Uncertainty-based Censoring Scheme in Distributed Detection Using Learning Techniques

Ricardo Santiago-Mozos and **Antonio Artés-Rodríguez**
Dept. of Signal Processing and Communications
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911, Leganés (Madrid) SPAIN
{rsmozos,antonio}@tsc.uc3m.es

Abstract

In this paper a novel approach to distributed detection is proposed. We use learning-based local classifiers and a likelihood ratio test (LRT) at the fusion centre. As the soft outputs of the local detectors are not restricted to have any probabilistic meaning, we estimate the a posteriori probabilities of the local classifiers outputs to formulate the LRT in the fusion centre.

The uncertainty of the estimated densities can produce a biased LRT. We propose to use confidence intervals for the conditional densities to determine the regions where the LRT is precise. This regions provide simple censoring schemes that only allow transmissions of reliable information for the decision.

Also, we develop the Neyman Pearson and sequential probability ratio test for this scheme. The proposed procedure is then applied to the automated infectious tuberculosis diagnosis.

Keywords: Distributed detection, learning-based classifiers, censoring, confidence intervals.

1 Introduction

Distributed detection common approach assumes precise probabilistic models for the local classifiers (sensors) and likelihood ratio test (LRT) for the fusion centre [21]. However, imprecisions or errors in the modelling can degrade significantly the performance of the test. Also, it is usually difficult to obtain good probabilistic models.

An statistical learning method that provides both local sensors and fusion centre rules using just a training set (each sample consist of all sensor outputs and the true decision) has been recently proposed [13]. This avoids the probabilistic modelling but it has two principal drawbacks: the obtention of the training set can be difficult and the failure of a single sensor provokes the hole system be retrained.

We proposed in previous works, [1, 2], the use of learning-based local detectors in the context of target detection in sensor networks. The probabilistic interpretation of the local classifier output is provided by the underlying physics of the sensed phenomenon. This approach does not suffer from the above drawbacks as all the sensors are identical and the data fusion is a LRT. In [18] we extended [1] in a more general setting considering the design of local classifiers and the probabilistic interpretation of their outputs. Also we proposed and ad-hoc censoring scheme.

In this work we generalise [18] by proposing a more principled approach to the censoring scheme. We consider censoring schemes like [16] not only to avoid uninformative transmissions to the fusion centre but also taking into account the uncertainty in the probabilistic interpretation of the sensor output. To manage that uncertainty we obtain

confidence intervals for the conditional probability density functions (pdfs). The information of the sensors is then transmitted if it is informative enough and is in a region where the confidence intervals of the conditional pdfs do not overlap.

The proposed method applicability goes beyond the typical application of sensor networks and, as an illustration, we address a medical image diagnosis problem: the detection of infectious tuberculosis patients.

The paper continues as follows: in Section 2 we address the problem of using a LRT at the fusion centre when the local detectors are designed using statistical learning methods. The confidence intervals calculation is described in Section 3. In Section 4 we develop the batch and sequential LRT and the asymptotic performance of the Neyman-Pearson (NP) test. We propose different censoring schemes in Section 5. The effectiveness of the methods is shown in the above mentioned medical diagnosis problem using real data in Section 6, and Section 7 concludes the paper.

2 Distributed Detection and Learning

Two hypothesis are considered: H_0 or null hypothesis, and H_1 or alternative hypothesis. The fusion centre inputs arrive from ℓ identical binary local classifiers (or sensors), each one providing a real soft output x_i. The local classifiers are designed using some generative or discriminative statistical learning method [9] for solving a classification problem related, but not necessarily identical, to the discrimination between H_0 and H_1.

The formulation of a LRT in the fusion centre needs the knowledge of the, in general not available, conditional densities $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$. If the learning method is a generative one, the generative model can provide these densities either directly or after some transformation. However, pure discriminative classifiers are simpler, have less computational requirements, and offer better performance [9] in terms of error classification.

