Multiple Reject Thresholds for Improving Classification Reliability

Giorgio Fumera, Fabio Roli[1], and Giorgio Giacinto

Dept. of Electrical and Electronic Engineering – University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{fumera, roli, giacinto}@diee.unica.it

Abstract. In pattern recognition systems, Chow's rule is commonly used to reach a trade-off between error and reject probabilities. In this paper, we investigate the effects of estimate errors affecting the a posteriori probabilities on the optimality of Chow's rule. We show that the optimal error-reject trade-off is not provided by Chow's rule if the a posteriori probabilities are affected by errors. The use of multiple reject thresholds related to the data classes is then proposed. The authors have proved in another work that the reject rule based on such thresholds provides a better error-reject trade-off than in Chow's rule. Reported results on the classification of multisensor remote-sensing images point out the advantages of the proposed reject rule.

1   Introduction

In statistical pattern recognition, the probability that a given pattern, characterized by a feature vector x, belongs to the i-th class, in a N-class problem, is provided by the a posteriori probability $P(\omega_i|x)$ through the Bayes formula:

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \quad i = 1, \ldots, N \ , \tag{1}$$

where $p(x|\omega_i)$ is the conditional probability density function for x in the i-th class, $P(\omega_i)$ is the a priori probability of occurrence of the i-th class, and $p(x)$ is the probability density function for x:

$$p(x) = \sum_{i=1}^{N} p(x \mid \omega_i)P(\omega_i) \ . \tag{2}$$

A classification algorithm is aimed to subdivide the feature space into N decision regions D_i, $i = 1, \ldots, N$, so that the patterns of the class ω_i belong to the region D_i. According to the Bayes theory, the decision regions are defined to maximize the following probability of correct recognition, commonly named "accuracy" of the classifier:

$$Accuracy = P(correct) = \sum_{i=1}^{N} \int_{D_i} p(x \mid \omega_i)P(\omega_i)dx \ . \tag{3}$$

---

[1] Corresponding author. Phone: +39-070-6755874 Fax: +39-070-6755900

To this end, each pattern x must be assigned to the class for which the $P(\omega_i|x)$ is maximum. This is the so called Bayes decision rule. The classifier that maximizes the above correct classification probability is named "optimal Bayes classifier". On the analogy of equation 3, it is easy to see that the classifier error probability can be computed as follows:

$$P(err) = \sum_{i=1}^{N} \int_{D_i} \sum_{\substack{j=1 \\ j \neq i}}^{N} p(x \mid \omega_j) P(\omega_j) dx \ , \qquad (4)$$

where $P(correct)+P(err)=1$. The minimum of the above error probability can be reached by the Bayes rule and it is named Bayes error.

Theoretically speaking, an error probability lower than Bayes error can be obtained using the so called "reject option". Namely, the patterns that are the most likely to be wrongly classified are "rejected", that is, they are not classified. Typically, they are then handled by more sophisticated procedures (e.g., a manual classification process is performed). In real applications, the aim of reject option is to safeguard against excessive errors in order to obtain the accuracy required by the end-user of the pattern recognition system. However, handling high reject rates is usually too time-consuming for application purposes. In addition, correct classifications may also be converted into rejects as the rejection rate increases. Therefore, a trade-off between error and reject is mandatory. The formulation of the best error-reject trade-off and the related optimal reject rule was given by Chow [1]. According to Chow's rule, a pattern x is rejected if the maximum of the a posteriori probabilities is lower than a given threshold value $T \in [0,1]$:

$$\max_{k=1,\ldots,N} P(\omega_k \mid x) = P(\omega_i \mid x) < T \ . \qquad (5)$$

On the other hand, the pattern x is "accepted" and assigned to the class ω_i, if:

$$\max_{k=1,\ldots,N} P(\omega_k \mid x) = P(\omega_i \mid x) \geq T \ . \qquad (6)$$

The rationale of Chow's reject rule becomes evident if one observes that $\max_i P(\omega_i \mid x)$ is the conditional probability of classifying a given pattern x correctly. Therefore, for a given threshold T and the related reject rate, the patterns with the highest probabilities to be wrongly classified are rejected. A detailed proof of the optimality of Chow's rule can be found in [4]. It is worth noting that, under the assumption that the a posteriori probabilities are exactly known, Chow proved that his decision rule provides the optimal error-reject trade-off [1].

