An Optimum Decision Rule for Pattern Recognition

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

November 30, 1995

Abstract

The concept of rejection is extended to that of class-selective rejection. That is, when an input pattern cannot be reliably assigned to one of the N classes in a N-class problem, it is assigned to a subset of classes that are most likely to issue the pattern, instead of simply rejecting the pattern. First, a new optimality criterion is appropriately defined to accommodate the newly introduced decision outcomes. Then, a new decision rule is derived and its optimality proven. Various upper-bounds on error rate are obtained. Finally, some examples are provided to illustrate the differences between the optimum rule and other heuristic rules, such as top-n ranking.

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 **3**

2 A Necessary Condition **6**

3 Optimal Decision Rule **7**

4 Upper-bounds **10**
4.1 An upper-bound of $e(t)$ 10
4.2 Upper-bounds of $n(x,t)$ and $\bar{n}(t)$ 11

5 Examples **11**
5.1 A Three-Class Problem 12
5.2 Comparison with Other Rules 12

6 Implications **15**

7 Conclusion **15**

1 Introduction

In statistical pattern recognition, the probability that a given sample 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:

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

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, and[1]

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

is the absolute probability density function [1, 2]. It follows immediately that the *posterior* probabilities sum up to 1, i.e.,

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

Based on the *posterior* probabilities, the Bayes *decision rule* assigns to sample 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:

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

The connection between classification and decision is illustrated in Fig. 1, for a three-class problem. 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.

The Bayes rule has also been modified by Chow to cope with a reject option [3, 4]. The reject option is desirable in those applications where it is more costly to make a wrong decision than to withhold making a decision. In such situations, the optimality espouses another meaning, that of a tradeoff between the error rate and the reject rate (reject probability). More specifically, the optimum rule minimises

[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.

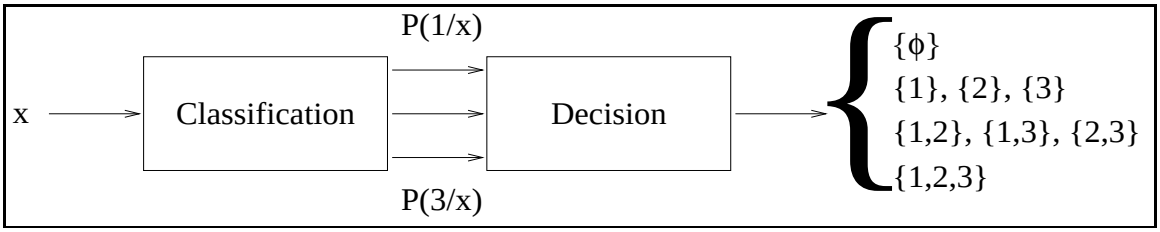

Figure 1: Relation between classification and decision. All possible outcomes of the decision process are shown on the right side.

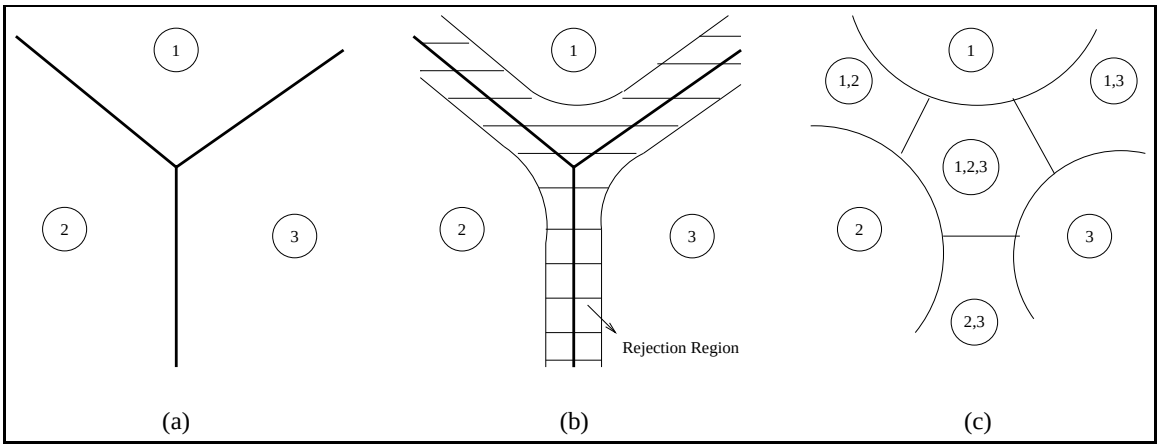

Figure 2: Three decision types.

the error rate for a given reject rate, or vice versa, and simply consists in rejecting the sample if its highest *posterior* probability is lower than some threshold. An adaptation of Chow's results to the k-nearest neighbour rule was achieved by Hellman leading to the (k,k')-nearest neighbor rule [5]. Interestingly, the outcomes of Chow's rule are also singletons, like in the Bayes rule, but augmented by the empty subset $\{\emptyset\}$, which represents the reject option; see Figs. 1 and 2b.

