Pattern Recognition Procedures with Nonparametric Density Estimates

WŁODZIMIERZ GREBLICKI

Abstract—**Modified class conditional density estimates for pattern classification obtained by replacing the sample sizes for particular classes by the overall sample size in expressions for the original estimates are presented, and their consistency is proved. Pattern recognition procedures derived from original and modified Rosenblaa-Parzen, Loftsgaarden-Quesenberry, and orthogonal series estimators are given, and Bayes risk consistency is established.**

I. Introduction

Pattern recognition algorithms with nonparametric density estimates have been studied by several authors. The Rosenblatt-Parzen estimate has been used by Van Ryzin [13], [14], while Devroye and Wagner have employed both the Rosenblatt-Parzen and the Loftsgaarden-Quesenberry estimates. Van Ryzin [14] has also examined procedures with orthogonal series estimates.

In this correspondence we present pattern recognition procedures with the Rosenblatt-Parzen, the Loftsgaarden-Quesenberry, and the orthogonal series class density estimates of their original forms, and we introduce modified class density estimates and derive appropriate procedures. The modified estimates are obtained by replacing the sample sizes for particular classes by the overall sample size in the expressions for the original estimates. We prove Theorems 4 and 7 on consistency of the modified class density estimates, and then using general the Greblicki [7] and the Wolverton-Wagner [16] theorems on Bayes risk consistency, we establish Theorems 5 and 8 on asymptotical optimality of so obtained procedures. We also show that procedures introduced by Van Ryzin [13], [14] and Devroye and Wagner [3] can be derived from either original or modified class density estimates.

II. Preliminaries

Let $\Theta = \{1, \cdots, M\}$; elements of Θ will be called classes. Let (θ, X) be a pair of random variables. θ takes values in Θ and $p_i = P\{\theta = i\}$. X takes values in R^p, and f_i is the class conditional density, i.e., the conditional density of X given the class i. $L(i, j)$ is the loss we incur in takin action $i \in \Theta$ when the class is j. We assume the 0-1 loss function. For a decision function ψ, i.e., for a function mapping R^p into Θ, the expected loss is

$$R(\psi) = \sum_{j=1}^{M} p_j \int L(\psi(x), j) f_j(x) d\mu(x),$$

where μ is the Lebesgue measure on R^p. A decision function ψ_0 which classifies every x as coming from any class i for which

$$p_i f_i(x) = \max_j p_j f_j(x)$$

is a Bayes decision function. All the class densities as well as the class prior probabilities are assumed to be unknown and will be estimated from the learning sequence

$$(\theta_1, X_1), \cdots, (\theta_n, X_n),$$

i.e., a sequence of n independent observations of the pair (θ, X). Let $\hat{p}_i = N_i/n$, where N_i is the number of observations from the class i, be an estimate of p_i, and let $\hat{f}_i(x)$ be an estimate of $f_i(x)$.

The Θ-valued function ψ_n, defined for all $x \in R^p$ and all realizations of the learning sequence is called an empirical decision function. Throughout this correspondence we are concerned with pattern recognition that are sequences $\{\psi_n\}$ of empirical decision functions classifying ever x among any class i for which

$$\hat{p}_i \hat{f}_i(x) = \max_j \hat{p}_j \hat{f}_j(x), n = 1, 2, \cdots.$$

We say that the procedure is Bayes risk consistent if

$$\lim_{n \to \infty} ER(\psi_n) = R(\psi_0),$$

where ψ_0 is any Bayes decision function. It is clear that the asymptotical optimality, i.e., Bayes risk consistency depends on properties of class density estimates. The next two theorems on the asymptotical optimality are due to Greblicki [7] and Wolverton and Wagner [16] respectively.

Theorem 1: If

$$\hat{f}_i(x) \xrightarrow{p} f_i(x), \text{ as } n \to \infty,$$

at almost all $(\mu)x \in R^p$, for $i = 1, \cdots, M$, then the procedure is Bayes risk consistent.

It should be mentioned that under some additional assumption, Bayes risk consistency of procedures derived from pointwise density estimates has also been obtained by Glick [4].

Theorem 2: If

$$\int (\hat{f}_i(x) - f_i(x))^2 d\mu(x) \xrightarrow{p} 0 \text{ as } n \to \infty,$$

for $i = 1, \cdots, M$, then the procedure is Bayes risk consistent.

This theorem is a generalization of Van Ryzin's [14] result which assumed that the space of observations is a subset of R^p and has a finite measure.

