive classifier and its enhanced version, has a huge impact on its performance with regard to both considered measures.

**F1 measure** In all cases but one the enhanced classifier works better (data sets `musk, credit-a, diabetes, adult`) or on par (`banknote`) with JOINT and MM estimators. In the case of *spam* it works marginally worse than JOINT and MM. This is interesting, especially in comparison with MM estimator which requires much more computing effort. It also outperforms TIcE and EN estimators on three data sets: `banknote, musk` and `spam`. On `adult` data set enhanced classifier works better than EN and on par with TIcE. Its excellent performance on `musk` data set is worth pointing out. The performance of enhanced estimator deteriorates for small values of $c$, possibly due to loss of accuracy of $\widehat{F1}_{PU}$ (note that the denominator of (14) becomes smaller for smaller $c$).

| Name | Size | Features | Fraction of positive observations |
|---|---|---|---|
| adult | 32561 | 57 | 0.24 |
| banknote | 1372 | 4 | 0.44 |
| credit-a | 690 | 38 | 0.44 |
| diabetes | 768 | 8 | 0.35 |
| musk | 6598 | 166 | 0.15 |
| spambase | 4601 | 57 | 0.39 |

TABLE I: Analysed datasets and their statistics

| Algorithm | Oracle | Enhanced | JOINT | MM | EN | TIcE |
|---|---|---|---|---|---|---|
| Time | 0.05s | 0.22s | 0.23s | 201s | 0.66s | 0.9s |

TABLE II: Mean training time in seconds on the largest dataset adult with $c = 0.5$.

**Balanced Accuracy** The performance of enhanced estimator with respect of Balanced Accuracy is similar to that with respect to F1 measure.

We have also analysed training times of the considered classifiers. Table II shows the training times for the largest data set `adult`. In the case of Enhanced and JOINT classifiers the times are approximately the same and 2-3 times shorter that the times for EN and TiCE classifiers. The most computation intensive is MM classifier as it requires inner loop of convex optimisation for each iteration of $\hat{\beta}$.

## V. CONCLUSION

We have studied a novel modification of naive classifier for Positive Unlabeled data under SCAR assumption. The classifier has strong theoretical underpinnings following from Ruud's theorem which are are established in Theorem 1. These indicate that the coefficients of logistic classifier corresponding to genuine predictors are consistently estimated based on observed $(X, S)$ data and the estimation problem boils down to consistent estimator of the intercept. We have proposed such an estimator based on maximisation of observable analogue of $F1$ measure. Moreover, we have shown analysing real data sets that the resulting enhanced naive estimator is a promising alternative to classifiers based on parametric models of posterior probability (JOINT and MM classifier) as well as nonparametric ones (TIcE and EN classifiers). Future research may include finding alternatives to the proposed method of estimating the intercept as well as extension of the considered method to the situation when SCAR assumption is violated. In particular, note that when posterior probability $y(x)$ satisfies (2) and $e(x)$ is an *arbitrary* function of $y(x)$, posterior probability $s(x)$ of $S = 1$ given $X = x$ is a function of $x^T\beta$ and it corresponds to misspecified logistic model. Thus the conclusion of Theorem 1 applies also to this more general situation which as its special case includes probabilistic gap assumption when $e(x)$ is an increasing function of $y(x)$.

## REFERENCES


[1] J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760, April 2020. http://dx.doi.org/10.1007/S10994-020-05877-5.

[2] Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2712–2719, April 2018. https://doi.org/10.1609/aaai.v32i1.11715.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, NY, 1991. http://dx.doi.org/10.1002/047174882X.

[4] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, August 2008. http://dx.doi.org/10.1145/1401890.1401920.

[5] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. *Journal of American Statistical Association*, 78:573–586, 1981. https://doi.org/10.2307/2288117.

[6] M. Łazecka, J. Mielniczuk, and P. Teisseyre. Estimating the class prior for positive and unlabelled data via logistic regression. *Advances in Data Analysis and Classification*, 15(4):1039–1068, June 2021. http://dx.doi.org/10.1007/S11634-021-00444-9.

[7] W. Lee and B. Liu. Learning with positive and unlabeled exampled using weighted logistic regression. In *Proceedings of the Twentieth International Conference on Machine Learning*, ICML '03, pages 448–455, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[8] K-C. Li and N. Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989. http://dx.doi.org/10.1214/aos/1176347254.