In this report, we generalise the rejection concept, define a new optimality criterion, propose a new decision rule, prove its optimality, and derive some important properties of the new rule. In Chow's works, a sample is rejected if its highest posterior probability is lower than a threshold, disregarding the probability distribution of the remaining classes. Instead, we propose a *class-selective* rejection scheme. That is, we do not reject the sample from all classes but only from those classes that are most unlikely to issue the sample. For instance, for a sample lying on the separation plane between classes 1 and 2, while being very far away from the center of the third class, we should reject only the third class and declare that it belongs to the group composed of the first and the second classes. In other words, we attempt to partition the pattern or feature space into regions each of which corresponds to a subset of classes. Since there are 2^N subsets in a set of N elements, we obtain $2^N - 1$ regions, excluding the empty subset, 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 $\{1,2,3\}$. However, there exists a trivial partition - that assigns the whole feature space to the group composed of all N classes - which nullifies the error rate. This partition would correspond to a no-decision rule, however.

To avoid the trivial partition, let us define the average number of classes by

$$\bar{n} = \int_X n(x)p(x)dx \tag{6}$$

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, and N_s is the total number of patterns involved in the experiment.

The optimality of the decision rule can now be 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

$$risk(x) = 1 - \sum_{i \in Selected\ Subset} P_i(x) = \sum_{i \in Rejected\ Subset} P_i(x) \tag{7}$$

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. (7) 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. (7) becomes Eq. (5).

Section 2 discusses a necessary condition for optimality. The optimum decision rule is then proposed and its optimality proven in Section 3. The next section

presents various upper-bounds on error rate and average number of classes. Some examples are given in Section 5 illustrating different aspects of the new rule. We discuss the implications of the selective rejection concept in Section 6 and conclude the study in Section 7.

2 A Necessary Condition

In this section, a necessary condition for optimality is given. This allows us to focus our attention only on those rules that satisfy the necessary condition in the establishment of the optimum rule. Some general properties of these rules are then presented as preliminaries to the derivation of the new optimum decision rule.

The key point in our decision rule is the choice of the Selected Subset, for a given pattern x, which uniquely specifies the *posterior* probabilities $\{P_i(x); i = 1, .., N\}$. For convenience, the choice is decomposed into two parts, namely, the cardinality $n(x)$ of the Selected Subset and the particular combination of $n(x)$ classes among the total number of N classes. The necessary condition concerns the choice of a particular combination of classes, assuming a given cardinality $n(x)$, whereas the determination of $n(x)$ will be subject of the next section.

Theorem 1 *A necessary condition for optimality is to select the $n(x)$ classes among N that have the highest posterior probabilities.*

Proof:

We shall prove that any other combination of n among N increases the error rate and therefore would not be optimum. Consider

$$e = \int_X risk(x)p(x)dx = \int_X [1 - \sum_{i \in Selected\ Subset} P_i(x)]p(x)dx \tag{8}$$

Since the cardinality is the same, any combination other than the one that has the highest posterior probabilities would have a smaller $\sum_{i \in Selected\ Subset} P_i(x)$, and therefore a larger error rate e, as $p(x) > 0$ (no compensation). If more than one combination of n among N classes have the same highest $\sum_{i \in Selected\ Subset} P_i(x)$, we can just pick any at random, since the others would give the same error rate.

To obtain an explicit expression of $risk(x)$, let us introduce the sequence $\{Q_i(x)\}$, which is simply an reordered sequence of $\{P_i(x)\}$ in decreasing order

$$\{P_i(x); i = 1, .., N\} \rightarrow \{Q_i(x); i = 1, .., N\} / Q_i(x) \geq Q_{i+1}(x); i = 1, .., N-1 \tag{9}$$

Taking the combination of n among N classes with highest posterior probabilities leads to

$$risk_n^*(x) = 1 - \sum_{i=1}^{n} Q_i(x) \tag{10}$$

where the superscript '*' denotes the optimality.

For convenience, let us also define the sequence of accumulated probabilities $F_k(x)$ by

$$F_k(x) = \sum_{i=1}^{k} Q_i(x) \tag{11}$$

It can readily be seen that $F_k(x)$ is non-decreasing since we only add non-negative values to the sequence, and it is convex upward because the added quantity decreases (or remains the same) with k ($Q_i(x)$ is a non-increasing sequence).

3 Optimal Decision Rule

Theorem 2 *The optimum cardinality is*

$$n^*(x,t) = \min_{k \in [1,N]} \{k / Q_{k+1}(x) \leq t\} \tag{12}$$

with the convention

$$Q_{N+1}(x) = 0 \tag{13}$$

and the domain of the pre-specified threshold

$$0 \leq t \leq \frac{1}{2} \tag{14}$$

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. (12) 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 cardinality equal to the minimum index, which is 1.

