Determination of the Optimum Error and Reject Validity Regions as Functions of threshold in the Case of Three Classes

Jorge Isidro Aranda[1], Arturo Baltazar[2]

[1] UMSNH, Facultad de Ciencias Físico-Matemáticas,
Ciudad Universitaria, Morelia, Michoacán,
chilo@fismat.umich.mx
[2] Cinvestav-Saltillo, División de Robótica y Manufactura Avanzada,
Carretera-Saltillo-Monterrey, Km. 13.5, Ramos Arizpe, Coah., 25900, México
arturo.baltazar@cinvestav.edu.mx

Abstract. The performance of a pattern recognition system is characterized by its error rate and the reject rate tradeoff. The error rate can be directly evaluated from the reject function assuming a threshold t, which is a parameter that limits the reject and acceptance regions. In this paper, optimum rejection rule of a recognition system for three classes is used to calculate the reject function and the error rate evaluated directly from the reject rate. A Bayes decision rule as a function of threshold t is used to determine the minimum risk. A simple parametrization that considers the distance between means of the normal distribution of classes is presented. Finally, the rejection rate, the error rate and conditional risk were estimated in terms of the threshold t to illustrate the effect of the proposed parametrization and Bayes decision rule for minimum risk.

Keywords: Bayes error; minimum risk; reject rate.

1 Introduction

It is known that the Bayes error provides the lowest error rate for a given pattern classification problem. An optimum rejection rule and a general relation between the error and reject probabilities are presented by Chow (1970). There are several classical approaches used to estimate or to find bounds of the Bayes error including those proposed by Thumar et al. (1966) and Doermann (2004) which considered the second order dependency between the class and decision, and found that a combined-based method renders better estimates than the classical methods of dependency-based product approximation (DBPA). Pierson (1998) used boundary methods for estimating class separability since it does not require knowledge of the posterior distributions. In this work, optimum rejection rule of a recognition system to calculate the reject function and the error rate evaluated directly from reject rate and the results are illustrated for a case involving three classes, commonly found in actual classification problems but hardly described in the literature. The Bayes decision rule for minimum error and reject option for n classes considering the minimum risk,

Advances in Artificial Intelligence.
Research in Computing Science 26, 2006, pp. 61-71
Received 03/06/06
Accepted 03/10/06
Final version 11/10/06

assuming a threshold t with a simple parametrization for a three class case is described.

2 Background

2.1 Decision Rule for Minimum Error and Reject Option

In actual classification problems where classes are not fully separable, it is unrealistic to expect absolute classification performance of the pattern recognition system. The object of a statistical classification problem is to reach the best possible performance. The question that arises is how to determine the optimum classification rate which can be answered by the determination of the Bayes error since the Bayes decision rule provides the lowest error rates (Tumer, 1996).
In general, to assign a pattern x to n classes wi (where i =1,…,n), a model for classification with a decision rule to partition the measurement space into n regions Ωi, i =1,…, n is needed. The boundaries between the regions Ωi are known as the decision boundaries or decision surfaces; usually it is near to these boundaries that the highest probability of misclassifications can occur. In such situations, the decision on the pattern may be withheld or rejected until further information is available, this option is known as the reject option (Webb, 2002).

According to the Bayes decision rule where assigning with minimum error a pattern x to a class wi, we have that:

$$p(w_i) > p(w_k) \quad k, i = 1,...,n;\ k \neq i\ , \tag{1}$$

where $p(w_1),\ldots, p(w_n)$,are known prior probabilities.

The optimum decision rule is to reject a pattern x if the maximum of the posterior probabilities does not exceed some predefined threshold t, which can take values between 0 and 1 ($0 \leq t \leq 1$). (Chow, 1970; Pierson, 1998). More explicitly, the optimum recognition rule is to accept the pattern x and to classify it as belonging to the kth class whenever the following is true:

$$p(w_k)\,p(x|w_k) \geq p(w_i)\,p(x|w_i) \quad , \tag{2}$$

and

$$p(w_k)\,p(x|w_k) \geq (1-t)\sum_{i=1}^{n} p(w_i)\,p(x|w_i)\,, \tag{3}$$

and to reject the pattern whenever:

$$\max_i \left[p(w_i)\, p(x|w_i) \right] < (1-t) \sum_{i=1}^{n} p(w_i)\, p(x|w_i) \quad . \tag{4}$$

