Science and Justice 55 (2015) 499–508

Contents lists available at ScienceDirect

Science and Justice

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

Sampling variability in forensic likelihood-ratio computation: A simulation study

Tauseef Ali [a,*], Luuk Spreeuwers [a], Raymond Veldhuis [a], Didier Meuwly [b]

[a] Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE, Enschede, The Netherlands
[b] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 2 September 2014
Received in revised form 10 April 2015
Accepted 8 May 2015

Keywords:
Likelihood-ratio
Forensics
Biometric recognition
Score
Sampling variability

ABSTRACT

Recently, in the forensic biometric community, there is a growing interest to compute a metric called "likelihood-ratio" when a pair of biometric specimens is compared using a biometric recognition system. Generally, a biometric recognition system outputs a score and therefore a likelihood-ratio computation method is used to convert the score to a likelihood-ratio. The likelihood-ratio is the probability of the score given the hypothesis of the prosecution, H_p (the two biometric specimens arose from a same source), divided by the probability of the score given the hypothesis of the defense, H_d (the two biometric specimens arose from different sources). Given a set of training scores under H_p and a set of training scores under H_d, several methods exist to convert a score to a likelihood-ratio. In this work, we focus on the issue of sampling variability in the training sets and carry out a detailed empirical study to quantify its effect on commonly proposed likelihood-ratio computation methods. We study the effect of the sampling variability varying: 1) the shapes of the probability density functions which model the distributions of scores in the two training sets; 2) the sizes of the training sets and 3) the score for which a likelihood-ratio is computed. For this purpose, we introduce a simulation framework which can be used to study several properties of a likelihood-ratio computation method and to quantify the effect of sampling variability in the likelihood-ratio computation. It is empirically shown that the sampling variability can be considerable, particularly when the training sets are small. Furthermore, a given method of likelihood-ratio computation can behave very differently for different shapes of the probability density functions of the scores in the training sets and different scores for which likelihood-ratios are computed.

1. Introduction

For a comparison of a biometric specimen from a known source and a biometric specimen from an unknown source, a metric called *score* can be computed using a biometric recognition system

$$s = g(x, y), \tag{1}$$

where x and y are the two biometric specimens, g is the biometric recognition algorithm (feature extraction and comparison) and s is the computed score. In general, a score quantifies the similarity between the two biometric specimens. The use of biometric recognition systems in applications such as access-control to a building and e-passport gates at some airports require the developer of the system to choose a threshold and consequently any score above the threshold implies a positive decision and vice versa [1]. This strategy works well in such applications; however, it presents several issues in forensic evaluation and reporting of the evidence from biometric recognition systems [2]. In forensics, the known-source biometric specimen can, for example, come from a suspect while the unknown-source biometric specimen can, for example, come from a crime scene and the goal is to give a degree of support for H_p or H_d. The selection of a threshold and therefore making a decision are not the province of a forensic practitioner. Furthermore, in most criminal cases, scientific analysis of the two biometric specimens provides additional information about the case at hand [3] and a threshold-based hard decision cannot be optimally integrated with other evidences in the case [2–4].

1.1. Likelihood-ratio (LR)

There is a growing interest among forensic practitioners to use biometric recognition systems to compare a pair of biometric specimens. The concept of LR can be used to present the output of such a comparison. It has been extensively used for DNA evidence evaluation [5]. In general, given two biometric specimens, one with a known source and another with an unknown source, it is the joint probability of the occurrence of the two biometric specimens given H_p divided by the joint probability of the occurrence of the two biometric specimens given H_d [6–8]. When the two biometric specimens x and y, are compared

[*] Corresponding author at: Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Building Zilverling 4061, PO Box 217, 7500 AE Enschede, The Netherlands.
E-mail addresses: T.Ali@utwente.nl (T. Ali), L.J.Spreeuwers@utwente.nl (L. Spreeuwers), R.N.J.Veldhuis@utwente.nl (R. Veldhuis), d.meuwly@nfi.minvenj.nl (D. Meuwly).

http://dx.doi.org/10.1016/j.scijus.2015.05.003
1355-0306/© 2015 The Chartered Society of Forensic Sciences. Published by Elsevier Ireland Ltd. All rights reserved.

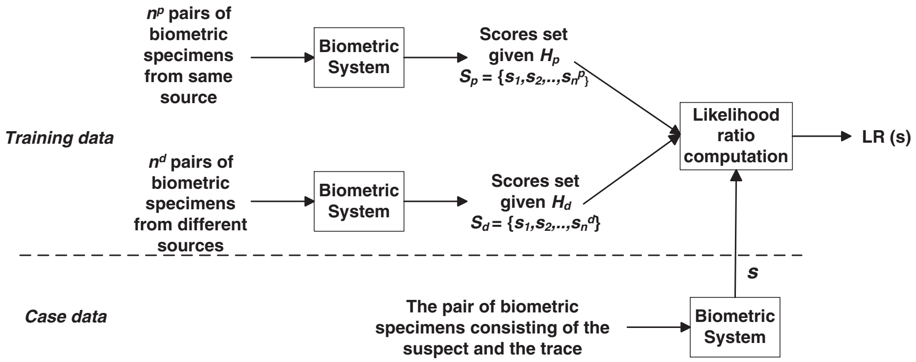
n^p pairs of biometric specimens from same source
Biometric System
Scores set given H_p $S_p = \{s_1, s_2, .., s_{n^p}\}$
Training data
n^d pairs of biometric specimens from different sources
Biometric System
Scores set given H_d $S_d = \{s_1, s_2, .., s_{n^d}\}$
Likelihood ratio computation
LR (s)
s
Case data
The pair of biometric specimens consisting of the suspect and the trace
Biometric System

