On Functional Relation between Recognition Error and Class-Selective Reject

Thien M. HA
University of Berne
Institut für Informatik und Angewandte Mathematik
Neubrückstr. 10, CH-3012 Berne, Switzerland
Phone: +41 / 31 / 631 86 99
Fax.: +41 / 31 / 631 39 65
E-Mail: haminh@iam.unibe.ch

March 11, 1996

Abstract

This report reviews various optimum decision rules for pattern recognition, namely, Bayes rule, Chow's rule (optimum error-reject tradeoff), and a recently proposed class-selective rejection rule. The latter provides an optimum tradeoff between the error rate and the average number of (selected) classes. A new general relation between the error rate and the average number of classes is presented. The error rate can directly be computed from the class-selective reject function, which in turn can be estimated from unlabelled patterns, by simply counting the rejects. Theoretical as well as practical implications are discussed and some future research directions are proposed.

CR Categories and Subject Descriptors: I.5.0 [Pattern Recognition]: General; I.5.1 [Pattern Recognition]: Models; I.5.2 [Pattern Recognition]: Design Methodology; I.5.m [Pattern Recognition]: Decision.
Key Words: classification, decision rule, Bayes rule, selective rejection, man-machine interface.

Contents

1 Introduction

Classification of an unknown pattern into one of the known classes is a common task for many pattern recognition systems. For such a task, the system performance is mainly characterised 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 [9]. 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 characterised by the error-reject tradeoff [2].

From an application point of view, characterising the system performance by the error-reject tradeoff is appropriate for many tasks, such as those involving optical character recognition (OCR). The reason is that when a pattern is rejected human correction can usually lower the classification error without excessive additional efforts (humans know quite well the set of characters being used). In contrast, for applications like face identification, humans may not know or remember all (maybe a huge amount of) reference faces. In such an application, a rejection would require the operator to compare the rejected pattern with hundreds, if not thousands, of reference faces [1]. 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 error-reject tradeoff becomes error-(number-of-classes) tradeoff.

Although the optimum error-reject tradeoff has been known for a long time [2], the optimum error-(number-of-classes) was discovered only recently [10]. Few theoretical results on these aspects are available [4, 13].

This report first gives an overview of optimum decision rules. A general relation between error rate and average number of classes is then presented. It will be shown that the error rate can be estimated directly from the empirical number of classes assigned to each unlabelled pattern. General properties of the optimum tradeoff curve are pointed out. Theoretical as well as practical implications are discussed and future research directions are proposed.

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[1]

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

is the absolute probability density function [5, 8]. It follows immediately that the *posterior* probabilities sum up to 1, i.e.,

$$\sum_{i=1}^{N} P_i(x) = 1 \tag{3}$$

The connection between classification and decision is illustrated in Fig. 1, for a three-class problem. Thus, the (soft) classifier is a device that computes the *posterior* probabilities, for a given x, which is then dichotomised into classes by the decision process. 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 [5, 8, 14]. In the following, we assume that $\{P_i(x); i = 1, .., N\}$ are known and concentrate our study on the decision process.

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.

[1] Without loss of generality, it will be assume that $p(x)$ is nonzero over the entire pattern space X, otherwise the region over which $p(x)$ is zero is first deleted.

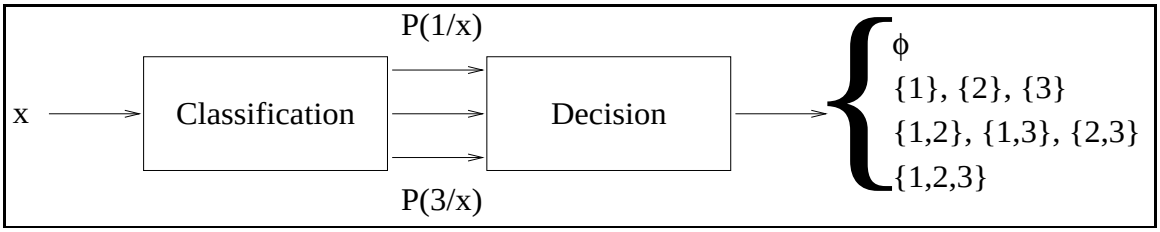

Figure 1: Relation between classification and decision. All possible outcomes of the decision process are shown on the right side.

