Robust Multihypothesis Discrimination of Controlled I.I.D. Processes

Stéphane HERBIN

ONERA
Département Traitement de l'Information et Modélisation
29, avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex
France
E-mail: Stephane.Herbin@onera.fr

Abstract

This paper describes a general framework for the robust discrimination of objects represented as a family of i.i.d. random distributions. Testing is based on accumulating evidences on the discrimination between all-pairs of hypotheses by sampling the family of distributions according to an optimal control law. The optimality criterion is built on constraint satisfaction issues.

An application on 2D rotation invariant shape recognition with noisy contours illustrates the approach.

1. Introduction

1.1. Motivation

We investigate in this paper the optimal design of a multihypothesis decision procedure based on sequentially querying the values of several features or measures.

The object tested are modelled as an i.i.d. (independent, identically distributed) controlled process, i.e. a family of probabilities that can be selectively sampled according to a given control law.

Such representations arise when dealing with marginal distributions over some space such as filter bank histograms which are used for instance in texture discrimination. We give a less conventional application on noisy shape recognition at the end of this paper.

In a modelling phase, i.i.d. controlled processes form a redundant or overcomplete feature description of the piece of data tested. Somehow, they can be seen as an intermediate representation between the actual data (object, shape or image) and a series of moments computed from specific statistics.

The question we address is the design of an optimal combination and selection schema for multiclass discrimination based on an empirical criterion. Since in many real applications available data is scarce because costly, learning should be stable even with few samples in each class.

We expect to fulfill this objective by exploiting the robustness of vector-based binary discrimination and the flexibility of active sampling strategies.

1.2. Problem formulation

At each time T, a query or action a_T is generated towards the environment which returns a feature value x_T. All those variables may be random and time dependent. However, in this paper, we make the following hypotheses:

- Action $a \in \mathcal{A}$ and feature $x \in \mathcal{X}_a$ values are finite.
- Action a_T is generated according to a fixed stationary sampling law $\mu(a)$.
- For a given environment, feature values x are stationary and independently distributed conditionaly to the sampling action a_T.

In this framework, objects are represented as a family of i.i.d. random distributions $\mathbf{P} = \{p(x \mid a)\}_{a \in \mathcal{A}, x \in \mathcal{X}_a}$.

The problem we address is the design of a decision $B(\Phi_T) \in \{0, 1, 2, \dots M\}$ among M hypotheses based on a given sequence of action/measurements $\Phi_T = (a_1, x_1, a_2, x_2, \dots a_T, x_T)$ where 0 is a rejection label. The underlying probability family $\mathbf{P}$ can only be known by sampling it using a stationary law μ.

1.3. Paper summary

Section 2 describes the features of the testing procedure and analyzes its asymptotic behavior. The specification of

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

the sampling law μ used to combine the various probability distributions describing an object is presented in section 3. An application on noisy shape recognition is described in section 4.

2. Testing

2.1. Multiple hypotheses

The testing procedure is based on accumulating elementary evidences generated by sampling the underlying family of probabilities $\mathbf{P}$. The evidences are measured by a series of one-vs-one hypothesis comparisons $\mathbf{R}_{ij} = \{r_{ij}(x \mid a)\}_{a \in \mathcal{A}, x \in \mathcal{X}_a}$ where $r_{ij}(x \mid a)$ is a real number. A positive value means that, if the sampling $a \to x$ is observed, hypothesis i is more favorable than j.

Let $L_{ij}(\Phi_T)$ be the average empirical evidence accumulation variable $\frac{1}{T}\sum_{t=1}^{T} r_{ij}(x_t \mid a_t)$. The testing procedure $B(\Phi_T)$ is a function of its sign configuration noted $\epsilon(\Phi_T)$ and defined as:

$$B(\Phi_T) = \begin{cases} i & \text{if } \epsilon(\Phi_T).\epsilon(i) \geq \mathbf{0} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\epsilon(i)$ is a specific sign configuration pattern such that $\epsilon_{ij} = 1$ if $i > j$, $\epsilon(i,j) = -1$ if $i < j$ and 0 otherwise. The multiplication $\epsilon(\Phi_T).\epsilon(i)$ is coordinate-wise.