The term (1-t) (see Figure 2) indicates the maximum values that can be assigned to the posterior probability p($wi|x$) to do a correct classification of a measurement pattern.Through Bayes's theorem, this posterior probability function is related to the class conditional density by:

$$p(w_i|x) = \frac{p(x|w_i)\, p(w_i)}{p(x)}, \tag{5}$$

where

$$p(x) = \sum_{i=1}^{n} p(w_i)\, p(x|w_i) \tag{6}$$

is the probability of pattern x. The class probability distributions $p(w_i)$ can be estimated using an iterative method with the patterns belonging to each class (Baram, 1999).

For any fixed value of t, the decision rule (this is used for the correct classification of samples) partitions the pattern space into two disjointed sets (or regions) A(t) and R(t) given by:

$$A(t) = \left\{ x \,\middle|\, \max_i p(w_i)\, p(x|w_i) \geq (1-t)\, p(x) \right\}, \tag{7}$$

$$R(t) = \left\{ x \,\middle|\, \max_i p(w_i)\, p(x|w_i) < (1-t)\, p(x) \right\}, \tag{8}$$

where A(t) is the acceptance region which implies that once the maximum posterior probability exceeds the threshold (1-t), a classification decision can be made; R(t) is the rejection region where the equations (3) and (4) hold. The integral of regions A(t) and R(t) defines the reject rate r(t) and correct classification c(t) expressed as:

$$r(t) = \int_{R(t)} p(x)\, dx, \tag{9}$$

which describes the unconditional probability of rejecting a measurement x and

$$c(t) = \int_{A(t)} \max_i \left[p(w_i)\, p(x|w_i) \right] dx, \tag{10}$$

is the probability of correct recognition of the patterns of the measurements. The probability $e(t)$ of accepting a pattern for classification and incorrectly classifying it is known as error rate given by:

$$e(t) = 1 - c(t) - r(t). \tag{11}$$

A correct recognition can be done if given an error rate (error probability) the reject rate (reject probability) is minimized. In this work, a parametrization to illustrate its effect on error and reject rate as well as the minimum risk for three classes is introduced following the work by Chow (1970) for the case of two classes.

Assuming two classes and a pattern x with equal prior probability of occurrence, $p(w_1) = p(w_2) = 1/2$, the condition for rejection (Eq. (4)) can never be satisfied when $t > ½$; this can be explained by the fact that the minimum value which $\max_i \left[p(w_i|x)\right]$ can attain is $1/n$ since $1 = \sum_{i=1}^{n} p(w_i|x) \le n \max_i p(w_i|x)$; using Eq. (4) the threshold rule $t \le 1 - 1/n$ can be obtained (Chow (1970); Webb (2002)) to activate the rejection option. The reject rate is always zero if t exceeds 1/2, therefore t only can have values in the range $0 \le t \le 1/2$. To estimate $r(t)$ and $e(t)$ two normal distributions $p(x|w_i) = (1/\sigma\sqrt{2\pi})\ \exp(-(x-\mu_i)^2/2\sigma^2)$ are assumed with means μ_1 and μ_2 ($\mu_1 > \mu_2$) and equal covariance σ^2. Chow (1970) used the following parametrization:

$$s_2 = \frac{\mu_1 - \mu_2}{\sigma}, \tag{12}$$

which describes the separation between the means of the distributions and is the only parameter of the distributions that $r(t)$ and $e(t)$ depend upon. The error and reject rates can be expressed using the standard cumulative distribution function $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-x^2/2)\,dx$ as: $e(t) = \Phi(a)$ and $r(t) = \Phi(b) - \Phi(a)$, where $a = -s_2/2 - \ln(1/t - 1)/s_2$ and $b = -s_2/2 + \ln(1/t - 1)/s_2$ (Chow, 1970). Figure 1, shows results of error and reject rate in terms of the threshold t and $s_2 = 1, 2, 3, 4$. All curves tend to zero when $t = 1/2$ and to 1 when $t = 0$ proving consistency with the threshold rule when n=2. The results for two classes are now compared with results in the case of three classes.

