MANCHESTER 1824

The University of Manchester

Statistical Hypothesis Testing in Positive Unlabelled Data

Konstantinos Sechidis*, Borja Calvo**, Gavin Brown*

***School of Computer Science**

University of Manchester, UK

****Department of Computer Science and Artificial Intelligence**

University of Basque Country, Spain

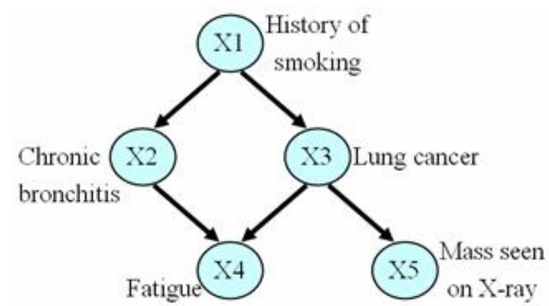
Scope of this work

We use hypothesis tests a lot.

G-test
Chi-squared test

strongly related

Mutual Information too…



e.g.

Bayesian network structure learning

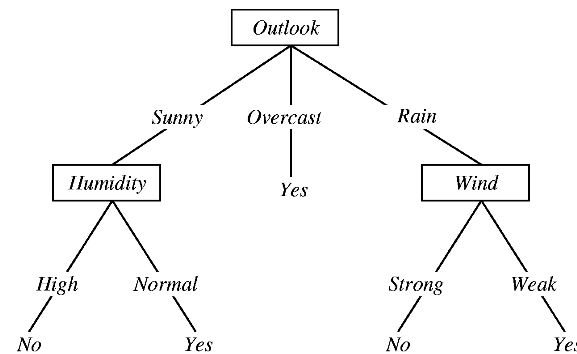
- should there be an arc between node X and Y?

Feature Selection

- should we select feature X?

Decision trees

- should we split on feature X?



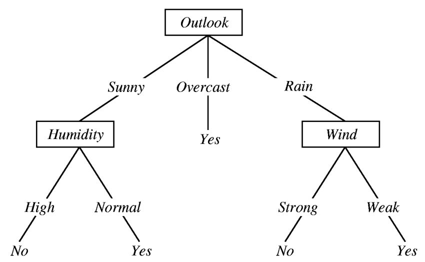
Contributions of this work

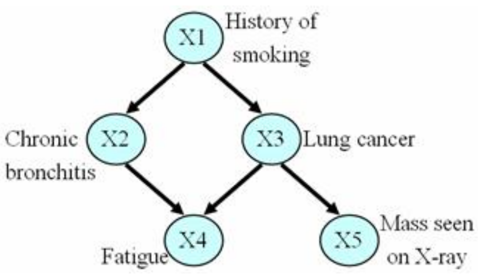
Explores the dynamics of hypothesis testing in "positive-unlabelled" data.

Special case of Semi-supervised data

1. Can we perform a valid hypothesis test in this situation?
2. Sample size determination: how many examples do I need?
3. "Supervision determination": how many labelled examples do I need?

NEW





“Positive-Unlabelled” data

Fully labelled

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

Semi supervised

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

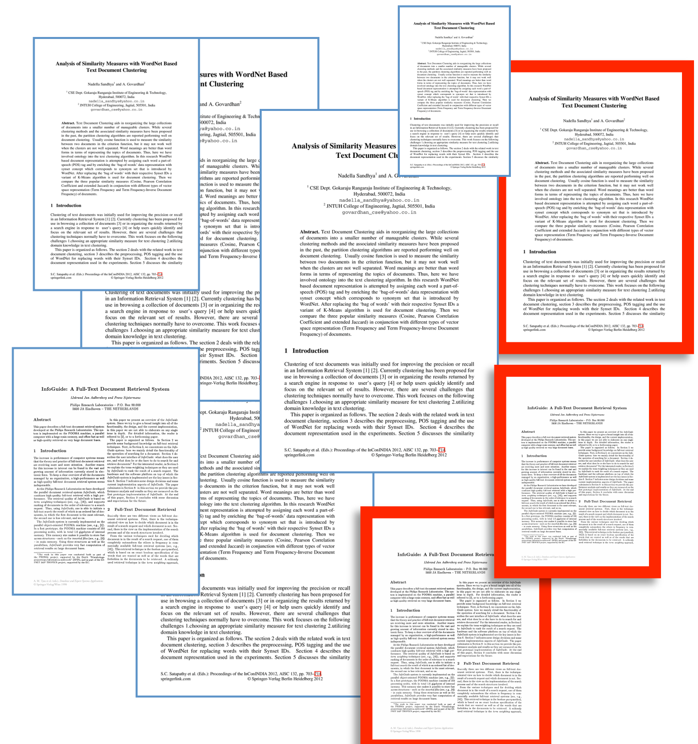
Positive Unlabelled

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

PU data applications

Many many applications…

Text Mining



Functional Genomics

“Gene 23 is associated with the disease.”