In words, the optimum decision rule assigns to pattern x all classes whose posterior probability is greater than the pre-specified threshold t. If there exist no such classes, the rule simply selects the (a) single best class.

Proof:

Let $e^*(t)$ and $\bar{n}^*(t)$ be the error rate and the average number of classes when the optimum decision rule is used.

$$e^*(t) = \int_X risk^*_{n^*(x,t)}(x)p(x)dx \tag{15}$$

$$\bar{n}^*(t) = \int_X n^*(x,t)p(x)dx \tag{16}$$

Consider any other rule, satisfying the necessary condition, such that

$$\bar{n} = \bar{n}^*(t) \tag{17}$$

We shall show that such a rule would produce an equal or higher error rate than the optimum error rate, i.e.

$$e \geq e^*(t) \tag{18}$$

Note that it is sufficient for the proof to take a rule that satisfies the necessary condition because for any rule that does not satisfies this condition, there exists another rule (obtained by permutation of classes) that does and that provides a lower error rate, for the same average number of classes.

Let us partition the feature or pattern space into three regions X_0, X_1, and X_2, such that

$$x \in X_0 : n = n^* \tag{19}$$

$$x \in X_1 : n > n^* \tag{20}$$

$$x \in X_2 : n < n^* \tag{21}$$

Then

$$\bar{n} - \bar{n}^*(t) = \int_X [n(x) - n^*(x,t)]p(x)dx \tag{22}$$

$$\bar{n} - \bar{n}^*(t) = \int_{X_1 \cup X_2} [n(x) - n^*(x,t)]p(x)dx \tag{23}$$

or, due to Eq. (17)

$$\int_{X_1} [n(x) - n^*(x,t)]p(x)dx + \int_{X_2} [n(x) - n^*(x,t)]p(x)dx = 0 \tag{24}$$

Similarly, we have

$$e - e^*(t) = \int_{X_1} [risk^*_n(x) - risk^*_{n^*}(x,t)]p(x)dx + \int_{X_2} [risk^*_n(x) - risk^*_{n^*}(x,t)]p(x)dx \tag{25}$$

But

$$risk^*_n(x) - risk^*_{n^*}(x,t) = [1 - \sum_{i=1}^{n} Q_i(x)] - [1 - \sum_{i=1}^{n^*} Q_i(x)] \tag{26}$$

$$risk_n^*(x) - risk_{n^*}^*(x,t) = \sum_{i=1}^{n^*} Q_i(x) - \sum_{i=1}^{n} Q_i(x) \tag{27}$$

Let us examine the risk variation in X_1 and X_2. For

$$x \in X_1 : n > n^* : risk_n^*(x) - risk_{n^*}^*(x,t) = -\sum_{i=n^*+1}^{n} Q_i(x) \tag{28}$$

We have

$$t \geq Q_{n^*+1} \geq Q_{n^*+2} \geq ... \geq Q_n \tag{29}$$

The first inequality is due to the application of the optimum decision rule (Eq. (12)) whereas the followings express the non-increasing nature of $\{Q_i\}$ (Expression (9)). Summing up all Q_is in the above expression leads to

$$t[n - n^*] \geq \sum_{i=n^*+1}^{n} Q_i(x) \tag{30}$$

and therefore

$$x \in X_1 : n > n^* : risk_n^*(x) - risk_{n^*}^*(x,t) \geq -t[n - n^*] \tag{31}$$

Similarly, for

$$x \in X_2 : n < n^* : risk_n^*(x) - risk_{n^*}^*(x,t) = \sum_{i=n+1}^{n^*} Q_i(x) \tag{32}$$

We have $(1 \leq) n < n^* \Rightarrow n^* > 1$. This, in turn, implies that we are not in *Case b* of *Remarks*, but must be in *Case a*. Therefore $Q_{n^*} > t$. By using again the non-increasing nature of $\{Q_i\}$ (Expression (9)),

$$Q_{n+1} \geq Q_{n+2} \geq ... \geq Q_{n^*} > t \tag{33}$$

Summing up all Q_is in the above expression leads to

$$\sum_{i=n+1}^{n^*} Q_i(x) > t[n^* - n] \tag{34}$$

Therefore

$$x \in X_2 : n < n^* : risk_n^*(x) - risk_{n^*}^*(x,t) > t[n^* - n] \tag{35}$$

Substituting Inequalities (31) and (35) into Eq. (25), and taking into account Eq. (24) result in

$$e - e^*(t) \geq -t \int_{X_1} [n(x) - n^*(x,t)]p(x)dx - t \int_{X_2} [n(x) - n^*(x,t)]p(x)dx = 0 \tag{36}$$

which is what we wanted to prove.

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 our 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.

4 Upper-bounds

In this section, we derive various upper-bounds on error rate and average number of classes, when using the optimum decision rule. For simplicity, the superscript '*' will be omitted.

