ON THE UNCERTAINTY IN SEQUENTIAL HYPOTHESIS TESTING

R. Santiago-Mozos[†]*, R. Fernández-Lorenzana*[†]*, F. Pérez-Cruz*[*] *and A. Artés-Rodríguez*[†]

[†]Dept. of Signal Processing and Communications
Universidad Carlos III de Madrid, Leganés (Madrid), SPAIN

[*]Department of Electrical Engineering
Princeton University, Princeton (NJ), USA

ABSTRACT

We consider the problem of sequential hypothesis testing when the exact pdfs are not known but instead a set of iid samples are used to describe the hypotheses. We modify the classical test by introducing a likelihood ratio interval which accommodates the uncertainty in the pdfs. The test finishes when the whole likelihood ratio interval crosses one of the thresholds and reduces to the classical test as the number of samples to describe the hypotheses tend to infinity. We illustrate the performance of this test in a medical image application related to tuberculosis diagnosis. We show in this example how the test confidence level can be accurately determined.

Index Terms— Uncertainty, sequential hypothesis testing, sequential probability ratio test, binary hypothesis test.

1. INTRODUCTION

The classical sequential hypothesis test [1] relies on the perfect knowledge of the probability density function (pdf) for each hypothesis. There are, however, many different practical scenarios, ranging from machine learning and information theory to neuroscience, in which each hypothesis is characterized by a set of samples and the actual pdfs are unknown. We have previously proposed several heuristics to manage uncertainty in sequential tests for continuous random variables that work well in practice [2, 3].

If the random variables are discrete, as we assume in this paper, the sequential test can use the maximum likelihood estimates of the pdfs. However, the theoretical performance guarantees of the sequential test, in terms of missdetection and false alarm probabilities, do not hold for estimates of the pdfs. In this paper we propose a sequential test that takes into account the uncertainty in the pdf and fulfills the specified missdetection and false alarm rates, as the classical sequential test does with perfect statistical knowledge.

We introduce a new framework that models the uncertainty in the pdfs estimation and carry it onto the sequential test. The uncertainty in the pdf estimation is measured by confidence intervals. The modified sequential test is based on the likelihoods confidence intervals, which are computed from the pdfs' confidence intervals. We also provide tight bounds for its performance. For simplicity, we develop the proposed framework for the binary random variables, which can be extended for any discrete random variable. In the binary case, both hypotheses are modeled by Bernoulli random variables, which are estimated using samples. Despite its simplicity this case is useful for any binary classifier application, which include, among others, medical and military applications, decision making and distributed detection.

We specifically apply this framework to the medical diagnosis of tuberculosis (TB) patients. A local binary detector analyzes microscopic auramine-stained images from patient sputum to detect TB bacillus. The outputs of this detector (bacillus/ not bacillus) are combined in our sequential test to determine whether the patient is infectious.

The rest of the paper is organized as follows. We first present our binary sequential probability ratio test in Section 2. Then, in Section 3 we show how this procedure can be effectively applied to the TB infectious diagnosis. Finally, in the conclusions we remark our main contributions and propose some future work.

2. SEQUENTIAL HYPOTHESIS TESTING

We propose a sequential test for deciding whether the samples from $S = \{z_1, z_2, \ldots\}$, $z_i \in \{0, 1\}$, were generated by:

H_1: Bernoulli(p).

H_0: Bernoulli(q).

The fixed probabilities p and q are unknown and a set of binary observations are given to characterize them; $S_{H_1} = \{x_1, \ldots, x_N\}$ for hypothesis H_1; and, $S_{H_0} = \{y_1, \ldots, y_M\}$ for H_0.

We first review the classical sequential test and then we

[*]F. Pérez-Cruz is supported by Marie Curie Fellowship AI-COM.

[†]This work has been partly supported by Ministerio de Educación y Ciencia of Spain (project 'MONIN', id. TEC2006-13514-C02-01)), and the Comunidad de Madrid (project 'PRO-MULTIDIS-CM', id. S0505/TIC/0223).

modify it for accounting for imperfect information about both hypotheses.

