Optimum Decision Rules in Pattern Recognition

Thien M. HA
University of Berne
Institut für Informatik und Angewandte Mathematik
Neubrückstr. 10, CH-3012 Berne, Switzerland
E-Mail: haminh@iam.unibe.ch

Abstract

This paper reviews various optimum decision rules for pattern recognition, namely, Bayes rule, Chow's rule (optimum error-reject tradeoff), and the recently proposed class-selective rejection rule. The last one provides the optimum tradeoff between the error rate and the average number of (selected) classes. The usage of each of these rules as well as their relationship are discussed. Some common properties to these rules are pointed out, e.g. the linear time complexity.

Key Words: classification, decision rule, Bayes rule, Chow's rule, class-selective rejection rule, nearest neighbour rules, man-machine interface, preselection, time complexity.

1 Introduction

Classification of an unknown pattern into one of a finite number of known classes is a common task for many pattern recognition systems. For such a task, the system performance is mainly characterized by its error rate. However, because of noise and other uncertain factors inherent in any real system, the error rate can be excessive for some applications, such as bank check reading [15]. Recognition with a reject option provides a means to reduce the error rate through a rejection mechanism, i.e., withhold making a decision if the confidence is not high enough and direct the rejected pattern to an exceptional handling, such as manual inspection. With a reject option, the system performance is characterized by the error-reject tradeoff [5, 6, 7].

However, for certain applications like computer–aided face identification, a rejection would require the operator to compare the rejected pattern with hundreds, if not thousands, of reference faces [3]. Therefore, a useful system should not make a simple rejection, but should provide a (preferably short) list of candidates or classes. For instance, the top-n ranking is such a mechanism. In this context, the rejection is *class-selective* in the sense that only the best classes are selected and the remaining classes are rejected. As a consequence, the error-reject tradeoff becomes the error-(number-of-classes) tradeoff.

More generally, classification is usually a first step to more sophisticated automatic processing, e.g. symbolic reasoning the complexity of which depends strongly on the number of classes provided by classification. In such a case, minimizing the number of classes

becomes a critical factor for the overall performance of the recognition system. Such concerns have emanated from typical applications in speech and character recognition [25, pp. 257–261][4, 1, 28]. In the simplest case, classification actually plays the role of preselection (or preclassification) and is usually carried out by a simple and fast classifier for selecting the most promising classes to be examined by a second, more accurate classifier, which usually consists in an exhaustive comparison of the input pattern with all selected classes. The second classifier can, for instance, be based on Hidden Markov models or nearest neighbour rule. Clearly, reducing the number of selected classes fastens the overall system.

Although the optimum error-reject tradeoff has been known for a long time [5], the optimum error-(number-of-classes) was discovered only recently [16, 18]. This paper aims at presenting some milestones along the way from the basic Bayes rule to the most recent results.

2 Optimum Decision Rules – An Overview

In statistical pattern recognition, the probability that a given sample or pattern x belongs to the i^{th} class, in a N-class problem, is provided by the *posterior* probability $P(i/x)$ through the Bayes formula:

$$P_i(x) \equiv P(i/x) = \frac{p(x/i) \cdot \pi_i}{p(x)}; i = 1, .., N \tag{1}$$

where $p(x/i)$ is the i^{th} class conditional probability density function (p.d.f.), π_i is the *a priori* probability of observing the i^{th} class, $\sum_{i=1}^{N} \pi_i = 1$, and

$$p(x) = \sum_{j=1}^{N} p(x/j) \cdot \pi_j \tag{2}$$

is the unconditional probability density function (also called mixture density or absolute probability density function) [11, 14]. If $p(x) = 0$, $\{P(i/x); i = 1, .., N\}$ are conventionally set to 0 [23]. Otherwise, the *posterior* probabilities sum up to 1, i.e.,

$$\sum_{i=1}^{N} P(i/x) = 1, \tag{3}$$

for $x \in X \equiv \{x : p(x) > 0\}$, which constitutes the main region of interest in this paper.

The connection between classification and decision is illustrated in Fig. 1, for a three-class problem. In most practical applications, $\{P(i/x); i = 1, .., N\}$ are unknown but can be estimated from a set of labelled patterns, called training set. Many estimation methods exist, e.g. Parzen estimate, nearest neighbour, potential functions, and neural networks [11, 14, 22, 28]. In the following, we assume that $\{P(i/x); i = 1, .., N\}$ are known and concentrate our discussions on the decision process.

