Asymptotically Optimal Pattern Recognition Procedures with Density Estimates

WŁODZIMIERZ GREBLICKI

Abstract— **The asymptotic optimality of pattern recognition procedures derived from pointwise consistent density estimates is established.**

I. Introduction

Pattern recognition procedures using density estimates have been studied by many authors. The asymptotic behavior of the risk for integratedly uniformly week or strong consistent estimates (for terminology, see [1]) has been examined by Wolverton and Wagner [2]. Van Ryzin has shown the asymptotic optimality of procedures applying integratedly uniformly consistent quadratic mean estimates [3]. Pointwise consistent estimates have been employed by Glick [4] and the author [5]. In this paper, the asymptotic optimality of pattern recognition procedure derived from pointwise consistent estimates is also established. Optimality conditions are not, however, as sharp as those of Glick.

II. Learning Procedures

Let $\Theta = \{1, 2, \cdots, M\}$; the elements of Θ will be called classes. Let (θ, X) be a pair of random variables, where θ takes values in $\{1, \cdots, M\}$, $p_i = P\{\theta = i\}$, and where the random vector X takes values in R^k. Let μ be a σ-finite measure on R^k, and let f_i be the class conditional density of X, i.e., the appropriate Radon-Nikodym derivative with respect to μ. $L(i, j)$ is the loss incurred in taking action $i \in \Theta$ when the class is j. We assume throughout this paper that $L(i, j) = 1$, for $i \neq j$, and $L(i, j) = 0$, for $i = j$. For a decision function $\psi : R^k \to \Theta$, the expected loss is

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

A decision function ψ_0 which assigns every $x \in R^k$ to any class for which

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

is an optimal decision function. Denote by A_x the set of all classes maximizing $p_i f_i(x)$.

It is assumed that neither prior probabilities $p_1, \cdots, p_M$ nor the class conditional densities $f_1, \cdots, f_M$ are known. We have, however, a learning sequence,

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

a sample of n independent observations of the pair (θ, X) from which we shall estimate all the unknown distributions.

Let $f_{in}(x)$ be an estimate of $f_i(x)$ and let $p_{in} = N_i/n$, where N_i is the number of observations from the class i, be an estimate of p_i. It should be noticed that the estimate f_{in} uses only N_i of the observations, and can be parametric as well as nonparametric.

The Θ-valued function ψ_n, defined for all $x \in R^k$ and all realizations of the learning sequence, is called an empirical decision function (see [6]), and the sequence $\{\psi_n\}$ is called a pattern recognition procedure. Hereafter, we are concerned with procedures which, for every $x \in R^k$, choose a class i for which

$$p_{in} f_{in}(x) = \max_j p_{jn} f_{jn}(x).$$

Let $p_i f_i(x) = g_i(x)$, and $p_{in} f_{in}(x) = g_{in}(x)$. Let also

$$\rho(i, \Theta') = \begin{cases} 0, & \text{for } i \in \Theta' \\ 1, & \text{otherwise} \end{cases}$$

be a distance between an action $i \in \Theta$ and a set of actions $\Theta' \subset \Theta$.

III. Pointwise Convergence of Procedures

We shall now give a theorem on the convergence of the procedure at a point $x \in R^k$, i.e., a theorem on convergence of $\{\psi_n(x)\}$ to the set of optimal actions A_x.

Theorem 1: If the probability density estimate is weakly or strongly consistent at a point x, that is, if

$$f_{in}(x) \underset{a.s.}{\overset{p}{\to}} f_i(x)$$

as $n \to \infty$ for $i = 1, \cdots, M$, then

$$\rho(\psi_n(x), A_x) \underset{a.s.}{\overset{p}{\to}} 0 \tag{1}$$

as $n \to \infty$.

Corollary 1: Convergence (1) implies

$$\lim_{n\to\infty} P\{\psi_n(x) \in A_x\} = 1.$$

In particular, the probability of an optimal action when x is observed tends to 1.

Corollary 2: If the optimal action is unique, i.e., if a set A_x consists of only one element, then (1) can be rewritten as

$$\psi_n(x) \underset{a.s.}{\overset{p}{\to}} \psi_0(x)$$

as $n \to \infty$, where ψ_0 is any optimal decision function.

Proof of Theorem 1: Obviously

$$g_{in}(x) \underset{a.s.}{\overset{p}{\to}} g_i(x)$$

as $n \to \infty$, for all $i \in \Theta$. Let

$$\varepsilon = \min_{j \neq A_x} (g_i(x) - g_{in}(x))/2, \tag{2}$$

where $i \in A_x$. The proof consists of two parts.

A. Weak version: Take $\delta > 0$. It is clear that there exists an N such that, for $n > N$,

$$P\{|g_i(x) - g_{in}(x)| < \varepsilon\} > 1 - \delta/2M \tag{3}$$

for all $i \in \Theta$. Fix $i \in A_x$. By (2), we get

$$\begin{aligned} &P\{g_{in}(x) > g_{jn}(x)\} \\ &\quad \geq P\{|g_i(x) - g_{in}(x)| < \varepsilon, |g_j(x) - g_{jn}(x)| < \varepsilon\} \end{aligned}$$

for any $j \notin A_x$. By (3), for $n > N$ and any $j \notin A_x$, the right side of the above inequality is greater than $1 - \delta/M$. Thus, for $n > N$,

$$P\left\{\bigcap_{j \notin A_x} (g_{in}(x) > g_{jn}(x))\right\} > 1 - \delta. \tag{4}$$

Clearly,

$$\begin{aligned} P\{\rho(\psi_n(x), A_x) = 0\} &= P\{\psi_n(x) \in A_x\} \\ &\geq P\left\{\bigcap_{j \notin A_x} (g_{in}(x) > g_{jn}(x))\right\}. \end{aligned}$$

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

By virtue of (4) and the last inequality, the convergence in (1) in probability is established.