2.1. Sequential test

A sequential test reads one sample at the time from S and decides if it has enough information to decide whether the set belongs to hypothesis H_1 or H_0 or it needs to process the next example. The sequential probability ratio test uses the likelihood of the hypotheses to take such decision [1, 4]. We can compute the likelihood of H_1 for the first k samples of S:

$$L^k_{H_1} = p^{k^*}(1-p)^{k-k^*}$$

where $k^* = \sum_{i=1}^{k} z_i$. And the likelihood of H_0:

$$L^k_{H_0} = q^{k^*}(1-q)^{k-k^*}$$

A sequential probability ratio test compares the Likelihood Ratio γ_k for $k = 1, \ldots$

$$\gamma_k = \frac{L^k_{H_1}}{L^k_{H_0}}$$

with two thresholds π_u, π_l, which depend on P_{FA} and P_D, the desired False Alarm and Detection probabilities, respectively.

$$\text{Test output} = \begin{cases} H_1, & \gamma_k \geq \pi_u \\ H_0, & \gamma_k \leq \pi_l \\ \text{continue}, & \pi_l \leq \gamma_k \leq \pi_h \end{cases}$$

in which $\pi_u = \frac{P_D}{P_{FA}}$ and $\pi_l = \frac{1-P_D}{1-P_{FA}}$ [4]. Wald proved this test finish with probability one [5]. When this sequential test finishes its performance is given by P'_D and P'_{FA} such that

$$\begin{aligned} P'_{FA} &\leq \frac{P_{FA}}{P_D} \\ 1 - P'_D &\leq \frac{1-P_D}{1-P_{FA}} \\ P'_{FA} + 1 - P'_D &\leq P_{FA} + 1 - P_D \end{aligned} \tag{1}$$

The obtained P'_{FA} and P'_D are close to the desired P_{FA} and P_D for most practical detection and false alarms probabilities.

2.2. Confidence Intervals for binary random variables

To estimate p we are given S_{H_1}. Now, suppose that n^* trials are equal to one. Thereby, the maximum likelihood (ML) estimate for p is $p_{ML}\frac{n^*}{N}$. However, other values of p could result the same n^*. Many methods have been proposed to estimate a confidence interval for p, see [6] and the references therein. Henderson and Mayer propose a Bayesian approach using a $Beta(a, b)$ distribution as a prior for p [7]. The posterior distribution is consequently a $Beta(n^* + a, N - n^* + b)$. The $100(1-\alpha)\%$ confidence interval $[p_l, p_h]$ for p can be computed by taking the $\alpha/2$ and $1-\alpha/2$ quantiles of this posterior distribution. If a pessimistic interval is needed the Clopper-Pearson Interval or the Blith-Still interval [8] can be used. Pessimistic intervals are preferred for critical applications.

2.3. Likelihood Ratios

When perfect statistical knowledge is available, the likelihood of each hypothesis is a number. We have model the uncertainty on p, as a confidence interval, which is the same as to model the hypothesis as a collection of pdfs. For each one of these pdfs we can obtain a likelihood value, therefore we get a likelihood interval which accounts the uncertainty on the hypothesis.

To construct such interval $[L^k_l, L^k_h]$ we need to find the maximum and minimum values of $L^k(\hat{p})$ for all $\hat{p}$ in $[p_l, p_h]$. $L^k(\hat{p})$ is maximized for $p_{\max} = \frac{k^*}{k}$. If $p_{\max}$ lies in the $[p_l, p_h]$ interval, then $L^k_h = L^k(p_{\max})$ and L^k_l is the minimum of $L^k(p_h)$ and $L^k(p_l)$, because the likelihood function is concave. Otherwise, when $p_{\max}$ is outside $[p_l, p_h]$, $L^k(p_h)$ and $L^k(p_l)$ define the likelihood interval. Therefore, the $100(1-\alpha)\%$ confidence interval for $L^k(\hat{p})$ is given by:

$$[L_l, L_h] = \begin{cases} [\min(L(p_l), L(p_h)), \max(L(p_l), L(p_h))] & p_{\max} \notin [p_l, p_h] \\ [\min(L(p_l), L(p_h)), L(p_{\max})] & p_{\max} \in [p_l, p_h] \end{cases}$$

where we have dropped the superscripts to avoid cluttering the notation.

