Handwriting Individualization Using Distance and Rarity

Yi Tang\textsuperscript{a}, Sargur Srihari\textsuperscript{a} and Harish Srinivasan\textsuperscript{b}

\textsuperscript{a}CEDAR, Department of Computer Science and Engineering, University at Buffalo, The State University of New York, USA

\textsuperscript{b}Janya Inc, 1408 Sweet Home Rd, Buffalo, NY 14228

ABSTRACT

Forensic individualization is the task of associating observed evidence with a specific source. The likelihood ratio (LR) is a quantitative measure that expresses the degree of uncertainty in individualization, where the numerator represents the likelihood that the evidence corresponds to the known and the denominator the likelihood that it does not correspond to the known. Since the number of parameters needed to compute the LR is exponential with the number of feature measurements, a commonly used simplification is the use of likelihoods based on distance (or similarity) given the two alternative hypotheses. This paper proposes an intermediate method which decomposes the LR as the product of two factors, one based on distance and the other on rarity. It was evaluated using a data set of handwriting samples, by determining whether two writing samples were written by the same/different writer(s). The accuracy of the distance and rarity method, as measured by error rates, is significantly better than the distance method.

Keywords: Document examination, likelihood ratio, rarity, forensic science, approximate inference

1. INTRODUCTION

The task of determining whether a known object is the source of observed evidence occurs in many forensic scenarios. As a measure of the strength of individualization, the likelihood ratio (LR) is increasingly used in the forensic sciences\textsuperscript{1–4}. The numerator of the LR is the likelihood of observing the evidence, as represented by a feature set, under the assumption that the evidence belongs to the object class. The denominator is a similar probability which is under the hypothesis that the evidence and object do not belong to the same class. In the forensic sciences the two hypotheses are referred to as the prosecution and defense hypotheses\textsuperscript{5}.

However the likelihood ratio based on full distribution of features has high data requirements. The number of parameters needed to specify the distributions is $O(n^2)$, where $n$ is the number of possible values for the object/evidence. This requires an impossibly large data set to determine parameters to a reasonable accuracy. Furthermore in forensic applications data sets are usually small making the naive approach infeasible. One solution is to use the likelihood of distance, or similarity, between the evidence and known instead of the likelihood of features\textsuperscript{3,6}. The distance based method, which has a fixed number of parameters independent of $n$, is simple to compute but there is a severe loss of information in going from a high-dimensional feature-space to a one-dimensional distance. This paper considers an intermediate method based on combining distance with a probabilistic measure of the evidence, where the number of parameters needed is $O(n)$. Since $n$ itself can be quite large (exponential with the number of features used to describe the object), further reduction in complexity is explored.

2. LIKELIHOOD RATIO

Let $O$ be the object produced from some source $s_1$, $E$ be the evidence that originated from some source $s_2$. The central problem is to determine whether or not $O$ and $E$ came from the same source or different sources. It is convenient to introduce a binary random variable $H \in \{h^0, h^1\}$ to denote two alternative hypotheses:

Further author information: (Send correspondence to Yi Tang)

Yi Tang: E-mail: yitang@buffalo.edu, Telephone: 1 716 645 6162

Sargur Srihari: E-mail: srihari@cedar.buffalo.edu, Telephone: 1 716 645 6162

Harish Srinivasan: E-mail: hsrinivasan@janya.com, Telephone: 1 716 565 0401
Figure 1. Distributions of sources from which known object \((O)\) and evidence \((E)\) are drawn: (a) Normally distributed sources, where each source \(s_i\) \((i=1, 2, \ldots)\) is \(\mathcal{N}(\theta_i, \sigma^2)\). \(O\) and \(E\) represent two sets of samples, which may come from a common source \(s_i\) or different sources \((s_i \neq s_j)\), and (b) source means \(\{\theta_1, \theta_2, \ldots\}\) are assumed to have distribution \(\mathcal{N}(\mu, \tau^2)\), with \(\tau \gg \sigma\).