This multi-hypothesis test is simply: "decide i if and only if $L_{ij}(\Phi_T) > 0$ for all $j \neq i$". It is a very selective form of multiclass test and other schemes based on voting or error correcting codes are possible [1, 4]. We restrict the studies to this simple test for the sake of both presentation conciseness and tractability of further computations.

2.2. Likelihood and linear decision function

So far, the accumulated evidences have not been defined. We introduce two standard ways to compute the functions $r_{ij}(x \mid a)$. The first one is the usual log-likelihood ratio $r_{ij}(x \mid a) = \log \frac{\hat{p}_i(x \mid a)}{\hat{p}_j(x \mid a)}$ where it has been assumed that the observations are sampled from known distributions $\{\hat{p}_i(x \mid a)\}_{i=1}^{M}$. This case is treated in [6]. Finding a set of probabilities $\hat{p}_i(x \mid a)$ equivalent to evidence functions $r_{ij}(x \mid a)$ can be done by using "coupling" techniques [5].

We are more interested in this article in the expression of evidence as a *linear* decision function value. Indeed, since the accumulating scheme is additive, we have a Law of Large Numbers property:

$$\lim_{T \to \infty} L_{ij}(\Phi_T) = \mathrm{E}_{\mathbf{P},\mu}[\mathbf{R}_{ij}] \quad (2)$$

where the expectation is defined as $\mathrm{E}_{\mathbf{P},\mu}[\mathbf{Z}] = \sum_a \mu(a) \sum_{x \in X_a} p(x \mid a) z(x \mid a)$.

Assume now that, for a fixed sampling parameter a, we have built a separating hyperplane with direction $\mathbf{W}_{ij}$ and bias $b_{ij}(a)$ between two populations of random distributions labeled i and j using your favorite linear classifier learning algorithm (SVM, discriminant analysis...). Define the evidences $\mathbf{R}_{ij}$ as:

$$r_{ij}(x \mid a) = w_{ij}(x \mid a) + b_{ij}(a).$$

Since $\sum_{x \in X_a} p(x \mid a) = 1$, we have:

$$\mathrm{E}_{\mathbf{P},\mu}[\mathbf{R}_{ij}] = \sum_a \mu(a) \left(b(a) + \sum_{x \in X_a} p(x \mid a) w_{ij}(x \mid a) \right)$$

meaning, thanks to (2), that accumulating an infinite number of evidence values is equivalent to computing the value of the linear decision function at point $\mathbf{P}$ with a combination law μ between features a. We expect that a discrimination using a combination of optimal binary decision functions will be more robust than a test based on ill-estimated likelihoods.

2.3. Asymptotics

In this section we examine the asymptotic properties of tests based on evidence accumulating schemes. We first start with a property stating the condition for good recognition using an active sampling scheme on an environment described by the probability family $\mathbf{P}$ labeled $l(\mathbf{P})$:

Proposition 1 *A necessary and sufficient condition for the active sampling test error to converge to zero as the number of samples goes to infinity is:*

$$\epsilon(\boldsymbol{P}, \mu).\epsilon(l(\boldsymbol{P})) \geq \mathbf{0} \quad (3)$$

where $\epsilon_{ij}(\boldsymbol{P}, \mu)$ is the sign of $E_{\boldsymbol{P},\mu}[\boldsymbol{R}_{ij}]$ and $l(\boldsymbol{P})$ is the label associated with $\boldsymbol{P}$.

Proof This is a direct consequence of the law of large numbers (2) and of the nature of the test (1).

We can be more precise in the way the active testing process converges. Define the discrimination error P_e by enumerating the empirical sign configurations $\epsilon(\Phi_T)$ giving a wrong answer. Let $\tilde{\mathbf{E}}(i)$ be the set of sign configuration patterns ϵ which give a wrong answer if hypothesis i is true, i.e. such that the condition $\epsilon.\epsilon(i) > \mathbf{0}$ is false. We have:

$$\mathrm{P}_e = \mathrm{P}[B(\Phi_T) \neq l(\mathbf{P})] = \sum_{\epsilon \in \tilde{\mathbf{E}}(l(\mathbf{P}))} \mathrm{P}[\epsilon(\Phi_T).\epsilon \geq \mathbf{0}]$$

If the condition (3) is satisfied, the error P_e is the sum of terms which converge exponentially fast to zero, with a logarithmic speed that can be computed exactly thanks to large

deviations techniques. Indeed one can show for all ϵ such that $\epsilon.\epsilon(\mathbf{P},\mu) \geq \mathbf{0}$ is false [1] :

$$\lim_{T\to\infty} -\frac{1}{T}\log \mathrm{P}[\epsilon(\Phi_T).\epsilon \geq \mathbf{0}] = \rho(\mathbf{P},\mu,\epsilon) > 0 \quad (4)$$

If convergence to zero occurs, the global convergence rate is therefore the minimal rate over all the terms (4). The convergence speeds can be used to estimate the number of samples needed to reach a given error level.