4.1 An upper-bound of $e(t)$

We shall show that using the optimum decision rule leads to

$$e(t) \leq t[N - \bar{n}(t)] \leq t(N-1) \tag{37}$$

where N is the total number of classes.

Since the *posterior* probabilities sum up to 1, the optimum risk (Eq. (10)) can also be expressed as

$$risk(x,t) = \sum_{i=n(x,t)+1}^{N} Q_i(x) \tag{38}$$

Using the optimum decision rule (Eq. (12)) and taking into account Expression (9), we get

$$t \geq Q_{n+1} \geq Q_{n+2} \geq ... \geq Q_N \tag{39}$$

Summing up all Q_is in the above expression leads to

$$t[N - n(x,t)] \geq \sum_{i=n(x,t)+1}^{N} Q_i(x) \tag{40}$$

The risk is therefore bounded by

$$risk(x,t) \leq t[N - n(x,t)] \tag{41}$$

Taking the expectation of both sides, we get

$$e(t) = \int_X risk(x,t)p(x)dx \leq \int_X t[N - n(x,t)]p(x)dx \tag{42}$$

or

$$e(t) \leq tN \int_X p(x)dx - t \int_X n(x,t)p(x)dx = t[N - \bar{n}(t)] \leq t(N-1) \tag{43}$$

since $\bar{n}(t) \geq 1$.

4.2 Upper-bounds of $n(x,t)$ and $\bar{n}(t)$

We will show that

$$n(x,t) < \frac{1}{t} \tag{44}$$

and

$$\bar{n}(t) < \frac{1}{t} \tag{45}$$

First, let us notice that

$$Q_i(x) \leq \frac{1}{i} \tag{46}$$

since otherwise $\sum_{i=1}^{N} Q_i(x) > 1$.

The optimum decision rule implies that, in *Case a*,

$$Q_{n(x,t)}(x) > t \tag{47}$$

Therefore

$$\frac{1}{n(x,t)} \geq Q_{n(x,t)}(x) > t \tag{48}$$

or

$$n(x,t) < \frac{1}{t} \tag{49}$$

In *Case b*, $Q_1 \leq t, n^* = 1$. Since $t \leq \frac{1}{2}$, we have $\frac{1}{t} \geq 2$, and the inequality (49) remains true.

Moreover, the average number of classes is also upper-bounded by the same bound

$$\bar{n}(t) = \int_X n(x,t)p(x)dx < \frac{1}{t}\int_X p(x)dx = \frac{1}{t} \tag{50}$$

5 Examples

We illustrate the theory by a few examples, where the pattern space is the two-dimensional Euclidean space. In Section 5.1, we consider a three-class problem with class conditional p.d.f.'s being gaussian. This example shows how the partition of the pattern space changes as the pre-specified threshold t varies from $\frac{1}{2}$ down to 0. It will also serve the illustration of the tradeoff between the error rate and the average number of classes, called $e - \bar{n}$ curve in the following, which is the counterpart of Chow's error-reject curve. In Section 5.2, we compare the optimum $e - \bar{n}$ curve with those of two other heuristic rules.

5.1 A Three-Class Problem

Consider the three-class problem. Each class p.d.f is a gaussian with unit standard deviation. The center coordinates of class 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. They are shown on the top of Fig. 3.

By varying the threshold t from $\frac{1}{2}$ down to 0, we obtain different partitions of the feature space as shown in Fig. 3. In case (a), $t = \frac{1}{2}$, the partition is identical to that of the Bayes rule. In case (b), $t = 0.4$ and thus $\frac{1}{t} = 2.5$, Inequality (44) predicts that $n(x,t) < 2.5$. Since n is an integer, the only values that it can take on are 1 and 2. This can be easily verified from Fig. 3b that there is no subset with three classes. As the threshold t decreases, see Figs. 3b, 3c, and 3d, we can observe the emergence of the subset $\{1,2,3\}$ at the center of the three classes. Its appearance happens precisely when $t = \frac{1}{3}$, i.e., $\frac{1}{t} = 3$, as can be expected from Inequality (44). At the other extreme, case (h), $t \to 0$, the quasi totality of the feature space is covered by the subset $\{1,2,3\}$, approaching the trivial partition mentionned in Section 1.

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. (4) and (6).

5.2 Comparison with Other Rules

Consider the seven-class problem. The class centers are shown in Fig. 5. We compare the $e - \bar{n}$ curve of the optimum decision rule with those of two other rules, namely, constant risk and top-n ranking.

The constant risk rule consists in taking, for each pattern x, as many classes as necessary to ensure that $risk(x)$ is lower than a pre-specified threshold t_1. This corresponds to choosing the smallest number of top-k classes such that the accumulated probability $F_k(x)$ exceeds $(1-t_1)$. See Eq. (11) for the definition of $F_k(x)$. The top-n ranking works as follows. For $n = 1$, take the single best class, i.e., using the Bayes rule. For $n = 2$, take the first and second best classes, no matter how small $Q_2(x)$ is, and so on. The optimum decision and constant risk rules are dynamic in that the number of classes varies with the input pattern x, whereas the top-n ranking rule is static.