Among the discriminative methods, the ones based on the Empirical Risk Minimisation (ERM) principle [20], such Neural Networks (NN) or Support Vector Machines (SVM), are the most widely used. Some cost functions in ERM provide solutions in which the soft output of the classifier is directly interpretable as a posterior probability [4] but at the cost of complex classifiers or suboptimum classifier architecture determination. Others, on the contrary, like the one used in SVM, offer excellent discrimination performances but its soft output has no probabilistic interpretation. Even more, different cost functions can provide the same classification boundary that tends to the Bayes optimum classifier but different soft output.

As the goal of the classifier is not necessarily the discrimination between H_0 and H_1, and we prefer pure discriminative methods we propose to estimate $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ from a training set (see [19] for different methods of density estimation).

The accuracy of the estimated densities $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ is very important for a LRT-based fusion centre. Uncertainty in the estimates can induce a bias in the LRT. This is very typical in the tails of the pdfs where $\frac{f_{\mathbf{X}|H_1}(\mathbf{x}|H_1)}{f_{\mathbf{X}|H_0}(\mathbf{x}|H_0)}$ can be artificially high (or low). To avoid biased (erroneous) decisions, confidence intervals must be calculated for the estimated pdfs allowing to measure quantitatively where the pdfs can provide reliable information.

3 Determination of confidence intervals

To obtain confidence intervals for the estimated pdfs we use Bootstrap methodology [7]. Let $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ a training set of the conditioned sensor output. Let $\hat{F}$ the empirical distribution of $\mathbf{Z}$. Bootstrap constructs B different datasets (bootstrap samples) of the same size than $\mathbf{Z}$ according to $\hat{F}$, that is, sampling with replacement from $\mathbf{Z}$. For each bootstrap sample $\mathbf{Z}^k$, $k = 1, \ldots, B$ a bootstrap pdf f_k is estimated.

To obtain the confidence interval for point x we evaluate all the bootstrap pdfs in that point $f_k(x)$, $k = 1, \ldots, B$. This produces B values for that point and a confidence interval can be obtained using the percentile, BCa or ABC methods [7].

We use two methods to estimate the pdfs. First, we use Gaussian Mixture Models (GMM) [12]. GMM are suitable for problems where the dataset is large and can provide coarse estimates. To determine the appropriate number of gaussians K we use a bayesian method [17]. Then we estimate the model using the EM method [6]. This results in B collections of the mixing parameters $\{(\pi_1, \boldsymbol{\mu}_1, \mathbf{C}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \mathbf{C}_K)\}$. Each collection generates a different pdf f_k.

Second, we use Parzen density estimation [14]. Parzen estimation main advantage is the accuracy and its main disadvantage is the computational cost when the number of points is very large. We propose it to get accurate confidence intervals of the pdf locally in the regions of interest. We use, as usual, a gaussian kernel for the Parzen estimate. The length of this kernel is obtained using a jackknife-maximum likelihood method [11]. Obviously, for each bootstrap sample the Parzen estimator obtains a different pdf. This pdfs are especially different in regions with very few points. By using this procedure, the regions where the accuracy of the pdf estimates is low can be easily determined.

4 Hypothesis Testing

Assuming conditionally independence between the local detector outputs, and being ℓ_t the number of sensors which are allowed to transmit, the conditional probabilities at the fusion centre are

$$f_{\mathbf{X}|H_1}(\mathbf{x}|H_1) = \prod_{i=1}^{\ell_t} f_{X|H_1}(x_i|H_1) \prod_{j=\ell_t+1}^{\ell} P(x_j \in \bar{\mathcal{R}}|H_1)$$

$$f_{\mathbf{X}|H_0}(\mathbf{x}|H_0) = \prod_{i=1}^{\ell_t} f_{X|H_0}(x_i|H_0) \prod_{i=\ell_t+1}^{\ell} P(x_i \in \bar{\mathcal{R}}|H_0)$$

where $\mathbf{x}$ are the observations and $\bar{\mathcal{R}}$ is complementary of $\mathcal{R}$, that is, the region where transmission is not allowed. We will denote $P(x_i \in \bar{\mathcal{R}}|H_1)$ as $P_{\bar{\mathcal{R}}|1}$ and $P(x_i \in \bar{\mathcal{R}}|H_0)$ as $P_{\bar{\mathcal{R}}|0}$ for short. Note that the sensors not transmitting also contribute to the log likelihood ratio (LLR). The LLR test for a sensing instant is

$$\gamma = \ln \frac{f_{\mathbf{X}|H_1}(\mathbf{x}|H_1)}{f_{\mathbf{X}|H_0}(\mathbf{x}|H_0)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{1}$$