Section 2.1 presents the Bayes rule. The extension of the Bayes rule to deal with rejection is summarised in Section 2.2, and that to deal with class-selective rejection in Section 2.3. A synthetic view of these three rules is presented in Table 1.

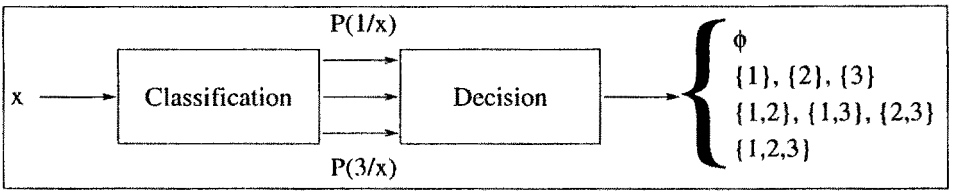

Figure 1: Relation between classification and decision. All possible outcomes of the decision process are shown on the right side.

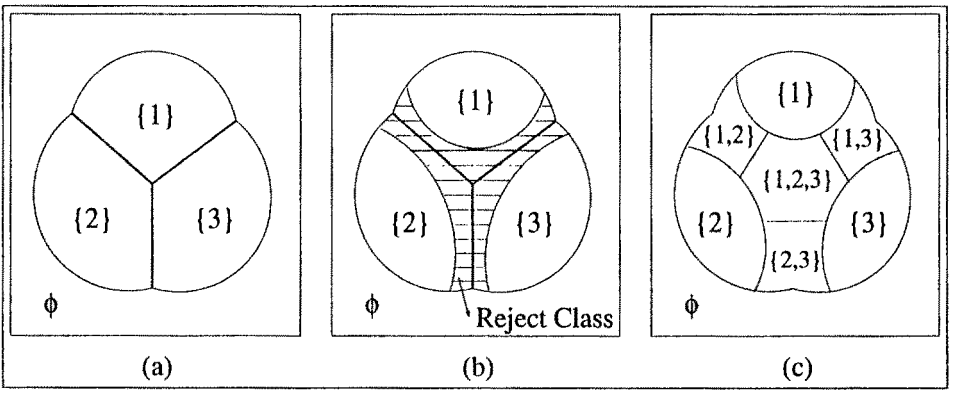

Figure 2: Three decision types. The empty set $\emptyset$ represents the region over which $p(x) = 0$.

2.1 Bayes Rule

Based on the *posterior* probabilities, the Bayes *decision rule* assigns to pattern x the class that has the highest *posterior* probability. It is known that this rule is optimal in the sense that no other rules can yield a lower error probability e, or error rate, given by

$$e = \int_X risk(x)p(x)dx \tag{4}$$

where $risk(x)$ is the (conditional) probability of making a wrong decision, for a given x. The (conditional) Bayes risk, i.e., the risk induced by using the Bayes decision rule is:

$$risk_{Bayes}(x) = 1 - \max_{i\in[1,..,N]}\{P_i(x)\} \tag{5}$$

In the Bayes decision rule, the possible outcomes of the decision process are limited to the singletons, i.e., subsets that are formed by exactly one class each. They are $\{1\}$, $\{2\}$, and $\{3\}$ for a three-class problem. Fig. 2a illustrates the partition of the pattern space X into three regions, each of which corresponds to a single class, when the Bayes rule is used.

2.2 Chow's Rule

The Bayes rule was modified by Chow to cope with a reject option [5, 6]. The idea is that when a pattern lies on or near a separation plane between two classes, the assignment to

one or the other class is merely a guess. In such a case, it may be better to withhold making the assignment (decision) and to reject the input pattern. The reject option is desirable in those applications where it is more costly to make a wrong decision than to withhold making a decision. With a reject option, the optimality espouses another meaning, that of a tradeoff between the error rate and the reject rate (reject probability). More specifically, Chow's optimum rule minimises the error rate for a given reject rate, or vice versa. The rule simply consists in rejecting the pattern if its highest *posterior* probability is lower than some threshold $(1-t), t \in [0, 1-\frac{1}{N}]$; otherwise, the decision is identical to Bayes' one, i.e. choosing the best class. Chow's rule is optimal is the sense that for the same reject rate specified by the threshold t, no other rules can yield a lower error rate. Interestingly, the outcomes of Chow's rule are also singletons, like in the Bayes rule, but augmented by the reject class; see Figs. 1 and 2b. The difference between Chow's reject class and the empty set $\emptyset$ is discussed in [10, 26, 28].

