Calculating a boundary for the Significance from the Equal-Error Rate

Heinz Hofbauer • Andreas Uhl

University of Salzburg, Department of Computer Sciences

UNIVERSITÄT SALZBURG
WAVELAB
Multimedia Signal Processing and Security Lab

Abstract

Given a common dataset, two methods operating on that dataset and reported equal-error rate (EER) for each method, then we can estimate whether the two methods differ significantly at the threshold leading to the EER. This enables the calculation of a boundary on the significance for methods where the significance was not reported in the original paper or to compare new methods to older ones by evaluating them on the same dataset.

Main Results

- Given a common dataset a bound for the EER can be calculated.
- The bound is coarse and, if possible, a proper significance analysis is preferable.
- However, the bound allows for:
  - Multiple comparisons at once (faster).
  - Comparison with other work when the implementation is not available.

General Remarks

- It should be noted that this work is not restricted to biometric research.
- There are common pitfalls when using significance tests (choice of critical values, etc.). Read

  S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–327, Sep. 1997

  where common pitfalls and recommended statistical methods for data mining are discussed. Especially the tutorial regarding critical values and the multiplicity effect.

Classification and McNemar

Given:

- A set of data $\mathcal{D}$.
- A dichotomous trait $\mathcal{T}$ over the elements of $\mathcal{D}$
- Two classification methods A and B, applying a dichotomous trait, T_A and T_B respectively, aiming to approximate $\mathcal{T}$.

The question then is:

- Given T_A and T_B, is the difference between method A and method B significant?

The Answer comes in the form of the McNemar test. Given the table

Method B \ Method A	C	W	
C	a	b	$a+b$
W	c	d	$c+d$
	$a+c$	$b+d$	N

we split the results, of T_A and T_B, into correctly (C) and wrongfully (W) classified, according to $\mathcal{T}$:

$$a = |\{x | x \in \mathcal{D} \wedge T_A(x) = \mathcal{T}(x) \wedge T_B(x) = \mathcal{T}(x)\}|, \quad (1)$$
$$b = |\{x | x \in \mathcal{D} \wedge T_A(x) \neq \mathcal{T}(x) \wedge T_B(x) = \mathcal{T}(x)\}|, \quad (2)$$
$$c = |\{x | x \in \mathcal{D} \wedge T_A(x) = \mathcal{T}(x) \wedge T_B(x) \neq \mathcal{T}(x)\}|, \quad (3)$$
$$d = |\{x | x \in \mathcal{D} \wedge T_A(x) \neq \mathcal{T}(x) \wedge T_B(x) \neq \mathcal{T}(x)\}|, \quad (4)$$

and $N = |\mathcal{D}|$.

- The McNemar test looks at the change between methods A and B, that is entries b and c in the table.
- If method A and B are similar then c and b should be distributed based on a coin flip ($\mathcal{B}(b+c, 0.5)$).
- For large $n = b + c$ we can instead use $\mathcal{B}(b+c, 0.5) \sim \mathcal{N}(np; np(1-p)) = \mathcal{N}(\frac{b+c}{2}; \frac{b+c}{2^2})$, or we can write:

$$\frac{X - \frac{b+c}{2}}{\sqrt{\frac{b+c}{2^2}}} = \frac{2X - (b+c)}{\sqrt{b+c}} \sim \mathcal{N}(0,1)$$

- We can then use this to apply a χ^2 test with one variable $X = b$ (one degree of freedom):

