Statistical Hypothesis Testing in Positive Unlabelled Data

Konstantinos Sechidis[1], Borja Calvo[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{sechidik,gavin.brown}@cs.manchester.ac.uk
[2] Department of Computer Science and Artificial Intelligence,
University of the Basque Country, Spain
borja.calvo@ehu.es

Abstract. We propose a set of novel methodologies which enable valid statistical hypothesis testing when we have only *positive and unlabelled* (PU) examples. This type of problem, a special case of semi-supervised data, is common in text mining, bioinformatics, and computer vision. Focusing on a generalised likelihood ratio test, we have 3 key contributions: (1) a proof that assuming all unlabelled examples are negative cases is sufficient for *independence* testing, but not for power analysis activities; (2) a new methodology that compensates this and enables power analysis, allowing sample size determination for observing an effect with a desired power; and finally, (3) a new capability, *supervision determination*, which can determine *a-priori* the number of labelled examples the user must collect before being able to observe a desired statistical effect. Beyond general hypothesis testing, we suggest the tools will additionally be useful for information theoretic feature selection, and Bayesian Network structure learning.

1 Introduction

Learning from Positive-Unlabelled (PU) data is a special case of semi-supervised learning, where we have a small number of examples from the positive class, and a large number of unlabelled examples which could be positive *or* negative. The objective in this situation is to perform standard machine learning activities despite this data restriction. The problem has been referred to in the literature under several names, including *partially supervised classification* [15], *positive example based learning* [19] and *positive unlabelled* learning [9]. A typical application has been text classification — given a number of query documents belonging to a particular class (e.g. academic articles about machine learning), plus a corpus of unlabelled documents, the task is to classify new documents as relevant to the query or not.

Most work in the PU area is concerned with classification, rather than theory. Denis [8] is an interesting exception, which generalised Valiant's PAC learning framework to PU data, and concluded that learning from positive and unlabelled examples is possible, but that we must have some additional prior knowledge

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-662-44845-8_5

about the underlying distribution of examples. We make use of this observation, exploring how *statistical hypothesis testing* manifests in PU data, and how such prior knowledge can be incorporated.

In this context, we focus on the G-test [18], a generalised likelihood ratio test used for testing independence of categorical variables, which is closely related to the mutual information (Sec. 3). The G-test and the mutual information are used extensively, for example in life sciences to test whether two observed natural processes are independent [16]. In machine learning and in data mining they also have a large number of applications, for instance in structure learning of a Bayesian Network [2] or in feature selection [4]. The main contributions[1] of our work are the following

- A proof that, the common assumption of all unlabelled examples being negative is sufficient for *testing independence* (Sec. 4.1), but *insufficient* for more advanced activities such as power analysis (Sec. 4.2).
- A methodology for a-priori power analysis in the PU scenario, enabling *sample size determination* for observing an effect with a desired power (Sec. 5).
- A novel capability: *supervision determination*, which can determine the minimum number of labelled data to achieve a desired power (Sec. 5).

In a general hypothesis testing scenario, our results make clear the implications of using the G-test under the PU constraint, and leads to more cost-effective experimental design. In wider machine learning activities, there are several applications. For example, constraint-based learning of Bayesian Network structures: the decision on whether to include an arc between two nodes is often made with a hypothesis test such as χ^2, or the mutual information, both of which are core to our work. Another example is information theoretic feature selection, Guyon et al. [13, pg 68] discuss how the statistical viewpoint on feature selection allows decision-making on the relevance/redundancy of a feature to be made in a principled manner. Our methods permit these activities under the PU data constraint.

2 Background on the positive unlabelled problem

Positive-Unlabelled data refers to situations where we have a small number of examples labelled as the positive class, and a large number of entirely unlabelled examples, which could be either positive *or* negative. Whilst classification is well explored in the PU scenario, an area in need of attention is *statistical hypothesis testing*: including independence tests, and more complex activities such as power analysis. We now introduce the formal framework of Elkan & Noto [10] for reasoning over PU data, which we build upon in our work.

2.1 Positive unlabelled framework

Assume that a dataset $\mathcal{D}$ is drawn i.i.d. from the joint distribution $p(X, Y, S)$, where the features X are categorical, the class Y is binary, and S is a further

[1] Matlab code for all methods and the supplementary material with all the proofs and extra results are available in `www.cs.man.ac.uk/~gbrown/posunlabelled/`.

random variable with possible values 's^+' and 's^-', indicating if the example is labelled (s^+) or not (s^-). Thus $p(x|s^+)$ is the probability of X takes the value x from its alphabet $\mathcal{X}$ conditioned on the labelled set. The same shorthand notation is used for Y, where the positive class is indicated by 'y^+', and the negative class by 'y^-'.

In this context, Elkan & Noto formalise the *selected completely at random* assumption, saying that the examples for the labelled set are selected completely at random from all the positive examples:

$$p(s^+|x, y^+) = p(s^+|y^+) \quad \forall \ \ x \in \mathcal{X}.$$

Thus, the probability of a positive example being labelled is *independent* of the input x. Perhaps most interestingly, Elkan & Noto proceed to show that this assumption has been followed either explicitly or *implicitly* in most research on PU data.