Figure 2. Bayesian network for computing likelihood ratio of forensic evidence. Node \(H\) denotes two alternative hypotheses: either \(s_1 = s_2\) or \(s_1 \neq s_2\). On observing \(O\) and \(E\), the central problem is to determine whether \(s_1 = s_2\) or not.

\(h^0\): \(O\) and \(E\) came from the same source, called the prosecution hypothesis;

\(h^1\): \(O\) and \(E\) came from different sources, called the defense hypothesis.

Obviously when \(h^0\) is true, \(s_1 = s_2\), otherwise \(s_1 \neq s_2\). The ratio \(\frac{P(h^0)}{P(h^1)}\) is the prior odds of the hypotheses. The strength of the evidence is quantified using the likelihood ratio \(LR = \frac{P(O, E|h^0)}{P(O, E|h^1)}\), which can be used to convert prior odds to posterior odds by \(\frac{P(h^0|O, E)}{P(h^1|O, E)} = \frac{P(h^0)}{P(h^1)} \times LR\).

We draw samples from both \(O\) and \(E\), and take measurements on these samples. Because of inherent variability, the measurements on each sample of \(O\) are not exactly the same even if they came from a single source \(O\). Similarly for \(E\). For the simplicity of discussion, let’s assume that there is only one feature of \(O/E\) for which the measurements follow a univariate Gaussian distribution. The distributions of the feature of samples from the same/different source(s) are shown in Fig. 1. Let us use random variable (r.v.) \(O\) represent the measurements made on the samples of \(O\), normally distributed about mean \(\theta_1\) with a constant variance \(\sigma^2\), i.e. \(O \sim \mathcal{N}(\theta_1, \sigma^2)\). Similarly let r.v. \(E\) denote the measurements made on the samples of \(E\), normally distributed about mean \(\theta_2\) with variance \(\sigma^2\), i.e. \(E \sim \mathcal{N}(\theta_2, \sigma^2)\). Meanwhile, because some sources are more commonly occurring than the others, it is useful to characterize the distribution of the samples across all different kinds
of sources. Let \( S = \{ s_1, s_2, \ldots \} \) be a complete list of sources, where \( \{ \theta_1, \theta_2, \ldots \} \) are the means of the samples drawn from each source respectively. These means are also assumed to be normally distributed with mean \( \mu \) and variance \( \tau^2 \). Because the variance across different sources is usually larger (sometimes much larger) than the variance within a single source, as exemplified by handwriting\(^7,8\) and trace evidence,\(^9\) we have \( \tau^2 \gg \sigma^2 \).

We use Bayesian network shown in Fig. 2 to represent the relationships between the variables. Applying the sum rule and product rule of probability theory, the LR can be expanded as

\[
\text{LR} = \frac{P(O, E|h^0)}{P(O, E|h^1)} = \frac{\int P(O|\theta)P(E|\theta)P(\theta)d\theta}{\int P(O|\theta_1)P(\theta_1)d\theta_1 \int P(E|\theta_2)P(\theta_2)d\theta_2},
\]

where \( \theta \) is the common mean of samples from \( O \) and \( E \) when \( h^0 \) is true, \( \theta_1 \) and \( \theta_2 \) are the respective means of \( O \) and \( E \) when \( h^1 \) is true, and the denominator has been factorized as the product of two marginals because \( O \) and \( E \) are conditionally independent given \( H = h^1 \).

### 3. EVALUATION OF LIKELIHOOD RATIO

Suppose \( E \) and \( O \) are represented by a set of features \( f(E) \) and \( f(O) \) over a feature space. Three methods for computing the likelihood ratio are: (1) feature distribution method, based on the probability distributions of features under \( h^0/h^1 \); (2) distance method, based on comparing the distributions of the distance given hypotheses \( h^0/h^1 \); and (3) method based on similarity (distance) and rarity.

