DISTRIBUTED HYPOTHESIS TESTING USING LOCAL LEARNING BASED CLASSIFIERS

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}@ieee.org

ABSTRACT

In this paper we propose a novel approach to distributed detection using learning-based local classifiers and likelihood ratio test (LRT) based fusion center. Local detector's soft outputs are not restricted to have any probabilistic meaning, so even pure discriminative training method can be used. We propose to estimate the conditional densities of the soft output of the local classifiers to formulate the LRT in the fusion center. Also, we suggest simple censoring schemes that take into account the learning-based approaches problem of the slow convergence of tails of learned distributions. The Neyman Pearson (NP) and the sequential probability radio tests are developed for this approach and NP performance is analyzed. The generality of the proposed procedure is illustrated in an example outside the typical field of sensor networks: the automated infectious tuberculosis (TB) diagnosis using local detections of TB bacilli in microscopic images.

1. INTRODUCTION

Distributed detection methods are traditionally based on optimum likelihood ratio tests (LRT) relying on precise probabilistic modelling [1]. Nevertheless, erroneous or imprecise knowledge on the model can substantially degrade the performance of the tests. Recently, a statistical learning method has been proposed for the simultaneous determination of both local and global detectors [2] that overcomes this difficulty by substituting the probabilistic modelling by a set of training examples (each example is an snapshot reading of the local detectors or sensors jointly with the ground truth). However, the main drawbacks of the last approach are the unrealistic setting of the problem (for example, in a sensor network application, the procurement of the training set may be problematic, and the failure of a single sensor provokes the whole system be re-trained), and the slow convergence of the method when extreme (i.e. very small) probability of error is required.

In a previous work, [3], we suggested the possibility of using learning-based local detectors in the context of target detection in sensor networks, although the probabilistic interpretation of the local classifier output is provided by the underlying physics of the sensed phenomenon. In comparison with [2], it does not suffer from the above mentioned drawbacks because all the sensor are assumed identical and the data fusion is performed by a LRT. In this work we extend [3] in a more general setting, considering specifically the design of local detectors and its probabilistic interpretation.

For simplicity, we only consider the binary decision problem in batch and sequential form, the first using the Neyman-Pearson (NP) criterion, and the second using a Sequential Probability Ratio Test (SPRT). We will also address censoring schemes similar to the one proposed in [4] not only with the purpose of avoiding uninformative transmission to the fusion center, but also for taking into account the uncertainty in the probabilistic interpretation of the local detector output. Also, we do not consider the problem of quantization of the (soft) local detector output.

The scope of the proposed methods goes further than the typical application of sensor networks and, as an illustration, we apply them to a medical image diagnosis problem: the detection of infectious tuberculous patients.

The rest of the paper is organized as follows: in Section 2 we address the problem of using a LRT at the fusion center when the local detectors are designed using statistical learning methods. In Section 3 we obtain the batch and sequential LRT and the asymptotic performance of the NP test. We propose different censoring schemes in Section 4. The effectiveness of the methods is shown in the above mentioned medical diagnosis problem using real data in Section 5, and Section 6 concludes the paper.

2. DISTRIBUTED DETECTION AND LEARNING

We consider two hypothesis, H_0 or null hypothesis, and H_1 or alternative hypothesis, and 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 [5] 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 center 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 [5] in terms of error classification.

Among the discriminative methods, the ones based on the Empirical Risk Minimization (ERM) principle [6], 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 [7] 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 their 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. There are some attempts to provide a probabilistic interpretation to the soft output of the SVM classifiers: in [8] the SVM is introduced

This work is partly supported by the MCyT under grant TIC2003-02602.

in a Bayesian framework; in [9] SVM is interpreted as the *maximum a posteriori* (MAP) solution of the inference in a Gaussian process (GP) with adequate priors; or [10], where the use of a different cost function in SVM gives an elegant formulation of the SVM as the MAP solution of the inference in a GP.

Even if the local classifier design allows a probabilistic interpretation of its soft output, its goal can not be the discrimination between H_0 and H_1, as mentioned above (more on this in Section 5). Accordingly, we propose, as a general rule, to estimate $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ from a training set (see [11] for different methods of density estimation).

We also propose not to transmit the sensor output if it is not informative enough (similarly to [4]) or if it is located in a region where the tails of $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ are not well characterized due to limitations on the size of the training dataset. We adopt a simple censoring scheme where x_i is transmitted if $x_i \in \mathcal{R}$ where $\mathcal{R}$ is the region where the transmission is allowed.