[9] P. Ruud. Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51:225–228, 1983. http://dx.doi.org/10.2307/1912257.

[10] S. Tabatabaei, J. Klein, and M Hoogendoorn. Estimating the F1 score for learning from positive and unlabeled examples. In *LOD 2020*. Springer, Cham, 2020. https://doi.org/10.1007/978-3-030-64583-0_15.

[11] P. Teisseyre, J. Mielniczuk, and M Łazecka. Different strategies of fitting logistic regression for positive and unlabeled data. In *Proceedings of the International Conference on Computational Science ICCS'20*, pages 3–17, Cham, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-50423-6_1.

[12] Q. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989. https://doi.org/10.2307/1912557.

[13] A. Wawrzenczyk and J. Mielniczuk. Strategies for fitting logistic regression for positive and unlabeled data revisited. *Int.J. Appl. Math. Comp. Sci.*, pages 299–309, 2022. https://doi.org/10.34768/amcs-2022-0022.

[14] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. https://doi.org/10.2307/1912526.

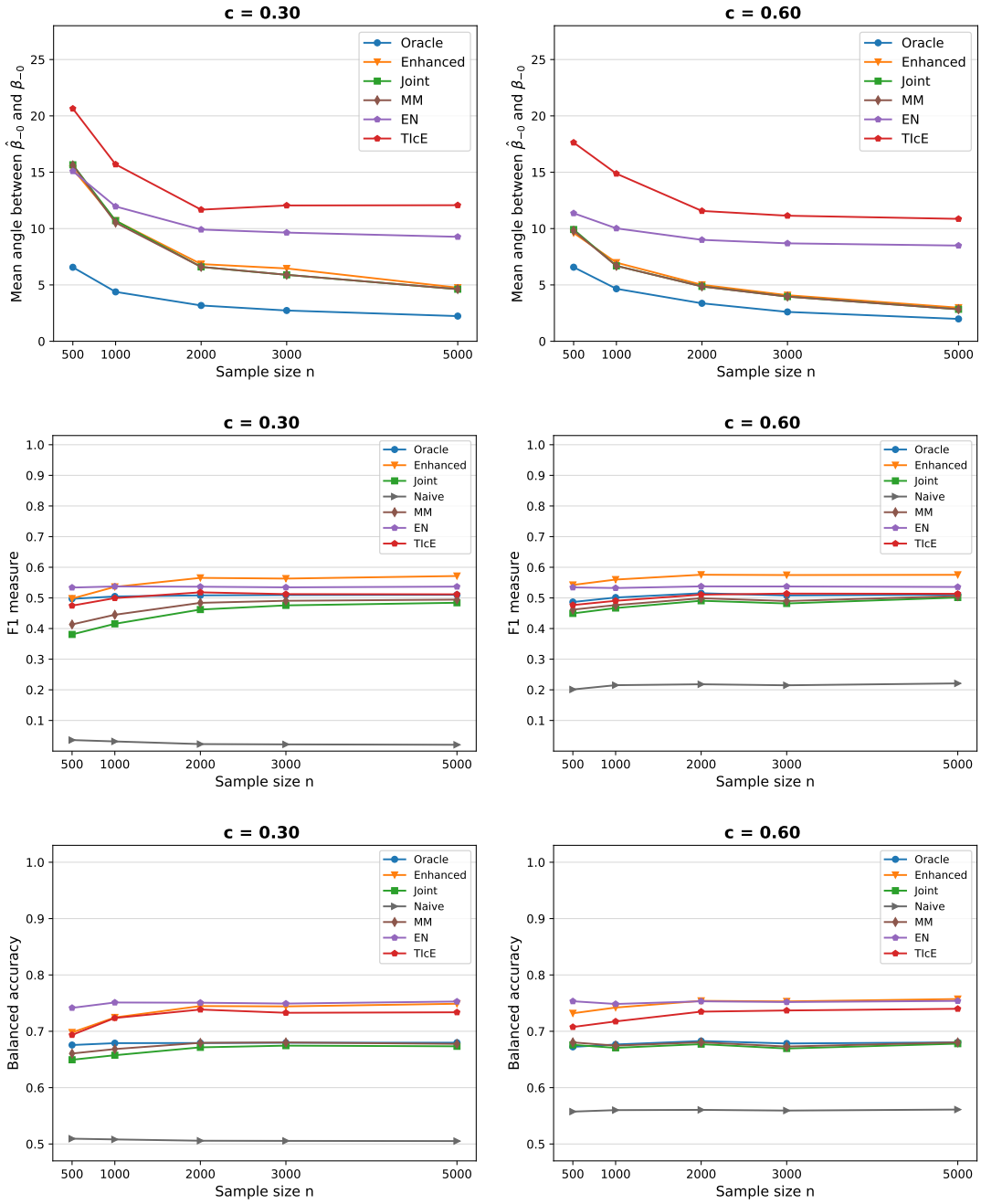


Fig. 1: Mean difference of angles, F1 and Balanced Accuracy against sample size for artificial data.

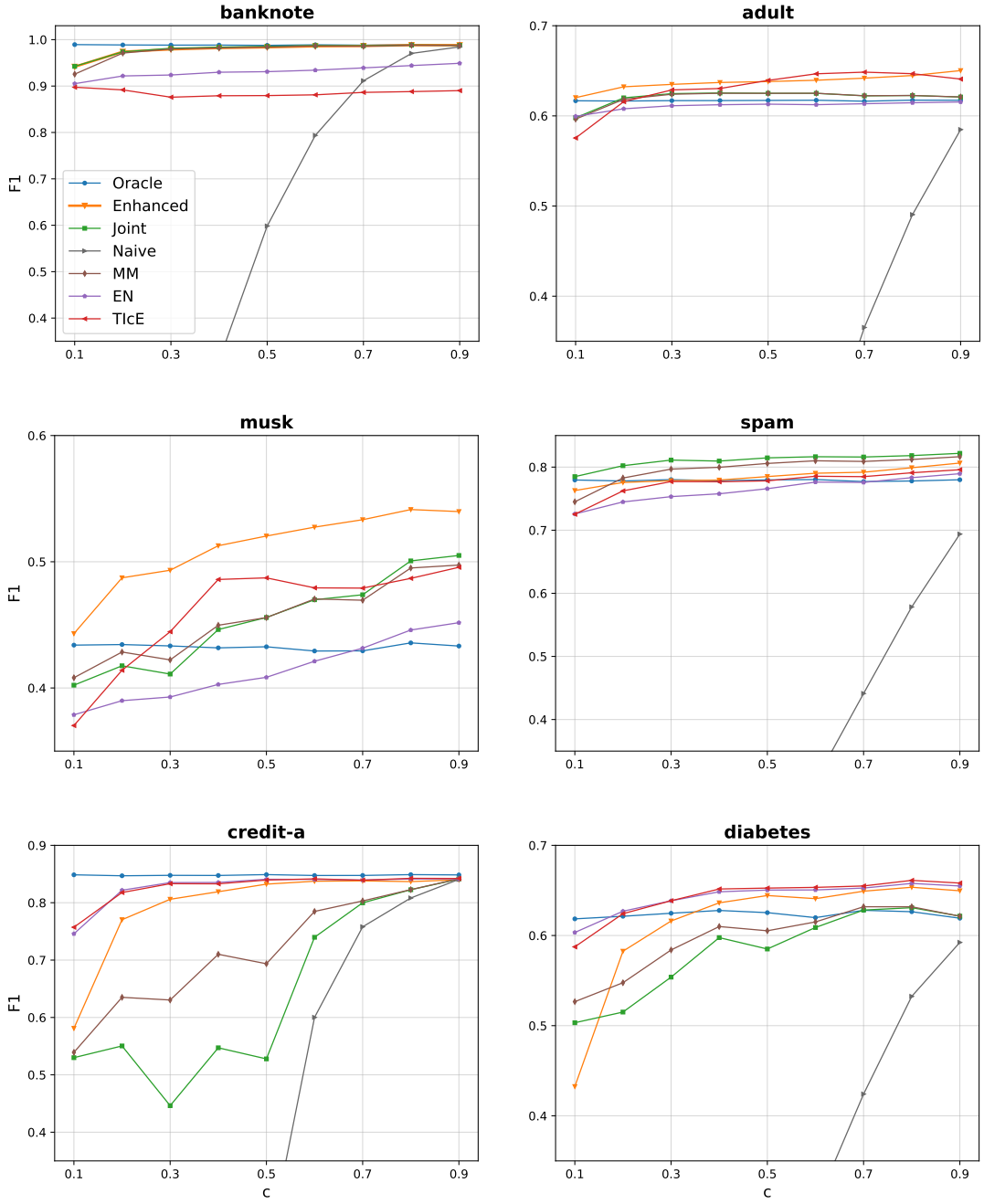


Fig. 2: F1 measure against values of c for the considered data sets.