For a given value of the threshold t, Chow's rule partitions the pattern space into a rejection region X_r, shaded in Fig. 2b, and an acceptance region X_a, unshaded.

The acceptance rate, $a(t)$, is the integral of the absolute p.d.f. $p(x)$ over the acceptance region. The reject rate, $r(t)$, is the integral of the same function over the (complementary) rejection region.

$$a(t) = \int_{X_a} p(x)dx \tag{6}$$

$$r(t) = \int_{X_r} p(x)dx \tag{7}$$

It follows that

$$a(t) + r(t) = 1 \tag{8}$$

which means that a pattern is either accepted or rejected. When it is accepted, the decision can either be correct or wrong.

The accuracy or correct recognition rate, $c(t)$, is the expected value of the maximum *posterior* probability, $\max_{i\in[1,..,N]}\{P_i(x)\}$, over the acceptance region.

$$c(t) = \int_{X_a} (\max_{i\in[1,..,N]}\{P_i(x)\})p(x)dx \tag{9}$$

The error rate, $e(t)$, is the expected value of the Bayes risk over the acceptance region

$$e(t) = \int_{X_a} (1 - \max_{i\in[1,..,N]}\{P_i(x)\})p(x)dx \tag{10}$$

Obviously,

$$a(t) = c(t) + e(t) \tag{11}$$

and therefore

$$c(t) + e(t) + r(t) = 1 \tag{12}$$

As t increases from 0 to $(1-\frac{1}{N})$, the rejection threshold $(1-t)$ decreases, and the reject rate $r(t)$ decreases whereas the error rate $e(t)$ increases. When $t = 1 - \frac{1}{N}$, the rejection threshold $(1-t)$ equals $\frac{1}{N}$, and Chow's rule becomes Bayes rule, also called recognition at zero rejection level or forced choice. Figure 3 shows a typical error-reject, $e(r)$, tradeoff curve.

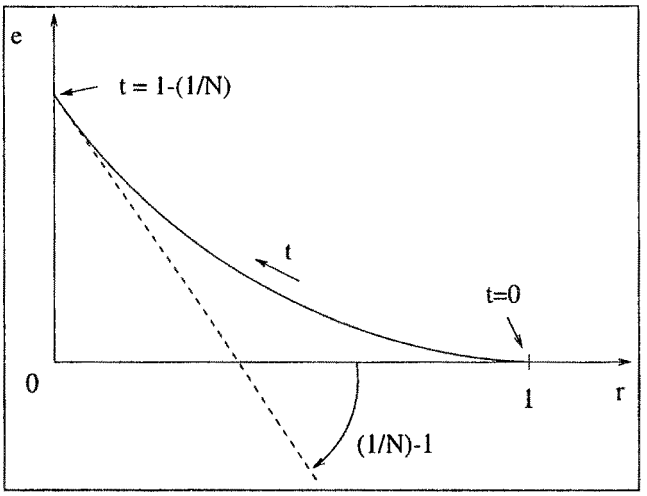

Figure 3: A typical $e(r)$ curve.

It turns out that it is possible to express the error rate directly as a function of the reject rate via the Stieltjes integral [6].

$$e(t_{ope}) = -\int_0^{t_{ope}} t \cdot dr(t) \tag{13}$$

where 'ope' stands for operating. (For an introduction to the Stieltjes integral, see [27].) The marvelous feature of the above equation is that it allows the computation of the error rate at any level t from $r(t)$ solely and that the latter can be estimated from unlabelled patterns, by just counting the rejects. In other words, the error rate at any level can be estimated without knowing the true classes of the patterns. For a more detailed discussion, see also [13]. In particular, the Bayes error rate is given by

$$e_{Bayes} = e(t_{ope} = 1 - \frac{1}{N}) = -\int_{t=0}^{1-\frac{1}{N}} t \cdot dr(t) \tag{14}$$

2.3 Optimum Class-Selective Rejection Rule

Recently, the optimum class-selective rejection rule was proposed [16, 18]. It differs from Chow's in that the outcomes of the decision process are extended to the power set of the set of classes, while excluding the empty set $\emptyset$. In Chow's rule, a pattern is rejected if its highest *posterior* probability is lower than a given threshold, disregarding the probability distribution of the remaining classes. Instead, the new rejection rule is *class-selective*. That is, it does not reject the pattern from all classes but only from those classes that are most unlikely to issue the pattern. For instance, for a pattern lying on the separation plane between classes 1 and 2, while being very far away from the center of the third class, the rule rejects only the third class and declares that the pattern belongs to the group composed of the first and the second classes. In other words, the pattern space is partitioned into regions each of which corresponds to a subset of classes. Since there are 2^N

subsets in a set of N elements, the resulting partition comprises $2^N - 1$ regions, excluding the empty set, in a N-class problem. In Fig. 2c, there are $2^3 - 1 = 7$ regions corresponding to the subsets $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, and $\{1,2,3\}$. It can readily be seen that there exists a trivial partition - that assigns the whole pattern space to the group composed of all N classes - which nullifies the error rate. This partition would correspond to a no-decision rule, however.