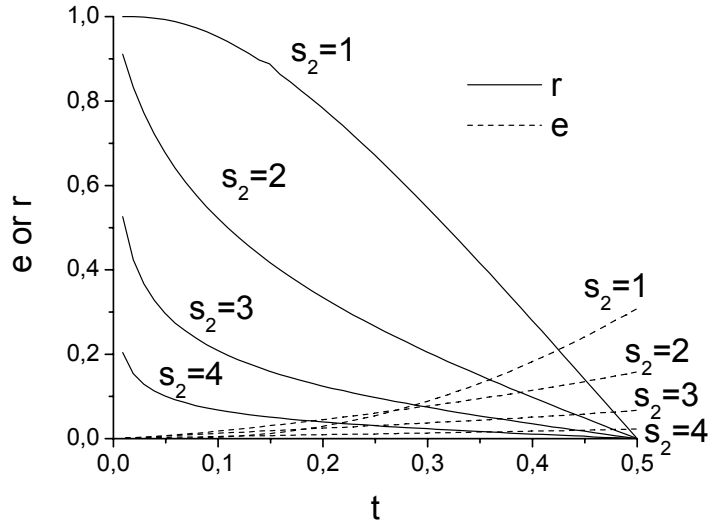


Fig. 1. Reject and error rate in terms of threshold *t* for two classes considering several values of the parameter s_2 are shown.

3. Estimation of Error and Reject Rate and Risk for Three Classes

3.1 Error and Reject rate

Accordingly with previous section, to calculate the error rate, Eq. (11), it is necessary to have the reject rate and the probability of correct recognition; to obtain such results the prior class-conditional probability density functions $p(x|w_i)$ for each class is needed. In the case of three classes, assuming a Gaussian distribution, these can be found as follows:

$$p(x|w_i) = \frac{1}{\sqrt{2\pi}\,\sigma}\left(e^{-(x-\mu_i)^2/2\sigma^2}\right). \tag{13}$$

In this case a parametrization is introduced after a change of variable in terms of the means and standard deviation as follows: $y = (x-\mu_1)/\sigma$, s_2 (given by Eq. 12) and $s_3 = (\mu_3 - \mu_1)/\sigma$. After some algebra and arranging terms in Eq. (13) (for *i*=1,2,3), the corresponding prior class-conditional probability density functions for three classes are:

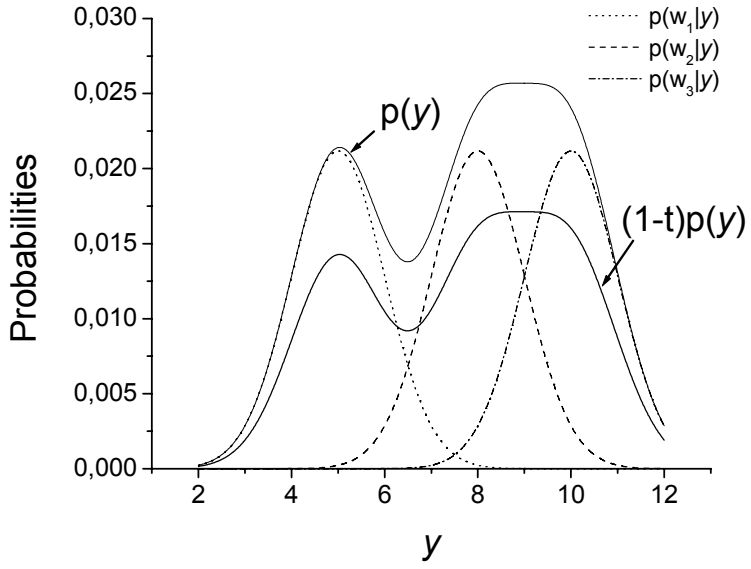
$$p(y|w_1) = \frac{1}{\sqrt{2\pi}\,\sigma}\left(e^{-y^2/2}\right),$$

$$p(y|w_2) = \frac{1}{\sqrt{2\pi}\,\sigma}\left(e^{-(y-s_2)^2/2}\right), \quad (14)$$