Fig. 1. Computation of a LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen.

using a biometric recognition system, the resultant score replaces the joint probability of the occurrence of the two specimens in a score-based LR computation [9,10]

$$LR(x,y) = \frac{P(x,y|H_p,I)}{P(x,y|H_d,I)} = \frac{P(s|H_p,I)}{P(s|H_d,I)}, \tag{2}$$

where I refers to the background information which may or may not be domain specific. Note that here the evidence x, y is redefined into the observation s. Once a forensic practitioner has computed a LR, one way to interpret it is as a multiplicative factor which updates the prior odds (before observing the evidence from a biometric system) to the posterior odds (after observing the evidence from a biometric system) using the Bayesian theorem:

$$\frac{P(H_p|s)}{P(H_d|s)} = \underbrace{\frac{P(s|H_p)}{P(s|H_d)}}_{\text{LR}} \times \frac{P(H_p)}{P(H_d)}, \tag{3}$$

where the background information, I, is omitted for simplicity. This is an appropriate probabilistic framework where the trier of fact is responsible for quantification of the prior beliefs about H_p and H_d while the forensic practitioner is responsible for computation of the LR.

The use of a LR is gradually becoming an accepted manner to report the strength of the evidence computed by biometric recognition systems. This is a more informative, balanced and useful metric than a score for forensic evidence evaluation and reporting [3,11]. A general description of the LR concept for evidence evaluation can be found in [2,3]. It is applied to several biometric modalities including speech [12–15] and fingerprint comparison [16]. Preliminary results of the evidence evaluation using the concept of the LR in the context of face and handwriting recognition systems are presented in [6,9,17,18].

1.2. Computation of a LR

In most cases, the conditional probabilities, $P(s|H_p)$ and $P(s|H_d)$, are unknown and they are computed empirically using a set of training scores under H_p, $s_p = \left\{s_j^p\right\}_{j=1}^{n^p}$ (a set of n^p number of scores given H_p) and a set of training scores under H_d, $s_d = \left\{s_j^d\right\}_{j=1}^{n^d}$ (a set of n^d number of scores given H_d) (see Fig. 1). The s_d scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from different sources whereas the s_p scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from the same source. These sets of training scores and the corresponding hypotheses should preferably be case-specific. To compute case-specific different-source scores, either the trace or the suspect's biometric specimen can be compared to the biometric specimens of the potential population (possible potential sources of the trace biometric specimen) [9,10,15,20]. The suspect's biometric specimen used for this purpose should be taken in conditions similar to the trace. Similarly, same-source scores can be obtained by comparing trace-like biometric specimens from the suspect to the reference biometric specimens of the suspect. The effect of using generic instead of the case-specific same-source and different-source scores in the training sets on a LR is studied in the context of handwriting, fingerprint and face recognition systems [9,19,20]. An important condition in LR computation is that the pairs of biometric specimens used for training should reflect the conditions of the pair of biometric specimens for which a LR is computed. Please refer to [21] for an overview of the biometric data set collection in forensic casework for LR computation.

1.3. Sampling variability

Statistically, the training biometric data sets are samples from large populations of biometric data sets. The training biometric data sets, when resampled, lead to slightly different values of the scores in the training sets due to the unavoidable sampling variability. This implies that the sets s_p and s_d consist of random draws from large sets of scores. When the resampling is repeated, slightly different LRs are computed for a given score. This is referred to as the "sampling variability" in a LR. It is desirable that a given LR computation method is less sensitive to the sampling variability in the training sets. If the probability density functions (PDFs) of the scores in the s_p and s_d sets are known, the LR computed using the given sets of training scores can be compared with the LR computed using the two PDFs. The closeness of the two values implies the suitability of a given LR computation method and in this article, we will refer to this performance indicator as "accuracy".

Note that in a given forensic case, the potential population, the trace biometric specimen and the suspect are deterministic inputs to a LR computation procedure. The sampling variability, however, is due to the training scores that are used to compute a LR from a score. This is because these training scores are finite and would vary from one to the next in repeated random sampling. In practice, generation of multiple realizations of the sets of training scores by resampling might not be

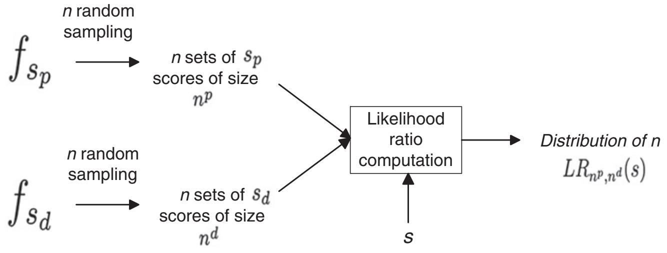
f_{s_p}
n random sampling
n sets of s_p scores of size n^p
f_{s_d}
n random sampling
n sets of s_d scores of size n^d
Likelihood ratio computation
Distribution of n $LR_{n^p,n^d}(s)$
s

Fig. 2. Generation of n realizations of the training sets by random sampling and computation of n LRs of a given score s. The standard deviation, minimum LR, maximum LR and mean LR follow from the set of n LRs of the score s.