When the number of sensor ℓ tends to infinity, the LLR γ tends to a normal random variable. Therefore, the threshold τ for NP test of level α can be obtained by asymptotic gaussianity (as in [1]) leading to

$$\tau = \sqrt{\ell(E\{\gamma_{H_0}^2\} - D^2(H_0||H_1))}Q^{-1}(\alpha) - \ell D(H_0||H_1)$$

where Q is the Marqum's function and Q^{-1} its inverse, D is the Kullback-Leibler (KL) divergence [5], $\gamma_{H_0} = \gamma|_{H=H_0}$ and $D(H_i||H_j) = D(f_{\mathbf{X}|H_i}(\mathbf{x}|H_i)||f_{\mathbf{X}|H_j}(\mathbf{x}|H_j))$.

To obtain the performance of the NP test we use the large deviation theory [5]. If ϵ_n is the probability of error (of some kind) obtained with n observations, the error exponent is defined as $\lim_{n\to\infty} -\frac{1}{n}\ln\epsilon_n$. In NP test, the best error exponent is given by Stein's lemma [5], that says that for any P_{FA} (Probability of False Alarm) $\alpha \in (0, 1)$

$$\lim_{n\to\infty} \frac{1}{n} \ln \beta_n = -D(f_{\mathbf{X}|H_0}(\mathbf{x}|H_0)||f_{\mathbf{X}|H_1}(\mathbf{x}|H_1)) \tag{2}$$

where β_n is the miss probability for a NP test with n observations. Accordingly, the purpose of the censoring scheme is maximise the KL divergence of the transmitted decisions.

A sequential test compares the accumulated LLR γ_k for $k = 1 \ldots \ell_t$

$$\gamma_k = (\ell - \ell_t) \ln \frac{P_{\bar{\mathcal{R}}|1}}{P_{\bar{\mathcal{R}}|0}} + \sum_{i=1}^{k} \ln \frac{f_{X|H_1}(x_i|H_1)}{f_{X|H_0}(x_i|H_0)}$$

with two thresholds π_u, π_l, which depend on P_{FA} and P_D (Probability of Detection).

$$\gamma_k \begin{cases} \geq \pi_u & \text{output } H_1 \\ \leq \pi_l & \text{output } H_0 \\ & \text{otherwise continue.} \end{cases}$$

By using Wald's approximations [15] $\pi_u \approx \ln \frac{P_D}{P_{FA}}$ and $\pi_l \approx \ln \frac{1-P_D}{1-P_{FA}}$. Note that sequential test does not necessarily use all the information available from the sensors.

For instance, if γ_k exceeds π_u or π_l for $k < \ell_t$ the test ends, no more information is needed. On the other hand, if γ_{ℓ_t} does not exceed π_u or π_l H_0 can be decided if $\gamma_{\ell_t} \leq 0$ or H_1 if $\gamma_{\ell_t} \geq 0$ but none of these decisions meets the quality constraints. In this situation, many applications made no decision and another sensing round is used to accumulate more evidence about the hypothesis. This procedure is repeated until a decision can be made.

The arbitrary precision is not the only nice property of Sequential Probability Radio Test (SPRT). Also, the average number of observations needed by an SPRT is not larger than the number of observations needed by a fixed-number of observations test (like NP). The expected number of observations needed by a sequential test to fulfill some requirements is analysed in [15].

