An Algorithm for Categorization of Documents on the Basis of Authentication Technique

Vladimir B. Balakirsky
Institute for Experimental Mathematics
Essen, Germany
e-mail: v_b_balakirsky@rambler.ru

Peter Trifonov
Saint-Petersburg State Polytechnic University
Saint-Petersburg, Russia
e-mail: petert@dcn.ftk.spbstu.ru

ABSTRACT

An authentication algorithm for binary vectors generated by a source having an unknown structure is described and applied to the problem of automatic characterization of documents.

1. INTRODUCTION

Automatic categorization is necessary to maintain large amounts of documents. However, its implementation is an extremely difficult task due to huge amount of data, vast number of different categories, and inherent difficulty of the categorization problem, which may confuse even a human expert. Furthermore, the set of categories is not fixed beforehand, and sometimes their meanings change. Nevertheless, one needs an automatic tool, which is able to classify an document as belonging to one of categories, specified by a pre–classified set of documents.

The first step of any automatic categorization algorithm is a transformation of the document being classified to a vector of features. Each category can be considered as a source with some multi-dimensional distribution over the space of feature vectors. The classifier has to decided if a particular feature vector is likely to be drawn from a given source or not. Similar problem is known as the biometrical authentication problem where the verifier has to check whether the presented vector is a reaction of the neuron system to a fixed stimulus, which is specified by records of the reactions to that stimulus received at the enrollment stage. Therefore, the algorithms that bring good results for the biometrical authentication can be considered as candidates for the text categorization as well. In the present correspondence, we propose the implementation of a special version of the BAr (Bernoulli Approximation) authentication scheme developed for spikes analysis [1].

2. AUTHENTICATION ON THE BASIS OF PAIRWISE COMPARISONS

2.1 Basic ideas

Suppose that there are L binary vectors of length n denoted by $\mathbf{x}_\ell = (x_{\ell,1}, \ldots, x_{\ell,n})$, $\ell = 1, \ldots, L$, characterizing the behavior of a complex system.

"Extracting the knowledge about the source" is understood as a formal description of an authentication procedure: given a binary vector $\mathbf{y}$ of length n, the authentication scheme has to decide whether this vector is generated by the same source as the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ or not. The error events, called the false acceptance and the false rejection, are possible in this case. The false acceptance event means that the vector generated by a different source received the positive answer. The false rejection event means that the vector generated by the fixed source received the negative answer. In signal detection theory, one tries to minimize the probabilities of both error events. The main difficulty with the setup describing complex systems is a proper introduction of the notion of "the probability" in a sense that processing of probabilities should reflect the requests. Nevertheless, without any formal introduction, we can notice that the quality of the data processing scheme can be measured by using the following experiments: (1) if the authentication is run L times when the given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ are substituted for the vector $\mathbf{y}$, then the most of the answers should be positive; (2) if randomly chosen vectors are substituted for the vector $\mathbf{y}$, then the most of the answers should be negative.

Let the verifier be given L binary vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ of length n generated by some source. He is also given another binary vector $\mathbf{y}$ of length n. The verifier makes either the acceptance decision (Y), which means that the vector $\mathbf{y}$ is generated by the same source as the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$, or the rejection decision (N), which means that this is not true.

The basic idea of our approach is the introduction of *an auxiliary source generating pairs of binary vectors of length* n. The general description of the data processing scheme can be presented as follows:

- *replace the given vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_L$ *with the sequence consisting of* L^2 *pairs of vectors* $(\mathbf{x}_\ell, \mathbf{x}_{\ell'})$, $\ell, \ell' = 1, \ldots, L$;
- *given a vector* $\mathbf{y}$, *form the sequence consisting of* L *pairs of vectors* $(\mathbf{x}_\ell, \mathbf{y})$, $\ell = 1, \ldots, L$;
- *compare statistical properties of two sequences.*

As the model for the source generating the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ is not defined, "statistical properties" mentioned above are understood in such a way that the acceptance/rejection decision is made by using the probabilities assigned to 2^{2n} pairs of vectors $(\mathbf{x}, \mathbf{x}')$. As we have access only to L^2 pairs, which is much less than 2^{2n}, the rules for assigning these probabilities have to be postulated. Let us denote the probability associated with the pair $(\mathbf{x}, \mathbf{x}')$ by $\Omega(\mathbf{x}, \mathbf{x}')$ and let

$$\Omega = (\,\Omega(\mathbf{x}, \mathbf{x}'), \mathbf{x}, \mathbf{x}' \in \{0,1\}^n\,)$$

denote the desired probability distribution. The following requirements have to be taken into account while specifying Ω:

(R1) the probability distribution Ω has to be symmetric, i.e., $\Omega(\mathbf{x},\mathbf{x}') = \Omega(\mathbf{x}',\mathbf{x})$ for all $\mathbf{x},\mathbf{x}' \in \{0,1\}^n$;

(R2) the length of description of Ω has to be small;

(R3) the probabilities $\Omega(\mathbf{x},\mathbf{x}')$ are large enough when $\mathbf{x},\mathbf{x}' \in \{\mathbf{x}_1,\ldots,\mathbf{x}_L\}$.

