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

Heinz Hofbauer[1] • Andreas Uhl[1]

[1]University of Salzburg, Department of Computer Sciences, {hhofbaue, uhl}@cosy.sbg.ac.at

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.

Contents

1 Introduction **2**

2 Classification and McNemar **2**

3 Estimating χ^2 from the Equal-Error Rate **2**

4 Some Practical Examples **3**

4.1 Minimum Equal-Error Rate Difference for Significance 3

4.2 Regarding the Coarseness Of the Bound 3

4.3 On the Usage of Direct and Maximum Comparisons 3

5 Conclusion **4**

1 Introduction

When a field of research is young the improvements are usually large and it is quite clear when an algorithm is better. As the field gets more mature improvements are achieved in smaller and smaller increments and the question of significance invariably arises. In recent years reviewers in biometrics based recognition frequently call for significance tests, and rightfully so.

Various statistics tests have been used, e.g., t-test in [1] or the McNemar test [2]. However, these tests require the underlying data and methods to be available. For the t-test the algorithm has to be run on a number of partitions of the dataset, for the McNemar test correctness of individual comparisons must be known. This presents a difficulty when comparing with older work, even in the rare cases where an implementation is available there is still an overhead in evaluation since all algorithms have to be rerun in oder to calculate a significance level.

It should be noted that this phenomenon is not one restricted to biometric research. It should also be noted that there are common pitfalls when using significance tests, especially the choice of critical values is often too lax. As an example see [3] where common pitfalls and recommended statistical methods for data mining are discussed. Especially the tutorial regarding critical values and the multiplicity effect [3, Sec. 3] are also applicable to biometric significance testing.

In this paper we will present an estimation of significance based on EER rates when two methods are evaluated on the same dataset, i.e., the experimental basis is the same. This not only allows for comparison with methods for which only the reported EER are available (but not an implementation of the algorithm), but also between two already published methods, given they were evaluated on the same dataset. Note that while we view the problem from the perspective of biometric recognition the assumptions are not specific to this field.

2 Classification and McNemar

Given a set of data $\mathcal{D}$ with a dichotomous trait $\mathcal{T}$. Two classification methods A and B also apply a dichotomous trait, T_A and T_B respectively, with the aim to approximate $\mathcal{T}$. Given T_A and T_B, is the difference between method A and method B significant?

To answer this question the McNemar test [4] can be used. Given table 1 we split the results into correctly (C) and wrongfully (W) classified. This means:

Table 1: The table for the McNemar test to compare correctly and incorrectly classified values between Methods A and B.

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

$$a = |\{x | x \in \mathcal{D} \wedge T_A(x) = \mathcal{T}(x) \wedge T_B(x) = \mathcal{T}(x)\}|, \quad (1)$$

$$b = |\{x | x \in \mathcal{D} \wedge T_A(x) \neq \mathcal{T}(x) \wedge T_B(x) = \mathcal{T}(x)\}|, \quad (2)$$

$$c = |\{x | x \in \mathcal{D} \wedge T_A(x) = \mathcal{T}(x) \wedge T_B(x) \neq \mathcal{T}(x)\}|, \quad (3)$$

$$d = |\{x | x \in \mathcal{D} \wedge T_A(x) \neq \mathcal{T}(x) \wedge T_B(x) \neq \mathcal{T}(x)\}|, \quad (4)$$

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, i.e. binomial with $p = q = 0.5$. Usually the Chi-squared approximation is used for the McNemar test, although for smaller $b+c$ an exact test can also be used. In the following we will use the χ^2 approximation without any continuity correction. The test statistic is then

$$\chi^2 = \frac{(b-c)^2}{b+c}. \quad (5)$$

Remark 2.1 (On the dataset). The dataset for the test is not comprised of the individuals of a database but rather the comparisons. The trait $\mathcal{T}$ is 'genuine comparison' and 'imposter comparison'. This requires to know the number of actual comparisons done in an experiment rather than the total number of individuals in the underlying database.

3 Estimating χ^2 from the Equal-Error Rate

Assuming on a dataset $\mathcal{D}$, for which we know the cardinality $N = |\mathcal{D}|$, two methods A and B report their respective equal-error rates EER_A and EER_B. In order to find whether or not the difference between A and B is significant we have to estimate Eq. (5).

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. In terms of the McNemar-test we can state

$$b + d = EER_A N \text{ and} \quad (6)$$

$$c + d = EER_B N. \quad (7)$$

In order to calculate χ^2 we have to calculate $b-c$ and $b+c$. One is easy: $b - c = b + d - d - c = (b+d) - (c+d) = (EER_A - EER_B)N$. The other we have to estimate since we do not know the ratios of $b : d$ and $c : d$.