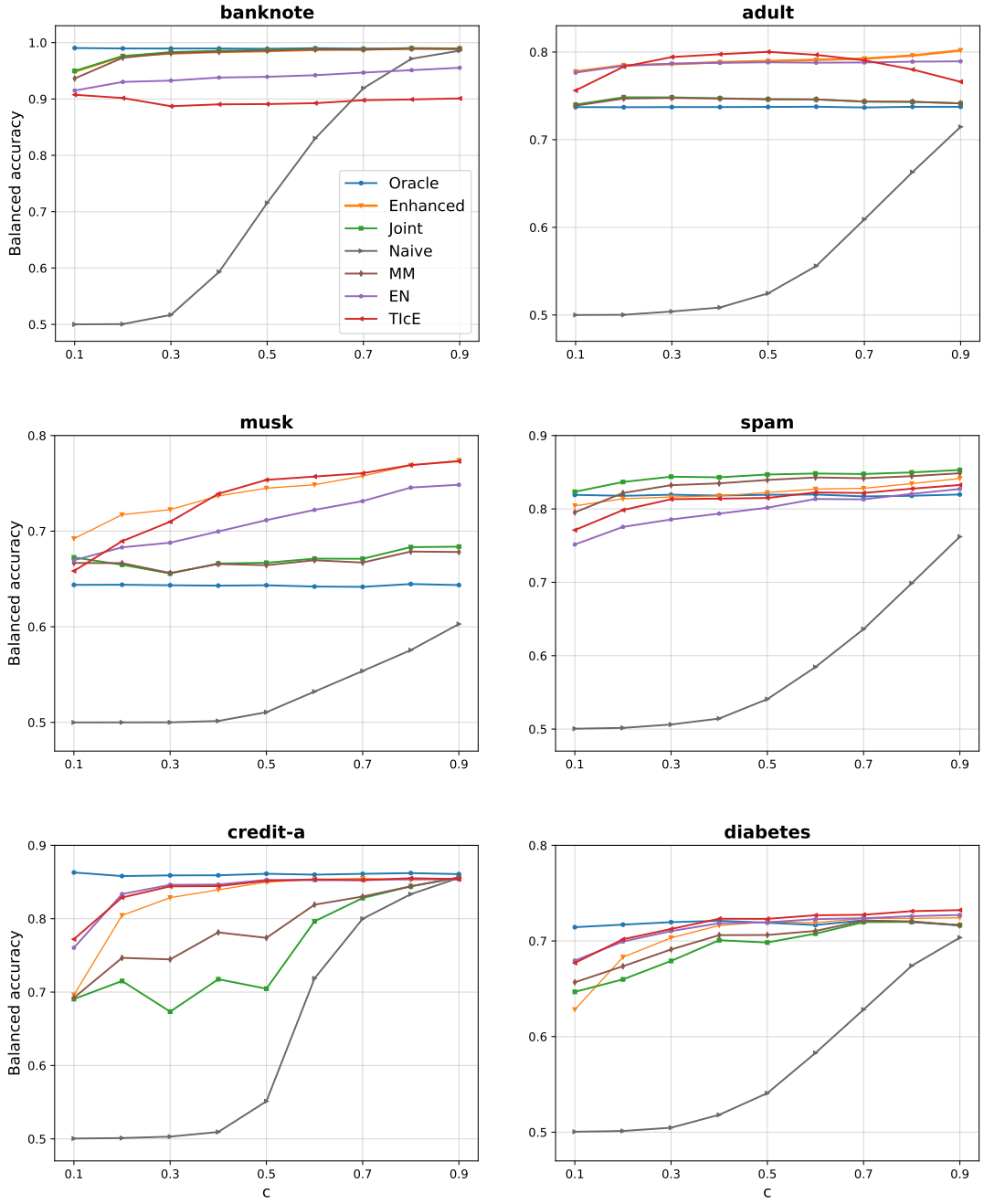


Fig. 3: Balanced Accuracy against values of c for the considered data sets.