The simplest algorithm for assigning the probability to any pair of binary vectors $(\mathbf{x},\mathbf{x}')$ is the non–stationary memoryless distribution when

$$\Omega(\mathbf{x},\mathbf{x}') = \prod_{t=1}^{n} \omega_t(x_t, x'_t). \tag{1}$$

Then the probability distribution Ω is specified by the 2×2 matrices

$$\mathbf{\Omega}_t = \left[\begin{array}{cc} \omega_t(0,0) & \omega_t(1,0) \\ \omega_t(1,0) & \omega_t(1,1) \end{array}\right]$$

such that

$$\left\{\begin{array}{l} \omega_t(0,0), \omega_t(1,0), \omega_t(1,0), \omega_t(1,1) \ge 0 \\ \omega_t(0,0) + \omega_t(1,0) + \omega_t(1,0) + \omega_t(1,1) = 1. \end{array}\right.$$

We also denote

$$\mathbf{\Lambda}_t \triangleq \left[\begin{array}{cc} -\log\omega_t(0,0) & -\log\omega_t(0,1) \\ -\log\omega_t(1,0) & -\log\omega_t(1,1) \end{array}\right]. \tag{2}$$

Notice that, by the symmetric requirement,

$$\omega_t(0,1) = \omega_t(1,0) = (1 - \omega_t(0,0) - \omega_t(1,1))/2 \tag{3}$$

and

$$\Omega(\mathbf{x},\mathbf{x}') = \prod_{t=1}^{n} \left\{\begin{array}{ll} \omega_t(0,0), & \text{if } x_t = x'_t = 0, \\ (1-\omega_t(0,0)-\omega_t(1,1))/2, & \text{if } x_t \ne x'_t, \\ \omega_t(1,1), & \text{if } x_t = x'_t = 1 \end{array}\right. \tag{4}$$

for all pairs of binary vectors $(\mathbf{x},\mathbf{x}')$.

If (4) holds, then the (R1)–(R2) requirements are satisfied, since the probability distribution Ω is symmetric and it is specified by $2n$ numbers $\omega_t(0,0), \omega_t(1,1)$, $t = 1,\ldots,n$. The (R3) requirement can be presented as assignment a measure to the given matrix $\mathbf{X}$ where all the data contribute to the obtained value and this value "has to be large". We have the geometric average and the arithmetic average of the entries $\Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})$ as candidates for such a measure and formalize the (R3) requirement as follows:

– *assign the numbers* $\omega_t(0,0), \omega_t(1,1)$, $t = 1,\ldots,n$, *in such a way that*

$$\Big(\prod_{\ell,\ell'=1}^{L} \Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})\Big)^{1/L^2} \;\to\; \max \tag{5}$$

or, equivalently,

$$\frac{1}{L^2}\sum_{\ell,\ell'=1}^{L} \Lambda(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) \;\to\; \min, \tag{6}$$

where

$$\Lambda(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) \triangleq -\log\Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'}). \tag{7}$$

Notice that the product at the left–hand side is the geometric average of the entries $\Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})$, $\ell,\ell' = 1,\ldots,L$, which is also a lower bound on the arithmetic average,

$$\frac{1}{L^2}\sum_{\ell,\ell'=1}^{L} \Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) \ge \Big(\prod_{\ell,\ell'=1}^{L} \Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})\Big)^{1/L^2}.$$

Proposition. *Let*

$$\gamma_t = \frac{1}{L}\Big|\Big\{\,\ell \in \{1,\ldots,L\} :\, x_{\ell,t} = 1\,\Big\}\Big| \tag{8}$$

denote the relative number of 1*'s in the* t*-th column of the matrix* $\mathbf{X}$*. If the entries of the probability distribution* Ω *are defined by* (4)*, then the product in* (5) *is maximized when*

$$\mathbf{\Omega}_t = \left[\begin{array}{cc} (1-\gamma_t)^2 & (1-\gamma_t)\gamma_t \\ \gamma_t(1-\gamma_t) & \gamma_t^2 \end{array}\right], \quad t = 1,\ldots,n. \tag{9}$$

Furthermore, for the optimum assignment,

$$\frac{1}{L^2}\sum_{\ell,\ell'=1}^{L} \Lambda(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) = 2\sum_{t=1}^{n} h(\gamma_t), \tag{10}$$

where

$$h(z) = -z\log z - (1-z)\log(1-z), \quad z \in (0,1),$$

is the binary entropy function.