#### 3.1 Feature Distribution Method

When evidence and object are represented by feature vectors the LR is defined by \( \text{LR} = \frac{P(f(O), f(E)|h^0)}{P(f(O), f(E)|h^1)} \). Suppose the number of possible values for \( O \) and \( E \) are \( l \) and \( m \) respectively, we need \( 2lm \) parameters to specify the joint distribution \( P(O, E|H) \) or \( P(f(O), f(E)|H) \), which is \( O(m^2) \) when \( l = m \) and can be potentially large. With handwriting, the number of types of handwriting formations \( n \) is over 5,000. Estimating so many parameters to a reasonable accuracy demands a large amount of data, and accordingly takes more time. So computing the LR directly is data demanding and computationally inefficient, for which some simplifications are necessary.

##### 3.1.1 Gaussian with one evidence and one object

Lindley\(^9\) derived a solution for computing the LR, where \( f(E) \) and \( f(O) \) are scalars (denoted as \( f_e \) and \( f_o \)), and the measurements were assumed to be normally distributed about its true value with a known, constant variance \( \sigma^2 \). Let \( f_e \) denote the measurement made on the evidence, normally distributed about the true value \( \theta_1 \) with variance \( \sigma^2 \), i.e., \( f_e \sim \mathcal{N}(\theta_1, \sigma^2) \). Similarly, let \( f_o \) be the measurement made on the object, i.e., \( f_o \sim \mathcal{N}(\theta_2, \sigma^2) \). The true values of the objects from all possible sources \( \theta_i \) \( (i = 1, 2, \ldots) \) were assumed to be normally distributed about the mean \( \mu \) with variance \( \tau^2 \), both of which are known, i.e., \( \theta_i \sim \mathcal{N}(\mu, \tau^2) \), \( i = 1, 2, \ldots \), as shown in Fig. 1(b). Typically \( \tau^2 \) (between-source variation) is much larger than the variance of the measurements \( \sigma^2 \) (within-source variation), i.e., \( \tau \gg \sigma \). Then the LR can be expressed as

\[
\text{LR} = \frac{\tau}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{(f_e - f_o)^2}{2\sigma^2} \right\} \exp \left\{ \frac{(W - \mu)^2}{2\tau^2} \right\},
\]

where \( W = (f_e + f_o)/2 \). The first exponential factor of Eq. (2) weighs the evidence by comparing the absolute difference \( |f_e - f_o| \) with the standard deviation \( \sqrt{2}\sigma \), and the second factor measures how typical \( f_e \) and \( f_o \) are. This has been generalized to the multivariate Gaussian case.\(^10\)

##### 3.1.2 Multinomial Distribution

Consider the case of a binary feature vector, e.g., handwriting with GSC features,\(^7,8\) which is of 1024-dimensions. This would involve learning \( 2^{1024} \) parameters, which is impossible. Thus independence or other assumptions become necessary making the LR computation suboptimal.
3.2 Distance Method

Let \( d(O, E) \) or \( d(f(O), f(E)) \) be the distance between \( E \) and \( O \) computed using some distance measure, which is denoted as \( d \) to avoid symbol clutters. In the distance method, the LR is defined by

\[
LR_d = \frac{P(d|h_0)}{P(d|h_1)}.
\]

Using the sum rule of probability and the \( d \)-separation principle of Bayesian networks, we have

\[
LR_d = \frac{\sum_{o,e} P(o,e|h_0)P(d|o,e)}{\sum_{o,e} P(o,e|h_1)P(d|o,e)},
\]

which involves enumerating all possible values of \( O \) and \( E \), that is practically impossible in most real applications as the number of values of \( O \) and \( E \) can be quite large. An approximating solution is to sample the conditional distribution \( P(o,e|h_i) \) \((i = 0, 1)\) to obtain a sufficient number of samples under each \( h_i \), which are expected to be representative of the whole population, then compute the distance \( d \) to build the distributions \( P(d|h_i) \). This method has been based on the simplification of the graph from Fig. 3(a) to Fig. 3(b), where intermediate nodes \( O \) and \( E \) are removed, so \( d \) is directly conditioned on \( H \). Due to its simplicity, this method has been widely used in fingerprint identification,\(^3\) handwriting analysis,\(^8\) pharmaceutical tablets comparison\(^11\) and etc. Below we present analytical results using examples of normal distributions for continuous scalars, and independent Bernoulli distributions for binary vectors.