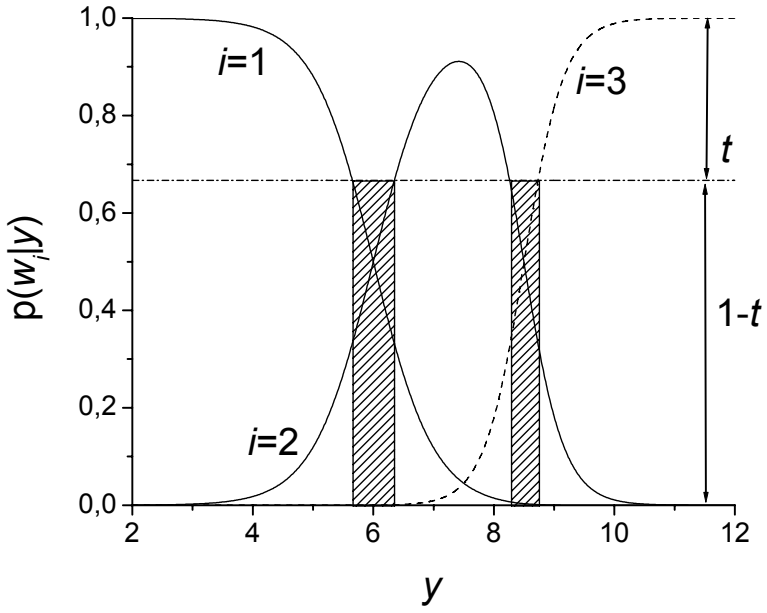
$$p(y|w_3) = \frac{1}{\sqrt{2\pi}\,\sigma}\left(e^{-(y-s)^2/2}\right),$$

The posterior probability densities for each class $p(w_1|\,y)$, $p(w_2|\,y)$ y $p(w_3|\,y)$ are now given using eq. (14) in Eq.(5).

It can be seen that Eq. (5) allows easy visualization of the intersection points for different values of s_3 and s_2 facilitating to calculate the area under the reject rate curve, as shown in Figure 2a. Two reject regions can be found, the first between the intersection points of $p(w_1|\,y)$ and $p(w_2|\,y)$ with (1-*t*), and the second between the intersection of $p(w_2|\,y)$ and $p(w_3|\,y)$ with (1-*t*) (Figure 2b).



(a)



(b)

Fig. 2. (a) Probability densities considering threshold *t* calculated from Eq. (3) assuming equal class probabilities; (b) a posterior probability densities estimated using Eq. (5); where the horizontal dashed line represents (1-*t*) the horizontal threshold; *i* = 1, 2, 3 indicates the three class-condition probabilities used in the calculations. The threshold *t* varies between the interval ($0 \le t \le 2/3$) in the case of three classes.

The error and reject rate for two cases $s_3 = 1$, $s_2 = 0.5$ and $s_3 = 2, s_2 = 1$ are shown in Figure 3; here it can be seen that the limits of error rate and reject rate corresponds to the maximum value of threshold t =2/3 for three classes and with a minimum value in t=0. Hence, this proves the consistency with the threshold rule $t \le 1 - 1/n$ when n=3. It can be observed that the curves for *e* and *r* in the case of three classes shown in Figure 3 are different of those for two classes described in Figure 1.

3.2 Decision Rule for the Minimum Risk.

In the previous section, the decision rule was such that the selected class has the maximum posterior probability $p(w_i|\, y)$ minimizing the probability of making an error. Therefore, a new decision rule that minimizes the expected *loss* or risk is