2.2 Description of the authentication algorithm

Our algorithm can be fixed as a procedure consisting of two steps, called the preprocessing and the authentication.

Preprocessing.

– *Compute* $\gamma_1,\ldots,\gamma_n$ *defined by* (5)*.*

– *For all* $\ell,\ell' = 1,\ldots,L$*, compute the probability* $\Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})$ *defined by* (4) *for* $(\mathbf{x},\mathbf{x}') = (\mathbf{x}_\ell,\mathbf{x}_{\ell'})$*.*

– *Define* $\Lambda_\alpha \ge 0$ *as the minimum number satisfying the inequality* $2^{-\Lambda_\alpha} \le \Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'})$ *for* αL^2 *pairs* $(\ell,\ell') \in \{1,\ldots,L\}^2$*, i.e.,*

$$\frac{1}{L^2}\Big|\Big\{\;(\ell,\ell') \in \{1,\ldots,L\}^2 : \Omega(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) \ge 2^{-\Lambda_\alpha}\,\Big\}\Big| = \alpha \tag{11}$$

or, equivalently,

$$\frac{1}{L^2}\Big|\Big\{\;(\ell,\ell') \in \{1,\ldots,L\}^2 : \Lambda(\mathbf{x}_\ell,\mathbf{x}_{\ell'}) \le \Lambda_\alpha\,\Big\}\Big| = \alpha. \tag{12}$$

Authentication.

– *Given a binary vector* $\mathbf{y}$*, compute the probabilities* $\Omega^{(\varepsilon)}(\mathbf{x}_\ell,\mathbf{y})$, $\ell = 1,\ldots,L$*, defined in* (2)*.*

– *Find the relative number of rows of the matrix* $\mathbf{X}$ *such that* $\Omega^{(\varepsilon)}(\mathbf{x}_\ell,\mathbf{y}) \ge 2^{-\Lambda_\alpha}$ *and denote it by*

$$\alpha^{(\varepsilon)}(\mathbf{y}) \triangleq \frac{1}{L}\Big|\Big\{\;\ell \in \{1,\ldots,L\} :\, \Omega^{(\varepsilon)}(\mathbf{x}_\ell,\mathbf{y}) \ge 2^{-\Lambda_\alpha}\,\Big\}\Big|.$$

– *Accept the claim that the vector* $\mathbf{y}$ *is generated by the same source, as rows of the matrix* $\mathbf{X}$*, if and only if* $\alpha^{(\varepsilon)}(\mathbf{y}) \ge \alpha$*.*

2.3 Numerical illustration

Suppose that $L = 5$, $n = 6$, and

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{13}$$

The computation of the ratios of the Hamming weights of the columns by L brings the following vector:

$$(\gamma_1, \ldots, \gamma_5) = (0.2, 0.2, 0.6, 0.4, 0.8, 0.6).$$

Thus, by (9),

$$\begin{aligned} \mathbf{\Omega}_1 = \mathbf{\Omega}_2 &= \begin{bmatrix} 0.64 & 0.16 \\ 0.16 & 0.04 \end{bmatrix}, \\ \mathbf{\Omega}_4 &= \begin{bmatrix} 0.36 & 0.24 \\ 0.24 & 0.16 \end{bmatrix}, \\ \mathbf{\Omega}_3 = \mathbf{\Omega}_6 &= \begin{bmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{bmatrix}, \\ \mathbf{\Omega}_5 &= \begin{bmatrix} 0.04 & 0.16 \\ 0.16 & 0.64 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 &= \begin{bmatrix} 0.6 & 2.6 \\ 2.6 & 4.6 \end{bmatrix}, \\ \mathbf{\Lambda}_4 &= \begin{bmatrix} 1.5 & 2.1 \\ 2.1 & 2.6 \end{bmatrix}, \\ \mathbf{\Lambda}_3 = \mathbf{\Lambda}_6 &= \begin{bmatrix} 2.6 & 2.1 \\ 2.1 & 1.5 \end{bmatrix}, \\ \mathbf{\Lambda}_5 &= \begin{bmatrix} 4.6 & 2.6 \\ 2.6 & 0.6 \end{bmatrix}. \end{aligned}$$

Let $\mathbf{\Lambda}$ denote the $L \times L$ matrix whose (ℓ, ℓ') entry is equal to $\Lambda(\mathbf{x}_\ell, \mathbf{x}_{\ell'})$, where $\ell, \ell' = 1, \ldots, L$. Then

$$\mathbf{\Lambda} = \begin{bmatrix} 8.7 & 7.5 & 12.1 & 10.1 & 8.7 \\ 7.5 & 6.4 & 10.9 & 8.9 & 7.5 \\ 12.1 & 10.9 & 15.5 & 13.5 & 12.1 \\ 10.1 & 8.9 & 13.5 & 11.5 & 10.1 \\ 8.7 & 7.5 & 12.1 & 10.1 & 8.7 \end{bmatrix}.$$