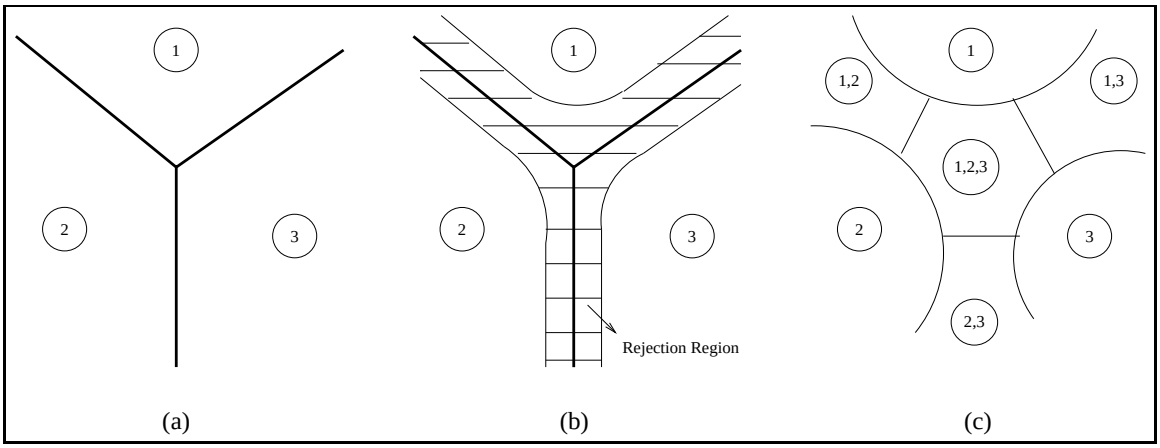

Figure 2: Three decision types.

2.2 Chow's Rule

The Bayes rule has also been modified by Chow to cope with a reject option [2, 3]. 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 empty set $\emptyset$, which represents the reject option; see Figs. 1 and 2b.

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. It turns out that it is possible to express the error rate directly as a function of the reject rate via the Stieltjes integral [3].

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

where 'ope' stands for operating. (For an introduction to the Stieltjes integral, see [17].)

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 [7]. 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, an optimum class-selective rejection rule was proposed [10]. 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 [10].

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

where $n(x)$ is the number of classes assigned to pattern x. The choice of $\bar{n} = E_X[n(x)]$ is natural, and more importantly, 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 [10]. 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.

In order to state the rule formally, let us introduce the sequence $\{Q_i(x)\}$, which is simply an reordered sequence of $\{P_i(x)\}$ in decreasing order of *posterior* probability

$$\{P_i(x); i = 1, .., N\} \rightarrow \{Q_i(x); i = 1, .., N\} / Q_i(x) \geq Q_{i+1}(x); i = 1, .., N-1 \tag{18}$$

Thus, $Q_1(x)$ is the maximum *posterior* probability of pattern x. The risk defined by Eq. (16) can then be expressed in a more explicit form as follows

$$risk(x,t) = 1 - \sum_{i=1}^{n(x,t)} Q_i(x) \tag{19}$$

The optimum class-selective rejection rule is formally given by

Decision Rule [10]: The optimum class-selective rejection rule assigns to pattern x the $n^*(x,t)$ best classes, where

$$n^*(x,t) = \min_{k \in [1,N]} \{k / Q_{k+1}(x) \leq t\} \tag{20}$$

with the convention

$$Q_{N+1}(x) = 0 \tag{21}$$

and the domain of the pre-specified threshold

$$0 \leq t \leq \frac{1}{2} \tag{22}$$

Remarks: It is instructive to distinguish two cases, depending on the relative value of $Q_1(x) = \max_{i\in[1,..,N]}\{P_i(x)\}$ with respect to t:

- *Case a:* $Q_1(x) > t$. We get $Q_{n^*}(x) > t$. This is obvious for $n^* = 1$. For $n^* > 1$, suppose that $Q_{n^*}(x) \leq t$, then $\exists k(= n^* - 1)$ such that $Q_{k+1(=n^*)} \leq t$, $k(= n^* - 1) \geq 1 \Rightarrow k \in [1, .., N]$, and $k(= n^* - 1) < n^*$, which means that the optimum decision rule given by Eq. (20) had not been used (n^* is not the minimum value possible).

- *Case b:* $Q_1(x) \leq t$. We have $Q_{n^*}(x) \leq t$ and $n^*(x) = 1$. The first statement is obvious because $Q_1(x) = \max_{i\in[1,..,N]}\{P_i(x)\} \leq t$ implies that $\forall i, Q_i(x) \leq t$, including $i = n^*$. The second statement is due to the fact that the optimum rule sets the cardinal number equal to the minimum index, which is 1.

Finally, let us consider the range of $t \in [0, \frac{1}{2}]$. 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. Indeed, in *Case a*, $Q_1 > t \geq \frac{1}{2}$ implies $Q_2 < \frac{1}{2} \leq t$, and thus $n^* = 1$. In *Case b*, $n^* = 1$. In any case, we choose only the best class. Only when t becomes smaller than $\frac{1}{2}$ does the rule provide the possibility of choosing more than one class.

3 Functional Relation Between e and $\bar{n}$

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. 3 shows a typical $e(\bar{n})$ curve. In this section we will show that there exists a functional relation between $e(t)$ and $\bar{n}(t)$, and that $\bar{n}(t)$ alone completely specifies $e(t)$ in the same manner as Eq. (13) does for the $e(r)$ curve.

Consider an incremental change of t by $\Delta t > 0$. For a given x, let us examine the optimum cardinal numbers and risks at levels t and $t + \Delta t$, respectively. Direct applications of Eqs. (20) and (19) yield[2]

[2]For simplicity, the superscript '*' will be omitted.

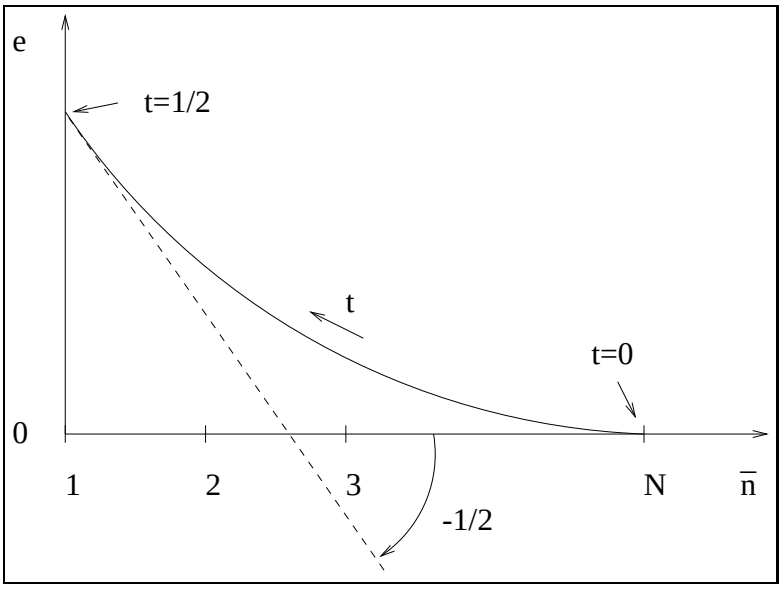

Figure 3: A typical $e(\bar{n})$ curve.