2.2 A naive approach – assuming unlabelled examples are negative

One approach to learn from this data is to simply assume that any unlabelled examples are negative. This approach, while seemingly naive, has proven to be useful for classification. Elkan & Noto [10] show that a probabilistic classifier trained on such data predicts posterior probabilities that differ from the true values by a constant factor; they suggest a number of ways to estimate this factor using a validation set. In a different context, Blanchard et al. [3] use the same assumption and prove that semi-supervised novelty detection can be reduced to Neyman-Pearson binary classification using the nominal and unlabelled samples as the two classes, in their terminology.

2.3 Incorporating prior knowledge

Another general approach follows the theoretical work of Denis [8], incorporating prior knowledge of the class distributions to augment the learning. For example, Calvo et al. [5] build PU naive bayes classifiers, and propose a Bayesian solution to deal with uncertainty in the distribution of the positive class.

At first glance, in the PU learning environment estimating $p(x|y^-)$ seems impossible without negative data. However, with a neat rearrangement of the marginal $p(x)$ and some extra information, it turns out to be possible. The marginal is $p(x) = p(x, y^-) + p(x, y^+)$, which can be rearranged:

$$p(x|y^-) = \frac{p(x) - p(x|y^+)p(y^+)}{1 - p(y^+)}. \tag{1}$$

Denis et al. [9] exploited this to construct a PU Naive Bayes classifier, estimating $p(x|y^+)$ by maximum likelihood on just the labelled set, i.e. assuming $\hat{p}(x|y^+) \approx \hat{p}(x|s^+)$, and estimating $p(x|y^-)$ using equation (1). The prior $p(y^+)$

was provided as a user-specified parameter, $\tilde{p}$. Thus, the missing conditional probability $p(x|y^-)$ is estimated as,

$$\hat{p}(x|y^-) \approx \frac{\hat{p}(x) - \hat{p}(x|s^+)\tilde{p}}{1-\tilde{p}},$$

where $\hat{p}(x)$, $\hat{p}(x|s^+)$ denote the maximum likelihood estimates of the respective probabilities. Although this seems a heuristic approach, we will now show with the following Lemma that under the *selected completely at random assumption* it is indeed valid.

Lemma 1. *Assuming data is selected completely at random, the conditional distribution of x given $y = 1$ is equal to the conditional distribution of x given that it is labelled.*

$$p(x|y^+) = p(x|s^+) \quad \forall \;\; x \in \mathcal{X}.$$

The proof of this Lemma and all the proofs of this work are available in the supplementary material.

2.4 Summary

In our work we will explore both approaches in the context of statistical hypothesis testing. By using the naive but common assumption that all the unlabelled examples are negative we can perform a test of independence between X against either the true labels (Y) or the assumed ones (S). In Section 4 we prove that these two cases have precisely the same false positive rate but different true positive rates (i.e. statistical power). While in Section 5, by using prior-knowledge, we derive a correction factor for the test that brings these into parity – identical true positive and false positive rates. As a consequence we can also perform positive unlabelled sample size determination, and determine the number of labeled examples needed to observe a desired statistical effect with a specified power. Before that, in Section 3 we review the likelihood ratio test that this work builds upon.

3 Hypothesis testing

3.1 The *G*-test of independence

In fully observed categorical data, the G-test can be used to determine statistical independence between categorical variables [18]. It is a generalised likelihood ratio test, where the test statistic can be calculated from sample data counts arranged in a contingency table. Denote by $O_{x,y}$ the observed count of the number of times the random variable X takes on the value x from its alphabet $\mathcal{X}$, while Y takes on $y \in \mathcal{Y}$; and by $O_{x,.}$ and $O_{.,y}$ the marginal counts. The estimated expected frequency of (x, y), assuming X, Y are independent, is given by

$E_{x,y} = \hat{p}(x)\hat{p}(y)N = \frac{O_{x,.}O_{.,y}}{N}$. The *G-statistic* can now be defined as

$$G = 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} O_{x,y} \ln \frac{O_{x,y}}{E_{x,y}} = 2N \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \hat{p}(x,y) \ln \frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)} = 2N\hat{I}(X;Y) \quad (2)$$

where $\hat{I}(X;Y)$ is the maximum likelihood estimator of the mutual information between X and Y [17]. Under the null hypothesis that X and Y are statistically independent, G is known to be asymptotically χ^2-distributed, with $\nu = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$ degrees of freedom [18]. For a given dataset, we calculate (2) and check to see whether it exceeds the critical value defined by a significance level α read from a standard statistical table giving the CDF of the χ^2-distribution — if the critical value is not exceeded, the variables are judged to be independent.

3.2 Power analysis

With such a test, it is common to perform a *power analysis* [6]. The *power* of a test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. This is also known as the *true positive rate*, or the probability of *not* committing a Type-II error. An a-priori power analysis would take a given sample size N, a required significance level α, an effect size ω, and would then compute the power of the statistical test. However, to do this we need a test statistic with a known distribution under the alternative hypothesis.

It is known that the G-statistic (2) has a large-sample *non-central* χ^2 distribution under the alternative hypothesis (i.e. when X and Y are dependent) as presented by Agresti [1, Section 16.3.5]. Agresti shows that the χ^2 non-centrality parameter (λ) has the same form as the G-statistic, but with sample values replaced by population values. In other words, the non-centrality parameter under the alternative hypothesis is given by $\lambda = 2NI(X;Y)$. Thus λ is a parameter, and G is a random variable following a distribution defined by λ.