It is easy to see that a classifier using reject option subdivides the feature space into $N+1$ decision regions $D_1,\ldots,D_N,D_0$, such that patterns belonging to D_0 are rejected, and patterns belonging to D_i are assigned to the class ω_i. The reject region D_0 is determined according to equation 5. Equation 6 is used for defining the decision regions $D_1,\ldots,D_N$. Using rejection option it makes sense to distinguish between rejected and accepted patterns. It is then useful to define the reject and acceptance probabilities. The probability that a pattern is rejected is computed as follows:

$$P(reject) = \int_{D_0} p(x) dx \ . \qquad (7)$$

On the other hand, the probability that a pattern is accepted is:

$$P(accept) = 1 - P(reject) = \sum_{i=1}^{N} \int_{D_i} p(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^{N} \int_{D_i} \sum_{j=1}^{N} p(\mathbf{x} \mid \omega_j) P(\omega_j) d\mathbf{x} \quad . \tag{8}$$

It is worth noting that only the accepted patterns are classified. Therefore, $P(correct)+P(err)<1$. It is also easy to see that $P(accept)=P(correct)+P(err)$, and $P(correct)+P(err)+P(reject)=1$.

For classifiers using rejection option the accuracy is defined as the conditional probability that a pattern is correctly classified given that it has been accepted:

$$Accuracy = P(correct \mid accept) = \frac{P(correct, accept)}{P(accept)} \quad .$$

Finally, according to equation 8 and taking into account that $P(correct,accept)=P(correct)$ (i.e., only the accepted patterns are correctly or wrongly classified), we can write the following equation:

$$Accuracy = P(correct \mid accept) = \frac{P(correct)}{P(correct) + P(err)} \quad . \tag{9}$$

As previously pointed out, Chow's reject rule provides the optimal trade-off between error and reject only if the a posteriori probabilities of the data classes are exactly known. However, in real applications, such assumption is not satisfied since the available a posteriori probabilities are affected by estimate errors. Therefore, approaches different from Chow's rule have been proposed to handle the error-reject trade-off [2,3]. However, to the best of our knowledge, no work theoretically addressed the problem of the optimal error-reject trade-off when a posteriori probabilities are affected by errors. In particular, the reject rules proposed in the literature were not theoretically compared with Chow's one.

In this paper, we investigate the effects of estimate errors affecting the a posteriori probabilities on the optimality of Chow's rule (Section 2). We show that the optimal error-reject trade-off is not provided by Chow's rule when the a posteriori probabilities are affected by errors. In section 3 the use of class-related reject thresholds is proposed. The authors have proved in [6] that the reject rule based on such thresholds provides a better error-reject trade-off than in Chow's rule. Section 4 reports results on the classification of multisensor remote-sensing images that point out the advantages of the proposed reject rule. Conclusions are drawn in Section 5.

2   Reject Option with Class-related Thresholds

As previously told, Chow's reject rule provides the optimal trade-off between error and reject, only if the posterior probabilities of the data classes are exactly known. This fact can be illustrated by an example. Figure 1 shows a simple one-dimensional classification task with two data classes ω_1 and ω_2 characterized by Gaussian distributions.
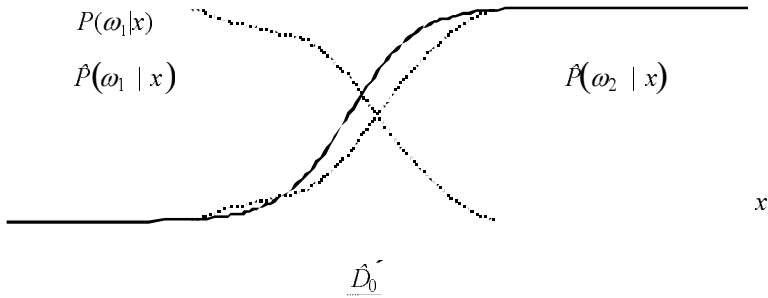
Fig. 1. A one-dimensional classification task with two data classes ω_1 and ω_2 characterized by Gaussian distributions. The application of Chow's rule with reject threshold T to the "true" and "estimated" a posteriori probabilities is shown.