$$n(x,t) = \min_{k\in[1,N]}\{k/Q_{k+1}(x) \leq t\} \tag{23}$$

$$n(x,t+\Delta t) = \min_{k\in[1,N]}\{k/Q_{k+1}(x) \leq t+\Delta t\} \tag{24}$$

$$risk(x,t) = 1 - \sum_{i=1}^{n(x,t)} Q_i(x) \tag{25}$$

$$risk(x,t+\Delta t) = 1 - \sum_{i=1}^{n(x,t+\Delta t)} Q_i(x) \tag{26}$$

Define the incremental changes of the cardinal number and risk by

$$\Delta n(x,t;\Delta t) = n(x,t+\Delta t) - n(x,t) \tag{27}$$

and

$$\Delta risk(x,t;\Delta t) = risk(x,t+\Delta t) - risk(x,t) \tag{28}$$

When t increases, the optimum decision rule selects less classes and thus $\Delta n \leq 0$ and $\Delta risk \geq 0$. More precisely,

$$\Delta risk = \sum_{i=1}^{n(t)} Q_i - \sum_{i=1}^{n(t+\Delta t)} Q_i \tag{29}$$

If $\Delta n = 0$, we have $\Delta risk = 0$. Otherwise, $\Delta n < 0$, i.e., $n(t+\Delta t) < n(t)$, and

$$\Delta risk = \sum_{i=n(t+\Delta t)+1}^{n(t)} Q_i = Q_{n(t+\Delta t)+1} + Q_{n(t+\Delta t)+2} + ... + Q_{n(t)} \tag{30}$$

Recall that $n(t)$ is the optimum cardinal number at t, and $n(t) > n(t+\Delta t) \geq 1$. This means that we are not is Case b of *Remarks*, but must be in Case a. Hence, $Q_{n(t)} > t$.

On the other hand, $n(t+\Delta t)$ is the optimum cardinal number at $t+\Delta t$, Eq. (24) implies $Q_{n(t+\Delta t)+1} \leq t + \Delta t$.

Taking into account the decreasing nature of $\{Q_i(x)\}$, we can set an upper-bound and a lower-bound on the summing terms of $\Delta risk$ as follows

$$t + \Delta t \geq Q_{n(t+\Delta t)+1} \geq Q_{n(t+\Delta t)+2} \geq ... \geq Q_{n(t)} \geq t \tag{31}$$

Summing up all Q_is leads to

$$(t + \Delta t) \cdot [n(t) - n(t + \Delta t)] \geq \sum_{i=n(t+\Delta t)+1}^{n(t)} Q_i > t \cdot [n(t) - n(t + \Delta t)] \tag{32}$$

or

$$-(t + \Delta t) \cdot \Delta n \geq \Delta risk > -t \cdot \Delta n \tag{33}$$

Taking the expectation of both sides with respect to x, we get

$$-(t + \Delta t) \cdot \Delta \bar{n} \geq \Delta e > -t \cdot \Delta \bar{n} \tag{34}$$

By varying t from 0 to t_{ope} ('ope' stands for operating) with constant increment Δt, and summing up all partial variations, we get

$$-\sum t \cdot \Delta \bar{n} - \sum \Delta t \cdot \Delta \bar{n} \geq \sum \Delta e > -\sum t \cdot \Delta \bar{n} \tag{35}$$

Letting $\Delta t \to 0$, we can drop the second order infinitesimal term $\Delta t \cdot \Delta \bar{n}$ and get the Stieltjes integral [17]

$$e(t_{ope}) = \int_{t=0}^{t_{ope}} de(t) = -\int_{t=0}^{t_{ope}} t \cdot d\bar{n}(t) \tag{36}$$

In particular, the Bayes error rate is given by

$$e_{Bayes} = e(t_{ope} = \frac{1}{2}) = -\int_{t=0}^{\frac{1}{2}} t \cdot d\bar{n}(t) \tag{37}$$

4 Properties of Optimum $e - \bar{n}$ Curves

Whenever the optimum class-selective rejection rule is used, the $e - \bar{n}$ curve exhibits the following properties.