Lemma 3.1. *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 $1 - \mathrm{cdf}(\chi^2) \leq 1 - \mathrm{cdf}(\chi^{2'})$. □

Consequently, 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 $b + c$. As a simplification let us assume that $EER_A + EER_B \leq 1$ which then allows

to maximize $b + c$ by assuming no overlap between $c + d$ and $b + d$, i.e. $d = 0$ and $b + c = (EER_A + EER_B)N$, resulting in

$$\chi^{2'} = \frac{(EER_A - EER_B)^2 N}{EER_A + EER_B}. \tag{8}$$

Remark 3.2. This is the lowest upper bound for p_V since $d = 0$ is possible (under the assumptions).

Now we come to the question which 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. We are interested in $\Delta EER = |EER_A - EER_B|$, and for a given p_V the critical ${\chi^{2^*}}_V = \text{ppf}(1 - p_V)$ can be calculated by using the percent point function, i.e. the inverse cdf.

From

$${\chi^{2^*}}_V = N\frac{\Delta EER^2}{2EER_M} < N\frac{\Delta EER^2}{EER_A + EER_B}, \tag{9}$$

with $EER_M = \max(EER_A, EER_B)$, we can calculate

$$\Delta EER = +\sqrt{\frac{2{\chi^{2^*}}_V}{N} EER_M}. \tag{10}$$

Remark 3.3. If $\Delta EER' \geq \Delta EER$ then ${\chi^{2^{*'}}}_V \geq {\chi^{2^*}}_V$ and consequently $p'_V = 1 - \text{cdf}(\chi^{2^{*'}}) \leq p_V$.

Remark 3.4. If $EER'_M \leq EER_M$ then ${\chi^{2^{*'}}}_V \geq {\chi^{2^*}}_V$ and consequently $p'_V = 1 - \text{cdf}(\chi^{2^{*'}}) \leq p_V$.

4 Some Practical Examples

In the following we provide examples to foster a better understanding of the bound.

The first is an example of a minimum ΔEER required for significance which lets us quickly screen a large number of tested algorithms for improvement. This example highlights that the required difference in EER can become quite small if the used dataset is big enough.

The second example examines the coarseness of this upper bound. From this example it will become clear that the estimation is rather coarse. This means that a proper significance analysis is always preferable, and the estimation should only be used when this is not possible, e.g., due to implementations not being available.

The following examples are based on the authors work with iris biometry, the methods described so far are however not limited to iris biometry.

4.1 Minimum Equal-Error Rate Difference for Significance

Within our assumptions, if we set $EER_M = 0.5$, which should really cover all reasonable cases, and given well known biometric databases, Casia v4–Interval [5] with $N_{C4} = 3480841$ and IIT Delhi [6] with $N_I = 2507680$, the EER differences in Table 2 would be significant with at least the given p-value.

One should note that this is a worst case scenario, and a smaller EER_M would lead to a smaller ΔEER as shown in Figure 1.

Table 2: Minimum difference in EER for the given significance levels on the given databases.

		$p_V = 0.05$	$p_V = 0.01$
		${\chi^{2^*}}_{0.05} = 3.84$	${\chi^{2^*}}_{0.01} = 6.64$
ΔEER	Casia v4I	$\approx 0.106\%$	$\approx 0.139\%$
	IIT Delhi	$\approx 0.124\%$	$\approx 0.163\%$

Reasonable methods for iris biometry are below the $EER_M = 4\%$ level. So a $\Delta EER_{C4} \approx 0.04\%$ reported on the Casia V4–Interval database would result in a 99% significance level. Likewise would a $\Delta EER_I \approx 0.05\%$, for IIT Delhi, result in a 99% significance level.

4.2 Regarding the Coarseness Of the Bound

The estimation of the $\chi^{2'}$ was done as an upper bound, this means in practice that the estimation is coarse and can falsely reject a significant difference. From a paper about iris segmentation fusion [2] for which we also have the underlying algorithms and data, allowing us to calculate the real χ^2, we take the following example.

Example 4.1. The evaluation is based on the Casia v4–Interval database with reported error rates for the algorithms OSIRIS (1.042698%), CAHT (1.224288%) and the fusion (1.031708%).

From the graph with $EER_M = 1.3\%$ we would get $\Delta EER_{p_V=0.05} \approx 0.02\%$. Clearly the fusion is better than CAHT, however with this coarse boundary the difference to OSIRIS is not significant.

When we use the better evaluation (8) we get:

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

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

However, running the actual McNemar test on the data gives the following result: $b = 26055$, $c = 26707$ (and $d = 10198$) resulting in $\chi^2 = 8.032$ and $p_V = 0.00459 \approx 0.46\%$. With this we see that the difference is actually significant and the fusion is a real improvement over both OSIRIS and CAHT.