3.2.1 Univariate Gaussian Distribution of Single Feature

In Lindley’s example, given \( h^0 \), \( f_e - f_o \) is normally distributed by \( \mathcal{N}(0, 2\sigma^2) \). Thus \( d = |f_e - f_o| \) is distributed by

\[
P(d|h^0) = \begin{cases} 
  \mathcal{N}(0, 2\sigma^2) & \text{if } d = 0 \\
  2\mathcal{N}(0, 2\sigma^2) & \text{if } d > 0,
\end{cases}
\]

which is called folded normal distribution\(^*\) or half-normal distribution when the mean is 0. Similarly, \( f_e - f_o|h^1 \sim \mathcal{N}(0, 2\tau^2 + 2\sigma^2) \), which leads to \( P(d|h^1) = \begin{cases} 
  \mathcal{N}(0, 2\tau^2 + 2\sigma^2) & \text{if } d = 0 \\
  2\mathcal{N}(0, 2\tau^2 + 2\sigma^2) & \text{if } d > 0.
\end{cases} \)

Two distributions are shown in Fig. 4(b). Then \( LR_d \) can be expressed by

\[
LR_d = \frac{P(d|h^0)}{P(d|h^1)} = \frac{\tau}{\sigma} \exp\left\{ -\frac{d^2}{4\sigma^2} \right\}.
\]

\(^*\text{Given a normally distributed random variable } X \text{ with mean } \mu \text{ and variance } \sigma^2, \text{ the random variable } Y = |X| \text{ has a folded normal distribution, which is also called half-normal distribution when } \mu = 0.\)
Comparing (4) with (2), we find that the distance method takes into account only similarity, and discards the rarity that is related to the distribution of features. The relationship between (4) and (2) can be expressed by

$$\text{LR} = \text{LR}_d \times \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(W-\mu)^2}{2\tau^2} \right\},$$

(5)

where the second factor can be regarded as un-normalized rarity. By going from the feature space (usu. high-dimensional) to a distance space (1-dimensional), the distance method loses information about the distribution of features, which has been shown\(^9\) to be relevant to the problem of inferring \(H\). Considering the complexity of the feature distribution method and the over-simplicity of the distance method, an intermediate solution would be to keep the distance/difference factor, while introducing a new factor on rarity to characterize the feature distribution, which will be discussed in Section 4.

3.2.2 Binary Features

Let binary vector \(f(E) = \{a_1, ..., a_n\}^T\), \(f(O) = \{b_1, ..., b_n\}^T\), where \(a_j, b_j \in \{0, 1\}, j = 1, ..., n\).

If the binary features are assumed to be independent and we use the normalized Hamming distance between \(f(E)\) and \(f(O)\), which is defined by \(d = \sum_{j=1}^n |a_j - b_j|\), and binomial distributions of \(d\) given hypotheses \(h^i, i = 0, 1\), then the distance method and the feature distribution method are equivalent.

Next we consider correlated binary features. Assuming each pair of bits \((a_i, b_i)(i=2, ..., n)\) is dependent on only the previous bits \((a_{i-1}, b_{i-1})\), we propose to use a correlation probability matrix \(T_{(4\times4)}\) to represent the conditional probability \(P(a_i, b_i|a_{i-1}, b_{i-1})\). The probabilities for the first two bits \(P(a_1, b_1)\) are also created by sampling. Construction of two binary vectors \(f(E)\) and \(f(O)\) starts by sampling from \(P(a_1, b_1)\), and continues repeatedly by sampling from the conditional probability \(P(a_i, b_i|a_{i-1}, b_{i-1})\) \((i = 2, ..., n)\). By controlling the parameters, we are able to sample the distribution given each hypothesis \(h^i(i = 0, 1)\), then compute the distance \(d\) using the correlation dissimilarity\(^{12}\) given by

$$d = \frac{1}{2} - \frac{S_{11}S_{00} - S_{10}S_{01}}{2((S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10}))^{1/2}},$$