In order to define the optimality of the class-selective rejection rule while avoiding the trivial partition, an additional constraint – the average number of classes $\bar{n}$ – was introduced [16].

$$\bar{n} = \int_X n(x)p(x)dx \tag{15}$$

where $n(x)$ is the number of classes assigned to pattern x. The choice of $\bar{n} = E_X[n(X)]$ is natural, and moreover, it can be directly estimated from experiments by the sample mean $\frac{1}{N_s}\sum_{i=1}^{N_s} n_i$, where n_i is the number of classes assigned to pattern x_i, and N_s is the total number of patterns involved in the experiment.

The optimality of the class-selective rejection rule is then defined as the rule that minimises the error rate for a given average number of classes. The error rate is still given by Eq. (4), but $risk(x)$, i.e., the conditional probability of making an error becomes

$$risk(x) = 1 - \sum_{i \in Selected\ Subset} P_i(x) = \sum_{i \in Rejected\ Subset} P_i(x) \tag{16}$$

For instance, if the Selected Subset for pattern x is $\{1,3\}$ in a three-class problem, then $risk(x) = 1 - [P_1(x) + P_3(x)] = P_2(x)$, due to Eq. (3). Notice that Eq. (16) is a general form of Eq. (5) in that if the Bayes rule is used, i.e., select only the single best class, then Eq. (16) becomes Eq. (5). Substituting Eq. (16) into Eq. (4), the error rate becomes

$$e = \int_X [1 - \sum_{i \in Selected\ Subset} P_i(x)]p(x)dx \tag{17}$$

The optimum class-selective rejection rule assigns to pattern x all classes whose *posterior* probability is greater than a pre-specified threshold t. If there exist no such classes, the rule simply selects the (a) single best class [16, 18]. Notice that the key point in this rule is the choice of the number of best classes, $n(x,t)$, to be assigned to pattern x. The rule is optimum in the sense that for a given average number of classes, no other rules can yield a lower error rate.

Let us consider the range of t. Since the decision rule involves the comparison between t and *posterior* probabilities, it makes sense only for $t \in [0,1]$. On the other hand, when $t \geq \frac{1}{2}$, it can be easily seen that the rule is identical to the Bayes rule, i.e., choose the single best class. Only when t becomes smaller than $\frac{1}{2}$ does the rule provide the possibility of choosing more than one class. In sum, the range of t is $[0, \frac{1}{2}]$

The tradeoff between error rate and average number of classes at all levels t is an important description of the performance of recognition systems. When t varies from $\frac{1}{2}$ to 0, the average number of classes $\bar{n}(t)$ increases due to the emergence of groups composed of more than one class each. At the same time, the error rate $e(t)$ decreases since assigning more classes to a pattern reduces the risk of making an error. Thus both $\bar{n}(t)$ and $e(t)$ are monotonic functions of t, and we can compute the tradeoff curve e versus $\bar{n}$ from $e(t)$ and $\bar{n}(t)$. Fig. 4 shows a typical $e(\bar{n})$ curve.

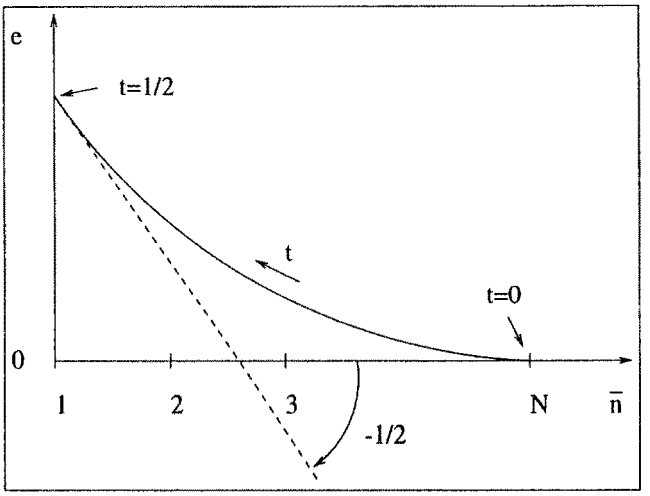

Figure 4: A typical $e(\bar{n})$ curve.