The terms $P(\omega_i|x)$ and $\hat{P}(\omega_i \mid x)$, $i=1,2$, indicate the "true" and "estimated" a posteriori probabilities, respectively. We hypothesized that estimate errors are negligible when the two classes are "well separated", that is, when the difference between the two a posteriori probabilities is large. Differently, significant errors affect the estimated probabilities in the range of feature values where the two classes are "overlapped". Other researchers share this assumption, which is in agreement with real experiments [5]. The optimal decision and reject regions provided by Chow's rule applied to the true probabilities are indicated by the terms D_1, D_2 and D_0. The term T indicates the reject threshold used in Chow's rule. Analogously, the terms $\hat{D}_1$, $\hat{D}_2$, and $\hat{D}_0$ stand for the decision and reject regions provided by Chow's rule applied to the estimated probabilities. It is easy to see that Chow's rule applied to the estimated probabilities never provides the optimal decision and reject regions D_1, D_2 and D_0. No value of the threshold T allows to obtain these regions. Therefore, the example in Figure 1 points out that Chow's rule cannot provide the optimal error-reject trade-off when the a posteriori probabilities are affected by errors. The authors proved the general validity of such conclusion. For the sake of brevity, the reader interested in such proof is referred to [6].

However, a careful analysis of Figure 1 suggests a different approach from Chow's rule for obtaining the optimal error-reject trade-off, even if the a posteriori probabilities are affected by errors. First of all, we can observe that the estimated decision regions $\hat{D}_1$ and $\hat{D}_2$ differ from the optimal ones in the ranges $(\hat{D}_1 - D_1)$ and $(D_2 - \hat{D}_2)$. Accordingly, non-optimal decisions are taken within these ranges by Chow's rule applied to the estimated probabilities. In particular, the patterns belonging to the range $(\hat{D}_1 - D_1)$ are erroneously accepted, since the a posteriori probability $\hat{P}(\omega_1 \mid x)$ takes higher values than the true ones within this range. However, it is easy to see that such patterns would be correctly rejected using a threshold value T_1 higher than T. Analogously, the patterns belonging to the range $(D_2 - \hat{D}_2)$ are erroneously rejected, since the a posteriori probability $\hat{P}(\omega_2 \mid x)$ takes lower values than the true ones within this range. Such patterns would be correctly accepted using a threshold value T_2 lower than T.

The above analysis suggests the use of multiple reject thresholds to obtain the optimal error-reject trade-off, even if the a posteriori probabilities are affected by errors. In particular, different thresholds for the different data classes should be used. Figure 2 shows the use of two different reject thresholds T_1 and T_2 for the classification task described in Figure 1.
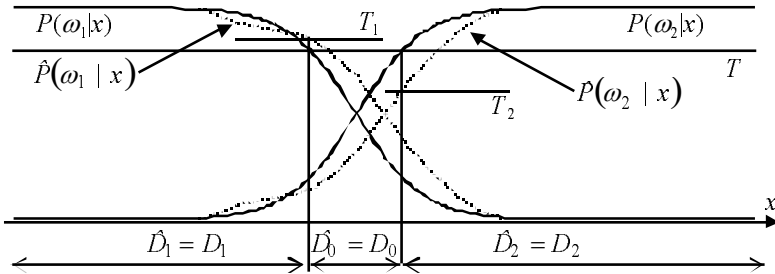

Fig. 2. Two different reject thresholds T_1 and T_2 are applied to the estimated class-posterior probabilities of the classification task in Figure 1. Such thresholds allow to obtain the optimal reject region corresponding to Chow's rule applied to the true class-posterior probabilities.