III. Pointwise Consistent Estimates

One must partition observations $\{X_1, \cdots, X_n\}$ of the learning sequence into M subsequences

$$\{X_1^1, \cdots, X_{N_1}^1\}, \cdots, \{X_1^M, \cdots, X_{N_1}^M\}$$

of observations from particular classes. It is clear that $\hat{f}_i(x)$ is evaluated from the random number N_i of observations from the ith subsequence. Because of this, consistency of $\hat{f}_i(x)$ signifies that for every $\varepsilon > 0$ and $\delta > 0$ there exists N such that

$$P\left\{\left|\hat{f}_i(x) - f_i(x)\right| < \varepsilon | N_i > N\right\} > 1 - \delta.$$

In order to connect consistency with n we shall give the following.

Lemma 1: If $\hat{f}_i(x)$ is consistent, then

$$\hat{f}_i(x) \xrightarrow{p} f_i(x) \text{ as } n \to \infty.$$

Proof: The proof is elementary and is a consequence of the fact that for every N, $P\{N_i > N\} \to 1$ as $n \to \infty$. ■

By Theorem 1 and Lemma 1 we obtain Theorem 3.

Theorem 3: If all the estimates of the class densities are pointwise consistent at almost every $(\mu)x \in R^p$, then the pattern recognition procedure is Bayes risk consistent.

Since a great deal of attention has been paid to nonparametric pointwise consistent density estimates, Theorem 3 is interesting from the practical viewpoint. We shall give now two examples of the most popular pointwise consistent nonparametric density estimates.

The author is with the Institute of Engineering Cybernetics, Technical University of Wrocław, Poland

Example 1: Let K be a Borel scalar function such that

$$K(x) \geq 0, \sup_x K(x) < \infty, \int K(x) d\mu(x) = 1,$$

$$\lim_{||x|| \to \infty} ||x||^p K(x) \to 0.$$

The Rosenblatt-Parzen (RP) estimate of $f_i(x)$ introduced by Rosenblatt [10] and deeply studied by Parzen [9] is of the following form:

$$\hat{f}_i(x) = \frac{1}{N_i h^p(N_i)} \sum_{j=1}^{N_i} K\left(\frac{x - X_j^i}{h(N_i)}\right). \qquad (1)$$

Cacoullos [1] showed that

$$h(n) > 0, \lim_{n\to\infty} h(n) = 0, \lim_{n\to\infty} nh^p(n) = \infty \qquad (2)$$

implies consistency of (1) at every point of continuity of f_i.

Example 2: For a fixed norm in R^p, let V be the Lebesgue measure of a set of all $x \in R^p$ for which $||x|| \leq 1$. Loftsgaarden and Quesenberry's (LQ) estimate of $f_i(x)$ has the following form:

$$\hat{f}_i(x) = \frac{k(N_i)}{V N_i R_i^p(k(N_i))}, \qquad (3)$$

where $R_i(k)$ is the distance from x to the kth nearest observation in the ith subsequence. If

$$k(n) > 0, \lim_{n\to\infty} k(n) = \infty, \lim_{n\to\infty} k(n)/n = 0, \qquad (4)$$

the LQ estimate is consistent at every point of continuity of f_i [8].

By Theorem 2, procedures derived from RP as well as LQ estimates are Bayes risk consistent if all the class densities are almost everywhere continuous.

We shall introduce now a modified estimate of $f_i(x)$. Both estimates given in the examples depend on some sequences of numbers and can be rewritten in the following form: $\hat{f}_i(x; a(N_i))$, where $\{a(n)\}$ is a sequence of numbers. An estimate $\hat{f}_i(x; a(n_i))$ has been proposed by Greblicki [6] (see also [3]) and will be called a modified estimate. The next theorem will be useful while proving Bayes risk consistence of procedures with modified class density estimates.

Theorem 4: If

$$a(n) > 0, \lim_{n\to\infty} a(n) = 0, \lim_{n\to\infty} a(n)n = \infty \qquad (5)$$

implies consistency of $\hat{f}_i(x; a(N_i))$, then (5) implies the following convergence:

$$\hat{f}_i(x; a(n_i)) \xrightarrow{p} f_i(x) \text{ as } n \to \infty.$$

Proof: By the assumed implication, for every $\varepsilon > 0$ and $\delta > 0$, there exist s and A such that

$$P\{|\hat{f}_i(x; a(n_i)) - f_i(x)| < \varepsilon | a(N_i)N_i > A, a(N_i) < a\} > 1 - \delta/2. \qquad (6)$$

Hence

$$P\{|\hat{f}_i(x; a(n_i)) - f_i(x)| < \varepsilon | a(n)N_i > A, a(n) < a\} > 1 - \delta/2.$$

Obviously,

$$P\{|\hat{f}_i(x; a(n_i)) - f_i(x)| < \varepsilon\}$$
$$\geq P\{|\hat{f}_i(x; a(n_i)) - f_i(x)| < \varepsilon | a(n)N_i > A\} P\{a(n)N_i > A\}. \qquad (7)$$