We denote the likelihood intervals of hypothesis H_1 and H_0 as $[L_{H_1,l}, L_{H_1,h}]$ and $[L_{H_0,l}, L_{H_0,h}]$ respectively. The minimum value of the confidence interval of the likelihood ratio $\frac{L_{H_1}}{L_{H_0}}$ is reached when L_{H_1} is minimum and L_{H_0} maximum, that is, $\frac{L_{H_1,l}}{L_{H_0,h}}$ and its maximum value is $\frac{L_{H_1,h}}{L_{H_0,l}}$. The $100(1-\alpha_{H_1})(1-\alpha_{H_0})\%$ confidence interval for the quotient $\frac{L_{H_1}}{L_{H_0}}$ results

$$[LR^l_{H_1,H_0}, LR^h_{H_1,H_0}] = \left[\frac{L_{H_1,l}}{L_{H_0,h}}, \frac{L_{H_1,h}}{L_{H_0,l}}\right]$$

2.4. Sequential hypothesis testing

The proposed statistical test finishes either when $LR^l_{H_1,H_0}$ is larger than π_u, and selects H_1, or when $LR^h_{H_1,H_0}$ is less than π_l, and selects H_0. This sequential test waits until the whole interval crosses one threshold.

With probability $P_c = (1-\alpha_{H_1})(1-\alpha_{H_0})$ the likelihood ratio is in the interval $[LR^l_{H_1,H_0}, LR^h_{H_1,H_0}]$ and the sequential test provides the performance guarantees shown in (1). In order to bound our sequential test performance, we assume $P_{D,n} = 0$ and $P_{FA,n} = 1$ for the other cases. The final sequential test performance is bounded by:

$$\begin{aligned} P_{D,\text{end}} &\geq P_c P'_D + (1-P_c)P_{D,n} = P_c P'_D \\ P_{FA,\text{end}} &\leq P_c P'_{FA} + (1-P_c)P_{FA,n} = P_c P'_{FA} + (1-P_c) \end{aligned}$$

This performance is dominated by P_c.

This test explicitly shows the relation among the global performance requirements, the uncertainty in the hypotheses and the number of test samples. When high detection probabilities and low false alarm rates are needed: large confidence intervals for p and q have to be used as P_c must be very close to 1; this makes a large confidence interval for the likelihood ratio and the sequential test needs many samples to finish. Low training sample sizes bring larger intervals for the same confidence and more test samples are needed to achieve the same performance. Furthermore, the maximum value of P_c is limited by the confidence intervals of p and q. If those intervals overlap the test never finish. Arbitrary performance is not possible with uncertainty in the hypotheses, even with an infinite number of samples.

3. EXPERIMENTS

Nowadays Tuberculosis (TB) is an important health problem [9]. To assist the human expert, automated machine-learning-based diagnosis techniques perform bacilli recognition in auramine-stained microscopic images of the sputum [10, 11]. Those *one-stage* approaches classify the patient as infectious if a bacillus is detected. However, the bacilli classifier needs to have high specificity (low false-alarm rate) to attain a good patient performance.

To overcome this difficulty we propose a *two-stages* approach to this problem. We add a patient classifier which combines the bacilli classifier outputs and makes the system more robust against the bacillus classifier false alarm rate. The system analyzes the patient sputum image by image. First, the image is divided in small pieces which are examined by the bacilli classifier. Then, the patient classifier, which is a sequential test, merges the bacilli classifier outputs and determines if its confidence is enough to make a decision taking into account its performance requirements. When more confidence is needed another image is analyzed. The system diagram is shown in Figure 1.

This approach tolerates the false alarm rate of the bacillus classifier much better than the one-stage ones because the performance requirements can be set in the patient classifier. It is also appropriate for this application because almost as images as desired are available to the automated classification system.

The experimental database has 44 Non-Infectious patients (NIP) and 6 infectious patients (IP). Each patient has about 300 1600x1200 RGB images which have been acquired with a 20x microscope. 29 NIP and, 3 IP have been used for training purposes and the rest for testing.

3.1. Bacilli classifier

The bacilli detector decides when a small piece of the image (region) contains a bacillus. A Support Vector Machine (SVM) classifier [12] has been selected for this task. Each region has dimension 37 pixel $\times$ 37 pixel $\times$ 3 colors = 4107, which is quite high. To reduce dimensionality principal component analysis (PCA) [13] has been applied to reduce the problem to 200 dimensions, which greatly simplifies the learning task and lowers its burden. Finally, the classifier complexity has been reduced from 4234 to 200 support vectors to speed up it by using the preimage method [14].