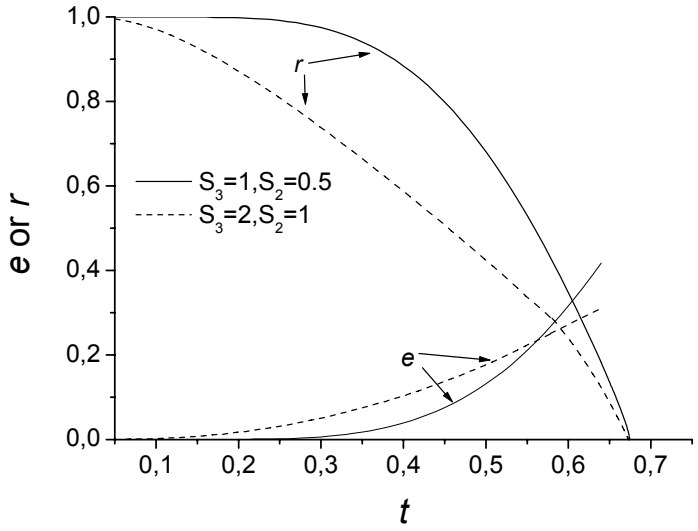


Fig. 3. Estimation of error (*e*) and reject (*r*) rate in terms of threshold *t* for three classes assuming $s_3 = 1$, $s_2 = 0.5$ (solid line) and $s_3 = 2$, $s_2 = 1$ (dash line).

described in Webb (2002); this is very important since in many applications the costs associated with misclassification depend upon the true class of the patterns as well as the class to be assigned. This loss is a measure of the cost of making the decision that a pattern belongs to class w_i when the true class is w_j.

The conditional risk of assigning a pattern y to class w_i can be defined as (Webb, 2002)

$$l^i(y) = \sum_{j=1}^{n} \lambda_{ji} p(w_j|y), \tag{15a}$$

where

$$\lambda_{ji} = \text{cost of assigning a pattern y to } w_i \text{ when } y \in w_j \ . \tag{15b}$$

In this case, a reject option can be introduced to establish a Bayes decision rule in terms of the conditional risk, the reject region R^* can be defined by $R^* = \left\{ \min_i l^i(y) > t \right\}$ (Webb, 2002); the decision rule is to accept a pattern y and assign it to a class w_i if $l^j(y) = \min_j l^j(y) \le t$ and reject y if $l^j(y) = \min_j l^j(y) > t$, this decision is equivalent to make a definition of a region

in which is valid a constant conditional risk $l^0(y) = t$, so that the Bayes decision rule is: to assing y to class w_i if $l^i(y) \leq l^j(y)$ with *j*=0, 1...,**n**. This implies that the Bayes decision rule for minimum risk (Webb,2002) gives the minimum risk r^* given by

$$r^* = \int_R t\, p(y)dy + \int_A \min_{i=1,\ldots,n} l^i(y)\, p(y)\, dy\,. \tag{16}$$

In this work, the expression in previous equation is used to calculate the minimum conditional risk in terms of the threshold *t* for the case of three classes. The posterior probability densities $p(w_i|\, y\,)$ and the threshold *t* are used to describe two regions which are disjoined and complete the reject region. The posterior probability densities contribute to the reject options used for the calculation of minimum conditional risk defined in Eq. (16).

Using the Eq. (15), the conditional risk for the case of three classes can be written as:

$$l^1(y) = \lambda_{11}\, p(w_1|y) + \lambda_{21}\, p(w_2|y) + \lambda_{31} p(w_3|y), \tag{17a}$$

$$l^2(y) = \lambda_{12}\, p(w_1|y) + \lambda_{22}\, p(w_2|y) + \lambda_{32} p(w_3|y), \tag{17b}$$

$$l^3(y) = \lambda_{13}\, p(w_1|y) + \lambda_{23}\, p(w_2|y) + \lambda_{33} p(w_3|y). \tag{17c}$$

A loss function that has been used extensively in practice because of its simplicity and sensibility is the symmetrical or zero-one loss function. Then, Eq. (15b) results in:

$$\lambda_{ij} = \begin{Bmatrix} 0 \text{ if } i = j \\ 1 \text{ if } i \neq j \end{Bmatrix}. \tag{18}$$

This function assigns no loss to a correct decision, and assign a unit loss otherwise, thus, all errors are equally costly. Therefore, minimizing the Bayes risk corresponds to maximizing the posterior probability (Webb, 2002). Considering Eq. (15b) or Eq. (18) it can be found that all the terms in the diagonal of Eq. (17) are zero.