5 Censoring

Once $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ are known (estimated) the region $\mathcal{R}$ where the transmission is allowed can be determined. $\mathcal{R}$ can be restricted to the values that contribute significantly to the LLR. All the x_i such that $\ln \frac{f_{X|H_0}(x_i|H_0)}{f_{X|H_1}(x_i|H_1)} \approx 0$ should not be transmitted. Note that if $\mathcal{R}$ has a small probability, no transmission will be allowed. In this case, $P_{\bar{\mathcal{R}}|1}$ and $P_{\bar{\mathcal{R}}|0}$ will determine the test output. For the sake of simplicity, let assume without loss of generalisation the following: the local classifier output is positive if it detects the target and its value is greater as its certainty increases; the same occurs for negative classifier output; assume that the a priori probability of target present is small. There are some simple possibilities for the region selection:

1. Only transmissions to confirm H_1 hypothesis are made. $\mathcal{R} = \{x_i \in (t_{h1}, t_{h2})\}$ is the interval where pdfs confidence intervals do not overlap and $\ln \frac{f_{X|H_1}(x_i|H_1)}{f_{X|H_0}(x_i|H_0)} > C$, where $C > 0$ is the threshold for non informative sensor outputs. The rational behind this strategy is that the a priori probability of H_1 is small in many practical applications. This allows a substantial reduction of the transmissions needed.

2. In addition to the previous criterion, when a sensor is quite confident about H_0 hypothesis its transmission is also allowed. $\mathcal{R} = \{x_i \in (t_{l1}, t_{l2}) \cup (t_{h1}, t_{h2})\}$ are intervals where pdfs confidence intervals do not overlap and $|\ln \frac{f_{X|H_1}(x_i|H_1)}{f_{X|H_0}(x_i|H_0)}| > C$; LLR lower than zero confirm H_0 and greater than zero confirm H_1. Transmitting information about H_0 can be useful if the sensors are distributed in such a way that a very confident negative decision can confirm H_0 with high probability and make the test finish quickly.

3. $\mathcal{R} = \{x_i \in (t_1, t_2) \cup (t_3, t_4) \cup \ldots\}$. This is a generalisation of the above schemes, by using two or more intervals where discrimination between hypothesis is high.

There is always a tradeoff between accuracy and power/bandwidth consumption. With perfect statistical information the best we can do for accuracy is to avoid censuring. But in practice, regions where the estimates of $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ are poor should not be used. This poor estimates can produce artificial LLR that may push the test towards the wrong direction. By using confidence intervals for the pdfs, poor regions are identified where the confidence intervals of the pdfs of both hypothesis overlap.

6 Experimental Results

To illustrate the usefulness of the proposed procedure, we consider an application outside the typical one in distributed detection: sensor networks. Here we address the medical diagnosis of patients who suffer tuberculosis (TB) from infectious (I) to non infectious (NI) at the fusion centre using a local detector which analyses microscopic images from patient's sputum to detect the TB bacillus. As mentioned in previous sections, the goal of local detector (the detection of TB bacilli) is different from the overall or data fusion goal

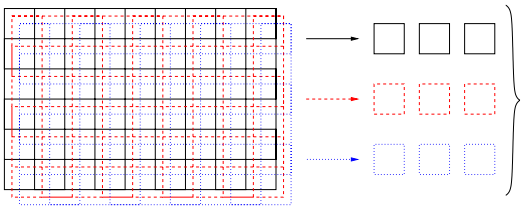


Figure 1: Regions obtained from an image.

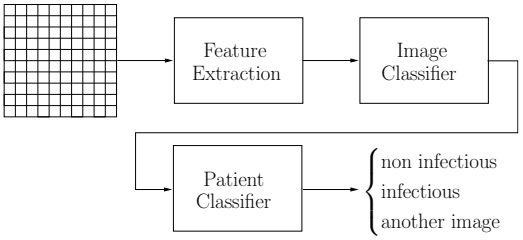


Figure 2: Automated detection of infectious patients system.

(the classification of the patients). This application requires a very low false alarm rate and TB bacillus labelling is a very expensive task. In this situation, it is very difficult to obtain a learning based classifier that fulfills the requirements. A statistical model of the target is very difficult too. On the other hand, as many images as required can be obtained from the sputum.

A decentralised detection system can model very well this problem: the images are divided in small overlapping regions[1], as shows Figure 1, and each region is presented to the local detector (it can be assumed that there is a sensor centred in each region). All the informative enough local detector outputs are sent to the fusion centre, which is an SPRT, who asks for more images until the requisites are fulfilled[2]. This way the performance of the local detector is not so critical because the performance requirements can be set in the fusion centre. The system diagram is showed in Figure 2.

