COMPARING SOFT CLASSIFICATION METHODS FOR THE RARE TYPE MATCH PROBLEM

Giulia Cereda [1], Fabio Corradi[1] and Cecilia Viscardi[1]

[1] Department of Statistics, Computer Science and Applications, University of Firenze, (e-mail: giulia.cereda@unifi.it, fabio.corradi@unifi.it, cecilia.viscardi@unifi.it)

ABSTRACT: In this work, we compare five different methods proposed in forensic statistics to cope with the rare type match problem. This problem arises when the DNA profile of a suspect coincides with the profile from a crime sample, but it is not present in the available database collected from the population of reference. We compare the methods designed to evaluate the likelihood ratio in this framework by using a set of supervised cases and by considering each method as a classifier that provides the posterior probabilities of two alternative hypotheses, those of the prosecution and the defense, starting from a grid of prior probabilities. We compare them using the value of the posterior cross entropy and decompose it into two terms quantifying their calibration and refinement loss.

KEYWORDS: Forensic statistics, Soft Decisions, Empirical Cross Entropy

1 Introduction

The rare type match problem is the challenging situation faced by a forensic statistician who has to provide the value of a match between the characteristic ($\tilde{y}$) of a crime stain and that of a suspect when $\tilde{y}$ is not in a database of reference of size n. The information provided by evidence, y, is evaluated through a likelihood ratio that can lead to the posterior odds of the hypotheses formulated by the prosecution and the defense, $H \in \{h_p, h_d\}$, for a grid of prior probabilities. Several methods have been designed in the literature to cope with this problem when evidence consists of Y-STR profiles. We aim to compare these methods according to Bayesian decision theory, evaluating the expected cost of decisions expressed as posterior probabilities for the two hypotheses. Using strictly proper scoring rules as cost functions, the expected cost can be decomposed into two components corresponding to calibration and refinement, features of a classifier useful to guide the choice among alternative methods.

Y-STR are polymorphic loci on the Y-chromosome containing a repeated sequence of nucleotides. Individuals differ by the number of times the se-

quence appears at each locus. A Y-STR profile is a list of the numbers of repetitions at a finite number (typically 7 to 23) of loci. The Y-chromosome is only contributed by the father so that there is no recombination; the loci are dependent and cannot be modeled separately. For this reason, a profile must be considered as a whole, and, in case of a rare type match, no frequencies are available from the database to estimate the rarity of the $\tilde{y}$ profile.

2 Proposals for the LR evaluation in case of a rare type match

We want to compare methods that address the rare type match problem differently. We restrict ourselves to five methods assuming that the observed profiles are an i.i.d. sample and not assuming any genetic model; other possibilities exist, e.g. (Andersen *et al.*, 2013), but are not directly comparable.

A first group of methods copes with the rare type match problem by including the suspect profile in the reference data base:

- Augmented Count (AC), is a frequentist method for which:

$$\mathrm{LR}_{AC} = (n+1)/(n_{\tilde{y}}+1) = n+1,$$

  with $n_{\tilde{y}}$ equal to the frequency of $\tilde{y}$ in the database.
- The Bayesian AC, B-AC, (Cereda, 2017a) assumes that the frequency of $\tilde{y}$ in the database is distributed according to a $\mathrm{Bin}(n, \phi_{\tilde{y}})$, with $\phi_{\tilde{y}}$, the unknown probability of $\tilde{y}$ in the population, distributed according to a $\mathrm{Beta}(1,1)$ distribution. These assumptions yield to:

$$\mathrm{LR}_{B-AC} \quad = \quad (n+3)/(n_{\tilde{y}}+2) = (n+3)/2.$$

A second group looks at the list of profiles in the data, including the suspect profile, as partitioned into subsets containing the same Y-STR profile. Building upon different assumptions, the methods evaluate the LR by summarizing the data through π_{n+1}, the vector containing the cardinality of the subsets.

- The two-parameter Poisson Dirichlet method (2PD) (Cereda *et al.*, 2023) assumes an infinite number of Y-STR profiles in the population and that the vector of their ordered relative frequencies follows a 2PD distribution with parameters $\alpha \in (0,1)$ and $\theta > -\alpha$. Thus, the LR becomes:

$$\mathrm{LR}_{2PD} \quad = \quad \left[\int \frac{1-\alpha}{n+1+\theta} \; p(\alpha, \theta \mid \pi_{n+1}) d\alpha d\theta \right]^{-1}.$$

- The Generalized Good (GG) method (Cereda, 2017b) evaluates the LR:

$$\mathrm{LR}_{GG} = n n_1 / 2 n_2,$$

  with n_1 and n_2 the number of singletons and doublets in the database.