It is also clear that there exists N such that, for $n > N$, both $a(n) < a$ and

$$P\{a(n)N_i > A\} = P\{a(n)n\hat{p}_i > A\} > 1 - \delta/2. \qquad (8)$$

Finally, by (6)–(8), for $n > N$,

$$P\{|\hat{f}_i(x; a(n_i)) - f_i(x)| < \varepsilon\} > 1 - \delta.$$

Since both ε and δ were arbitrary, the proof is completed. ■

Theorem 4 together with Theorem 7 in Section IV are crucial results of this correspondence. They make possible the demonstration of Bayes risk consistence of procedures derived from modifications of the widest spread types of density estimates, namely of RP and LQ, as well as orthogonal series estimates.

We shall now give modified versions of RP and LQ estimates, namely,

a) the modified RP estimate

$$\frac{1}{N_i h^p(n)} \sum_{j=1}^{N_i} K\left(\frac{x - X_j^i}{h(n)}\right),$$

b) the modified LQ estimate

$$\frac{k(n)}{V N_i R_i^p(k(n))}.$$

By Theorem 4, the modified RP estimate as well as the modified LQ estimate converge to a class density as $n \to \infty$ at every point of continuity of the density if the sequences $\{h(n)\}$ and $\{k(n(\}$ satisfy (2) and (3), respectively. Moreover, by Theorem 3, pattern recognition procedures with those estimates are Bayes risk consistent if all the class densities are almost everywhere (μ) continuous. This can be summarized in the next theorem.

Theorem 5: Pattern recognition procedures derived from modified either RP or LQ class density estimates are Bayes risk consistent if either (2) or (3) is satisfied, respectively, and all the class densities are almost everywhere (μ) continuous.

A similar result has been shown by Devroye and Wagner [3] who proved that for procedures with modified RP or LQ estimates,

$$P\{\psi_n(x) \neq \psi_0(x) | X_1, \theta_1, \cdots, X_n, \theta_n\} \to 0$$

as $n \to \infty$, in probability or almost surely.

IV. Integrated Uniformly consistent Estimates

Beside pointwise consistency, integrated uniform consistency has been studied by many authors. The latter is natural for orthogonal series (OS) estimates.

An estimate $\hat{f}_i$ of f_i is said to be integrated uniformly consistent, if for any $\varepsilon > 0$ and $\delta > 0$ there exists N such that

$$P\left\{\int (\hat{f}_i(x) - f_i(x))^2 d\mu(x) < \varepsilon | N_i > N\right\} > 1 - \delta.$$

It is easy to show the following:

Lemma 2: If $\hat{f}_i$ is an integrated uniformly consistent estimate, then

$$\int (\hat{f}_i(x) - f_i(x))^2 d\mu(x) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

By Theorem 2 and Lemma 2 we obtain Theorem 6.

Theorem 6: If all the class densities are square integrable and if all the estimates of the class densities are integrated uniformly consistent, then the pattern recognition procedure is Bayes risk consistent.

The most popular estimate, which has the property under consideration, is the OS estimate. Let f_i be square integrable and $\{g_i; i = 0, 1, 2, \cdots\}$ be a complete set of orthonormal functions on R^p. The OS estimate has the following form:

$$\hat{f}_i(x) = \sum_{k=0}^{q(N_i)} \hat{a}_k g_k(x), \tag{9}$$

where

$$\hat{a}_k = \frac{1}{N_i} \sum_{j=1}^{N_i} g_k(X_j^i)$$

and $\{q(n)\}$ is a sequence of numbers. Schwartz [12] has shown that (9) is integrated uniformly consistent if

$$q(n) > 0, \lim_{n\to\infty} q(n) = \infty, \lim_{n\to\infty} q(n)/n = 0. \tag{10}$$

By rewriting (9) as $\hat{f}_i(x; q(N_i))$, we can get the modified estimate $\hat{f}_i(x; q(n))$, namely,

$$\sum_{k=0}^{q(n)} \hat{a}_k g_k(x).$$

Applying a method used in the proof of Theorem 4, we can obtain the following:

Theorem 7: If (10) implies integratedly uniform consistency of $\hat{f}_i(.; q(N_i))$, the (10) implies the following convergence:

$$\int (\hat{f}_i(x; q(n)) - f_i(x))^2 d\mu(x) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

As a consequence of Theorems 6 and 7, we obtain finally Theorem 8.

Theorem 8: If all the class densities are square integrable, then the pattern recognition procedure with the modified OS class density estimate is Bayes risk consistent.

