Likelihood ratio estimation in forensic identification using similarity and rarity

Yi Tang, Sargur N. Srihari *

University at Buffalo, The State University of New York, 338, Davis Hall, Buffalo, NY 14260, USA

A B S T R A C T

Forensic identification is the task of determining whether or not observed evidence arose from a known source. It is useful to associate probabilities with identification/exclusion opinions, either for presentation in court or to evaluate the discriminative power of a given set of attributes. At present, in most forensic domains outside of DNA evidence, it is not possible to make such a statement since the necessary probability distributions cannot be computed with reasonable accuracy, although the probabilistic approach itself is well-understood. In principle, it involves determining a likelihood ratio (LR) – the ratio of the joint probability of the evidence and source under the *identification* hypothesis (that the evidence came from the source) and under the *exclusion* hypothesis (that the evidence did not arise from the source). Evaluating the joint probability is computationally intractable when the number of variables is even moderately large. It is also statistically infeasible since the number of parameters to be determined from the data is exponential with the number of variables. An approximate method is to replace the joint probability by another probability: that of distance (or similarity) between evidence and object under the two hypotheses. While this reduces to linear complexity with the number of variables, it is an oversimplification leading to errors. We consider a third method which decomposes the LR into a product of two factors, one based on *distance* and the other on *rarity*. This result, which is exact for the univariate Gaussian case, has an intuitive appeal – forensic examiners assign higher importance to rare feature values in the evidence and low importance to common feature values. We generalize this approach to more complex data such as vectors and graphs, which makes LR estimation computationally tractable. Empirical evaluations of the three methods, done with several data types (continuous features, binary features, multinomial and graph) and several modalities (handwriting with binary features, handwriting with multinomial features and footwear impressions with continuous features), show that the distance and rarity method is significantly better than the distance only method.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Forensic identification concerns whether observed evidence arose from a known source. The probabilistic approach is to determine the likelihood ratio positive (LR+) [17,2,38,24,29] whose numerator is the joint probability of the evidence and source under the null hypothesis that the evidence arises from the source and the denominator is the joint probability under the alternate hypothesis that the evidence does not arise from the object. The evidence is deemed to have arisen from the source if $LR + > 1$ and not from the source otherwise.

Determining the joint probability has high data requirements. Assume that the evidence and object are both characterized by n binary features, the joint distribution requires 2^{2n} probabilities or parameters. Even for small n this requires extremely large data sets for estimating parameters. Furthermore in forensic applications data sets are usually small making the approach infeasible.

A common solution is to use the probability of distance, or similarity, between the evidence and known instead of the joint probability [24,33]. The distance based method, which has a *constant* number of parameters, is simple to compute but there is a severe loss of information in going from a high-dimensional joint probability space to a one-dimensional distance space. This paper considers a third method based on a result of Lindley [21] for univariate Gaussian samples which combines the probability of distance with the probability of the mean of evidence and object, called *rarity*. Computing rarity exactly has the complexity of 2^n still, for which probabilistic graphical models (e.g., Bayesian networks) and mixture models are used to further simplify the computation.

1.1. Forensic evidence

Several forensic modalities can benefit from a characterisation of uncertainty, particularly *impression* evidence where "objects or materials retain characteristics of other objects or materials impressed against them", e.g., latent fingerprints, footwear and

tire impressions [5], tool marks and bite marks. Handwriting can also be regarded as impression evidence.

The characteristics need an appropriate representation for both human and computer processing. Examples of such representations are feature vectors or graphs providing a structural description. Features in forensics are usually further divided into *class* and *individual* characteristics. Examples of class characteristics in the case of footwear evidence are those that result from the mass machine manufacturing and can associate the evidence only to one class, whereas *individual characteristics* can potentially associate the evidence to a specific instance of that class. Another example from latent print analysis where class characteristics are type 1 detail (whorls, arches, etc.) and individualizing characteristics are type 2 detail (minutiae).

1.2. Organization of paper

The rest of the paper is structured as follows. Section 2 defines the likelihood ratio (LR) under the identification scenario, Lindley's result and its generalization and the computational complexities of the three methods. Sections 3–5 discuss the application of the three methods to data with feature distributions that are continuous, discrete and graphs, respectively. A theoretical analysis using K-L divergence is given in Section 6. Concluding remarks are made in Section 7.

2. Likelihood ratio

Let $S = \{s_i\}$ be a set of sources. They correspond to, say, footwear outsoles. We denote the *evidence* by E, and the *object* by O. Let $\mathbf{o}$ represent an object drawn from a source s_i, e.g., impression of a known footwear. Let $\mathbf{e}$ represent the evidence drawn from a source s_j, e.g., a crime scene impression. The task is to determine the odds of whether $\mathbf{o}$ and $\mathbf{e}$ came from the same or different source. We can state two opposing hypotheses:

h^0 : $\mathbf{o}$ and $\mathbf{e}$ are from the *same* source ($i{=}j$); and
h^1 : $\mathbf{o}$ and $\mathbf{e}$ are from *different* sources ($i \neq j$), which are the *identification* and *exclusion* hypotheses of forensics; some forensic statistics literature also refers to them as *prosecution* and *defense* hypotheses [2].