- $e(\bar{n})$ is non-increasing.

- The slope varies from $-\frac{1}{2}$ to 0, as $\bar{n}$ increases from 1 to N.
- $e(\bar{n})$ is concave upward.

Fig. 3 shows a typical $e(\bar{n})$ curve.

The non-increasing nature of $e(\bar{n})$ can easily be verified by considering Eq. (34) and letting $\Delta t \to 0$.

$$\frac{de}{d\bar{n}} = \lim_{\Delta\bar{n}\to 0} \frac{\Delta e}{\Delta \bar{n}} = -t \tag{38}$$

Since $t \in [0, \frac{1}{2}]$, we get $\frac{de}{d\bar{n}} \leq 0$, confirming that $e(\bar{n})$ is non-increasing.

When t decreases from $\frac{1}{2}$ down to 0, $\bar{n}$ increases from 1 (Bayes rule) to N (the trivial partition), and Eq. (38) shows that the slope varies from $-\frac{1}{2}$ to 0. Interestingly, the value of the slope at the origin is totally independent of any other system parameters (e.g. various probabilities and total number of classes), and always equals $-\frac{1}{2}$.

By taking the derivative of Eq. (38) with respect to $\bar{n}$, we get

$$\frac{d^2e}{d\bar{n}^2} = -\frac{dt}{d\bar{n}} \tag{39}$$

When t increases ($dt > 0$), the optimum rule selects less classes ($d\bar{n} \leq 0$), and thus $\frac{d^2e}{d\bar{n}^2} \geq 0$. Hence $e(\bar{n})$ is concave upward.

5 Examples

Consider a three-class problem. Each class p.d.f is a Gaussian with unit standard deviation. The center coordinates of classes 1, 2, and 3 are $(0.0, 1.0)$, $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, respectively. The *a priori* probabilities of all three classes are equal to $\frac{1}{3}$.

The error rate versus the average number of classes, $e - \bar{n}$ curve, obtained by using the optimum decision rule is plotted in Fig. 4. This curve is obtained by varying t from $\frac{1}{2}$ down to 0, and numerically integrating Eqs. (15) and (17).

The dotted line in Fig. 4, representing the tangent at the origin ($\bar{n} = 1, t = \frac{1}{2}$), shows that the slope is equal to $-\frac{1}{2}$, according to Eq. (38).

6 Implications and Future Research

The main results presented in this report fill a hole in the theory of optimum decision rules for pattern recognition. For the optimum class-selective rejection rule, the functional relation between error rate and average number of classes, Eq. (36), is established. It takes the same form as Chow's optimum error-reject tradeoff curve, Eq. (13).

The optimum $e - \bar{n}$ curve shares many properties with Chow's optimum error-reject, $e - r$, curve [3]. Since the slope of the $e - \bar{n}$ curve is $-t$, the tradeoff, i.e.,