One important usage of a-priori power analysis is *sample size determination*. In this prospective procedure we specify the significance level of the test (e.g. $\alpha = 0.05$), the desired power (e.g. a false negative rate of 0.01) and the desired effect size – from this we can determine the minimum number of examples required to detect that effect.

It turns out that the effect size of the G-test can be naturally expressed as a function of the *mutual information*. More specifically, the effect size (ω) is the square root of the non-centrality parameter divided by the sample size [6], thus we have $\omega = \sqrt{2I(X;Y)}$. Therefore, to understand hypothesis testing in PU data, we must understand the properties of mutual information in such data.

4 Hypothesis testing in positive unlabelled data

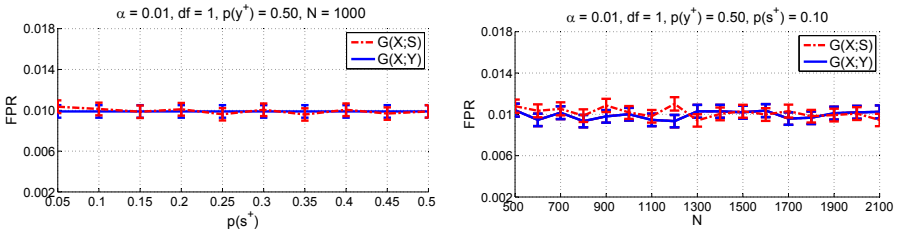
In this section we will focus in PU data by adopting the very common assumption that all unlabelled examples are negative, and exploring the consequences for hypothesis testing.

4.1 Testing for independence in positive unlabelled data

In positive unlabelled data, it is not immediately obvious how to apply the G-test described in the previous section, since the variable Y is only partially observed. The 'naive' approach would be to assume all unlabelled examples as negatives, and test for independence in the usual manner. This is in effect testing independence between X and S, the variable describing whether an example is labelled. While this is arguably a rather significant assumption, it is in fact sufficient to answer the question of whether X, Y are *independent*. This is proved formally with the following simple theorem.

Theorem 1. *In the positive unlabelled scenario, under the selected completely at random assumption, a variable X is independent of the class label Y if and only if X is independent of S, so it holds $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp S$.*

Intuitively, we can describe variable S as a noisy copy of Y, with no false positives but potentially a large number of false negatives. So instead of checking the independence with the actual variable Y we can check with the noise version S. The proof of the Theorem 1 is available in the supplementary material, though the theorem can be also experimentally verified with a simple 'sanity check' experiment. We generated data as so: X, Y are independent Bernoulli variables each with $p = 0.5$ and take $N = 1000$ observations. For the PU case, all negative examples have their labels removed, and we randomly remove a fraction $(1 - c)$ of the positive labels, where c is the fraction of all positive examples that are labelled, also written as $c = p(s^+)/p(y^+)$. To test independence we apply the G-test with a significance level of $\alpha = 0.01$ to test the assertion that $X \perp\!\!\!\perp Y$ in the supervised case, and $X \perp\!\!\!\perp S$ in the PU case. Since the null hypothesis $(X \perp\!\!\!\perp Y)$ is true, we expect a false positive rate of 1% in both cases — this is verified below in Figure 1a (over *100,000* repeats) holding for all supervision levels $p(s^+)$ and in Figure 1b holding for different sample sizes when we fix the supervision level to be $p(s^+) = 0.1$. The slight fluctuation comes from the limited



(a) (b)

Fig. 1: Figure for Type-I error. (a) Type-I error changing as a function of the probability of an example being labeled $p(s^+)$, for fixed $\alpha = 0.01$, $N = 1000$. (b) Type-I error changing as a function of N, for fixed $\alpha = 0.01$, $p(s^+) = 0.10$.

sample size. While Theorem 1 tells us that the two possible tests, $G(X; Y)$ and $G(X; S)$, are equivalent for observing independencies, it says nothing about how

well the naive $G(X;S)$ test will perform when the null hypothesis is *false*. In this case we must compare the tests in terms of their *power* to detect a given effect.

4.2 Comparing the power of the tests

In order to compare the power of the two tests we must examine their *non-centrality parameters*. Section 3.2 presents that the non-centrality parameter for the G-test is $\lambda = 2NI(X;Y)$. Therefore, the power of the tests depends on the population values of the mutual informations $I(X;Y)$ and $I(X;S)$. With the following theorem we prove an inequality between these two quantities

Theorem 2. *In the positive unlabelled scenario, under the selected completely at random assumption, when X and Y are dependent random variables ($X \not\perp\!\!\!\perp Y$) we have $I(X;Y) > I(X;S)$.*

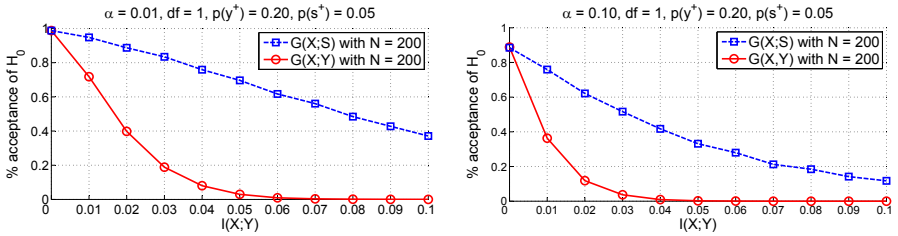
A direct consequence of the theorem is the following corollary.