3. 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 centers 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 to $\mathcal{R}$, 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 [3]) 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 [12], $\gamma_{H_0} = \gamma|_{H=H_0}$ 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 [12]. If ϵ_n is the probability of error (of some kind) obtained from 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 [12], that says that for any $\alpha \in (0,1)$

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

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

A sequential test compares the accumulated LLR γ_k

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

for $k = 1 \ldots \ell_t$ with two thresholds π_u, π_l, which depend on desired false alarm probability (P_{FA}) and detection probability (P_D).

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

By using Wald's approximations [13] $\pi_u \approx \ln\frac{P_D}{P_{FA}}$ 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 make 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 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 analyzed in [13].

4. 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 generalization 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). Also assume that the a priori probability of target present is small. In this situation, there are some simple possibilities for the region selection:

1. $\mathcal{R} = \{x_i \in (t_h, \infty)\}$ where $t_h \in \mathbb{R}$. Only transmissions are made to confirm H_1 hypothesis. The rational behind this strategy is that the a priori probability of H_1 is small in many practical applications.
2. $\mathcal{R} = \{x_i \in (-\infty, t_h') \cup (t_h, \infty)\}$ where $t_h' \leq t_h \in \mathbb{R}$. Now, information is added when a sensor is quite confident about H_0 hypothesis. 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. $\mathcal{R} = \{x_i \in (a,b) \cup (c,d) \cup \ldots\}$ where $a \leq b \leq c \leq d \in \mathbb{R}$. If KL divergence has two or more intervals where the discrimination between the 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 many applications huge

power/bandwidth savings can be attained with small losses in accuracy. Another important practical point is not to use regions where the estimates of $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ are poor. This poor estimates can produce artificial LLR that may push the test towards the wrong direction. This always happens when using samples of the tails of the estimated densities. Therefore, the intervals showed above should be restricted for high positive or negative values.

5. APPLICATION EXAMPLE

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) using a local detector which analyzes 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 (the classification of the patients). This application requires a very low FA rate. 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 decentralized detection system can model very well this problem: the images are divided in small overlapping regions and each region is presented to the local detector. All the informative local detector outputs are sent to the fusion center, which is an SPRT that asks for more images until the requisites are fulfilled. This way, the performance of the local detector is not so critical because the performance requirements can be set in the fusion center.

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.

5.1. Local detector

The training set for the local detector contains 9987 regions labelled as bacillus, which have been obtained from regions centered 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 $\times$ 41 pixel $\times$ 2 bands = 3362 pixels, which is quite high. To reduce dimensionality we perform a feature extraction procedure. We apply principal component analysis (PCA) [14], linear discriminant analysis (LDA) [15] and maximization of mutual information (MMI) [16] methods to this problem. Table 1 shows a comparison among these 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. Also note the high accuracy obtained by LDA using just 1 feature.

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

Table 1. Classification performance using different feature extraction methods.

5.2. Fusion center

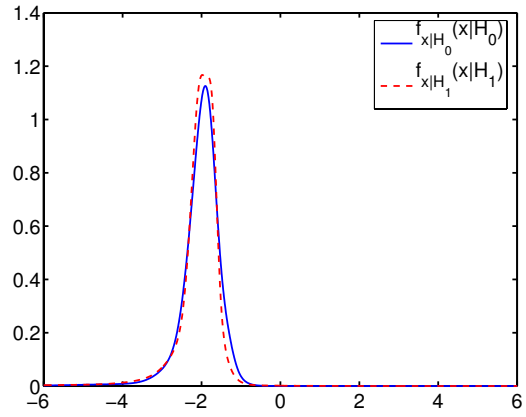
Let H_0 be the hypothesis that the patient is NI and H_1 the hypothesis that is I. To obtain the SPRT $f_{X|H_j}(x_i|H_j)$ need to be estimated. First, the output of the local detector is obtained for all the regions of the training images. Then, a Gaussian mixture model is used to estimate $f_{X|H_0}(x_i|H_0)$ and $f_{X|H_1}(x_i|H_1)$ using the images of NI and I patients respectively. The estimates are plotted in Figure 1 (a).

As most of the regions in NI and I patients do not contain bacilli the negative part dominates both $f_{X|H_0}(x_i|H_0)$ 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 densities 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 where x_i is positive as can be seen in Figure 1 (b). Note that a detected bacillus corresponds to a positive output. The censoring scheme is clear in this case, we can select $\mathcal{R} = \{x_i \in (0, \infty)\}$ because 0 is the threshold of the local detector to decide that a bacillus is present. As an alternative to 0 it is also reasonable to choose the point where $f_{X|H_1}(x_i|H_1) > f_{X|H_0}(x_i|H_0)$ in Figure 1 (b). But, as $f_{X|H_j}(x_i|H_j)$ are just estimates and the local detector outputs fulfill $x_i \leq 4$ then the LR can be artificially very high for $x_i \gg 4$ and may confuse the test. So we set $\mathcal{R} = \{x_i \in (0, 6)\}$ to avoid this situation. 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 $P_{\bar{\mathcal{R}}|0} \approx 0.99985$ and $P_{\bar{\mathcal{R}}|1} \approx 0.99837$.