For the experiments we use a database with 11 I patients, and 35 NI patients. There are 897 images, 424 belonging to I patients and the rest to NI patients. The images are 1600x1200 pixels RGB, but we only use RG bands as we do not expect to find information in blue band. We employ 9 I patients and 20 NI for training purposes and 2 I patients and 15 NI patients for testing. This results in 569 images for training and 328 images for testing.

6.1 Local detector

The training set for the local detector contains 9987 regions labelled as bacillus, which have been obtained from regions centred on the bacillus that include real bacillus and rotations and/or displacements of them as virtual ones. We selected 10515 regions for the background (regions where the bacillus is not present). The test set contains 1179 regions with bacillus presence and 28295 regions from the background. Each region has dimension 41 pixel × 41 pixel × 2 bands = 3362 pixels, which is quite high. To reduce dimensionality we perform a feature extraction procedure. We apply principal component analysis (PCA) [3], linear discriminant analysis (LDA) [8] and maximisation of mutual information (MMI) [10] methods to this problem. After feature extraction a SVM classifier was trained. Table 1 shows a comparison among the accuracy obtained by these feature extraction methods. The last column in this table shows a measure of the pseudo Receiver Operation Characteristic (ROC) curve that is called Area Under the Curve (AUC). This pseudo-ROC is obtained by sweeping the bias of the SVM to achieve pairs (P_D, P_{FA}) from P_{FA}=0 to P_{FA}=1. The first row represents no feature extraction. According to this table, PCA seems the best option, but is valuable to note the high accuracy obtained by

Table 1: Classification performance using different feature extraction methods.

Method	final dimension	Accuracy	AUC
-	3362	99.8371%	0.99347
PCA20	20	99.8914%	0.99985
LDA1	1	98.7379%	0.992
MMI20	20	99.6811%	0.99836

[1] Overlapping increase system's performance and simplifies classifier design. However, if regions overlap then classifier outputs are not independent. The test for correlated outputs will be addressed in future work.

[2] We expect non independence has little influence in the final decision but it may affect quality requisites.

LDA using just 1 feature.

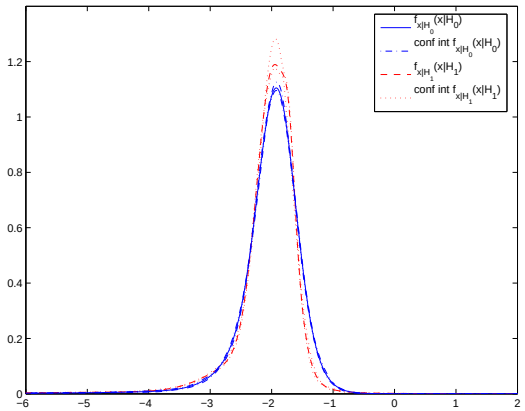
Using the pseudo-ROC the local detector has been slightly biased to lower the false alarm rate.

6.2 Fusion centre

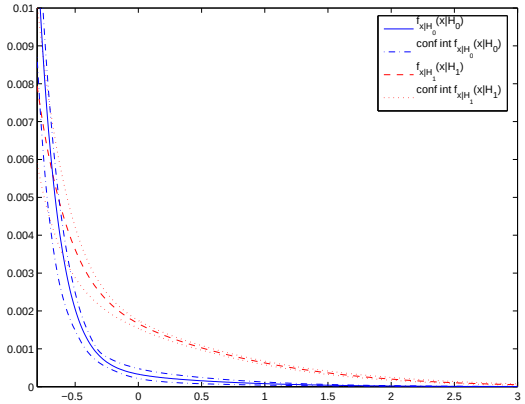
Let H_0 be the hypothesis that the patient is NI and H_1 the hypothesis that is I. To obtain the SPRT $f_{X|H_j}(x_i|H_j)$ $j = 0, 1$ need to be estimated. First, the output of the local detector (a SVM without feature extraction in this case) is obtained for all the regions of the training images. Then, a Gaussian mixture model is used to estimate $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ using the images of NI and I patients respectively. These estimates are plotted in Figure 3 (a) for a confidence value of 99%. To obtain the confidence intervals 1600 bootstrap samples where generated.