In Fig. 6, the three $e - \bar{n}$ curves are plotted, where each class p.d.f. is a gaussian with unit standard deviation. In Fig. 7, the same comparison is performed, where each class p.d.f. is a uniform and circular distribution (a flat-top cylinder). Notice that the $e - \bar{n}$ curve of the top-n ranking rule is defined only for integer values of $\bar{n}$. However, to ease the visual comparison, linear interpolation is used in Figs. 6 and 7.

Figs. 6 and 7 reveal that, depending on the class p.d.f.s, heuristic rules may be largely sub-optimal.

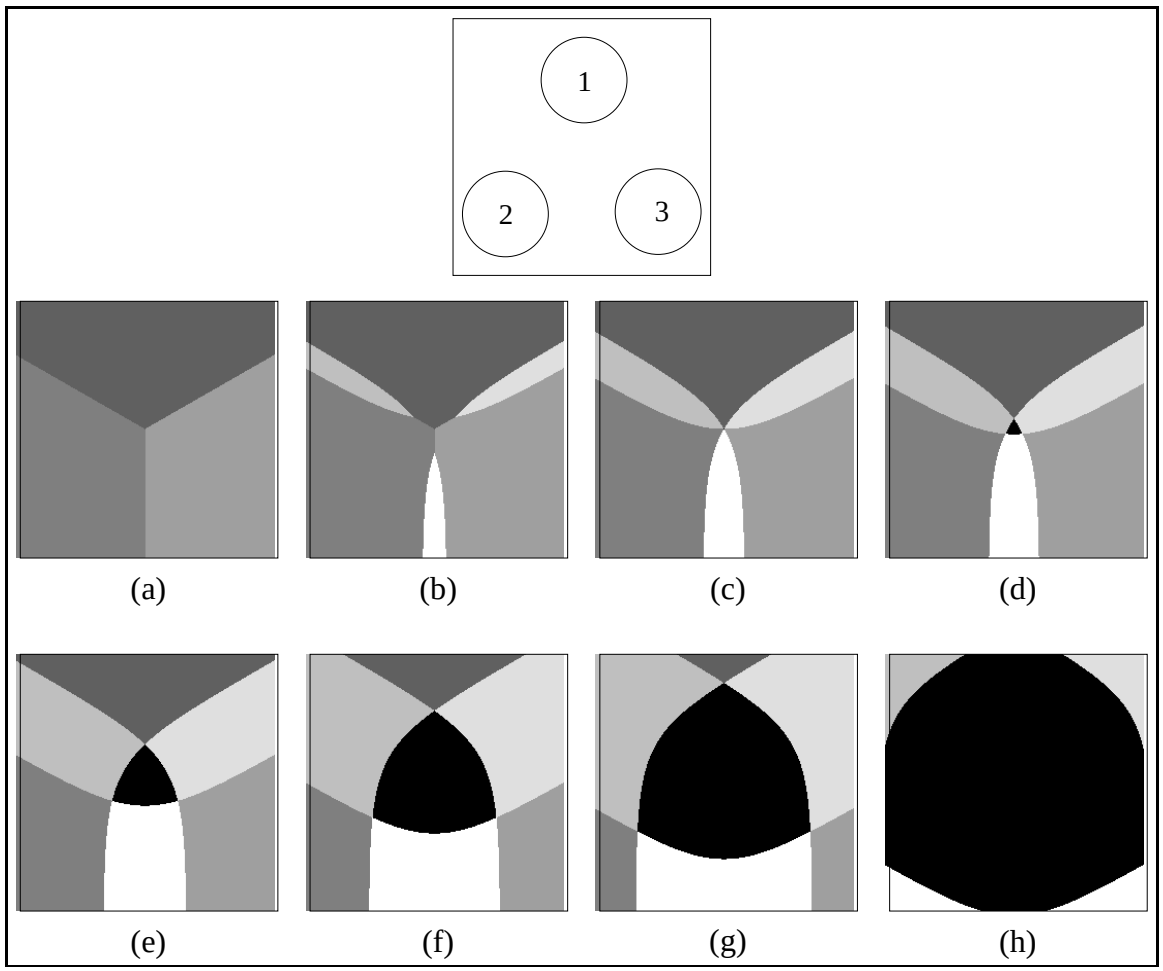

Figure 3: Decision regions; (a) t=0.5; (b) t=0.4; (c) t=1/3; (d) t=0.3; (e) t=0.2; (f) t=0.1; (g) t=0.05; and (h) t=0.01.

