Forensic Science International 230 (2013) 156–169

Contents lists available at SciVerse ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Reliable support: Measuring calibration of likelihood ratios☆

Daniel Ramos *, Joaquin Gonzalez-Rodriguez

Research Institute on Forensic Science (ICFS), ATVS, Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/ Francisco Tomas y Valiente 11, E-28049 Madrid, Spain

ARTICLE INFO

Article history:
Available online 10 May 2013

Keywords:
Calibration
Empirical Cross-Entropy
Accuracy
Likelihood ratio
Performance
Evidence evaluation

ABSTRACT

Calculation of likelihood ratios (LR) in evidence evaluation still presents major challenges in many forensic disciplines: for instance, an incorrect selection of databases, a bad choice of statistical models, low quantity and bad quality of the evidence are factors that may lead to likelihood ratios supporting the wrong proposition in a given case. However, measuring performance of LR values is not straightforward, and adequate metrics should be defined and used. With this objective, in this work we describe the concept of calibration, a property of a set of LR values. We highlight that some desirable behavior of LR values happens if they are well calibrated. Moreover, we propose a tool for representing performance, the Empirical Cross-Entropy (ECE) plot, showing that it can explicitly measure calibration of LR values. We finally describe some examples using speech evidence, where the usefulness of ECE plots and the measurement of calibration is shown.

1. Introduction

Despite the increasing acceptance of the likelihood ratio (LR) approach of evidence evaluation in forensic science [1], computation of LR values still remains a challenge. There are many factors that may lead to values of the LR supporting the wrong proposition in a case, an effect known as *misleading evidence* [2]. If this happens, the LR values are said to present bad performance. Those factors may include sparsity of the databases used as populations [3,4], mismatch in the conditions of the elements in the population databases and in the evidence [5,6], degraded quality or quantity of the evidential materials [7–9], and so forth.

Good performance of the LR is essential in casework. Otherwise, misleading LR values in court may lead fact finders to wrong decisions. This idea is the main motivation behind the establishment of validation procedures for evidence evaluation methods, as a way to establish procedures to control and allow the use of LR models in casework. These validation procedures of evidence evaluation methods should be based on a careful process of performance measurement.

Motivated by this critical problem, in this work we adopt a methodology for the measurement of performance of LR methods in forensic science based on so-called Strictly Proper Scoring Rules (SPSR) [10–12] that has solid grounds on Bayesian statistics. The main contribution of this work is highlighting the importance of a property of a set of LR values called *calibration*, and its relationship with the desirable behavior that the LR should have. Also, although the SPSR methodology is not new, we adapt it to the LR framework for forensic evaluation inference; and we describe a useful representation of the performance of LR values in terms of SPSR and calibration: the Empirical Cross-Entropy (ECE) plot. This methodology for measuring calibration is not intended to replace other methods for measuring performance of the LR, based on *e.g.* Tippett plots or other measurements over the numerator and the denominator of the LR separately. Conversely, we show in this article that measuring the calibration of the LR is an excellent complement to all those methods, in order to have a deep analysis of the performance of the LR with views to a validation process of LR computation in forensic science. In this sense, the example shown in this article illustrate the adequacy and complementarity of using ECE plots in addition to Tippett plots.

Calibration is understood here as a property of a set of LR values, which can be measured. Although the term *calibration* has been recently used to denote a process for obtaining likelihood ratios, we do not follow that meaning in this article. Therefore, our proposal in this article is not about methods to compute the LR, but a methodology to measure the performance and the calibration of a set of LR values, no matter how they were computed. Thus, LR values can be computed using *e.g.* widely accepted models which assign probabilities separately to the numerator and the denominator of the LR (such as the ones described in [13]), and the calibration of the LR values can be measured for those LR values using the methodology proposed in this article.

The article is organized as follows. First, we present the performance assessment methodology based on SPSR, particularizing

☆ This paper is part of the special issue entitled: 6th European Academy of Forensic Science Conference (EAFS 2012), Guest-edited by Didier Meuwly.

* Corresponding author. Tel.: +34 91 4976206; fax: +34 91 4972235.
E-mail address: daniel.ramos@uam.es (D. Ramos).

0379-0738/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.forsciint.2013.04.014

in the classical example of weather forecasting. Then, we intuitively define and describe the concept of calibration. After that, we give reasons that reveal that it is not straightforward to directly apply this methodology to forensic science, and we describe the ECE plot as a solution to overcome those difficulties. Finally, we present experimental examples in forensic speaker recognition where the properties of well-calibrated likelihood ratios are highlighted, after which we draw some conclusions.

2. Measuring performance of probabilistic assessments

In this work, we start by adopting a methodology for measuring performance based on Strictly Proper Scoring Rules (SPSR) [10,12], which is not new and has been studied for decades in Bayesian statistics. We begin with a classical example that has motivated abundant research: the elicitation of probabilistic assessments for weather forecasting [14,11].

2.1. Probabilistic weather forecasting

Consider an unknown variable, say θ, whose value we want to know. Let θ be binary, which means that it only can take one out of 2 values: either $\theta = \theta_p$ or $\theta = \theta_d$.[1] In the weather forecasting example we are going to assume that the unknown variable θ refers to a particular day in the future. We therefore denote $\theta^{(i)}$ as the corresponding variable θ for day i. Thus, in that context the values of $\theta^{(i)} \in \{\theta_p, \theta_d\}$, with the following meaning for day i:

- θ_p: it rains in day i.
- θ_d: it does not rain in day i.