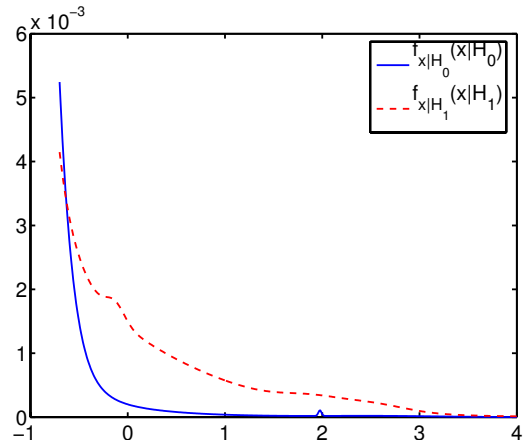
In Table 2 we show the performance of the patient classifier. The experiments where carried out for a β = 1-P_D = 1e-8 and a $\alpha = P_{FA}$ = 1e-8. In the table, *class* is the correct hypothesis of the patient; *id* means the patient's id; *reg#* the number of regions available for that patient; *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 that the patient is infectious and, finally, *reg. used* is the number of regions analyzed by the fusion center 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 are 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.

6. CONCLUSIONS

In these paper a novel approach to distributed detection has been proposed. The main novelty is the use of learning-based local detectors with no statistically interpretable outputs. This enables the use of simple and pure discriminative methods from machine-learning in distributed detection. We show how to construct log-likelihood ratio tests using these local detectors and develop the NP and SPRT tests. Also, we suggest simple censoring schemes that take into account the learning-based problem of the slow convergence of the tails of



(a)



(b)

Fig. 1. (a) Conditional probability densities of the local detector output. (b) Zoom of the probability densities in the positive part.

class	id	reg#	dec.	done	LLR	reg. used
H_1	675	536052	H_1	yes	18.9525	56721
H_1	738	237765	H_1	yes	18.9182	135035
H_0	139	43230	H_0	yes	-18.4212	30207
H_0	210	43230	H_0	yes	-18.4209	34103
H_0	930	43230	H_0	no	-2.02386	43230
H_0	931	43230	H_0	yes	-18.4208	30446
H_0	944	43230	H_0	yes	-18.4209	33443
H_0	945	38907	H_0	yes	-18.4212	35702
H_0	950	43230	H_0	yes	-18.4209	30084
H_0	986	43230	H_0	yes	-18.4211	28634
H_0	820	43230	H_0	no	-3.9428	43230
H_0	855	43230	H_0	no	-13.8922	43230
H_0	857	38907	H_0	no	-17.0282	38907
H_0	858	43230	H_0	no	-12.8873	43230
H_0	859	38907	H_0	yes	-18.4207	33200
H_0	860	34584	H_0	yes	-18.4208	33339
H_0	861	43230	H_0	yes	-18.4207	33350

Table 2. Fusion center decisions and confidences.

learned distributions, which is very important in practical applications. We have successfully applied the suggested framework to the automated tuberculosis (TB) diagnosis.

As future work, using the bounds we have obtained, and examining the relation of the probability densities before and after censoring it is possible to obtain the optimal censoring region taking into account both the performance and the stability of the tests.

7. REFERENCES

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

[2] X. Nguyen, M.J. Wainwright, and M.I. Jordan, "Nonparametric decentralized detection using kernel methods," *IEEE Trans. Signal Processing*, vol. 53, no. 11, pp. 4053–4066, November 2005.

[3] Antonio Artés-Rodríguez, "Decentralized detection in sensor networks using range information," in *Proceedings of the ICASSP 2004*, Montreal, Canada, May 2004, vol. II, pp. 265–268.

[4] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: a low communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, April 1996.

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

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

[7] Jesús Cid-Sueiro and Aníbal R. Figueiras-Vidal, "On the structure of strict sense bayesian cost functions and its applications," *IEEE Trans. Neural Networks*, vol. 12, no. 3, pp. 445–455, May 2001.

[8] J. T. Kwok, "The evidence framework applied to support vector machines," *IEEE Trans. Neural Networks*, vol. 11, no. 5, pp. 1162–1173, September 2000.

[9] Peter Sollich, "Bayesian methods for Support Vector Machines: Evidence and predictive class probabilities," *Machine Learning*, vol. 46, pp. 21–52, 2002.

[10] Wei Chu, S. Sathiya Keerthi, and Chong Jin Ong, "Bayesian trigonometric support vector classifier," *Neural Computation*, vol. 15, no. 9, 2003, 2227–2254.

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

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

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

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

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

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