Analogously to the error-reject tradeoff, there also exists a functional relation between $e(t)$ and $\bar{n}(t)$, and $\bar{n}(t)$ alone completely specifies $e(t)$ in the same manner as Eq. (13) does for the $e(r)$ curve.

$$e(t_{ope}) = -\int_{t=0}^{t_{ope}} t \cdot d\bar{n}(t) \tag{18}$$

In particular, the Bayes error rate is given by

$$e_{Bayes} = e(t_{ope} = \frac{1}{2}) = -\int_{t=0}^{\frac{1}{2}} t \cdot d\bar{n}(t) \tag{19}$$

Full details about this functional relation can be found in [17].

3 Some Remarks

Inspecting the three rules summarised in Table 1, one can see that their time complexity is linear in the total number of classes. This common property is important in those applications where the number of classes is very large, e.g. Chinese character recognition.

The functional relation between error rate and average number of classes, Eq. (18), takes the same form as Chow's optimum error-reject tradeoff curve, Eq. (13). Likewise, the optimum $e(\bar{n})$ curve shares many properties with Chow's optimum error-reject, $e(r)$, curve. It has been shown in [17] that the slope of the $e(\bar{n})$ curve is $-t$. That is, the ratio of error reduction to additional average number of classes is most effective near the origin ($\bar{n} = 1, t = \frac{1}{2}$). This is common in our practical experience: excessive additional classes are generally required to reduce residual errors. In fact, Chow already observed this behaviour in the error-reject curve: excessive rejection is generally required to reduce residual errors [6]. Moreover, the non-decreasing nature and the upward concavity are common properties to both the $e(\bar{n})$ and $e(r)$ curves.

Table 1: A synthetic view of optimum decision rules. MAP stands for maximum a posteriori.

OPTIMALITY		
minimise c	minimise $e/r; \forall r$	minimise $e/\bar{n}; \forall \bar{n}$
RULE		
MAP.	IF MAP $< (1-t)$ REJECT ELSE USE MAP. $0 \leq t \leq 1 - \frac{1}{N}$	All classes whose Post. Prob. $> t$. IF NONE USE MAP. $0 \leq t \leq \frac{1}{2}$
ESTIMATION - FUNCTIONAL RELATION		
$e = E_X[risk(X)]$	$e(t_o) = -\int_0^{t_o} t \cdot dr(t)$	$e(t_o) = -\int_0^{t_o} t \cdot d\bar{n}(t)$
Cost(t) =		
$C \cdot e$	$C_e \cdot e(t) + C_r \cdot r(t)$ $t_{opt} = \frac{C_r}{C_e}$	$C_e \cdot e(t) + C_{\bar{n}} \cdot \bar{n}(t)$ $t_{opt} = \frac{C_{\bar{n}}}{C_e}$

In some sense, when nearest neighbour (NN) estimates are used instead of posterior probabilities, Bayes rule can be replaced by the $k-NN$ rule [8], and Chow's rule by the $(k,k')-NN$ rule [21]. Nearest neighbour version of the functional relation between $e(t)$ and $r(t)$ was discovered by Devijver [9].

4 Conclusion

We have reviewed various optimum decision rules for pattern recognition, namely, Bayes rule, Chow's rule, and class-selective rejection rule. It can be said that the theory of optimum decision rules for pattern recognition is well understood. To the author's view, future research should be concentrated on the estimation of posterior probabilities which remains a difficult problem for complex models such as Hidden Markov models.

References

[1] H.S. Baird and C.L. Mallows, "Bounded-Error Preclassification Trees," in *Shape, Structure and Pattern Recognition*, D. Dori and A. Bruckstein (Eds.), World Scientific, 1995, pp. 343-349.

[2] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer-Verlag, 1985.

[3] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proceedings of the IEEE*, Vol. 83, No. 5, pp. 705-740, May 1995.

[4] W. Cho, S.W. Lee, and J.H. Kim, "Modeling and Recognition of Cursive Words with Hidden Markov Models," *Pattern Recognition* ***28***, pp. 1941-1953, 1995.

[5] C.K. Chow, "An Optimum Character Recognition System Using Decision Functions," *Institute of Radio Engineers (IRE) Transactions on Electronic Computers*, Vol. EC-6, No. 4, pp. 247-254, December 1957.