“The other ones… we don’t know.”

Background: Hypothesis Tests

Step 1.

Calculate G-statistic…

$$G = 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} O_{x,y} \ln \frac{O_{x,y}}{E_{x,y}}$$

Step 2.

Compare to critical value, obtained from lookup table.

Tabulated values of cumulative chi-squared distribution

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

if G > critical value…

REJECT NULL HYPOTHESIS

i.e. assume X and Y are related

else

ACCEPT NULL HYPOTHESIS

i.e. assume X and Y are INDEPENDENT

Hypothesis Testing in PU data?

<u>Step 1.</u>

Calculate G-statistic…

$$G = 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} O_{x,y} \ln \frac{O_{x,y}}{E_{x,y}}$$

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

S
1
1
1
0
0
0
0
0
0
0

Assume all negative?

S = "surrogate"

Then use test for G(X;S) instead of G(X;Y) …..

Selected completely at random assumption

Positive examples labelled with probability $p(s^+|y^+)$

Lemma 1

$$p(x|y^+) = p(x|s^+)$$

Theorem 1

$$X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp S$$

Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (2008)

Theorem 1 implies…

Testing with the surrogate G(X;S) will have an
identical FALSE POSITIVE rate
to the original (unobservable) test.

However….

Testing for G(X;S) will have a higher **FALSE NEGATIVE** rate.

That is… If X;Y are truly related,
…… we may **miss** that dependency by using the surrogate test G(X;S).

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

S
1
1
1
0
0
0
0
0
0
0

Technical explanation for this….

Theorem 2

If $X \not\perp\!\!\!\perp Y$ then $I(X;Y) > I(X;S)$

Contributions of this work

1. Can we perform a valid hypothesis test in this situation? ✓

2. **Sample size determination: how many examples do I need?**

3. "Supervision determination": how many labelled examples do I need?

Sample Size Determination

~~G(X;Y)~~ G(X;S)

I want my statistical test to detect an "effect" as small as

$I(X;Y) = 0.053$

and, have

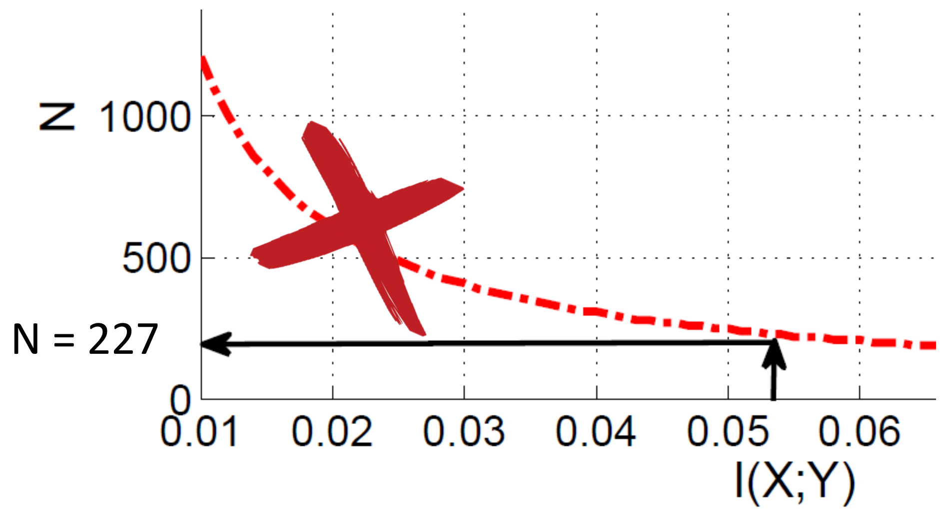
.....a FALSE POSITIVE rate of <u>0.01</u>

.....a FALSE NEGATIVE rate of <u>0.01</u>

How many examples (N) do I need?