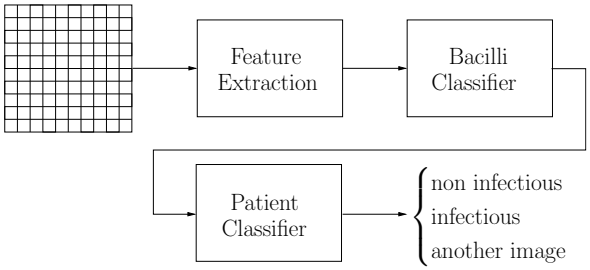

Fig. 1. Automated detection of infectious patients system.

The training set contains roughly 10000 regions labeled as bacillus, which include real bacillus, labeled by an expert, and rotations and/or displacements of them. About 20000 regions have been selected for background (regions where the bacillus is not present) from the training NIP.

3.2. Patient Classifier

The patient classifier is the sequential test proposed in Section 2. First this problem must be modeled as a binary hypothesis testing one. IP and NIP classes are very rich, therefore the training patients in each class can not be mixed in a single model. Hypothesis H_1 is modeled by the worst IP, which is the one with less bacilli detections; and hypothesis H_0 by the worst NIP, which is the one with more bacilli detections. For H_1 hypothesis it has been selected an IP with 36 bacilli detections in 441696 regions; and, for H_0 a NIP with 26 bacilli detections in 392160 regions. The bacillus probability of H_0 ($q_{ML} = 6.62 \times 10^{-5}$) is just the false alarm rate of the bacilli classifier in the worst NIP. The corresponding probability for H_1 ($p_{ML} = 8.15 \times 10^{-5}$) is the sum of false alarms plus detections rates in the worst IP.

The test patients were analyzed starting with $\alpha_{H_1} = \alpha_{H_0} = 10^{-5}$ and $P_{FA} = 1 - P_D = 10^{-6}$ by examining their available images. The quality constraints were lowered each time more samples were needed to meet them. The results of the comparison between our method and the classical one are showed in Tables 1 and 2. The first three columns show the true hypothesis of the patient, an identifier and the available number of regions. The five next columns show the decision made by the patient classifier; when the initial constraints were met; the final $P_{FA,end}$ and $P_{D,end}$ achieved by the test with the available samples; and, the final number of samples used by the test. Both test perform well. But our test is more

class	id	reg#	dec.	done	$P_{FA,end}$	$P_{D,end}$	reg. used
H_1	IP1	404544	H_1	no	0.0054	0.99461	319000
H_1	IP2	390784	H_1	no	0.0003	0.99968	350000
H_1	IP3	434816	H_1	no	0.0305	0.96952	434816
H_0	NIP1	434816	H_0	no	0.0013697	0.99863	434000
H_0	NIP2	401792	H_0	no	0.0042189	0.99578	401000
H_0	NIP3	410048	H_0	no	0.069017	0.93098	366000
H_0	NIP4	456832	H_0	no	0.0010217	0.99898	453000
H_0	NIP5	426560	H_0	no	0.0023449	0.99766	426000
H_0	NIP6	377024	H_0	no	0.023723	0.97628	369000
H_0	NIP7	467840	H_0	no	0.0015859	0.99841	465000
H_0	NIP8	415552	H_0	no	0.0018363	0.99816	412000
H_0	NIP9	408672	H_0	no	0.015165	0.98484	398000
H_0	NIP10	410048	H_0	no	0.0020249	0.99798	408000
H_0	NIP11	412800	H_0	no	0.0065601	0.99344	412800
H_0	NIP12	410048	H_0	no	0.076633	0.92337	332000
H_0	NIP13	437568	H_0	no	0.0029946	0.99701	437000
H_0	NIP14	443072	H_0	no	0.0018363	0.99816	442000
H_0	NIP15	448576	H_0	no	0.0013697	0.99863	448000

Table 1. Classical sequential test decisions and confidences.