(6)

where \(S_{ij} (i,j \in \{0,1\})\) is the number of occurrences of matches with \(i\) in the first vector \(f(E)\) and \(j\) in the second vector \(f(O)\) at the corresponding positions. Fig. 5 plots the histograms of \(d\) given \(h^i(i = 0, 1)\). Depending on the type of features used to represent \(O/E\), we can always find a proper way to go from feature space to distance space.
Figure 5. Histograms of within/between-source distance between correlated binary feature vectors. For binary vectors \( f(E) = \{a_1, ..., a_n\}^T \) and \( f(O) = \{b_1, ..., b_n\}^T \), the cross correlations between features are captured by the conditional probabilities \( P(a_i, b_i | a_{i-1}, b_{i-1}) \), and the correlation dissimilarity \(^{12} \) is used to compute the distance. Each histogram is computed from 20,000 samples, where each sample is independently drawn from the initial distribution \( P(a_1, b_1) \) and the correlation probability distribution \( P(a_i, b_i | a_{i-1}, b_{i-1}) \).

### 3.2.3 Two Improved Distance Methods

Srihari and Ball\(^{13} \) proposed an improved distance method in which the denominator \( P(d|h^1) \) is modeled based on the distances between one specific object/evidence under consideration and all other samples, not the distances between all possible pairs of samples from different sources. For example, suppose there are \( n \) known writers \( \{w_1, ..., w_i, ..., w_n\} \) in total, the improved distance method computes \( P(d|h^1) \) for each specific writer \( w_i \) by comparing the samples written by \( w_i \) against the samples from all writers including \( w_i \), computing the pair-wise distance values, from which \( P(d|h^1) \) is estimated.

We have made further improvement based on the fact that the numerator \( P(d|h^0) \) also differs from writer to writer, which indicates that some writer writes quite consistently and has a relatively fixed pattern of writing the same letter (e.g. “th”), while other writers may write quite differently each time and have a large within-writer variance. Therefore \( P(d|h^0) \) is computed for each known writer and used in the LR computation.

### 4. DISTANCE AND RARITY METHOD

By rewriting Eq. (2) as

\[
    LR = \frac{1}{\sqrt{2\pi(2\hat{\sigma})^2}} \exp \left\{ -\frac{(f_e - f_o)^2}{2(2\hat{\sigma})^2} \right\} * \frac{1}{\sqrt{2\pi\tau}} \exp \left\{ -\frac{(W - \mu)^2}{2\tau^2} \right\} = \mathcal{N}(f_e - f_o | 0, 2\sigma^2) * \frac{1}{\mathcal{N}(W | \mu, \tau^2)}, \tag{7}
\]

we find that the LR can be interpreted as the product of two factors, i.e. distance and rarity, where the distance factor is quantified by the probability density of the distribution of the difference \( f_e - f_o \) given \( h^0 \), i.e. \( f_e - f_o | h^0 \sim \mathcal{N}(0, (2\hat{\sigma})^2) \), and rarity is quantified by the reciprocal of the probability density of a Gaussian random variable \( W \sim \mathcal{N}(\mu, \tau^2) \) at known value of \( W \), which gives an estimate of how unusual \( W \) is, i.e. lower probability density indicating higher rarity.

It is interesting to note that the rarity is based on neither \( f_e \) nor \( f_o \), but on the average of \( f_e \) and \( f_o \), i.e. \( W = (f_e + f_o)/2 \), because given \( H = h^0 \), \( W \) is a better estimate of the true value \( \theta \) than either \( f_e \) or \( f_o \), as evidenced by its variance \( \sigma^2/2 \) being less than the variance of \( f_e \) or \( f_o \). In a word, the distance factor captures within-source variation through a normal distribution of the difference as compared to its variance, and the rarity
exploits the distribution of true values of the object/evidence originating from all possible sources, i.e. making use of between-source distribution of features.