Corollary 1. *The derived test under the the naive assumption, $G(X;S)$, is less powerful than $G(X;Y)$. In other words using the noisy copy S of Y, will result in a test $G(X;S)$ which will have a higher false negative rate than $G(X;Y)$.*

The proof of the Theorem 2 is available in the supplementary material, here we will give an experimental verification. As a sanity check we should explore how the two tests perform when we have an actual effect to observe. To create pairs of X and Y with a specific effect we generate data as follows. Firstly, generate a random sample $\mathbf{x} = \{x_1, .., x_N\}$, where each $x_i \in \{0, 1\}$ and $p(x = 1) = 0.2$. Then create an identical copy of this sample as $\mathbf{y}_{i=1}^{N}$. This creates a dataset where $\mathbf{x}, \mathbf{y}$ are by definition completely dependent. We then corrupt this dependency by picking a random fraction of the examples, and setting a new value for each selected x_i by drawing a binary random variable with parameter $p = 0.5$. It is clear that by varying the number of examples which are corrupted by noise, we generate random variables with different mutual informations. For example when we corrupt 60% of the examples with noise, we can calculate analytically that the resulting variables have $I(X;Y) = 0.053$ (in this work the effect sizes are written to 3 decimal places).

In order to observe the power of the two tests we will plot figures similar to the figures in Gretton & Györfi [12]. In the x-axis we have different values for the effect size, while in the y-axis is the acceptance rate of the null hypothesis $\mathcal{H}_0$ (over *10,000* independent generations of the data, each of size $N = 200$). The y-intercept represents $1 - (\textit{Type I error})$, and should be close to $1 - \alpha$, while elsewhere the plots indicate the Type II error. As we observe from the Figure 2 the test between X and S is less powerful than the test between X and Y, and this result verifies the Corollary 1. Furthermore the intercepts are at the same value (close to the design parameter $1 - \alpha$), which again verifies that the tests have the same Type-I error, but as can be seen different Type-II error.

Given this corollary, it is interesting to ask how we might modify our practice with $G(X;S)$ to achieve a desired power, in spite of the partially observed variable. In the next section we will show how we can incorporate prior knowledge to address this question.



(a) (b)

Fig. 2: Figure for comparing the Type-II error of the two tests using (a) $\alpha = 0.01$ and (b) $\alpha = 0.10$.

5 Incorporating prior knowledge for power analysis

In order to use $G(X;S)$ for a-priori power analysis activities, we should quantify the amount of power that we lose by adopting the naive assumption that all the unlabelled examples are negative. In this section we show how to incorporate prior knowledge to calculate this quantity.

Cohen [6] proposed (with appropriate caution) several conventional effect sizes, facilitating cross-experiment comparison. In the case of χ^2 tests, a "medium" effect is $\omega = 0.3$. Since $\omega = \sqrt{2I(X;Y)}$, this translates to $I(X;Y) = 0.045$. With PU data, the key problem emerges here in that the standard effect size is naturally expressed in terms of $I(X;Y)$, whereas our test $G(X;S)$ in terms of $I(X;S)$. In order to deal with this problem, we will incorporate a user's prior knowledge of $p(y^+)$, and correct the non-centrality parameter of the test in such a way that we can use it for a-priori power analysis.

Theorem 3. *The non-centrality parameter of the G-test between X and S takes the form:*

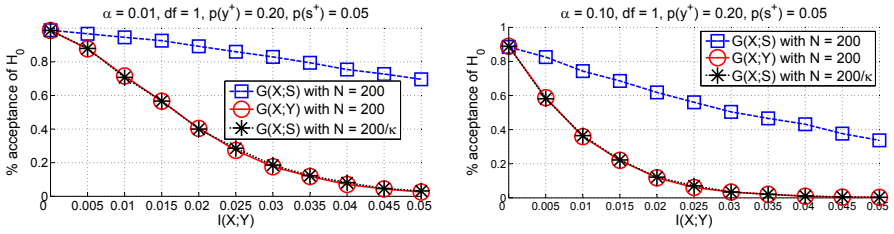
$$\lambda_{G(X;S)} = \kappa\lambda_{G(X;Y)} = \kappa 2NI(X;Y),$$

where $\kappa = \frac{1-p(y^+)}{p(y^+)}\frac{p(s^+)}{1-p(s^+)} = \frac{1-p(y^+)}{p(y^+)}\frac{N_{S^+}}{N-N_{S^+}}$.

Again the proof is in the supplementary material, and here we will give an empirical verification following the same experimental setup as the one described in Section 4.2. Thus as a sanity check in Figure 3 we observe that if we increase the sample size of the test between X and S by a factor κ, the two tests have the same power, and this result verifies Theorem 3. No matter what the sample size is, the intercepts are always at the same value (close to the design parameter $1-\alpha$), which again verifies that the tests have the same Type-I error.

We see that the non-centrality parameter $\lambda_{G(X;S)}$ is a function of: the sample size, the desired effect size and additionally a *correction factor*, κ, which depends on the number of labelled examples that we have (N_{S^+}). When we have full supervision, in other words when $p(s^+) = p(y^+)$, the κ takes the maximum value 1. In any other PU case, where $p(s^+) < p(y^+)$, the value is $\kappa < 1$.