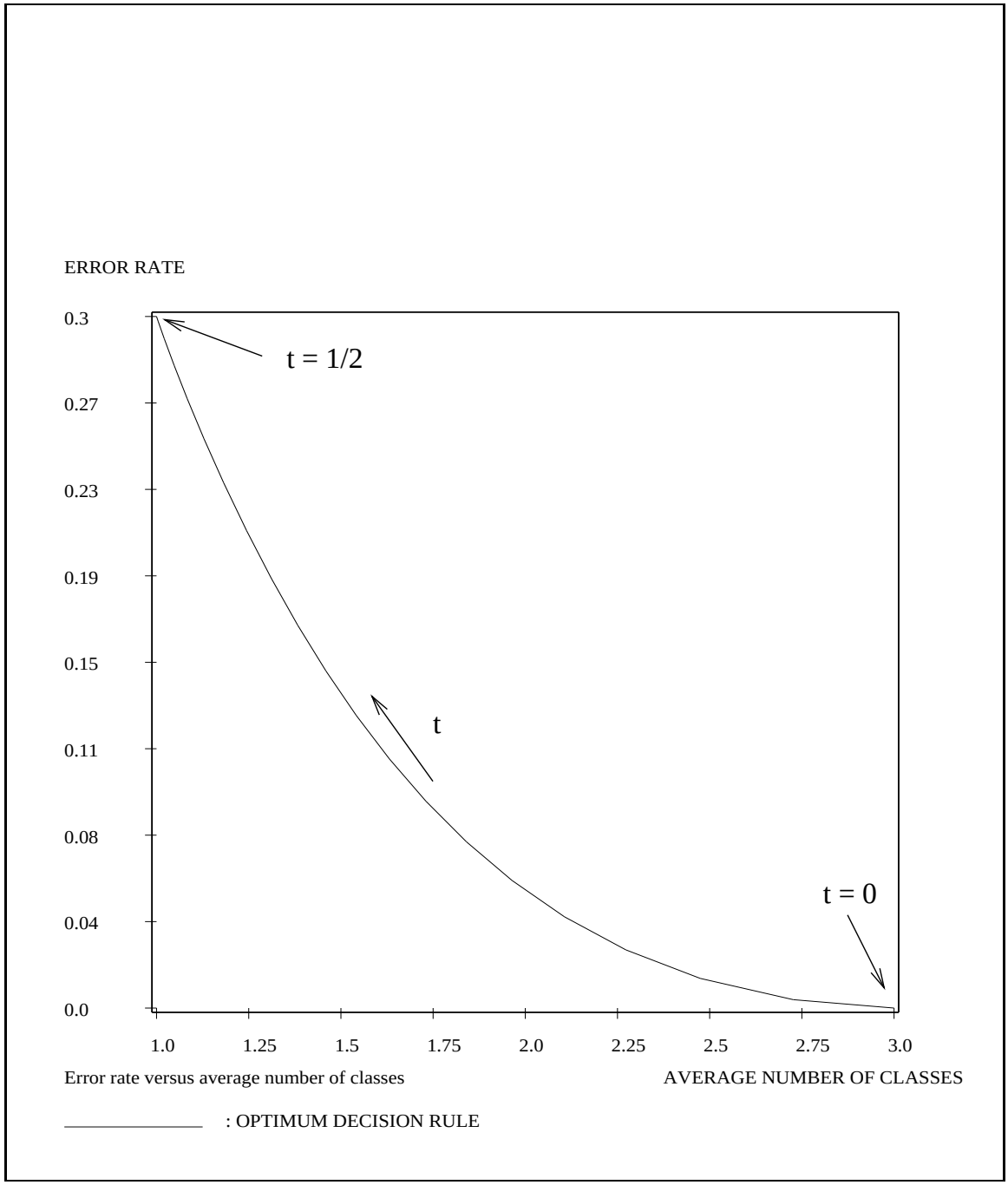

Figure 4: Relation between error rate and average number of classes for a three class problem.

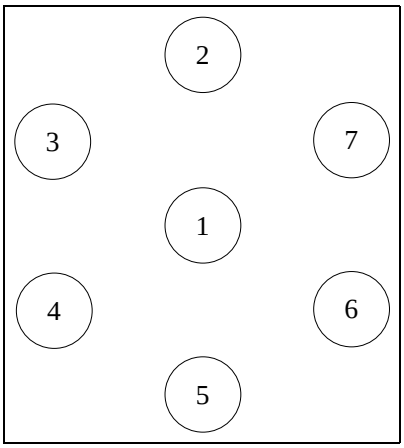

Figure 5: Statistical distributions of a seven-class problem. The distances between the center of class 1 and the others are 1.

6 Implications

On the theoretical level, the formulation of the problem constitutes an extension of the concept of rejection to that of selective rejection. The latter allows of avoiding situations where "frustrating" decisions have to be made. For instance, when $Q_1(x) = Q_2(x) = \frac{1}{2}$ and $Q_3(x) = .. = Q_N(x) = 0$, the Bayes rule takes one of the two best classes at random, whereas Chow's rule simply rejects the pattern x, although it is compelling to reject all classes but the best two. This is precisely what the new rule provides. On the other hand, the commonly used top-n ranking rule needs a long list of candidates (large and fixed value of n) to lower the error rate. Often, it provides candidates with quasi-null *posterior* probabilities, simply because it has to fill-in the list.