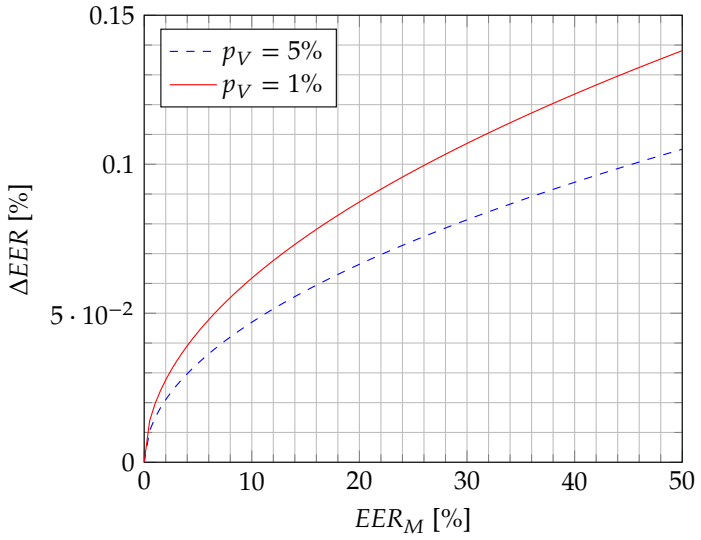
4.3 On the Usage of Direct and Maximum Comparisons

Frequently authors give a list of results to compare to, as an example take Bastys *et al.* [7].

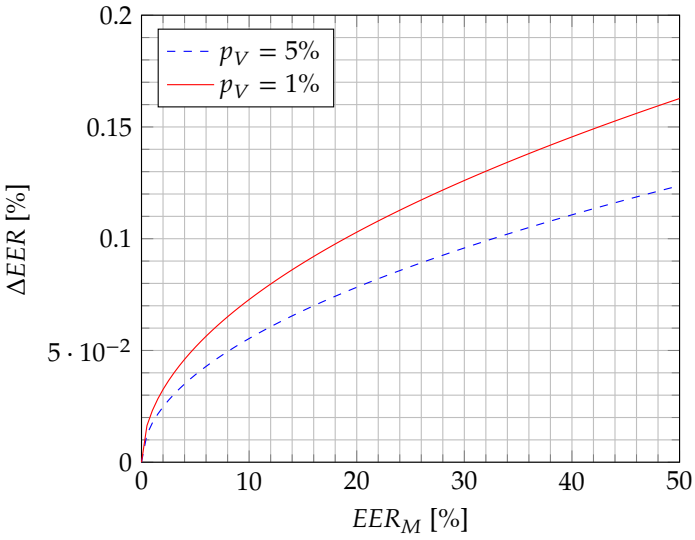
Example 4.2 (Single Comparison)**.** Bastys *et al.* lists EERs for their method (0.13%) and another method from literature (EER=0.58%) on the Casia V2.0 (device 1) database, but do not give a significance analysis.

In this case, where only two methods are compared we can directly use Eq. (8) for the best result. With $N = 719400$, $EER_A = 0.0013$ and $EER_B = 0.0058$ this gives us $\chi^2 = 2051$ and $p_v < 10^{-6}$. A significant difference.

Subsequently we give another example from the same paper to highlight the use of EER_M to simplify multiple comparisons.



(a) Casia v4–Interval $N_{C4} = 3480841$



(b) IIT Delhi $N_I = 2507680$

Figure 1: Required difference in EER, from a maximum EER, too achieve significance to the given level.

Example 4.3 (Multiple Comparison)**.** Bastys *et al.* report the following EERs on the CASIA v1.0 database: TAN (0.57%), DAUGMAN (0.08%), MA (0.07%) and YAO (0.28%) and their own (0.00%). In this case the ΔEER from Eq. (10) 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.

5 Conclusion

We have shown how to calculate a boundary on the significance for a method based on reported EER rates acquired over the same dataset. While the estimation is coarse, and proper significance analysis is always preferable, this methods gives us a tool for comparing two, or more, methods when an actual significance analysis is not possible.

The most important application is that Eq. (8) and Eq. (10) allow us to compare a new method with a method from literature for which we do not have an implementation. Using the bounds presented in this paper we 'only' need to repeat the same experiment on the same dataset with the new method and based on the EERs can calculate a bound on the significance.

Acknowledgements

This work was partially supported by the Austrian Science Fund, project no. P27776.

References

[1] K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn, "Improved iris recognition through fusion of hamming distance and fragile bit distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2465–2476, 2011 (cit. on p. 2).

[2] 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 (cit. on pp. 2, 3).

[3] 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 (cit. on p. 2).

[4] Q. McNemar, "Note on the sampling error of the difference between correlated proportions of percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947 (cit. on p. 2).

[5] Chinese Academy of Sciences' Institute of Automation, *Casia iris image database v4.0 — interval*, `http://biometrics.idealtest.org`, 2002 (cit. on p. 3).

[6] IIT Delhi, *Iris database (version 1.0)*, `http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Iris.htm`, 2007 (cit. on p. 3).

[7] 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 (cit. on p. 3).