Comparing Eq. (7) with Eq. (4), we find that the distance factor in (7) is similar to the numerator of the distance method, while the denominator of the distance method is replaced by the rarity factor. By making an analogy from (7), we can formulate the LR as the product of two factors, distance \( T_d \) and rarity \( T_r \), i.e.

\[
\text{LR} = T_d \ast T_r
\]

where \( T_d \) evaluates the closeness of \( E \) and \( O \), and \( T_r \) assesses how unusual the matched features between \( E \) and \( O \) are, i.e. how likely these characteristics can be duplicated in another object randomly selected from that class.

4.1 Rarity and Its Evaluation

**Def.1.** Given a probability space \( (\Omega, \mathcal{F}, P) \), the rarity of a random event \( e \in \mathcal{F} \) is given by \( R(e) = \frac{1}{P(e)} \), \( P(e) \neq 0 \).

4.1.1 Evaluation of Rarity \( T_r \)

We generalize feature representation from a vector to a set, i.e. \( f(E) = \{f_1, ..., f_i, ..., f_{n_1}\} \), where each feature \( f_i (i = 1, ..., n_1) \) is a \( n \)-dimensional vector \((f_{i1}, ..., f_{in_1})^T\). Similarly, \( f(O) = \{g_1, ..., g_j, ..., g_{n_2}\} \), and \( g_j = (g_{j1}, ..., g_{jn_2})^T, j = 1, ..., n_2 \).

Suppose \( M = \{\overline{f_1}, ..., \overline{f_i}, ..., \overline{f_{n_1}}\} \) (1 \( \leq i_1 < ... < i_k \leq n_1 \)) are the set of matched features between \( f(E) \) and \( f(O) \), where \( \overline{f_i} (r = 1, ..., k) \) is the mean of two matched features \( f_i \) and \( g_j \). Two features \( f_i \) and \( g_j \) are said to match if \(|f_i - g_j| \leq \epsilon \) where \( \epsilon = (\epsilon_1, ..., \epsilon_n)^T \) is a given tolerance. Then the rarity \( T_r \) is evaluated by

\[ T_r = R(M) = \prod_{r=1}^{k} R(\overline{f_i}) = \prod_{r=1}^{k} \frac{1}{P(\overline{f_i}) \prod_{i=1}^{\epsilon} \epsilon_i} \]

if the features are independent.

4.2 Evaluation of Distance Factor \( T_d \)

The distance factor \( T_d \) is evaluated by the p.d.f. of the distribution of the absolute difference \( d = |f(O) - f(E)| \) given \( H = h^0 \), i.e.

\[
T_d = P(d|h^0) = P(|f(O) - f(E)| |h^0).
\]

4.2.1 Binary Features

Let binary vector \( f(E) = \{a_1, ..., a_n\}^T \), \( f(O) = \{b_1, ..., b_n\}^T \), where \( a_j, b_j \in \{0, 1\}, j = 1, ..., n \).

**Independent Features**

Individual bits of \( f(E) \) (or \( f(O) \)) are assumed to be independent. The distance between \( f(E) \) and \( f(O) \) can be evaluated by the normalized Hamming distance. Note that \( d \) is a discrete variable with domain \( \{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1\} \). Based on the previous discussion, we may assume that, given \( H = h^0 \), the distribution of \( d \) is modeled by a binomial distribution, i.e. \( d|h^0 \sim \text{Bin}(n, p) \), where \( p \) is the probability of two bits at corresponding positions being different given \( h^0 \).

**Correlated Features**

When independence assumption doesn’t hold, we take into account the correlation between features by adopting a more appropriate distance measure called correlation dissimilarity, defined in (6).