A probabilistic weather forecaster, or simply a forecaster, is defined as someone who assigns probabilities for $\theta^{(i)} = \theta_p$ or $\theta^{(i)} = \theta_d$ *before* the value of $\theta^{(i)}$ is known, aiming at predicting its value. The mechanism by which the forecaster assigns probabilities does not need to be known, but it can be said that, as any other probabilistic assignment, it must consider all the knowledge available to the forecaster, say K [15]. The probability that $\theta^{(i)} = \theta_p$ given K is then denoted as $P(\theta^{(i)} = \theta_p|K)$ which, in words of the forecaster, should be read *the probability that it rains in day i in the future, given all my available knowledge K*. We denote K, the available knowledge, as an *observed* value, in the sense that it is known and fixed. It may include the education, experience and preferences of the forecaster; some data in which the forecaster is basing their assessment; a statistical model; etc. All the resources that are known to the forecaster and used in some way for the elicitation of the probabilistic forecast are included in K, no matter their origin.

For simplicity and convenience, we will eliminate the reference to the day i from the notation when it is clear from the context. Therefore, in those cases we will denote $\theta^{(i)} \equiv \theta$ and $P(\theta^{(i)} = \theta_p|K) \equiv P(\theta_p|K)$. Moreover, by definition of θ_p and θ_d, both values have to be complementary, *i.e.*, $P(\theta_p|K) = 1 - P(\theta_d|K)$.

We assume that at the end of day i the actual value of $\theta^{(i)}$ in day i and all past days will be known. In other words, at the end of the current day the fact of whether it rained or not in any day in the past will be known. Thus, the forecaster will elicit forecasts for future days from day i, when θ is actually unknown.

Notice that $P(\theta^{(i)} = \theta_p|K)$ denotes a probability of the value of the variable of interest (θ) given all the available, *observed* knowledge K. In Bayesian inference, this is known as a *posterior* probability, and therefore probabilistic weather forecasters assign posterior probabilities.

[1] We adopt this notation intentionally, because we ultimately aim at the forensic inference problem.

2.2. Performance of probabilistic assessments: strictly proper scoring rules

During decades, Bayesian statisticians have been seriously concerned about the elicitation of probabilistic assessments [10,16,17], which can be understood given the Bayesian interpretation of probability as a degree of belief [18,19]. In this topic of research, one of the main questions under study has always been the performance of the probabilistic assessments, that can be summarized as follows: if someone is eliciting probability assessments (according to a given model and data, or based on personal experience), how can we evaluate how they perform?

Contextualizing to our weather forecasting example, we can get some intuition about how to evaluate the performance of one single probabilistic assessment of the forecaster. Imagine that the forecaster assigns a probability of raining for tomorrow (day i) as $P(\theta_p|K) = 0.9$. Then, after two days it turns out that it did actually rain in day i, i.e. $\theta = \theta_p$. As the probability given by the forecaster to the value of θ that actually occured (θ_p) is fairly high, then for that particular probabilistic assessment the forecaster did a good work. Therefore, if an external evaluator would assign a cost (or penalty) to that particular forecast, that penalty should be low. However, if the forecaster would have assigned $P(\theta_p|K) = 0.1$, then that forecast would not have been a good one, since the probability for what it actually happened (it rained, θ_p) would have been low. These examples suggest that, in order to evaluate a single forecast, two elements are needed: the probability distribution of θ as assigned by the forecaster (the probability of rain in day i, $P(\theta_p|K)$), and the actual value of the variable θ, that was unknown by the forecaster, but it is known when performance is to be measured.

According to this intuition in Bayesian statistics the performance of probabilistic assessments has been classically addressed by the use of Strictly Proper Scoring Rules (SPSR) [10–12]. A SPSR is a function both of a probability distribution assigned to a given unknown variable, *and* the actual value of the variable. The value of the SPSR will be interpreted as a *loss* or a *cost* given to the probability distribution depending on the actual value of the variable. In this work we will use the *logarithmic* SPSR, which is defined as follows[2]:

$$\mathcal{C}(P(\theta_p|K), \theta) = \begin{cases} -\log_2(P(\theta_p|K)) & \text{if} \quad \theta = \theta_p; \\ -\log_2(1 - P(\theta_p|K)) & \text{if} \quad \theta = \theta_d. \end{cases} \tag{1}$$

where $\mathcal{C}(P(\theta_p|K), \theta)$ represents the SPSR as a function of $P(\theta_p|K)$ and the actual value of θ. The intuition behind SPSR will be exemplified with the representation of the logarithmic SPSR in Fig. 1. The figure shows the two possible values of the logarithmic SPSR depending on the actual value of θ, as a function of $P(\theta_d|K)$. According to Eq. (1), if $\theta = \theta_p$ (it actually rained in day i), the SPSR assigns a high penalty to low values of $P(\theta_d|K)$, and vice-versa. This corresponds to the fact that, if the weather forecaster expressed a high probability of rain in day i (high $P(\theta_d|K)$), and it actually rained ($\theta = \theta_p$), then the penalty should be low, and vice-versa. In the limit, if the forecaster expressed a categorical probability of $P(\theta_d|K) = 0$ (*i.e., it is impossible that tomorrow it will rain*), and it actually rained, the penalty will be *infinite* for the logarithmic SPSR.[3] From Fig. 1, an analogous reasoning can be followed for the case where $\theta = \theta_d$ (it did not rain in day i), where forecasts

[2] There are strong reasons to prefer the logarithmic scoring rule to other SPSR, but they are out of the scope of this work, see [20,21] for details. The base of the logarithm is irrelevant for the expositions. We use base-2 logarithms for information-theoretical reasons, that are explained in [22].

[3] This is, in fact, one desirable property of the logarithmis SPSR, if it is assumed that someone who categorically expresses a wrong judgement should be the worst possible forecaster.