As most of the regions in NI and I patients do not contain bacilli the negative part that dominates both pdfs $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$. Also, the number of regions with bacilli is very small compared with the region without bacilli and the mass probability of both pdfs for positive outputs of the local detector is very small. However, the discrimination is possible because $f_{X|H_1}(x_i|H_1) > f_{X|H_0}(x_i|H_0)$ in the region $x_i \in (-0.54, 7.91)$ is positive as can be see in Figure 3 (b). Note that a detected bacillus corresponds to a positive output. As we note in the above section we biased the local detector to lower the false alarm rate. This is not important because we can see clearly that negative outputs greater than -0.54 increase the evidence for an ill patient.

To improve the accuracy of the confidence intervals in the selected region using GMM we estimate the conditional pdfs of the samples greater than -0.9 using Parzen method. We use samples from -0.9 instead from -0.54 to avoid border effects. The result using 1500 bootstrap samples is showed in Figure 4. Note than the region where the conditional pdfs confidence intervals do not overlap has been reduced to $x_i \in (-0.48, 3.14)$.



(a)



(b)

Figure 3: (a) GMM estimated conditional pdfs of the local detector output and confidence intervals. (b) Zoom of the pdfs and confidence intervals in the discriminative region. The confidence value selected is 99%.

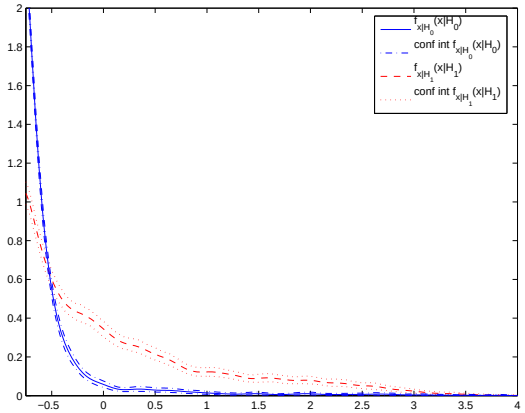


Figure 4: Parzen estimated conditional pdfs and confidence intervals of the local detector output greater than -0.9. The confidence value selected is 99%.

The censoring scheme is clear in this case, we select $\mathcal{R} = \{x_i \in (-0.33, 3.14)\}$. The lower limit has been slightly increased until $\frac{f_{X|H_1}(x_i|H_1)}{f_{X|H_0}(x_i|H_0)} \geq 2$, due to limit the transmissions to informative enough ones. Although $f_{X|H_1}(x_i|H_1) > f_{X|H_0}(x_i|H_0)$ for all points greater than zero, $f_{X|H_j}(x_i|H_j)$ are just estimates and the local detector outputs used in the estimate of the pdfs fulfil $x_i \leq 4$. Therefore the LR for $x_i > 3.14$ can be artificially very high and may confuse the test. Finally, the probabilities of not transmitting $P_{\bar{\mathcal{R}}|0}$ and $P_{\bar{\mathcal{R}}|1}$ are calculated by integrating $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ in $\bar{\mathcal{R}}$

resulting $P_{\bar{\mathcal{R}}|0} \approx 0.99975$ and $P_{\bar{\mathcal{R}}|1} \approx 0.99781$.

In Table 2 we show the performance of the patient classifier. The experiments where carried out for a $\beta = 1\text{-}P_D = 10^{-8}$ and a $\alpha = P_{FA} = 10^{-8}$[3]. In the table, *class* is the correct hypothesis of the patient; *id* means the patient's id; *reg#* the number of regions available for that patient; *dec.* the decision made, H_1 stands for I and H_0 for NI; *done* informs if the sequential test has enough confidence to take the decision; *LLR* is the log-likelihood ratio, if it is greater than 0 is more probable than the patient is infectious and, finally, *reg. used* is the number of regions analysed by the fusion centre to make the decision. The first thing to note in the table is that all the patients are well classified. However, it has not enough confidence in 5 decisions, more images where needed in these cases. If more images are not available, these cases should be supervised by a human expert. Another point is that the number of images required varies from patient to patient, we think that this depends on the number and the characteristics of the bacilli that appear in each image.