$$\frac{2b - (b+c)^2}{\sqrt{b+c}} = \frac{(b-c)^2}{b+c} \sim \mathrm{X}^2(1)$$

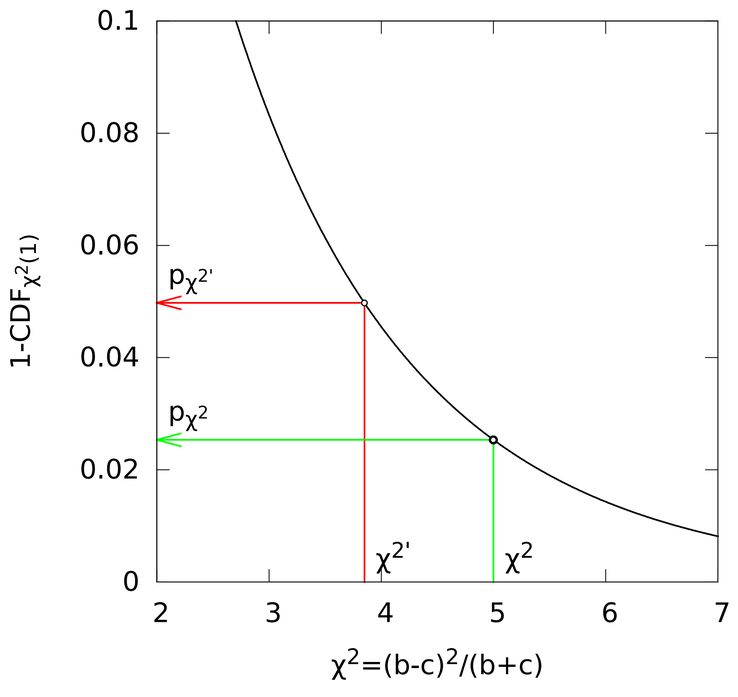

Figure 1: Plot of $p_v = 1 - \mathrm{cdf}(\chi^2)$ over χ^2 and direction of estimation.

Estimating χ^2

Given:

- A dataset $\mathcal{D}$ with known cardinality $N = |\mathcal{D}|$.
- Two methods A and B and report equal-error rates EER_A and EER_B.

The EER is the error rate where false non-match rate and false match rate are equal: That is a total of $EER \times N$ elements of $\mathcal{D}$ are wrongfully classified:

- $b + d = EER_A N$.
- $c + d = EER_B N$.

We have to estimate

$$\chi^2 = \frac{(b-c)^2}{b+c}.$$

or more specifically:

- $b - c$ (see below)
- $b + c$ (see below)

which results in

$$\chi^{2'} = \frac{(EER_A - EER_B)^2 N}{EER_A + EER_B}. \quad (5)$$

Estimating $b - c$

This one is straightforward and follows directly from EER_A and EER_B:

$$b - c = b + d - d - c = (b+d) - (c+d) = (EER_A - EER_B)N.$$

Estimating $b + c$

This we can not directly calculate since we do not know the ratios of $b : d$ and $c : d$.

Lemma 1. *For the Chi-squared distribution the p-value of $\chi^{2'}$ is $p' = 1 - \mathrm{cdf}(\chi^{2'})$ and $\forall \chi^2 : \chi^2 \geq \chi^{2'} \implies p \leq p'$ with $p = 1 - \mathrm{cdf}(\chi^2)$.*

Proof. Since the cdf is monotonic increasing and $\chi^2 \geq \chi^{2'}$ we know that $\mathrm{cdf}(\chi^2) \geq \mathrm{cdf}(\chi^{2'})$. The codomain of the cdf is $[0:1]$, thus $1 - \mathrm{cdf}(\chi^2) \leq 1 - \mathrm{cdf}(\chi^{2'})$. □

See also Fig. 1.

- If we minimize our estimation $\chi^{2'}$ then any realization of b, c and d will at least be as significant as the estimation.
- In essence, the p-value calculated based on this estimate is a upper boundary for the real p-value.
- To minimize $\chi^{2'}$ we have to maximize $b + c$.

Assumption:

- As a simplification let us assume that $EER_A + EER_B \leq 1$.

This allows to maximize $b + c$ by assuming no overlap between $c + d$ and $b + d$, i.e. $d = 0$ and

$$b + c = (EER_A + EER_B)N.$$

Example and Coarseness Of the Bound

- The estimation of the $\chi^{2'}$ is an upper bound.
- This means it can falsely reject a significant difference!