V. Pattern Recognition Procedures

Application of the RP estimate leads to a procedure which classifies x as coming from any class for which

$$\frac{1}{h^p(N_i)} \sum_{j=1}^{N_i} K\left(\frac{x - X_j^i}{h(N_i)}\right) \tag{11}$$

is maximal. For

$$K(x) = \begin{cases} \text{constant}, & \text{for } ||x|| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

we get a simple algorithm assigning x to any class for which

$$\frac{\text{the number of samples from class } i \text{ in } S_x(h(N_i))}{h^p(N_i)},$$

(where $S_x(h)$ is a sphere centered at x and h is the radius) is maximal.

The modified procedure classifies x among any class which maximizes

$$\sum_{j=1}^{N_i} K\left(\frac{x - X_j^i}{h(n)}\right). \tag{13}$$

For the kernel of the form (12) we get a very simple rule which recognizes x as belonging to any class for which the number of samples from class i in $S_x(h(n))$ is maximal.

Applying the LQ estimate we classify x among any class which minimizes

$$\frac{R_i^p(k(N_i))}{k(N_i)}. \tag{14}$$

The modified algorithm is even simpler, since it recognizes x as coming from any class for which

$$R_i(k(n)) \tag{15}$$

takes the minimal value. For $M = 2$, this procedure is equivalent to the k_n-NN (nearest neighbor) rule, where $k_n = 2k(n) - 1$ (see Cover and Hart [2] and Goldstein [5]). Stone [11] has shown that this procedure is Bayes risk consistent for all, not only almost everywhere continuous, class densities.

The algorithm with the OS estimate maximizes

$$\sum_{k=0}^{q(N_i)} \sum_{j=1}^{N_i} g_k(x) g_k(X_j^i), \tag{16}$$

whereas its modification maximizes

$$\sum_{k=0}^{q(n)} \sum_{j=1}^{N_i} g_k(x) g_k(X_j^i). \tag{17}$$

VI. Final Remarks

Procedures (11) and (13) have been studied by Van Ryzin [13], [14] and Devroye and Wagner [3]; (14) and (15) have been examined by Devroye and Wagner [3], whereas (16) has been used be Van Ryzin [14]. The k_n-nearest neighbor rule (15) and its various modifications have been extensively treated by Stone [11]. All these algorithms are derived from either original or modified class density estimates. We have shown that modifying RP, LQ, and OS estimates does not make them lose their basic property, i.e., consistency, and have established Bayes risk consistency of procedures with modified class density estimates.

Acknowledgment

The author thanks the referee for comments which helped to present the ideas of this work more precisely.

References

[1] T. Cacoullos, "Estimation of multivariate density", *Inst. Statist. Math.*, vol. 18, pp. 179-189, 1966.

[2] T.M. Cover and P.E. Hart, "Nearest neighbor pattern recognition classification", *IEEE Trans. Inform. Theory.*, vol. IT-13, pp. 21-27, Jan. 1967.

[3] L.P. Devroye and T.J. Wagner, "Nonparametric discrimination and density estimation", Rep. 183, Electronics Research Center, University of Texas, Austin, TX, Dec. 1976.

[4] N. Glick, "Sample-based classification procedures derived from density estimators" *J. Amer. Statist. Assoc.*, vol. 67, pp. 116-122, 1972.

[5] M. Goldstein, "k_n-nearest neighbor classification", *IEEE Trans. Inform. Theory.*, vol. IT-18, pp. 627-630, Nov. 1972.

[6] W. Greblicki, "Asymptotically optimal probabilistic algorithms for pattern recognition and identification", Scientific Papers of the Institute of Technical Cybernetics of Wrocław Technical University No. 18, Series: Monographs No. 3, Wrocław 1974 (in Polish)

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

[8] D.O. Loftsgaarden and C.P. Quesenberry, "A nonparametric estimation of multivariate density function", *Ann. Math. Statist.*, vol. 36, pp. 1049-1051, 1965.

[9] E. Parzen, "On estimation of a probability density and mode", *Ann. Math. Statist.*, vol. 33, pp. 1065-1076, 1962.

[10] M Rosenblatt, "Remarks on some nonparametric estimates of a density function", *Ann. Math. Statist.*, vol. 27, pp. 823-837, 1957.

[12] S.C. Schwartz, "Estimation of probability density by orthogonal series", *Ann. Math. Statist.*, vol. 38, pp. 1261-1265, 1967.

[13] J. Van Ryzin, "Non-parametric Bayesian decision procedure for (pattern) classification with stochastic learning", in *Trans. 4th Prague Conf. Information Theory, Statistical Decision Functions, and Random Processes*, Prague, 1965, pp. 479-494.

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

[15] E.J. Wegman, "Nonparametric probability density estimation; I. A summary of available methods", *Technometrics*, vol. 14, pp. 533-546, 1972.

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