Practically, the optimum decision rule can be used for pre-classification purpose. In man-machine interface, the new rule provides a convenient way for computer to preselect the most promising candidates to be examined by human operators in a later stage. Finally, the upper-bound of error rate allows an automatic setting of the threshold t, according to a pre-specified tolerable error rate (Inequality (37)).

7 Conclusion

The concept of rejection has been extended to that of class-selective rejection. We started by reviewing various decision rules for statistical pattern recognition, namely, Bayes and Chow's rules. The outcomes of such decision rules were considered as particular cases in a more general framework, where possible decision outcomes are extended to all subsets of the power-set of classes. We defined a new optimality criterion, for accommodating the newly introduced decision outcomes, to be the rule that minimises the error rate for a given average number of classes. Then, a new decision rule is proposed and its optimality proven. The optimum decision

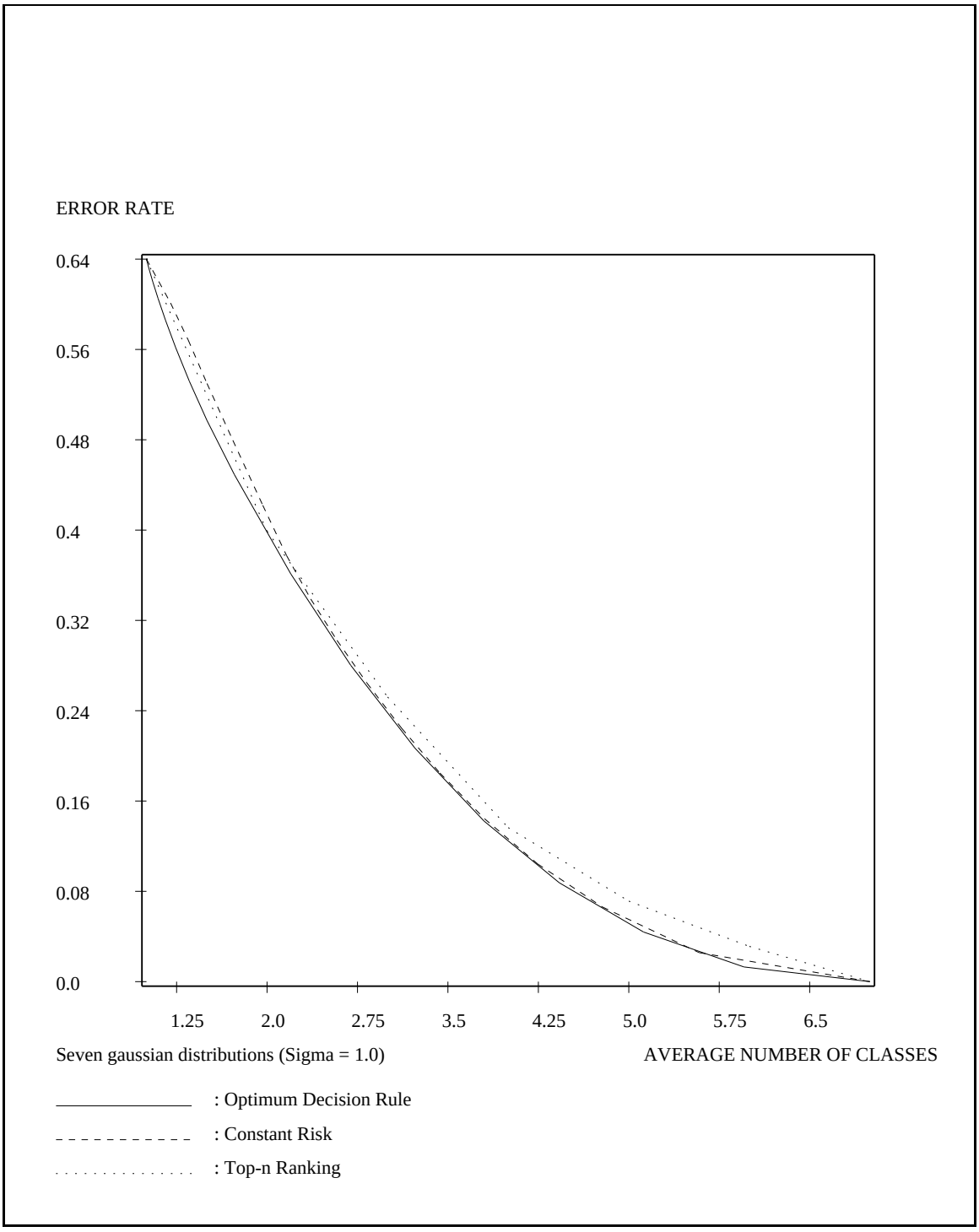

Figure 6: Comparison of the optimum decision rule with the constant risk and top-n ranking rules for a seven-class problem with gaussian distributions.