X	Y
2	1
3	1
1	1
2	1
1	1
3	0
1	0
3	0
3	0
2	0

S
1
1
1
0
0
0
0
0
0
0



Standard sample size determination does not apply....

So, we derive a correction factor, kappa....

The non-centrality parameter of the test:

Probability of p(s=1) i.e. probability of being labelled.

Theorem 3

$$\lambda_{G(X;S)} = \kappa 2NI(X;Y), \kappa = \frac{1-p(y^+)}{p(y^+)} \frac{p(s^+)}{1-p(s^+)}$$

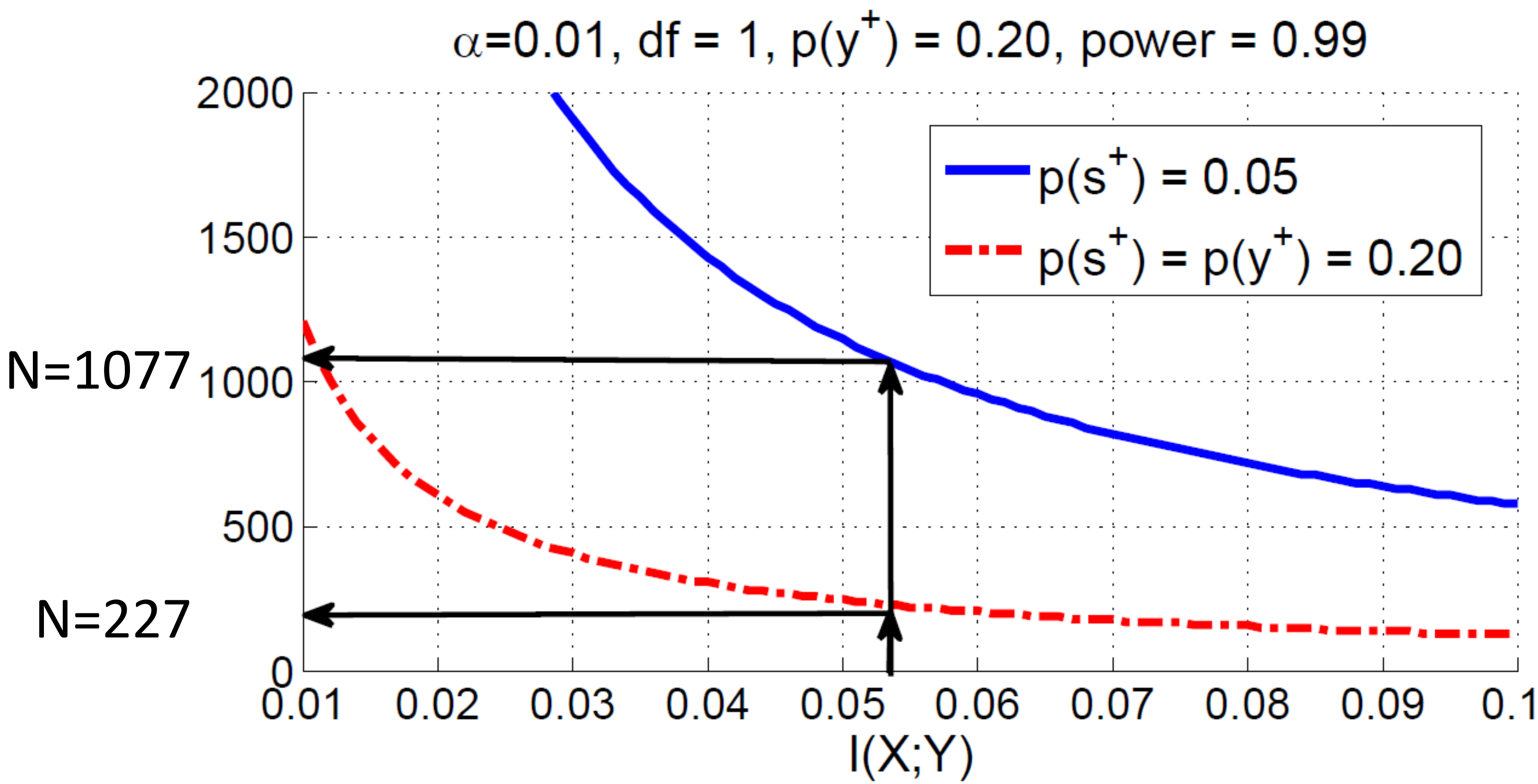
Prior probability of p(y=1)

G(X;S) test with sample size N/k

....