class	id	reg#	dec.	done	$P_{FA,end}$	$P_{D,end}$	reg. used
H_1	IP1	404544	H_1	no	0.085157	0.91484	312000
H_1	IP2	390784	H_1	no	0.076633	0.92337	350000
H_1	IP3	434816	H_1	no	0.099921	0.90008	398000
H_0	NIP1	434816	H_0	no	0.080774	0.91923	396000
H_0	NIP2	401792	H_0	no	0.085157	0.91484	348000
H_0	NIP3	410048	H_0	no	0.10544	0.89456	354000
H_0	NIP4	456832	H_0	no	0.080774	0.91923	396000
H_0	NIP5	426560	H_0	no	0.080774	0.91923	423000
H_0	NIP6	377024	H_0	no	0.094712	0.90529	319000
H_0	NIP7	467840	H_0	no	0.080774	0.91923	436000
H_0	NIP8	415552	H_0	no	0.080774	0.91923	412000
H_0	NIP9	408672	H_0	no	0.089797	0.9102	387000
H_0	NIP10	410048	H_0	no	0.080774	0.91923	396000
H_0	NIP11	412800	H_0	no	0.085157	0.91484	401000
H_0	NIP12	410048	H_0	no	0.10544	0.89456	314000
H_0	NIP13	437568	H_0	no	0.085157	0.91484	375000
H_0	NIP14	443072	H_0	no	0.080774	0.91923	423000
H_0	NIP15	448576	H_0	no	0.080774	0.91923	410000

Table 2. Our sequential test decisions and confidences.

pessimistic as we see in the achievable $P_{D,end}$ and $P_{FA,end}$. On the other hand, the classical test gives overconfident probabilities because it have few samples to estimate the bacillus probability, which is very low.

4. CONCLUSIONS AND FUTURE WORK

We have proposed a novel sequential probability ratio test that models the uncertainty when the exact pdfs are unknown. We also obtained bounds of the maximum attainable performance of the test. We have developed it for the interesting case of binary hypotheses; the extension to general discrete models will be presented in a future work. We have compared the performance of this test with the classical test in the diagnosis of tuberculosis and have shown how the uncertainty in the pdfs estimates leads to more samples for the same quality constraints in terms of false alarm and detection probabilities.

5. REFERENCES

[1] A. Wald, *Sequential Analysis*, John Wiley and Sons, New York, 1947.

[2] Ricardo Santiago-Mozos and Antonio Artés-Rodríguez, "Distributed hypothesis testing using local learning based classifiers," in *ICASSP*, 2006, vol. 4, pp. 861–864.

[3] Ricardo Santiago-Mozos and Antonio Artés-Rodríguez, "Uncertainty-based censoring scheme in distributed detection using learning techniques," in *IPMU*, July 2006, pp. 2027–2034.

[4] H. Vincent Poor, *An introduction to signal detection and estimation*, Springer-Verlag New York, Inc., New York, NY, USA, second edition, 1994.

[5] Wald, A. and Wolfowitz, J., "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, sep 1948.

[6] L. D. Brown, T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical Science*, vol. 16, pp. 101–133, 2001.

[7] Henderson, Michael and Meyer, Mary C., "Exploring the confidence interval for a binomial parameter in a first course in statistical computing," *The American Statistician*, vol. 55, no. 4, pp. 337–344, nov 2001.

[8] Blyth, Colin R. and Still, Harold A., "Binomial confidence intervals," *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 108–116, mar 1983.

[9] Council of the Infectious Disease Society of America, "Diagnostic standards and classification of tuberculosis in adults and children," *Am. J. Respir. Crit. Care Med.*, vol. 161, no. 4, pp. 1376–1395, 2000.

[10] K. Veropoulos, G. Learmonth, C. Campbell, B. Knight, and J. Simpson, "The automated identification of tubercle bacilli in sputum: A preliminary investigation.," *Analytical and Quantitative Cytology and Histology*, vol. 21, no. 4, pp. 277–281, Aug. 1999.

[11] M.G. Forero, G. Cristobal, and M. Desco, "Automatic identification of mycobacterium tuberculosis by gaussian mixture models," *Journal of Microscopy*, vol. 223, no. 2, pp. 120–132, 2006.

[12] Vladimir N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[13] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[14] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, USA, 2001.