In PU data, the prior probability $p(y^+)$ is in general unknown. Elkan & Noto [10] suggest an estimator for this parameter, which could potentially be used



(a) (b)

Fig. 3: Figure for comparing the Type-II error of the two tests and the G-test between X and S with corrected sample size , using (a) $\alpha = 0.01$. (b) $\alpha = 0.10$.

before an a-priori power analysis. A different way is to introduce a prior belief over that parameter; we will represent our prior belief as $\tilde{p}$, and the correction factor is re-written as

$$\kappa = \frac{1-\tilde{p}}{\tilde{p}} \frac{p(s^+)}{1-p(s^+)} = \frac{1-\tilde{p}}{\tilde{p}} \frac{N_{S^+}}{N - N_{S^+}}.$$

This correction factor enables us to use the $G(X;S)$ test in place of the $G(X;Y)$ for power analysis activities, such as sample size determination. Taking advantage of the extra degree of freedom in $p(s^+)$, we can also determine the *required level of supervision* (i.e. number of labelled examples), following the same procedure as in sample size determination. These capabilities will be empirically evaluated in the next section.

6 Experiments for a-priori power analysis

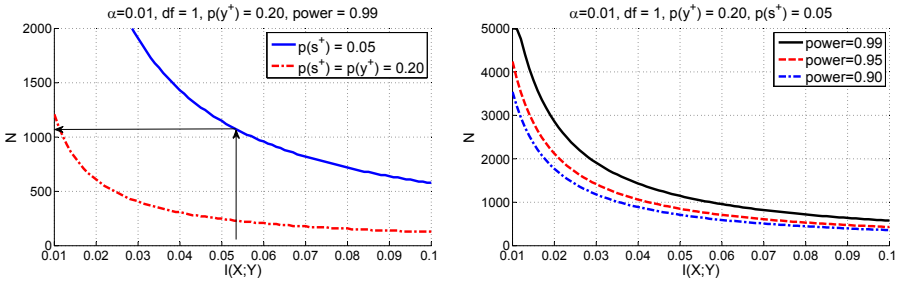
In this section we will show the capabilities of the G-test between X and S when the non-centrality parameter is corrected with the κ presented in Theorem 3, including sample size determination under the PU constraint, and a novel capability — determining the minimum number of labelled examples necessary to achieve statistical significance. We separate these experiments in two parts, the first one where we have perfect prior knowledge and the second where we use uncertain prior knowledge.

6.1 Perfect prior knowledge

In this section firstly we will provide some theoretical predictions for sample size and supervision determination, and then we will verify them empirically.

Theoretical predictions for sample size determination
Figure 4a shows how classical power analysis changes under the PU constraint. The illustration is for significance level $\alpha = 0.01$, a required power of 0.99, $p(y^+) = 0.2$, and binary features (degrees of freedom $\nu = 1$). For the reader's interest, all the figures and tables of this work are reproduced in the supplementary material with $\nu = 9$, meaning $|\mathcal{X}| = 10$.

In Figure 4a we see the dashed line, which shows classical sample size determination – this is a standard result. The solid line shows the PU case, when we can obtain labels only for 5% of the examples (i.e. $p(s^+) = 0.05$).



(a) (b)

Fig. 4: Figures for sample size determination. (a) Contrasting classical power analysis ($p(s^+) = p(y^+)$) with PU power analysis to determine the minimum sample size. Arrows show that with 5% supervision ($p(s^+) = 0.05$), we need $N \geq 1077$ examples to achieve the desired power in order to observe a supervised effect $I(X;Y) = 0.053$. (b) Sample size determination under the PU constraint. Given a required statistical power, this illustrates the minimum total number of examples (N) needed, assuming we can only label 5% of the instances. For example, if we wish to detect a mutual information as low as 0.04, we need $N \geq 1430$ to have a power of 99%.

The figure can be interpreted as follows: if we wish to detect a dependency with mutual information as low as $I(X;Y) = 0.053$, with power 99%, in the fully supervised case (dashed line) we require $N \geq 227$. However in the PU scenario (solid line) with $p(s^+) = 0.05$, this a-priori power analysis indicates we need $N \geq 1077$. Note that the required increase is not a simple multiple of the supervision level: with only 1/4 of the positive examples being labelled one might assume we need a sample 4× larger, which would be 908, however this is insufficient for the required power as shown by the figure. In this case, $\kappa = 0.2105$, and the required increase is a multiple of that factor: $227 \times (1/\kappa) \approx 1078$. The above results are expanded upon in Figure 4b, showing the required N to obtain different power levels.

Theoretical predictions for supervision determination

For power analysis in the PU constraint, we are able to use the same methodologies as in sample size determination to *determine the necessary level of supervision*, i.e. the number of labelled examples. This may have implications in active learning [7], where we can request the labels of particular examples — this methodology allows us to predict when we have sufficient labels to have statistically significant results.

Figure 5 presents the a-priori PU power analysis, allowing us to determine the minimum level of supervision to achieve a certain statistical power. The y-

axis is N_{S^+}, the number positive examples that have labels. This shows just one scenario, with $\alpha = 0.01$, $N = 1000$, when the true prior is $p(y^+) = 0.2$. As an illustration, the solid line predicts that to detect a dependency as low as $I(X;Y) = 0.053$, with power greater than 99%, we will need to label at least 54 examples or in other words the probability of an example being labeled should be $p(s^+) \geq 0.054$.