For example,

$$\Lambda(\mathbf{x}_1, \mathbf{x}_2) = 0.6 + 0.6 + 1.5 + 2.1 + 0.6 + 2.1 = 7.5.$$

As

$$(h(\gamma_1), \ldots, h(\gamma_5)) = (0.72, 0.72, 0.97, 0.97, 0.72, 0.97),$$

the ratio of the sum of entries of the matrix $\mathbf{\Lambda}$ and $L^2 = 25$ is equal to

$$2(0.72 + 0.72 + 0.97 + 0.97 + 0.72 + 0.97) = 10.16,$$

as it follows from (10).

Suppose that $\alpha = 0.44$, i.e., the threshold Λ_α has to be determined as the minimum number such that there are $\alpha L^2 = 0.44 \cdot 25 = 11$ entries of the matrix $\mathbf{\Lambda}$, which are not greater than Λ_α. Then $\Lambda_\alpha = 8.9$, and the matrix $\mathbf{\Lambda}$ is rewriteen below, where we show these entries in the bold font,

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{8.7} & \mathbf{7.5} & 12.1 & 10.1 & \mathbf{8.7} \\ \mathbf{7.5} & \mathbf{6.4} & 10.9 & \mathbf{8.9} & \mathbf{7.5} \\ 12.1 & 10.9 & 15.5 & 13.5 & 12.1 \\ 10.1 & \mathbf{8.9} & 13.5 & 11.5 & 10.1 \\ \mathbf{8.7} & \mathbf{7.5} & 12.1 & 10.1 & \mathbf{8.7} \end{bmatrix}.$$

Some examples of the authentication with this value of the threshold are given in Table 1.

3. APPLICATION TO THE CATEGORIZATION OF DOCUMENTS

In applied the authentication algorithm to categorize documents belonging to the Reuters-21578 test collection. The supervised learning scenario assumes that one is provided with a training set. Each document in the training set is manually assigned to zero or more categories. The first step of classifier training is construction of a vocabulary for each category. The vocabulary consists of a number of terms (keywords), which are used to decide if a document belongs to a given category or not. We used the stemming algorithm [2] to group different grammatic word forms. We also assume the multivariate binomial data model [3] when each document is represented as a binary vector x_i of length n, where $x_{ij} = 1$ if and only if the document contains the j-th term (keyword). The training of the pairwise classifier was performed as follows. For each category, documents of the training set are used to construct the vocabulary according to the mutual information criterion. Then the training documents assigned to each category are represented by binary vectors, and the threshold values Λ_α are computed according to (11), where α is the parameter used to balance the probability of false acceptance and false rejection. To classify a document, it was first transformed to a binary vector $y^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)})$, where $y_t^{(i)} = 1$ if and only if the t-th term in the vocabulary of the i-th category appears in the document. For each category the authentication test described above is applied, and the document is assigned to a category if it passes this test with positive decision.

Results of simulation are assumed to be presented in the final version.

REFERENCES

[1] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Estimating the entropy of the probability distributions for memoryless sources and spikes analysis". Invited paper to the *IEEE Special Issue in Molecular Biology and Neuroscience*, 2009.

[2] M. F. Porter, "An algorithm for suffix stripping", *Program*, vol. 14, pp. 130–137.

[3] C. D. Manning, P. Raghavan, H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Table 1: Results of processing several vectors by the verifier for the matrix $\mathbf{X}$ defined in (13) when $\varepsilon = 0$, $\alpha = 0.4$, $\Lambda = 8.9$, where $\Lambda_\ell = \Lambda(\mathbf{x}_\ell, \mathbf{y})$, $\ell = 1, \ldots, L$. If the vector $\mathbf{y}$ is equal to the vector at the ℓ-th row of the matrix $\mathbf{X}$, then the value of ℓ is given in the first column.

ℓ	$\mathbf{y}$	Λ_1	Λ_2	Λ_3	Λ_4	Λ_5	Y/N
1	001110	**8.7**	**7.5**	12.1	10.1	**8.7**	Y
2	001011	**7.5**	**6.4**	10.9	**8.9**	**7.5**	Y
5	000111	**8.7**	**7.5**	12.1	10.1	**8.7**	Y
	000010	**8.7**	**7.5**	12.1	10.1	**8.7**	Y
	000011	**8.1**	**6.9**	11.5	9.5	**8.1**	Y
	001010	**8.1**	**6.9**	11.5	9.5	**8.1**	Y
	001111	**8.1**	**6.9**	11.5	9.5	**8.1**	Y
3	101000	12.1	10.9	15.5	13.5	12.1	N
4	010011	10.1	**8.9**	13.5	11.5	10.1	N
	$\vdots$						N