....obtains identical FPR and FNR to the (<u>unobservable</u>) test G(X;Y)

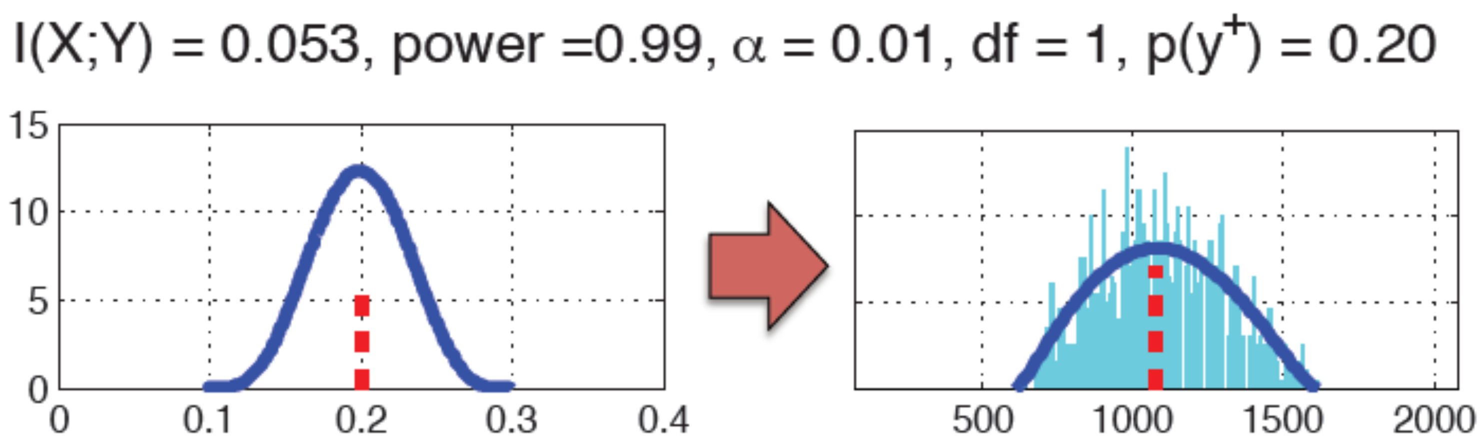
PU sample size determination… assuming we know p(y=1)



PU sample size determination… uncertainty over p(y=1)

Place Beta distribution over p(y=1), use Monte Carlo simulation to determine posterior over sample size N.

Analytical solution seems to be intractable.



Belief over p(y=1)

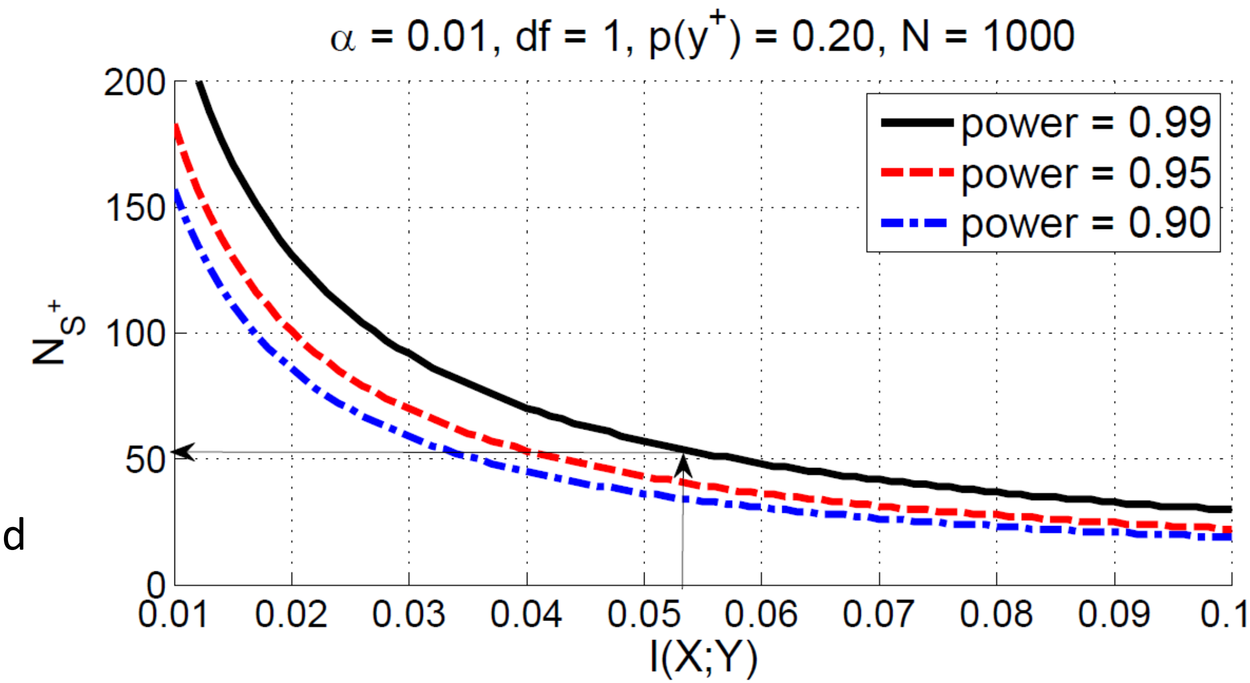
Minimum N for specified test when p(s=1) = 0.05