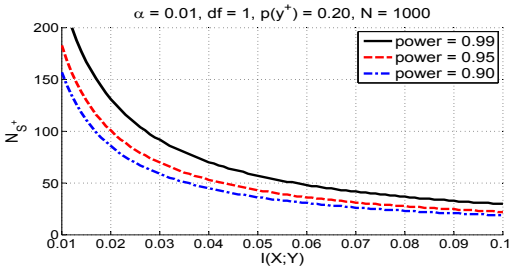


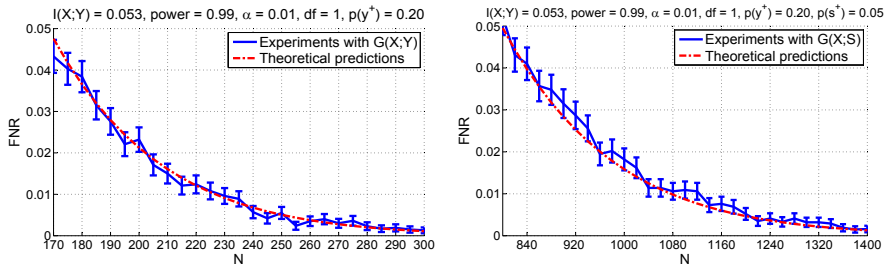
Fig. 5: Determining the required number of labelled examples. This illustrates the required number of labelled examples (N_{S^+}), assuming $N = 1000$. For example, to detect a mutual information dependency as low as 0.02, in order to have a power of 95%, we need labels for 101 examples, which means that we need to label at least half of the positive examples.

Verifying the theoretical predictions

To verify the theoretical predictions of required sample size and supervision level, we generate binary variables with a very small dependency (i.e. very small effect size) and observe the ability of a test to reject the null hypothesis — or in other words the False Negative Rate (Type-II error). Since the power is given by $1 - FNR$, any prediction of required N to achieve a particular power will translate directly to a corresponding FNR.

As a sanity check, we first verify the classical sample size determination for the G-test. Figure 4a (dashed line) predicts that we will need $N \geq 227$ to detect an underlying effect size of $I(X;Y) = 0.053$, with $\alpha = 0.01$ and power 99%. Figure 6a shows the FNR over *10,000* repeats. Note that the FNR crosses below the 1% rate when $N \approx 225$. The next experiment verifies the PU sample size prediction. As before, the negative examples all have their labels removed, and we randomly remove a fraction $(1-c)$ of the positive labels. Figure 4a (solid line) predicted that to detect an effect as small as $I(X;Y) = 0.053$, with $\alpha = 0.01$ we would require $N \geq 1077$ to achieve an FNR below 1%. The FNR again over *10,000* repeats is shown in Figure 6b, supporting the theory as the FNR crosses 1% when $N \approx 1080$.

Finally we verify the predictions from Figure 5. We generate PU data as before, introducing noise such that the true underlying variables have $I(X;Y) = 0.053$. Figure 7 shows the FNR, verifying that when we provide labels to the example with probability less than 0.054, or in other words when we label less than 54 examples the Type-II error is greater than 1%, and agrees with Figure 5.



(a) (b)

Fig. 6: Figures for FNR. (a) Full supervision, when the true mutual information is $I(X;Y) = 0.053$. This verifies the theoretical prediction from Fig. 4a. (b) Supervision level $p(s^+) = 0.05$, supporting the predictions of Fig. 4a.

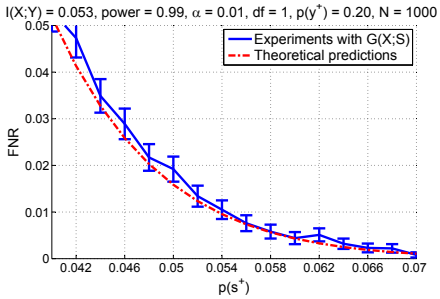


Fig. 7: FNR for varying levels of supervision in the PU constraint, with required power 99%, verifying Figure 5 (solid line), which predicted we would need $p(s^+) \geq 0.054 \Leftrightarrow N_{s^+} \geq 54$ to get $FNR < 0.01$.

6.2 Uncertain prior knowledge

The previous section assumed we somehow knew the exact value of $p(y^+)$. In a more realistic scenario prior knowledge may be provided as a *distribution* over possible values. We model $\tilde{p}$ as a generalised Beta distribution, between a minimum and a maximum value [14], and use Monte-Carlo simulation to explore the resultant uncertainty in the required sample/supervision sizes.

Figure 8 presents sample size determination when we have uncertain prior knowledge. The dashed vertical line indicates the perfect prior knowledge situation from the previous section. If we use a sample size less than this, we have an increased false negative rate. On the other hand, choosing a larger size will achieve at least the desired power, but at the cost of collecting more data.

Figure 9 presents how this uncertainty would translate to the required number of labeled examples. The same principle of choosing a value over/under the dashed line applies: here if we select $N_{s^+} > 54$ we are unnecessarily increasing our cost of label collection. In Figure 10 we observe the behavior when we underestimate (first row) or overestimate (second row) the $p(y^+)$. A general conclusion is that the uncertainty in the prior translates quite directly to an uncertainty of a similar form over the minimum number of samples and a minimum amount of supervision.