We can define two joint probability distributions $P(\mathbf{o}, \mathbf{e}|h^0)$ and $P(\mathbf{o}, \mathbf{e}|h^1)$ which specify as to how often each instance of the object and evidence occur together when they belong to the same source or to different sources. The relative strengths of evidence supporting the two hypotheses is quantified by the likelihood ratio

$$LR_J = LR(\mathbf{o}, \mathbf{e}) = \frac{P(\mathbf{o}, \mathbf{e}|h^0)}{P(\mathbf{o}, \mathbf{e}|h^1)}. \tag{1}$$

The corresponding log-likelihood ratio, $LLR(\mathbf{o}, \mathbf{e}) = \ln P(\mathbf{o}, \mathbf{e}|h^0) - \ln P(\mathbf{o}, \mathbf{e}|h^1)$, has representational advantages: its sign is indicative of same or different source, it has a smaller range than LR, and additivity of contributions of independent features.[1]

It is useful to convert LRs into probabilities of identification and exclusion using a Bayesian formulation.[2] Let the *prior* probabilities of the hypotheses be $P(h^0)$ and $P(h^1)$ with $P(h^0) + P(h^1) = 1$.

---

[1] In performing LLR additions, since LR values in the interval $(1, \infty)$ convert to positive LLRs and LR values in the interval $(0,1)$ convert to negative LLRs, the precisions of LRs < 1 must be high, otherwise the ranges of positive and negative LLRs will not be symmetric.

[2] There has been recent discussion in the legal community regarding whether the terms identification and individualization, commonly used in expressing forensic opinion, are appropriate in view of the fact that there is uncertainty in identification [26]. In response the forensic community has proposed that probabilities can be associated with identification, and that when the probability approaches 1 it is referred to as individualization [39].

Defining the prior odds as $O_{prior} = P(h^0)/P(h^1)$, we can express the prior probability of the same source as

$$P(h^0) = O_{prior}/(1 + O_{prior}). \tag{2}$$

The prior odds can be converted into posterior odds as

$$O_{posterior} = \frac{P(h^0|\mathbf{o}, \mathbf{e})}{P(h^1|\mathbf{o}, \mathbf{e})} = O_{prior} \times LR(\mathbf{o}, \mathbf{e}). \tag{3}$$

Thus we can write the posterior probability of the same source as $P(h^0|\mathbf{o}, \mathbf{e}) = O_{posterior}/(1 + O_{posterior})$. The particular case of equal priors is of interest in forensics, as opinion without prior bias, i. e., no information other than the evidence is used. In this case we get a simple form for the probability of identification as

$$P(h^0|\mathbf{o}, \mathbf{e}) = \frac{LR(\mathbf{o}, \mathbf{e})}{1 + LR(\mathbf{o}, \mathbf{e})} = \frac{\exp(LLR(\mathbf{o}, \mathbf{e}))}{1 + \exp(LLR(\mathbf{o}, \mathbf{e}))}. \tag{4}$$

The probability of exclusion is $P(h^1|\mathbf{o}, \mathbf{e}) = 1 - P(h^0|\mathbf{o}, \mathbf{e}) = 1/[1 + LR(\mathbf{o}, \mathbf{e})] = 1/[1 + e^{LLR(\mathbf{o}, \mathbf{e})}]$.

Thus the key to determining the probability of identification is to determine *LR* defined by Eq. (1), which in turn requires the distributions $P(\mathbf{o}, \mathbf{e}|h^i)$ ($i = 0, 1$), defined over all possible values of objects and their evidential forms. If $\mathbf{o}$ and $\mathbf{e}$ are n-dimensional multinomial vectors with each feature taking K possible values, then $2K^{2n}$ parameters are needed to specify the joint distribution. Determining these distributions is computationally and statistically infeasible. Computationally, kernel density estimation [3] and finite mixture models [23] have been proposed, but they have limitations as well.[3] More important is the statistical limitation of having a sufficient number of samples for so many parameters. Today, objects and evidence can be represented by ever finer features due to higher camera resolution and automatic feature extraction methods and their possible evidential forms are infinite.

One method of simplification is to use a (dis)similarity function between object and evidence. The approach is to define $d(\mathbf{o}, \mathbf{e})$ as a scalar *distance* between object and evidence and define another likelihood ratio as follows:

$$LR_D = \frac{P(d(\mathbf{o}, \mathbf{e})|h^0)}{P(d(\mathbf{o}, \mathbf{e})|h^1)}. \tag{5}$$

The number of parameters needed to evaluate LR_D is constant, or $O(1)$, and is independent of the number of features n. Due to its simplicity, this method has been proposed for fingerprint identification [24], handwriting analysis [32], pharmaceutical tablet comparison [6], etc.

For certain feature spaces and distance measures, e.g., continuous features with Euclidean distance, this approach is equivalent to a *kernel* method [28]. The scalar distance d is just the magnitude of the vector difference $\mathbf{d}$. However, because it maps two distributions of $2n$ variables each into two scalar distributions there is severe loss of information (many pairs of $\mathbf{o}$ and $\mathbf{e}$ can have the same distance). A natural extension is to use vector difference $\mathbf{d}$, which quantifies the distribution of both the magnitude and the orientation of the difference between $\mathbf{o}$ and $\mathbf{e}$, giving a much fine-grained characterization of the difference between $\mathbf{o}$ and $\mathbf{e}$. While this likelihood ratio LR_{VD} provides the simplification of mapping two distributions of $2n$ variables each into two distributions of n variables each, there is still a loss of information in the many to one mappings.

---

[3] Kernel density estimation is expensive in memory as it needs to store the entire training data set, and the cost of evaluating the density grows linearly with size of data set. For mixture models, an important issue is to select the number of components, also, the training algorithm such as EM may converge to a local optimum or the boundary of the parameter space [18].