Density-Free Bayes Risk Consistency of Nonparametric Pattern Recognition Procedures

WŁODZIMIERZ GREBLICKI and LESZEK RUTKOWSKI

***Abstract*—Pattern recognition procedures based on the Cesàro mean of orthogonal series are presented and their Bayes risk consistency is establised. No restrictions are put on the class conditional densities.**

I. Introduction

Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be a sequence if i.i.d. pairs of random variables; Y takes values in $S = \{1, \ldots, M\}$, whereas X is A valued, where A is a Borel set of a real line R. The problem is to estimate Y from X and V_n, where $V_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ is the learning sequence. We consider a pattern recognition procedure classifying every $x \in A$ to any class i, $i \in S$, which maximizes $\hat{p}_i \hat{f}_i(x)$; $\hat{p}_i$ is an estimate of the prior class probability $p_i = P\{Y = i\}$, $\hat{f}_i(x)$ is an estimate of $f_i(x)$. Let us denote by $\hat{Y}_n$ the decision of so defined procedure. Let, moreover, $R_n = P\{\hat{Y}_n \neq Y | V_n\}$. The procedure is said to be weakly or strongly Bayes risk consistent (BRC) with respect to the class densities $f_1, \ldots, f_M$ if $R_n \overset{n}{\to} R_0$ in probability or w.p.1; R_0 is the Bayes probability of error.

Van Ryzin [8], Wolverton and Wagner [9] and Greblicki [3] examined procedures which are BRC if the class densities are square integrable or continuous almost everywhere in A. Recently, Devroye and Wagner [2] have showed that the procedures with Rosenblatt-Parzen estimates are BRC for all class densities. Thus our result, like that of Devroye and Wagner [2], is density-free.

II. Pattern Recognition Procedures

The estimator $\hat{p}_i$ is of the following form: $\hat{p}_i = n_i/n$, where n_i is the number of observations of V_n form the class i. The estimates of the class densities are based on a complete orthonormal system $\{g_k\}$, $k = 0, \ldots, 1$, defined on A. In this note

$$\hat{f}_i(x) = \sum_{k=0}^{N(n)} \left(1 - \frac{k}{N(n)+1} \hat{a}_{ik} g_k(x)\right), \tag{1}$$

where

$$\hat{a}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n} t_j^i g_k(X_j), \tag{2}$$

where t_j^i is the indicator of the class i, i.e., t_j^i is 1 or 0 according to Y_j is or differ from i.

If A is an interval, say $[-\pi, \pi]$, one can use the trigonometric system

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos kx, \frac{1}{\sqrt{\pi}} \sin kx, \ k = 1, 2, \ldots.$$

If $A = [0, \infty)$, we can apply a system

$$g_k(x) = \frac{1}{\sqrt{k+1}} x^{1/2} e^{-x/2} L_k(x),$$

where

$$L_0(x) = 1, L_1(x) = \frac{1}{k! x} e^x \frac{d^k}{dx^k} x^{k+1} e^{-x}, \ k = 1, 2, \ldots,$$

are Laguerre polynomials.

If $A = (-\infty, \infty)$, the Hermite orthonormal system

$$g_k(x) = \frac{1}{\sqrt{2^k k! \sqrt{\pi}}} e^{-x^2/2} H_k(x),$$

is applicable, where

$$H_0(x) = 1, H_k(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2}, \ k = 1, 2, \ldots.$$

For the trigonometric, the Laguerre, the Hermite systems, the following inequality is satisfied:

$$\sup_{x \in A} |g_k(x)| \leq c_1 k^r,$$

where $r = 0, 1/4, -1/12$, respectively (see Szegö [7, pp. 241, 242]).

We shall now give conditions under which the class density estimate (1) is weakly or strongly consistent.

Theorem 1: Let

$$N(n) \overset{n}{\to} \infty. \tag{3}$$

The class density estimate (1) with the trigonometric, the Laguerre and the Hermite orthonormal systems is weakly or strongly consistent at almost every $x \in A$ if

$$\frac{N^{2(1+2r)}(n)}{n} \overset{n}{\to} 0, \text{ or } \frac{N^{2(1+2r)}(n) \log n}{n} \overset{n}{\to} 0. \tag{4}$$

Proof: Observe that the conditional variance of $\hat{f}_i$ given $n_i = m$ is not greater than

$$\frac{c_1^4}{m} \left(\sum_{k=0}^{N(n)} k^{2r} \right)^2 \leq \frac{c_2}{m} N^{2(1+2r)}(n),$$

some c_2. Thus the variance of $\hat{f}_i$ converges to zero as n tends to infinity (see, e.g., Greblicki [4]). The asymptotical unbiasedness of the estimate derived from the trigonometrc system is implied by the Lebesgue theorem (see Sansone [7, p. 104]) on the asymptotical properties of the Fejèr kernel. The same prorperty for the estimates derived from the Laguerre and the Hermite orthonormal systems is a consequence of the equiconvergence theorem in Szegö [8]. In order to establish the strong consistency, it suffices to consider $\hat{f}_i(x) - E\hat{f}_i(x)$ and use Bernstein's inequality (see, e.g., Bennett [1]). ■

III. The Main Result

From Theorem 1 and that of Greblicki [3] on the Bayes risk consistency of pattern recognition procedures, we get the main result of the note.

Theorem 2: Let (3) hold. The pattern recognition procedure is weakly (strongly) BRC for all class densities if (4) is satisfied.

It should be mentioned that similar results for sequential pattern recognition procedures derived from the Haar orthonormal system have been obtained by Rutkowski [4].

W. Greblicki is with the Institute of Engineering Cybernetics, Technical University of Wrocław, Janiszewskiego 11/17, 50-370 Wrocław, Poland.

L. Rutkowski is with the Department of Electrical Engineering, Technical University of Czestochowa, Zawadzkiego 17, 42-200 Czestochowa, Poland.

References

[1] G. Bennett, "Probability inequalities for the sum of independent random variables," *J. Amer, Star. Assoc.*, vol. 57, pp. 33-45, 1962.

[2] L. P. Devroye and T. J. Wagner, "On the L_1 convergence of kernel estimators of regression functions with applications in discrimination," *Z. Wahrscheinlichkeirstheorie and Verw. Gebiete*, vol. 51, pp. 15-25, 1980.

[3] W. Greblicki, "Asymptotically optimal pattern recognition procedures with density estimates," *IEEE Trans, Inform. Theory*, vol. IT-24, pp. 250-251, 1978.

[4] W. Greblicki, "Pattern recognition procedures with nonpararactric density estimates," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-8, pp. 809-812, 1978.

[5] Rutkowski, "Sequential estimates of probability densities by orthogonal series and their application in pattern classification," *IEEE Trans. Syst. Man. Cybern.*, to be published.

[6] G. Sansone, Orthogonal Functions. New York: Interscience, 1959,

[7] G. Szegö, "Orthogonal Polynomials," *Amer. Math. Soc. Coll. Publ.*, vol. 23, 1959.

[8] J. Van Ryzin, "Bayer risk consistency of classification procedures using density estimates," *Sankhyā, Ser. A*, vol. 28, pp. 161-170, 1966.

[9] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal dis-criminant functions for pattern recognition," *IEEE Trans, Inform, Theory*, vol. IT-15, pp. 258-265, 1969.