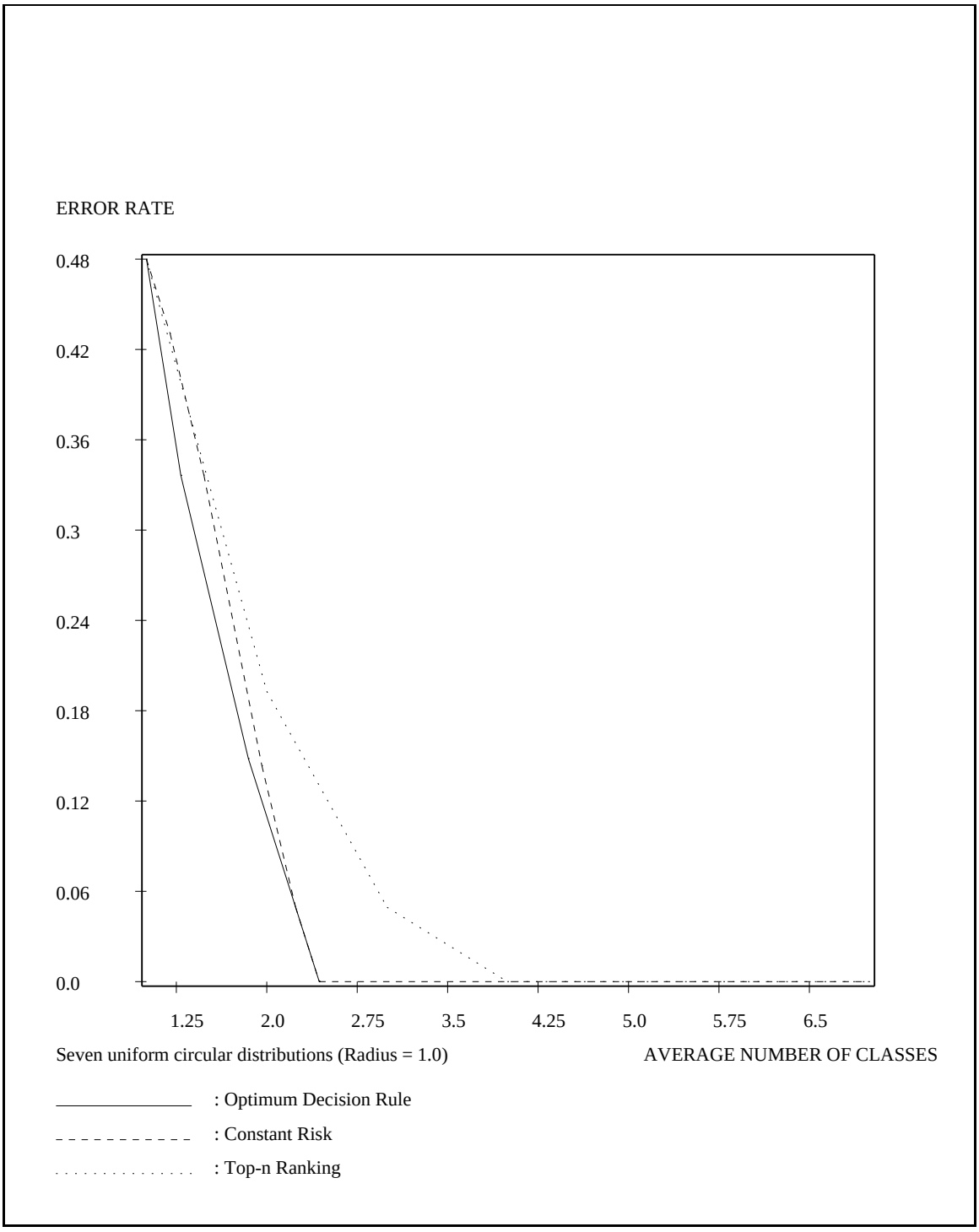

Figure 7: Comparison of the optimum decision rule with the constant risk and top-n ranking rules for a seven-class problem with uniform, circular distributions.

rule consists in assigning to the input pattern all classes whose *posterior* probability exceeds a pre-specified threshold; if there exist no such classes, the rule simply selects the (a) single best class. Various upper-bounds on error rate and average number of classes were obtained. Examples were provided to illustrate the various partitions of pattern space according to the pre-specified threshold, and the optimum tradeoff curve between the error rate and the average number of classes. Comparisons with some heuristic rules, such as the top-n ranking, showed that the latters may be largely sub-optimal for some configurations of class distributions. Finally, potential applications of the new rule were discussed, including pre-classification and man-machine interface.

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

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

[3] 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.

[4] 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.

[5] 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.