CONCLUSIONS

- Gold standards in classification are often more like pyrite standards
- We have created a method to quantify the uncertainty caused by not having a gold standard
- Method only requires a quantification of how good the pyrite standard is
- Method has statistical performance

29TH EUROPEAN SAFETY AND RELIABILITY CONFERENCE
HANNOVER, GERMANY

Uncertainty in Classification Without Gold Standards

NICK GRAY
INSTITUTE FOR RISK AND UNCERTAINTY, UNIVERSITY OF LIVERPOOL
NICKGRAY@LIV.AC.UK

Introduction

Classifications occur in many different fields from patient diagnosis in medicine, machine learning in computer science to structural health in engineering. However, naively interpreting the results of a test could be misleading, they are often imperfect yielding false positives and negatives. It not uncommon for the majority of positive tests to be false positives. For example under 10% of positive breast cancer test are true positives. [1]

In order to understand how good a test is a set of sample objects must be used for which the true classification has been established using some gold standard test. In medicine, biopsies and autopsies are often the gold standard test whereas in machine learning, humans are often used to provide manually labeled data. However, gold standards are often not as gold as they seem, they are more like pyrite standards (*fool's gold*). For example, if the gold standard test for a particular disease is a biopsy then it is possible that the affected tissue has been missed during surgery or the histopathologists may disagree on the true classification. Humans are notorious for making mistakes, in *prove you are a human* tests that appear on websites, most of which are used to provide training data for machine learning purposes, humans make errors approximately 1% of the time. [2]

Although sometimes multiple tests can be combined together in order to reduce the uncertainty about a diagnosis. [3] It may also be the case that there is simply no good gold standard possible, or it may also be impractical or prohibitively expensive to perform, or simply non-existent forcing the use of a pyrite standard.

Test Statistics

A confusion matrix can be used in order to tabulate the results of trial set of data, Table 1 shows an example confusion matrix.

	GOLD POSITIVE	GOLD NEGATIVE
TEST POSITIVE	a	b
TEST NEGATIVE	c	d

Table 1: An example confusion matrix

From this statistics about the new test can be calculated. The sensitivity is given by

$$s = \frac{a}{a+b}. \tag{1}$$

the specificity by

$$t = \frac{d}{b+d}. \tag{2}$$

The prevalence *could* be calculated as

$$p = \frac{a+c}{a+b+c+d}, \tag{3}$$

however this depends of the sample tested being representative of the overall population, in a medical context this is often not the case. The positive predicted value (PPV), the probability that the result is a true positive after a positive test, can be calculated using Bayes' rule

$$PPV = \frac{ps}{ps + (1-p)(1-t)}. \tag{4}$$

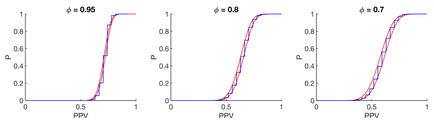

Figure 1: Plot of the gilding c-box (red) against the empirical distribution of simulated trials

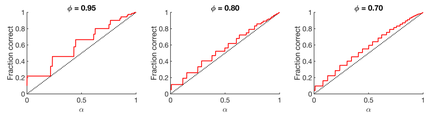

Figure 2: Demonstration that the gilding method maintains confidence interpretation. The red line is a plot of the fraction that the method got correct against α value. The black line is Fraction Correct = α

Uncertainty

If we let, ϕ be a representation of how good the gold standard test is, then we can consider at what ϕ values the uncertainty is at its greatest and smallest values

- **When $\phi = 1$, the Pyrite standard is perfect (i.e Gold):**
  - Every positive/negative is a true positive/negative, there is *no uncertainty*
- **When $\phi = 0$, the Pyrite standard is perfectly wrong:**
  - Every positive/negative is a false positive/negative, there is *no uncertainty*
- **When $\phi = 0.5$, the Pyrite standard has no predictive value:**
  - No way of knowing if a positive/negative is a true positive/negative, this is where there is *maximum uncertainty*

The value of ϕ could be chosen in several different ways, it could be based on the accuracy of the pyrite standard, the PPV or NPV, or if there is no true gold standard to compare it to then simple heuristics could be used instead.

Gilding

We have developed a new method in order to take account of information when there is no gold standard. Instead of using the value from the basic confusion matrix we instead *gild* the confusion matrix by transforming the values into confidence boxes (c-boxes) using the measure of how good the gold standard is, ϕ. The use of c-boxes ensures that the method characterises the inferential uncertainty of the sample size

If we start with the confusion matrix in Table 1 but instead of using a true gold standard instead we use a pyrite standard for reference, then we can replace the values with c-boxes that have been created using Equations (5) and (6) as shown in Table 2

$$f(x,y,\phi) = (x+y)\left[\text{beta}\left(g(x,y,\phi), g(y,x,\phi)+1\right), \text{beta}\left(g(x,y,\phi)+1, g(y,x,\phi)\right)\right] \tag{5}$$

$$g(x,y,\phi) = (x\phi + y(1-\phi))\sqrt{\frac{|\phi - 0.5|}{\phi(1-\phi)}} \tag{6}$$

	GILDED POSITIVE	GILDED NEGATIVE
TEST POSITIVE	$f(a,b,\phi)$	$f(b,a,\phi)$
TEST NEGATIVE	$f(c,d,\phi)$	$f(d,c,\phi)$

Table 2: Gilded confusion matrix

Once we have made these replacements then it is possible to use equations (1), (2) and (4) in order to calculate the values of s and t and subsequently the *PPV*. Although care must be taken due to the fact that dependance has been added within the confusion matrix values.

Figure 1 shows that the PPV calculated using the gilding method closely matches PPV's calculated by simulating the effect of using a pyrite standard with accuracy ϕ. Using Figure 2 we can also see that gilding has the confidence interpretation, that among all confidence intervals computed by the same method, a proportion α will contain the true value, for different values of ϕ.

References

[1] Gigerenzer, G. (2011) 'What are natural frequencies?', BMJ (Online), 343(7829), pp. 1–2. doi: 10.1136/bmj.d6386.
[2] von Ahn, L. et al. (2008) 'reCAPTCHA : Human-Based Recognition via Web Character', Science, 321(5895), pp. 1465–1468.
[3] Joseph, L., Gyorkos, T. W. and Coupal, L. (1995) 'Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard', American Journal of Epidemiology, 141(3), pp. 263–272. doi: 10.1093/oxfordjournals.aje.a117428.
[4] Ferson, S. et al. (2013) 'Computing with Confidence', in Proceedings of the Eighth International Symposium on Imprecise Probability: Theory and Applications. Compiègne, France. doi: doi:10.1103/PhysRevLett.113.266406.

UNIVERSITY OF LIVERPOOL | Institute for Risk and Uncertainty

SCAN FOR MORE INFORMATION