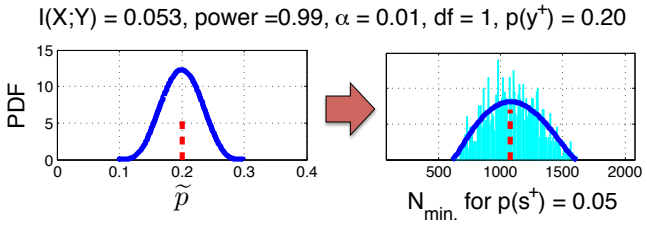


Fig. 8: Sample size determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of $p(y^+)$. The dashed line shows the *true* (but unknown) value in the data. RIGHT: The resultant uncertainty in the required sample size when we have only 5% of the examples being labeled, we plot both the histogram and the generalized beta distribution best fits to the data.

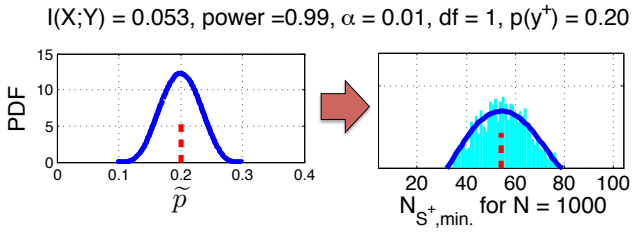


Fig. 9: Supervision determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of $p(y^+)$. The dashed line shows the *true* (but unknown) value in the data. RIGHT: The resultant uncertainty in the minimum number of required labeled examples when we have only $N = 1000$. The dashed line indicates the the true value with no uncertainty in $p(y^+)$.

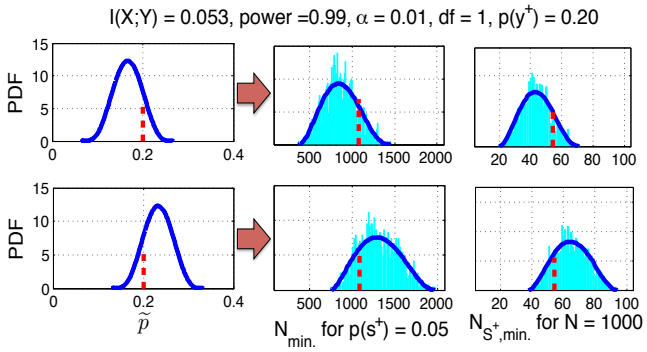


Fig. 10: A-priori power analysis under uncertain prior knowledge, when we underestimate (first row) and overestimate (second row) the prior.

7 Guidance for practitioners

Guidance for practitioners depends on the conditions in a given application. To ensure an effect would not be missed when indeed present, one should *overes-*

timate the value of $p(y^+)$, hence leading to a larger number of examples/labels being collected. Conversely, if collection of examples/labels is a costly matter, one can take a more risky, but informed decision, using less examples/labels. Furthermore, to achieve a desired statistical power, choosing to fix the amount of supervision or the sample size is application-dependent.

Under our framework we can generate tables for sample size and supervision determination under the PU constraint similar with that used in the literature, e.g. Table 1(a). For a given effect size (ω), degrees of freedom (df), significance level (α), level of desired power, prevalence ($p(y^+)$) and fixed supervision level ($p(s^+)$), in Table 1(b) we can observe the minimum sample size is needed in order to observe the effect with the given power. For the effect we followed the three levels of Cohen [6] small ($\omega = 0.10 \Leftrightarrow I(X;Y) = 0.005$), medium ($\omega = 0.30 \Leftrightarrow I(X;Y) = 0.045$) and large ($\omega = 0.50 \Leftrightarrow I(X;Y) = 0.125$).

Table 1: Sample size required for $df = 1$ and $\alpha = 0.01$.

(a) Traditional

	Effect sizes		
Power	Small	Medium	Large
0.70	962	107	39
0.80	1168	130	47
0.90	1488	166	60
0.95	1782	198	72
0.99	2404	268	97

(b) PU with $p(y^+) = 0.20$, $p(s^+) = 0.05$

	Effect sizes		
Power	Small	Medium	Large
0.70	4566	508	183
0.80	5548	617	222
0.90	7068	786	283
0.95	8462	941	339
0.99	11415	1269	457

A new type of table can be generated when we fix the sample size and we want to determine the minimum amount of supervision (or in other words, the minimum number of labelled examples) that we need in order to observe a specific effect with a desired statistical power. Table 2 presents the minimum number of labelled positive examples that we need when we have similar conditions as before but now we fix the sample size to be $N = 3000$.

Table 2: Labelled positive examples required for a PU test with $p(y^+) = 0.20$, $N = 3000$, $df = 1$ and $\alpha = 0.01$.