- The Brenner's kappa method (Bk) (Brenner, 2010) evaluates the LR as:

$$\mathrm{LR}_{Bk} = (n+1)^2/(n-n_1).$$

3 The posterior cross entropy and its decomposition

We use tools developed by Bayesian decision and information theory to evaluate the five reviewed proposals. Starting from an LR provided by a method $m \in \{\text{AC, B-AC, 2PD, GG, Bk}\}$, LR_m, the evaluation concerns the distribution $p_m(H \mid y)$ with $p_m(h_p \mid y) = \frac{LR_m O(H)}{1+LR_m O(H)}$, where $O(H) = \frac{p(h_p)}{p(h_d)}$ is the prior odds. We consider the log cost function, acting when H is known:

$$C[p_m(h|y)] = \begin{cases} -\log_2(p_m(h|y)) & \text{if } H = h \\ -\log_2(1-p_m(h|y)) & \text{if } H \neq h\,. \end{cases}$$

Since H is usually unknown, we must consider the expected cost corresponding to Shannon's Entropy of $H|y$.

In comparing methods, the mixing distribution of the costs, $p(\cdot|h)$, can be thought of as how Nature expresses the uncertainty on $Y|h$ and, consequently, via Bayes' theorem, on $H|y$. Moreover, we are interested in an average over all the possible evidence y which could arise from the population. This leads to the posterior cross entropy:

$$\mathcal{CE}_{p,p_m}(H \mid Y) = -\sum_{h\in\{h_p,h_d\}} p(h) \sum_{y\in\mathcal{Y}} p(y \mid h)\log(p_m(h|y)) = \mathcal{D}_{p,p_m}(H \mid Y) + \mathcal{E}_p(H \mid Y).$$

As a result, $\mathcal{CE}_{p,p_m}(H \mid Y)$ is the primary criterion of evaluation. The two other criteria are a) $\mathcal{D}_{p,p_m}(h \mid Y)$, the Kullback-Leibler divergence that quantifies the calibration loss, i.e., how the method puts forward posteriors on H in agree with Nature; b) $\mathcal{E}_p(H \mid Y)$, the posterior entropy that quantifies the refinement loss, i.e., the degree of sharpness provided in discriminating hypotheses. We denote the evidence generically by y, but different methods provide probability distributions based on different statistics. Unfortunately, we cannot directly compute the two terms in the decomposition since we have no access to $p(y|H)$, so we provide empirical estimates that require a strategy for building a database of supervised cases starting from a large sample from the population. The proposed solution is based on a Monte Carlo approach and relies on a Pool-Adjacent-Violators (PAV) algorithm that provides an approximate solution. Our results can be presented as the so-called $\mathcal{ECE}$-plot, showing each method's empirical posterior cross-entropy evaluated for different prior probabilities $p(h)$. An example is in Fig 3, where we can compare the

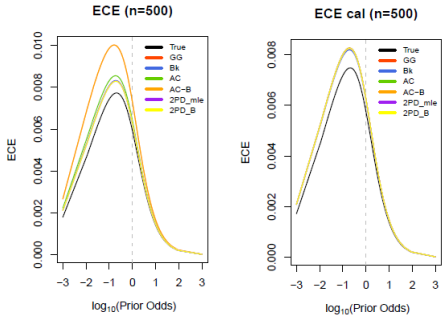

Figure 1. *$\mathcal{ECE}$ before (LHS) and after (RHS) the application the PAV algorithm.*

$\mathcal{ECE}$ of the five methods before and after applying the PAV algorithm. Fig 3 (left) shows that 2PD-B, exploiting π_{n+1}, achieves the smallest $\mathcal{ECE}$; while, the worst method is AC-B which uses only the size of the data and makes the lazy assumption of a flat prior distribution on the probability of the "rare" characteristic. Fig 3 (right) shows that, once recalibrated, all the methods have almost the same refinement so that the main differences attain calibration.

References

ANDERSEN, M. M., ERIKSEN, P. S., & MORLING, N. 2013. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**(7), 39–51.

BRENNER, C. H. 2010. Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics.*, **4**, 281–291.

BRÜMMER, N. 2010. *Measuring, refining and calibrating speaker and language information extracted from speech*. Ph.D. thesis, Stellenbosch University.

CEREDA, G. 2017a. Bayesian approach to LR in case of rare type match. *Statistica Neerlandica.*, **71**, 141–164.

CEREDA, G. 2017b. Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach. *Scandinavian Journal of Statistics.*, **44**, 230–248.

CEREDA, G., CORRADI, F., & VISCARDI, C. 2023. Learning the two parameters of the Poisson-Dirichlet distribution with a forensic application. *Scandinavian Journal of Statistics.*, **50**(1), 120 – 141.