3. Learning

The above analysis was made for a fixed stationary law μ. The final and original goal of an active sampling approach is to specify a control law that is optimal in some sense.

We have used a min-max approach based on the maximization of the $M-1$ constraint margins that have to be satisfied for each probability family $\mathbf{P}$ to fulfill condition (3). Given a learning population of i.i.d. probability families $\{\mathbf{P}_k\}_{k=1}^N$ labeled $l(k)$, we want to find the law μ^* that satisfies:

$$\mu^* = \arg\max_{\mu} \min_{k} \min_{j\neq l(k)} \mathrm{E}_{\mathbf{P}_k,\mu}[\mathbf{R}_{l(k)j}] \quad (5)$$

The solution is the result of a linear min-max optimization on linear constraints which can be solved using standard libraries. Due to the constraint $\sum_a \mu(a) = 1$, the optimal μ is usually sparse and produces an implicite feature selection.

Since the constraints (3) are linear in $\mathbf{P}$, optimizing (5) is equivalent to assert that any convex combination of the learning samples in each class can be correctly labelled by the sampling process.

Good recognition occurs when the value of $\min_k \min_{j\neq l(k)} \mathrm{E}_{\mathbf{P}_k,\mu}[\mathbf{R}_{l(k)j}]$ is strictly positive: in general, this can not be readily achieved on the original learning population. In the global learning process, we adopt a "robust" strategy by removing iteratively from the learning population part of the examples that do not satisfy condition (3). We stop when we find a solution of (5) which satisfies all the constraints on the remaining examples.

[1] The exact computation of the convergence rate is a direct application of Cramér theorem [3] and some elementary convex analysis. Due to lack of space, we only give the formula without proof:

$$\rho(\mathbf{P},\mu,\epsilon) = -\inf_{\lambda>0} \log M(\lambda;\mathbf{P},\mu,\epsilon)$$

with the Laplace transform defined as:

$$M(\lambda;\mathbf{P},\mu,\epsilon) = \mathrm{E}_{\mathbf{P},\mu}\left[\exp\left(\sum_{i>j}\lambda_{ij}\epsilon_{ij}r_{ij}(x\mid a)\right)\right].$$

4. Noisy shape recognition

4.1. Shape as family of random bipoint features

The approach described above was tested on a 2D object recognition problem based on noisy images (Figure 1). The problem is to discriminate between six types of planes observed from above in any orientation. The typical applicative context is remote sensing from an aerial camera where the distance to the object, hence the scale, is known. In this context, the nuisance parameters are clutter, object rotation, small affine distortion and sub-pixel location.

Figure 1. Three examples of images.

Figure 2 shows some examples of contours produced by a standard algorithm (Canny) on several noisy images of the same object but with various sub-pixel positions, affine distortion and noise outcomes. The images tested were synthesized using a sensor model characterized by a gaussian transfer function, additive white noise (SNR = 15dB) and uniform random spatial sub-sampling to account for aliasing. In the sensor model studied, shapes are between 25 to 60 pixels wide depending on the size of the plane observed.

Figure 2. Influence of noise on contours.

The shape representation exploit the distribution of specific features attached to pairs of contour points.

We use bipoint feature value distributions conditioned on specific orientations to model shapes as a controlled i.i.d. process. Orientation conditioning will be function of local grey-level gradients computed using standard algorithms. Figure 3 shows three examples of bipoint distributions conditionned by an angle with the local gradient equal to 0, 180 and 270 degrees.

A shape will be represented by two kinds of bipoint feature distributions: their length and the difference between local gradient angles at the two bipoint ends.