	Effect sizes		
Power	Small	Medium	Large
0.70	223	27	10
0.80	267	33	12
0.90	331	41	15
0.95	388	49	18
0.99	501	66	24

So in practical terms: if we assume we had 3000 examples, and we know that approximately 600 of them are positive – if we wish to detect a "medium" sized effect (in Cohen's terminology), then, in order to achieve a false negative rate of 5% (i.e. power 0.95), we only need to identify correctly 49 from those 600 examples, according to Table 2. A different way to read the results is the

following: imagine that we want to design an experiment in order to observe a medium effect with a statistical power of 80%, and the prevalence is $p(y^+) = 0.20$. If we could label both positive and negative cases, we would need 130 examples according to Table 1(a). So we would need to label 26 positives and 104 negatives. Instead of this we can use the results of Table 1(b) and collect 617 examples out of which we will label only 5%; in other words, we will label only 31 examples as positive and keep the rest as unlabelled. Thus, instead of labelling 104 negative examples, we can label 5 more positive examples and keep 586 as unlabelled. This approach can be useful when it is expensive or difficult to label examples, while it is cheaper to collect unlabelled. Since in the PU context labelling samples is expensive this methodology can be used to save resources.

Our results can be used in any research involving hypothesis testing in PU data. Our framework has been described in terms of the G-test, and the mutual information as an effect size. We can use the same framework to derive similar expressions for the χ^2-test, and the ϕ-coefficient as an effect size. Since both G and χ^2 are used extensively in behavioral sciences and biology, our work may have strong relevance in experimental design for partially supervised data [6, 11]. The proposed methods can be used in several machine learning applications. Structure learning of a Bayesian network or Markov Blanket discovery in PU data, would use our corrected G-test to decide whether we add an arc or not, since the same correction factor κ can be derived for the conditional independence test. Furthermore our power analysis methodology would provide guidance in controlling the FNR, preventing potential underfitting of the model; a recent work for the fully supervised case is Bacciu et al. [2] – our framework generalises this to PU data. Another potential application area is information theoretic feature selection. We can apply a wide variety of feature selection criteria in PU data; a recent work for fully supervised data is Brown et al. [4].

8 Conclusions and future work

In this work we developed a set of novel methodologies, enabling statistical hypothesis testing activities in PU data. We proved that a very common assumption, of all unlabelled examples being negative, is sufficient for detecting *independence*. However, a G-test using this assumption is less powerful than the fully supervised version, indicating the assumption is invalid for more complex power analysis activities. We solve this problem by deriving a *correction factor* for the test, incorporating prior knowledge from the user. Using this, we can perform sample size determination, and have a novel capability: determining *the required number of labelled examples*. Experimental evidence supports all theoretical predictions. As a future work we will investigate how our framework can be extended to fully semi-supervised data and how the principles can apply to other types of hypothesis test.

Acknowledgments. The research leading to these results has received funding from EPSRC Anyscale project EP/L000725/1 and the European Union's Sev-

enth Framework Programme (FP7/2007-2013) under grant agreement nº 318633. This work was supported by EPSRC grant [EP/I028099/1]. Sechidis gratefully acknowledges the support of the Propondis Foundation.

References

1. Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics, Wiley-Interscience, 3rd edn. (2013)
2. Bacciu, D., Etchells, T., Lisboa, P., Whittaker, J.: Efficient identification of independence networks using mutual information. Computational Statistics 28(2), 621–646 (2013)
3. Blanchard, G., Lee, G., Scott, C.: Semi-Supervised Novelty Detection. Jour. of Mach. Learn. Res. 11 (Mar 2010)
4. Brown, G., Pocock, A., Zhao, M., Lujan, M.: Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Jour. of Mach. Learn. Res. 13, 27–66 (2012)
5. Calvo, B., Larrañaga, P., Lozano, J.: Learning Bayesian classifiers from positive and unlabeled examples. Patt. Rec. Letters 28, 2375–2384 (2007)
6. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Routledge Academic (1988)
7. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning 15(2), 201–221 (1994)
8. Denis, F.: PAC learning from positive statistical queries. In: Proceedings of the 9th International Conference on Algorithmic Learning Theory (ALT). pp. 112–126. Springer-Verlag, London, UK, UK (1998)
9. Denis, F., Laurent, A., Gilleron, R., Tommasi, M.: Text classification and co-training from positive and unlabeled examples. In: International Conf. on Machine Learning, Workshop: The Continuum from Labeled to Unlabeled Data (2003)
10. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (2008)
11. Ellis, P.: The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Camb. Univ. Press (2010)
12. Gretton, A., Györfi, L.: Consistent nonparametric tests of independence. The Journal of Machine Learning Research 99, 1391–1423 (2010)
13. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.: Feature Extraction: Foundations and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
14. Hahn, G., Shapiro, S.: Statistical Models in Engineering. Wiley Series on Systems Engineering and Analysis Series, John Wiley & Sons (1967)
15. Liu, B., Lee, W., Yu, P., Li, X.: Partially supervised classification of text documents. In: International Conf. on Machine Learning. pp. 387–394 (2002)
16. Nielsen, F.G., Kooyman, M., Kensche, P., Marks, H., Stunnenberg, H., Huynen, M., et al.: The pinkthing for analysing chip profiling data in their genomic context. BMC research notes 6(1), 133 (2013)
17. Paninski, L.: Estimation of entropy and mutual information. Neural Computation 15(6), 1191–1253 (Jun 2003)
18. Sokal, R., Rohlf, F.: Biometry: The principles and practice of Statistics in Biological data. W. H. Freeman & Co., third edn. (1995)
19. Yu, H., Han, J., Chang, K.: PEBL: positive example based learning for web page classification using svm. In: SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (2002)