B. Strong version: Let $\delta > 0$. By the definition of almost sure convergence [7], there exists an N such that

$$P\left\{\sup_{n>N}|g_i(x)-g_{in}(x)|<\varepsilon\right\}>1-\delta/2M \tag{5}$$

for all $i \in \Theta$. Fix $i \in A_x$. For any $j \notin A_x$, we get from (2) that

$$P\left\{g_{in}(x)>g_{jn}(x),\ \text{for all } n>N\right\}$$
$$\geq P\left\{\sup_{n>N}|g_i(x)-g_{in}(x)|<\varepsilon,\ \sup_{n>N}|g_j(x)-g_{jn}(x)|<\varepsilon\right\}.$$

By (5), for any $j \notin A_x$, the right side of the above inequality is greater than $1-\delta/M$. Thus

$$P\left\{\bigcap_{j\notin A_x}\left(g_{in}(x)>g_{jn}(x),\ \text{for all } n>N\right)\right\}>1-\delta.$$

It is easy to see that

$$P\left\{\sup_{n>N}\rho\left(\psi_n(x),A_x\right)=0\right\}=P\left\{\psi_n(x)\in A_x,\ \text{for all } n>N\right\}$$
$$\geq P\left\{\bigcap_{j\notin A_x}\left(g_{in}(x)>g_{jn}(x),\ \text{for all } \dot{n}>N\right)\right\}.$$

Thus

$$P\left\{\sup_{n>N}\rho\left(\psi_n(x),A_x\right)=0\right\}>1-\delta.$$

Since δ is arbitrary, the strong version of (1) is proved and the proof of the theorem is complete. ■

IV. Asymptotic Optimality of the Risk

We shall consider now the convergence of the risk to the minimal risk, denoted by R_0.

Theorem 2: If the density estimate is weakly or strongly pointwise consistent almost everywhere (μ) in R^k, then

$$R(\psi_n)\ \underset{a.s.}{\overset{p}{\to}}\ R_0 \tag{6}$$

as $n \to \infty$. Moreover

$$\lim_{n\to\infty} ER(\psi_n)=R_0. \tag{7}$$

Remark 1: Under an additional constraint,

$$\textstyle\int f_{in}(x)d\mu(x)\ \underset{a.s.}{\overset{p}{\to}}\ 1$$

as $n \to \infty$, the same has been shown by Glick [4]. Unfortunately, this additional assumption cannot always be satisfied (e.g., for the Loftsgaarden-Quesenberry density estimate [8]).

Remark 2: The risk convergence in mean (7) signifies that the probability of misclassification tends to the minimal one.

Proof of Theorem 2: It is clear that

$$\textstyle\sum_{j=1}^{M}p_jf_j(x)[L(\psi_n(x),j)-L(\psi_0(x),j)]\leq\rho\left(\psi_n(x),A_x\right)f(x),$$

where ψ_0 is an optimal decision function and $f(x) = \Sigma_{j=1}^{M}p_jf_j(x)$. Therefore

$$0\leq R(\psi_n)-R_0\leq\textstyle\int\rho\left(\psi_n(x),A_x\right)f(x)d\mu(x). \tag{8}$$

By Theorem 1, we have

$$\rho(\psi_n(x),A_x)\ \underset{a.s.}{\overset{p}{\to}}\ 0$$

as $n \to \infty$ at almost all (μ) $x \in R^k$. Since the distance is bounded, Glick's modification [9] of the Lebesgue bounded convergence theorem for random functions yields (6). The convergence (7) is a consequence of the Lebesgue bounded convergence theorem and thus the theorem is established. ■

V. Conclusion

Pointwise consistency is a property of nonparametric density estimates that has received a great deal of attention (see [1]). The theorems given herein establish asymptotic optimality of pattern recognition procedures applying such estimates. For instance, Parzen's estimate can be weakly [10], [11], or strongly [12] consistent at every point of continuity of a density. For that estimate, the pattern recognition procedure considered here converges to the set of optimal actions in probability or almost surely at every point of continuity of all the class densities. Thus, if all the densities are continuous almost everywhere, the risk tends to the minimal risk in probability or almost surely, respectively. Applying other estimates, such as the Loftsgaarden-Quesenberry estimate [13] or the orthogonal series estimate [14], leads to other procedures (for details, see [2], [3], [5], [15], [16]).

It is clear that all the results given in this paper can be generalized to an arbitrary loss function with no difficulty. Finally, it should be mentioned that similar results have been independently obtained by Devroye and Wagner [15].

Acknowledgment

The author thanks the reviewers for their helpful suggestions.

References

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

[2] 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.

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

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

[5] 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)

[6] H. Robbins, "The empirical Bayes approach to statistical decision problems", *Ann. Math. Statist.*, vol. 35, pp. 1-20, 1964.

[7] M. Loève, *Probability Theory*, Princeton, NJ: Van Nostrand, 1963, Ch. III.

[8] T. J. Wagner, "Nonparametric estimates of probability densities", *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 438-440, Sept. 1975.

[9] N. Glick, "Consistency conditions for probability estimators and integrals for density estimators", *Utilitas Math.*, vol. 5, pp. 61-74, 1974.

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

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

[12] J. Van Ryzin, "On strong consistency of density estimates", *Ann. Math. Statist.*, vol. 40, pp. 1765-1772, 1969.

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

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

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

[16] 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.