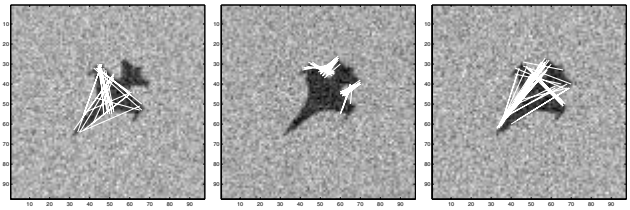

Figure 3. Examples of bipoint distributions.

Given a set of contour points, we define a control, i.e. an action, as the choice of an orientation coupled with a type of feature — gradient angle difference or bipoint length. Active sampling is the operation of generating a series of controls and collecting the random bipoint feature values.

Accumulation of graphical features has a long history in pattern recognition and computer vision. Geometric hashing [7] is one of the most famous. More recently, local statistics of graphical features have been defined as *shape contexts* and were used for multiple model matching [8]. Combination of various decision tree interpretation of random bipoint features have been described in [2].

4.2. Experiment

The shapes were modelled using 36 control orientations. To differentiate local details and long range contour point configurations, we added another control which selected, according to their length, the bipoints used for the computation of the gradient angle difference: the selection was made using two length intervals $[3, 10]$ and $[11, 70]$. The overall control space $\mathcal{A}$ resulted in 102 different types of queries.

The gradient angles were uniformely quantized with a 15° step resulting in 24 values. The bipoint lengths were quantized with a 4 pixel length step resulting in 21 values.

The learning phase – linear separating directions and feature sampling law μ – was performed using 40 samples per class resulting in $40 * 6 * 5 = 1200$ constraints to be satisfied by the min-max learning algorithm (5). The resulting sampling law was found to select 45 features.

The testing was performed on 400 sample images per class. In all the experiments we made by varying the set of examples used for learning, all the 2400 testing images were correctly recognized, i.e. the corresponding probability families satisfy the condition (3) for error convergence to zero as the number of bipoints sampled goes to infinity. Figure 4 shows the error decrease to zero for each hypothesis as the number of samples goes to infinity. On this experiment, all the test examples were correctly recognized using 450 sampled bipoints.

Compared with the approach presented in [6], the optimal sampling strategy obtained with (5) achieves a better control of the number of good recognitions with a much faster learning phase. However it can not guarantee that the error speeds of convergence will be optimal.

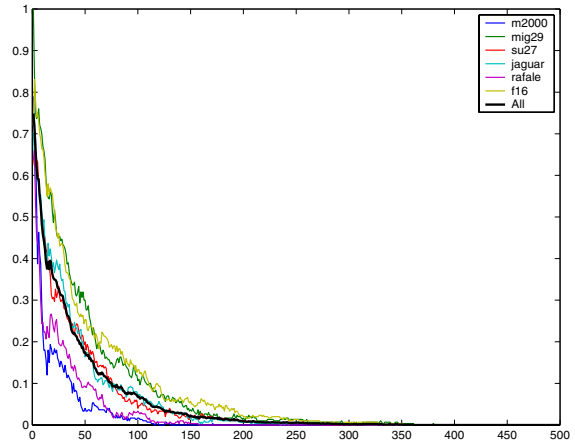

Figure 4. Recognition error.

5. Conclusion

This paper has described a framework for combining one-vs-one hypothesis testing in a global multiclass optimization framework when objects to recognize are modelled as i.i.d. controlled processes. Testing is an empirical decision produced from accumulating evidences on randomly sampled feature values. The sampling law optimizes an empirical robust criterion based on constraints that have to be satisfied for error convergence to zero of the sampling procedure.

Results on noisy shape recognition have shown that the approach performs well even with a small number of learning examples.

References

[1] E. Allwein, E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. of Mach. Learn. Res.*, 1:113–141, 2000.

[2] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neur. Comp.*, 9(7):1545–1588, 1997.

[3] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers, Boston, 1993.

[4] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *J. of Art. Int. Res.*, 2:263–286, 1995.

[5] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Ann. of Stat.*, 26(2), 1998.

[6] S. Herbin. Active sampling strategies for multihypothesis testing. In *EMMCVPR 2003*, volume 2683 of *Lect. N. on Comp. Sci.*, pages 97–112, Berlin, 2003. Springer Verlag.

[7] H. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Computational Science and Engineering*, 4(4):10–21, October 1997.

[8] H. Zhang and J. Malik. Learning a discriminative classifier using shape context distances. In *CVPR'03*, pages I: 242–247, 2003.