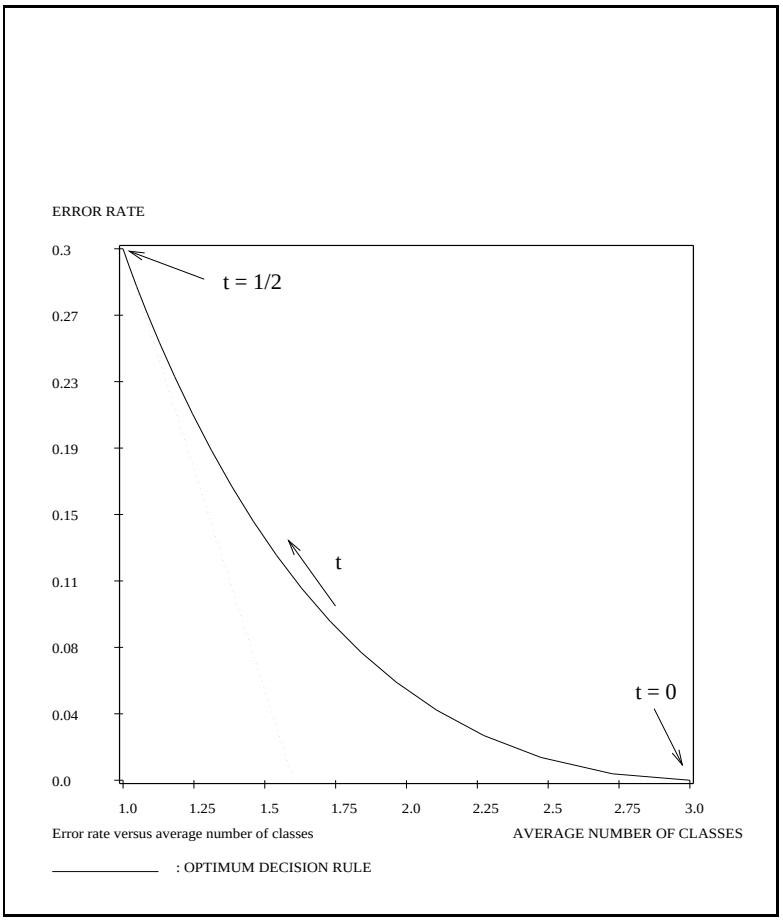

Figure 4: Relation between error rate and average number of classes for a three class problem. The dotted line represents the tangent at the origin and has a slope equal to $-\frac{1}{2}$.

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. For Chow's error-reject curve, the same behaviour can be observed: excessive rejection is generally required to reduce residual errors. Moreover, the non-decreasing nature and the upward concavity are common properties to both $e - \bar{n}$ and $e - r$ curves.

In practice, all properties (Section 4) derived from the optimum recognition system can serve as checking criteria of the classifier design. Indeed, these properties are necessary conditions for an optimum classifier. If empirical data do not fit these properties, we can conclude that there is a flaw in the design of the classifier, although no more specific information is available as to where the flaw(s) may be. In contrast, a well fitting of empirical data to these properties does *not* imply that the classifier is optimum, although it may constitute a strong support.

The use of Eq. (36) in estimating the error rate without having recourse to the true labels of testing patterns should be further investigated. The study should take into account the observations made by Fukunaga and Kessel in [7] on the Stieltjes integral, Eq. (13), proposed by Chow [3]. See also [18, 11] for reviews of error estimation methods. Other interesting related papers are [19, 12, 6, 16, 15].

7 Conclusion

We have reviewed various optimum decision rules for pattern recognition, namely, Bayes rule, Chow's rule (optimum error-reject tradeoff), and a recently proposed class-selective rejection rule. The latter provides an optimum tradeoff between the error rate and the average number of classes. A new general relation between the error rate and the average number of classes is presented. It is expressed by a Stieltjes integral and takes a similar form as Chow's optimum error-reject curve. This integral allows the error rate to be computed from the class-selective reject function, which can be estimated from unlabelled patterns, by simply counting the rejects. Further investigation on the use of this integral for error estimation should be pursued. Many general properties of the optimum tradeoff curve are derived. They constitute a set of necessary conditions for an optimum recognition system, and thus can serve as checking criteria of the classifier design.

Acknowledgements: This work was partly supported by the Union Bank of Switzerland Information Technology Laboratory (UBILAB) and the Swiss National Science Foundation. The author would like to thank Prof. H. Bunke for his constant encouragement.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] K. Fukunaga and D.L. Kessel, "Application of Optimum Error-Reject Functions," *IEEE Transactions on Information Theory,* Vol. IT-18, No. ??, pp. 814-817, November 1972.

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

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

[10] Thien M. Ha, "An Optimum Decision Rule for Pattern Recognition," Technical Report IAM-95-009, Institute of Computer Science and Applied Mathematics, University of Berne, Switzerland, November 1995.

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] G.T. Toussaint, "Bibliography on Estimation of Misclassifications," *IEEE Transactions on Information Theory,* Vol. IT-20, No. ??, pp. 472-479, July 1974.

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