This example is from a paper about segmentation fusion:

P. Wild, H. Hofbauer, J. Ferryman, and A. Uhl, "Segmentation-level fusion for iris recognition," in *Proceedings of the International Conference of the Biometrics Special Interest Group*, Sep. 2015, p. 12

Example 1. The evaluation is based on the Casia v4–Interval database with reported error rates

- $EER_A = 1.0426\%$
- $EER_B = 1.0317\%$.

When we use equation (5) we get:

$$\chi^{2'} = \frac{(0.010426 - 0.010317)^2 3480841}{0.010427 + 0.010317} = 2.02668,$$
$$p'_V = 1 - \mathrm{cdf}(2.02668) = 0.1546 \approx 15.5\%,$$

which would suggest that differences are by chance and thus not significant.

Running the actual McNemar test on the data gives:

- $b = 26055$, $c = 26707$ (and $d = 10198$)
- $\chi^2 = 8.032$ and $p_V = 0.00459 \approx 0.46\%$

which indicates that the difference is indeed significant.

Multiple Comparisons

For multiple comparisons we can either use Eq. (5) multiple times or we can look at differences:

- What minimum EER difference is necessary, such that the difference between method A and B is significant with at least a p-value of p_V for all realizations.
- That is, what $\Delta EER = |EER_A - EER_B|$ is required for a given p_V (and critical $\chi^{2^*}{}_V = \mathrm{ppf}(1 - p_V)$).

Using Eq.(5) we can replace:

- $EER_A - EER_B$ with ΔEER
- $EER_A + EER_B$ with $2EER_M = 2\max(EER_A, EER_B)$ (Lemma 1),

resulting in

$$\chi^{2^*}{}_V = N\frac{\Delta EER^2}{2EER_M} < N\frac{\Delta EER^2}{EER_A + EER_B}. \quad (6)$$

Then we directly get

$$\Delta EER = +\sqrt{\frac{2\chi^{2^*}{}_V}{N} EER_M}. \quad (7)$$

Directly from Fig. 1 we can see that if we use

- $\Delta EER' \geq \Delta EER$, and/or
- $EER'_M \leq EER_M$

then $p'_V \leq p_V$.

Example of Direct and Maximum Comparisons

Frequently authors give a list of results to compare to, examples from Bastys *et al.*

A. Bastys, J. Kranauskas, and R. Masiulis, "Iris recognition by local extremum points of multiscale taylor expansion," *Pattern Recognition*, vol. 42, no. 9, pp. 1869–1877, 2009

Direct Comparison

Example 2 (Single Comparison)**.** The evaluation is based on the Casia v2–Device 1 database with reported error rates

- $EER_A = 0.13\%$
- $EER_B = 0.58\%$

They did not do a significance analysis, so let us do it here. There are only two EERs to compare so we use Eq. (5) for the best result. With $N = 719400$, based on the database, this gives us

- $\chi^2 = 2051$ and
- $p_v < 10^{-6}$,

a significant difference.

Multiple Comparisons

From the same paper we give this example to highlight the use of EER_M to simplify multiple comparisons.

Example 3 (Multiple Comparison)**.** The evaluation is based on the Casia v1.0 with reported methods and error rates

- TAN (0.57%)
- DAUGMAN (0.08%)
- MA (0.07%)
- YAO (0.28%)
- BASTYS (0.00%)

In this case the ΔEER from Eq. (7) is more useful (one calculation, multiple comparisons):
Setting

- $p_v = 1\%$ and
- $EER_M = 0.0058$ (we don't know the exact number for TAN).

For Casia v1.0, $N = 285390$, we can calculate

$$\Delta EER = 0.052\%.$$

From this bound all differences, except between MA and DAUGMAN are significant to the given p_v.

For the difference between MA and DAUGMAN we can use Eq. (5) to take a closer look, resulting in

- $\chi^2 = 1.9026$ and $p_v = 16.78\%$.

Without the implementations we can not perform a proper significance analysis so we have to assume that the difference is not significant.