5. EXPERIMENTS AND RESULTS

The proposed method has been evaluated with a data set of handwriting, which consists of 3125 samples of “th” written by 499 writers, where each writer wrote “th” 3 to 10 times. Each sample is a gray-scale image from which binary features were extracted. Among 3125 samples, 2125 samples were used for training and the remaining 1000 samples were used for testing. Below we discuss the binary GSC features, the computation of distance factor \( T_d \) and rarity \( T_r \), then finally present the results.
For computational feasibility, we assume that individual binary features are independent, therefore we propose a further improved method which estimates the distribution of the absolute difference \( |x - y| \) between two binary feature vectors \( f(O) \) and \( f(E) \) (denoted as \( x \) and \( y \) in this section) is computed using the correlation dissimilarity \( d \) given in Eq. (6).

\[
P(d|h^i) = \frac{1}{\Gamma(\alpha)\beta} (1 - d)^{\alpha - 1} \exp \left\{ -\frac{1}{\beta} d \right\}, \quad \text{where shape parameter } \alpha \text{ and scale parameter } \beta \text{ were estimated to be } \alpha = 125.0318, \beta = 0.004827.
\]

The improved distance method proposed by Srihari and Ball uses the same numerator as in the basic distance method, and computes the denominator \( P(d|h^1) \) for each writer, as described in subsection 3.2.3. We propose a further improved method which estimates \( P(d|h^1) \) in the same way as Srihari and Ball’s method, and in addition, computes \( P(d|h^0) \) for each known writer, and uses it in the computation of LRd.

5.3 Implementation of the Feature Distribution Method

For computational feasibility, we assume that individual binary features are independent, therefore \( P(x, y|h^i) = \prod_{j=1}^{n} P(x_j, y_j|h^i) \), where \( n = 1024 \). For each \( j \in \{1, ..., n\} \), there are four possible values for \( (x_j, y_j) \), whose probabilities were estimated using maximum likelihood.

5.4 Implementation of the Distance and Rarity Method

Modeling the Distribution of the Absolute Difference \( |x - y| \). The absolute difference \( d \) between two writing samples \( x \) and \( y \), i.e. \( d = |x - y| = (d_1, ..., d_n)^T \) (where \( d_i = |x_i - y_i| \), for \( i = 1, ..., n \)), is modeled by a naive Bayes model: given hypothesis \( H = h^0 \), the absolute difference \( d \) between individual bits are assumed to be independent, so \( P(d|h^0) = \prod_{i=1}^{n} P(d_i|h^0) \). Generally, let us consider a set of binary random variables...
\[d = (d_1, ..., d_i, ..., d_n)^T,\] where each \(d_i\) (i = 1, ..., n) is governed by a Bernoulli distribution with parameter \(\mu_i\), i.e. \(P(d_i|h^0) = \mu_i^{d_i}(1 - \mu_i)^{1-d_i}\), thus the distance factor \(T_d\) is computed by

\[T_d = P(d|h^0) = \prod_{i=1}^{n} P(d_i|h^0) = \prod_{i=1}^{n} \mu_i^{d_i}(1 - \mu_i)^{1-d_i} ,\]  \hspace{1cm} (10)

where \(\mu_i\) is estimated to be the fraction of the samples in the data set in which the \(i^{th}\) bit is 1.

**Modeling the Binary Features using Mixture of Bernoulli Distributions.** Although different writers write differently, some writers write “th” in a similar way, i.e. they are more likely in the same cluster if we consider the binary vector of each “th” sample as a point in a 1024-dimensional feature space. Therefore we use mixture of Bernoulli distributions\(^{14}\) to model the distribution of binary features. We implemented the mixture of Bernoulli distributions, and learnt its parameters via E-M algorithm. We introduced *Bayesian Information Criterion* (BIC) to penalize more complex models, i.e. the models with more mixture components. Finally we found that the number of mixtures \(K = 100\) gives the best fit. Fig. 7 shows one sample cluster.