[6] C.K. Chow, "On Optimum Recognition Error and Reject Tradeoff," *IEEE Transactions on Information Theory*, Vol. IT-16, No. 1, pp. 41-46, January 1970.

[7] C.K. Chow, "Recognition Error and Reject Trade-off," *Third Annual Symp. on Document Analysis and Information Retrieval*, April 11-13, 1994, University of Nevada, Las Vegas, U.S.A., pp. 1-8.

[8] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, Vol. 13, pp. 21-27, Jan. 1967.

[9] P.A. Devijver, "Error and Reject Tradeoff for Nearest Neighbor Decision Rules," in G. Tacconi (Ed.) *Aspects of Signal Processing*,, Part 2, D. Reidel Publishing Company, Dordrecht-Holland, pp. 525-538, 1977.

[10] B. Dubuisson and M. Masson, "A Statistical Decision Rule with Incomplete Knowledge about Classes," *Pattern Recognition* ***26***, pp. 155-165, 1993.

[11] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[12] G.M. Fitzmaurice and D.J. Hand, "A Comparison of Two Average Conditional Error Rate Estimators," *Pattern Recognition Letters*, Vol. 6, pp. 221-224, 1987.

[13] K. Fukunaga and D.L. Kessel, "Application of Optimum Error-Reject Functions," *IEEE Transactions on Information Theory*, Vol. IT-18, pp. 814-817, November 1972.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition, Academic Press, 1990.

[15] Thien M. Ha, D. Niggeler, H. Bunke, and J. Clarinval, "Giro Form Reading Machine," *Optical Engineering*, Vol. 34, No. 8, pp. 2277-2288, 1995.

[16] Thien M. Ha, "An Optimum Class-Selective Rejection Rule for Pattern Recognition," *Proceedings of the* 13^{th} *International Conference on Pattern Recognition, Vol. II,* Aug. 25-30, 1996, Vienna, Austria, pp. 75-80.

[17] Thien M. Ha, "On Functional Relation between Class-Selective Rejection Error and Average Number of Classes," *IEEE International Symposia on Intelligence and Systems,* Nov. 4-5, 1996, Rockville, Maryland, U.S.A., pp. 282-287.

[18] Thien M. Ha, "The Optimum Class-Selective Rejection Rule," *IEEE Trans. on Pattern Analysis and Machine Intelligence.* Vol. 19, No. 6, pp. 608-615, June 1997.

[19] D.J. Hand, "Recent Advances in Error Rate Estimation," *Pattern Recognition Letters,* Vol. 4, pp. 335-346, 1986.

[20] D.J. Hand, "An Optimal Error Rate Estimator Based on Average Conditional Error Rate: Asymptotic Results," *Pattern Recognition Letters,* Vol. 4, pp. 347-350, 1986.

[21] M.E. Hellman, "The Nearest Neighbor Classification Rule with a Reject Option," *IEEE Transactions on Systems, Science, and Cybernetics,* Vol. SSC-6, No. 3, pp. 179-185, July 1970.

[22] C.G.Y. Lau (Editor), *Neural Networks: Theoretical Foundations and Analysis,* IEEE Press, 1992.

[23] G. Lugosi and M. Pawlak, "On the Posterior-Probability Estimate of the Error Rate of Nonparametric Classification Rules," *IEEE Transactions on Information Theory,* Vol. IT-40, No.2, pp. 475-481, March 1994.

[24] M. Pawlak, "On the Asymptotic Properties of Smoothed Estimators of the Classification Error Rate," *Pattern Recognition,* Vol. 21, No. 5, pp. 515-524, 1988.

[25] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition,* Prentice-Hall, 1993.

[26] B.D. Ripley, *Pattern Recognition and Neural Networks,* Cambridge University Press, 1996.

[27] S.M. Ross, *A First Course in Probability,* third edition, Macmillan Publishing Company, 1988.

[28] J. Schürmann, *Pattern Classification: A Unified View of Statistical and Neural Approaches,* John Wiley & Sons, 1996.

[29] G.T. Toussaint, "Bibliography on Estimation of Misclassifications," *IEEE Transactions on Information Theory,* Vol. IT-20, pp. 472-479, July 1974.

[30] G.E. Tutz, "Smoothed Additive Estimators for Non-Error Rates in Multiple Discriminant Analysis," *Pattern Recognition,* Vol. 18, No. 2, pp. 151-159, 1985.