After calculations, the minimum risk *r** for three classes can be calculated with Eq. (16) as shown in Figure 4 in terms of the threshold *t*, the parameters considered here are $s = 1$, $s_2 = 0.5$ and $s_3 = 2$, $s_2 = 1$. In this Figure 4 it is shown the minimum risk in terms of threshold. The results are consistent with fact that the risk goes to zero when *t*=0 and is maximum when *t*=2/3. The best performance of a pattern recognition system is realized when this minimum risk is reached.

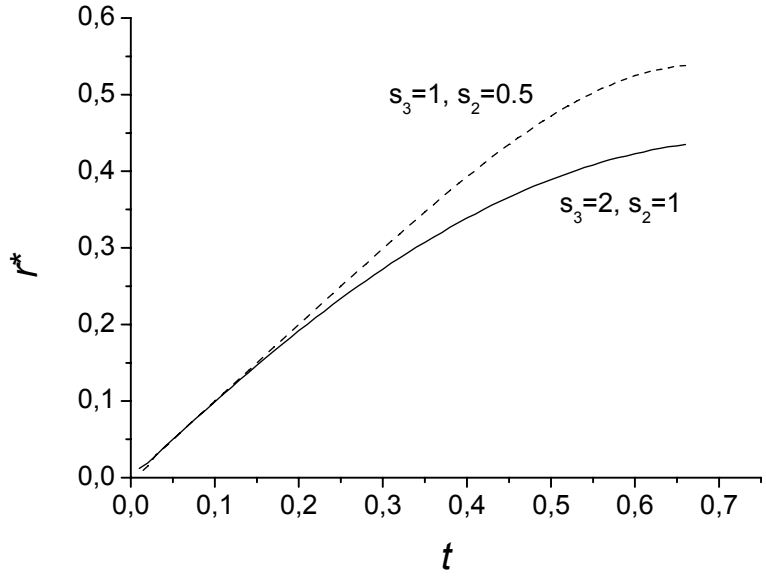


Fig. 4. Minimum risk r^* in terms of threshold t for three classes, with parametrization $s_3 = 1$, $s_2 = 0.5$ (dash) and $s_3 = 2, s_2 = 1$ (solid).

4 Conclusion

The Bayes decision rule to obtain the reject rate for a pattern recognition system for the case of three classes was estimated. A simple parametrization to illustrate the error rate as function for a given threshold was proposed. It was showed the validity of the threshold rule $t \leq 1 - 1/n$ in the case of three classes. It was discussed that when a threshold t, which partitions the measurement space, is fixed, determination of the minimum risk and error rate is allowed. It was showed that through the definition of the threshold t a more efficient pattern recognition system can be reached if the minimum risk is known.

Acknowledgments. The authors appreciate the financial support of CONACYT-SEP through the grant #48085. It is also appreciated to the CIC of the UMSNH for their economic support by means of the projects 9.11 and 9.23.

References

1. Pierson W. E. "Using Boundary Methods for Estimating Class Separability", Ph.D. Thesis, The Ohio State University, 1998.
2. Webb, A. "Statistical pattern recognition", Ed. John Wiley & Sons, LTD, England, 2002.
3. Chow, C. K., "On Optimum Recognition Error and Reject Tradeoff", IEEE Transactions on Information Theory, Vol. IT-16, No. 1, January, 41-46, 1970.
4. Specht, D., "Probabilistic Neural Networks and the Polynomial Adaline as Complementary Techniques for Classification", IEEE Transactions on Neural Networks Vol. 1, No. 1, 525-532, 1990.
5. Baram, Y., "Bayesian classification by iterated weighting", Neurocomputing, 25, 73-79, 1999.
6. Doerman D. and Kang H., "Product Approximation by Minimizing the Upper Bound of Bayes Error Rate for Bayesian Combination of Classifiers", Proceedings of the 17th ICPR'04, IEEE, 1051-4651, 2004.
7. Tumer K. and Ghosh J. "Estimating the Bayes Error Rate Through Classifier Combining", Proceedings of ICPR'96, IEEE, 1015-4651, 1996.