Table 2: Fusion centre decisions and confidences.

class	id	reg#	dec.	done	LLR	reg. used
H_1	675	536052	H_1	yes	18.8224	74792
H_1	738	237765	H_1	yes	18.7393	131416
H_0	139	43230	H_0	yes	-18.4208	27122
H_0	210	43230	H_0	yes	-18.4207	29614
H_0	930	43230	H_0	no	-6.28833	43230
H_0	931	43230	H_0	yes	-18.4208	23232
H_0	944	43230	H_0	yes	-18.4211	26049
H_0	945	38907	H_0	yes	-18.4214	29088
H_0	950	43230	H_0	yes	-18.4211	22956
H_0	986	43230	H_0	yes	-18.4209	21849
H_0	820	43230	H_0	no	-2.45052	43230
H_0	855	43230	H_0	no	-11.8288	43230
H_0	857	38907	H_0	no	-17.933	38907
H_0	858	43230	H_0	no	-16.5651	43230
H_0	859	38907	H_0	yes	-18.4209	27355
H_0	860	34584	H_0	yes	-18.4209	28471
H_0	861	43230	H_0	yes	-18.4211	25237

7 Conclusions

In these paper a novel approach to distributed detection has been proposed. The main novelty is the use of learning-based local detectors without statistically interpretable outputs. This opens the use of simple and pure discriminative methods from machine-learning in distributed detection. We show how to construct log-likelihood ratio tests using those local detectors and develop the NP and SPRT tests. Also, we suggest simple censoring schemes that take into account both the contribution to the LLR and the precision of the conditional pdfs estimates. This way we avoid biased LLR test due to imprecise pdf estimation and provide energy savings, which is very important in practical applications. We have successfully applied the suggested framework to the automated tuberculosis (TB) diagnosis.

[3]Note than this quality constraints may be reached too early because of the non independence of the local classifiers outputs.

Acknowledgements

This work has been partly supported by Ministerio de Educación y Ciencia of Spain (project 'DOIRAS', id. TIC2003-02602), and the Comunidad de Madrid (project 'PROMULTIDIS-CM', id. S0505/TIC/0223).

References

[1] A. Artés-Rodríguez. Decentralized detection in sensor networks using range information. In *Proceedings of the ICASSP 2004*, volume II, pages 265–268, Montreal, Canada, May 2004.

[2] A. Artés-Rodríguez, M. Lázaro, and M. Sánchez-Fernández. Decentralized detection in dense sensor networks with censored transmissions. In *Proceedings of*

the ICASSP 2005, volume IV, pages 817–820, Philadelphia, USA, March 2005.

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

[4] J. Cid-Sueiro and A. R. Figueiras-Vidal. On the structure of strict sense bayesian cost functions and its applications. *IEEE Trans. Neural Networks*, 12(3):445–455, May 2001.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc.*, 39(1):1–38, 1977.

[7] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1990.

[9] T. Jebara. *Machine Learning: Discriminative and Generative*. Kluwer Academic Publishers, 2004.

[10] J. M. Leiva and A. Artés. A gaussian mixture based maximization of mutual information for supervised feature extraction. In *5th International Conference ICA 2004*, pages 271–278, Granada, Spain, 2004.

[11] J. M. Leiva-Murillo and A. Artés-Rodríguez. A fixed-point algorithm for finding the optimal covariance matrix in kernel density modeling. Submitted to ICASSP 2006.

[12] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[13] X. Nguyen, M. Wainwright, and M. Jordan. Nonparametric decentralized detection using kernel methods. *IEEE Trans. Signal Processing*, 53(11):4053–4066, November 2005.

[14] E. Parzen. On the estimation of probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

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

[16] C. Rago, P. Willett, and Y. Bar-Shalom. Censoring sensors: a low communication-rate scheme for distributed detection. *IEEE Trans. Aerosp. Electron. Syst.*, 32(2):554–568, April 1996.

[17] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to gaussian mixture modelling. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(11):1133–1142, November 1998.

[18] R. Santiago-Mozos and A. Artés-Rodríguez. Distributed hypothesis testing using local learning based classifiers. Submitted to ICASSP 2006.

[19] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[20] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[21] P. K. Varshney. *Distributed Detection and Data Fusion*. Springer-Verlag, 1997.