**Compute the Rarity \(T_r\).** The common features between two writing samples are the common 1’s and 0’s between their binary vectors. For two binary vectors \(x = (x_1, ..., x_n)^T\) and \(y = (y_1, ..., y_n)^T\), the vector of common features \(c\) is defined as follows: \(\forall i \in \{1, ..., n\}, c_i = 1\) iff \(x_i = y_i = 1\), \(c_i = 0\) iff \(x_i = y_i = 0\), otherwise \(c_i\) is undefined. The rarity factor \(T_r\) is evaluated as the rarity of common features \(c\), given by

\[T_r = R(c) = \frac{1}{\sum_{k=1}^{K} \pi_k p(c|\mu_k)},\]  \hspace{1cm} (11)

where \(\pi_k\) (k = 1, ..., K) are the mixing coefficients, \(\mu_k\) are the parameters for each mixture component. Finally, Eq. (8) is used to compute the LR.

### 5.5 Five Methods and Their Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Type-I error(%)</th>
<th>Type-II error(%)</th>
<th>Average error(%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature distribution method</td>
<td>35.96</td>
<td>22.08</td>
<td>29.02</td>
<td>1261.5</td>
</tr>
<tr>
<td>Distance method</td>
<td>38.30</td>
<td>19.46</td>
<td>28.88</td>
<td>0.4</td>
</tr>
<tr>
<td>Improved distance method by Srihari et al.(^{13})</td>
<td>40.70</td>
<td>15.94</td>
<td>28.32</td>
<td>504.1</td>
</tr>
<tr>
<td>Improved distance method proposed by authors</td>
<td>34.14</td>
<td>21.84</td>
<td>27.99</td>
<td>228.2</td>
</tr>
<tr>
<td>Distance &amp; rarity method</td>
<td>18.69</td>
<td>11.12</td>
<td>14.90</td>
<td>141.7</td>
</tr>
</tbody>
</table>

We evaluated five methods: (1) feature distribution method; (2) distance method; (3,4) two improved distance methods,\(^{13}\) and (5) distance and rarity method. When each pair of test samples are compared with each other, the *log-likelihood ratio* was computed by \(LLR = \log(LR)\), where LR is the likelihood ratio, then a classification is made according to the following criterion: if \(LLR > 0\), then the two samples were written by the same writer, else they were written by different writers. The accuracy of the methods was measured by the probability of misclassification,\(^{7}\) in which *Type-I error* is the probability of misclassifying two samples as coming from different sources when they actually came from the same source, whereas *Type-II error* is the probability of misclassifying two samples as coming from the same source when they actually came from different sources, and the results on accuracy and efficiency are shown in Table 1.

Table 1 shows that the feature distribution method has the lowest accuracy and speed. This may be because that the assumption of independence between individual binary bits does not hold, therefore the model for feature distribution is quite inaccurate. The distance method performs better than the feature distribution method because the model for distributions of the distance is accurate although there is loss of information when we go from high-dimensional feature space to one-dimensional distance space. Srihari and Ball’s method captures the uniqueness of each writer by computing the denominator \(P(d|h^1)\) for each known writer, which has slightly better results; we extend Srihari and Ball’s method by computing the numerator \(P(d|h^0)\) for each known writer, and the results shows further improvement. Finally, the proposed distance and rarity method makes full use of the rarity of the writing of each writer, and achieves the highest accuracy.
6. SUMMARY AND CONCLUSION

The likelihood ratio (LR) is a quantitative measure of the degree of uncertainty in forensic individualization, in which the central problem is to determine whether the observed evidence came from a particular source/object. Evaluation of the LR based on the feature distribution needs a large number of parameters for the model, which is data demanding and computationally expensive. One simplification is to compute the ratio of the likelihoods of the distance. Although being simple to compute, there is a severe loss of information in going from a high-dimensional feature space to a one-dimensional distance. This paper proposed an intermediate solution, in which the LR is decomposed as the product of two factors, distance and rarity. The proposed method was evaluated using a data set of handwriting samples, and compared with the feature distribution method and distance methods. The results showed that the proposed method outperforms the distance methods and the feature distribution method in terms of accuracy, with acceptable efficiency.

Acknowledgment

This work was supported by the DoJ under NIJ grant 2010-DN-BX-K037. The opinions expressed in the paper are those of the authors and not of the DoJ.

REFERENCES