Contributions of this work

1. Can we perform a valid hypothesis test in this situation? ✓
2. Sample size determination: how many examples do I need? ✓
3. **"Supervision determination": how many labelled examples do I need?**

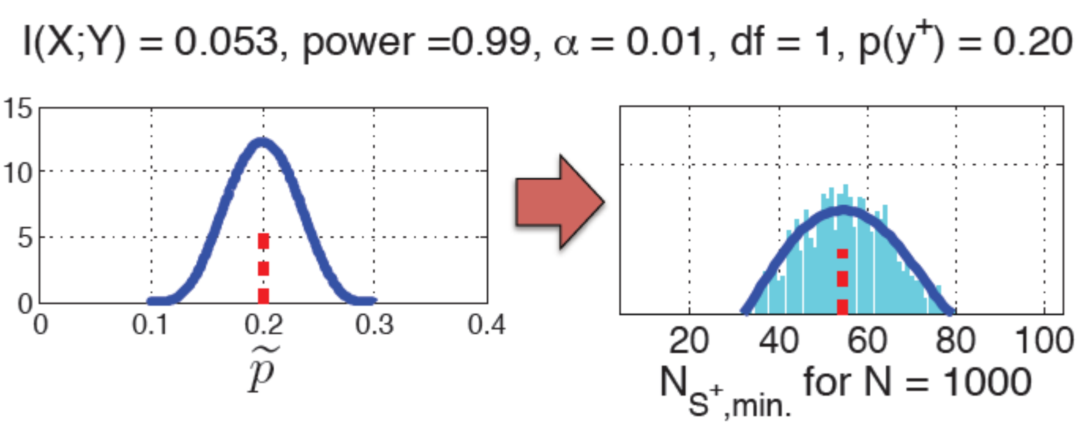
PU "supervision determination"

Use similar methodology to before, except we fix N, and solve for p(s+)



From the total N=1000,
…you need only 57 labels.
…the rest can be unlabelled

If uncertain p(y=1) ….



Example for Practitioner
Supervision determination

Design an experiment to observe …

… a medium effect of I(X;Y) = 0.045 nats
… with False Positive Rate = 0.01
… with False Negative Rate = 0.05
…. having N = 3000 examples

Positive Unlabelled
knowing prior to be p(y=1) = 0.20

3000 examples

- **49 positives**
- **2951 unlabelled**

Conclusions

1. We perform a valid hypothesis test in PU data by unlabelled as negatives. ✓
2. Sample size determination: How many examples we need to collect. ✓
3. "Supervision determination": How many labelled examples we need to collect. ✓

Thank you for your attention!

www.cs.man.ac.uk/~gbrown/posunlabelled/

Example for Practitioner
Sample size determination

Design an experiment to observe …

… a medium effect of I(X;Y) = 0.045 nats
… with False Positive Rate = 0.01
… with False Negative Rate = 0.20

Traditional, fully supervised

130 labelled examples

Positive Unlabelled, with labelling 5%
… and knowing prior to be p(y=1) = 0.20

617 examples

- **31 positives**
- **586 unlabelled**