It is easy to see that such thresholds applied to the estimated probabilities allow to obtain the optimal reject region corresponding to the single-threshold Chow's rule applied to the true probabilities. It is worth remarking that Chow's rule applied to the estimated probabilities is not able to provide this optimal reject region. Therefore, under the assumption that the a posteriori probabilities are affected by errors, the use of multiple thresholds can provide a better error-reject trade-off than Chow's one.

The general validity of the above conclusion has been proved in [6]. In particular, under the assumption that the a posteriori probabilities are affected by significant errors, we have proved that, for any reject rate R, such values of the thresholds $T_1,...,T_N$ exist, that the corresponding classifier's accuracy $A(T_1,...,T_N)$ is equal or higher than the accuracy $A(T)$ provided by Chow's rule.

Therefore, we propose the following reject rule for a classification task with N data classes that are characterized by "estimated" posterior probabilities $\hat{P}(\omega_i \mid x)$, $i=1,...,N$. A pattern x is rejected if:

$$\max_{k=1,...,N} \hat{P}(\omega_k \mid x) = \hat{P}(\omega_i \mid x) < T_i \; , \tag{10}$$

while x is accepted and assigned to the class ω_i, if:

$$\max_{k=1,...,N} \hat{P}(\omega_k \mid x) = \hat{P}(\omega_i \mid x) \geq T_i \; . \tag{11}$$

The above thresholds $T_1,...,T_N$ are named "class-related reject thresholds" (CRTs), and take on values in the range [0,1]. Accordingly, the proposed rule is named CRT rule. It is worth noting that, analogously to Chow's rule, in real applications, the values of the CRTs have to be estimated according to the classification task at hand. In the next section, we describe the basic concepts of an algorithm devoted to estimate

such values. For the sake of brevity, we refer the reader interested in more details about this algorithm to [6].

3   An Algorithm for Estimating Class-related Reject Thresholds

In [6] the authors have proved that the following proposition is true:

$$\forall R \;\; \exists \, T_1, T_2, ..., T_N : \;\; A(T_1, T_2, ..., T_N) \geq A(T) \;\; . \tag{12}$$

Namely, we have proved that, for any given reject rate R, and the corresponding Chow's threshold T, values of the CRT thresholds exist such that the accuracy provided by the CRT rule is equal or higher than in Chow's rule. It is easy to see that such CRT values can be estimated by evaluating the maximum of the function $A(T_1,...,T_N)$ for a given reject rate R. Accordingly, the CRT values that satisfy equation 12 are estimated by solving the following maximization problem:

$$\begin{cases} \displaystyle \max_{T_1,...,T_N} A\big(T_1, ..., T_N\big) \\ R\big(T_1, ..., T_N\big) \leq R_{MAX} \end{cases} \tag{13}$$

It is worth noting that the inequality constraint in the above equation is aimed to take into account the error-reject requirements of real pattern recognition applications. The end-user of a pattern recognition system usually wishes to obtain the highest classification accuracy and a reject rate below a fixed threshold R_{MAX}.

According to the CRT rule, the accuracy and the reject probabilities $A(T_1,...,T_N)$ and $R(T_1,...,T_N)$ are functions of the CRTs. For given values of the CRTs, such probabilities can be estimated according to equations 7 and 9 using a validation set. Since the functions $A(T_1,...,T_N)$ and $R(T_1,...,T_N)$ are computed using a finite data set, they take on a finite number of values in the range $[0,1]$. Therefore, equation 13 corresponds to a constrained maximization problem, where the "target" and the "constraint" functions $A(T_1,...,T_N)$ and $R(T_1,...,T_N)$ are discrete-valued functions of continuous variables. Unfortunately, to the best of our knowledge, no algorithm reported in literature fits well the characteristics of the above maximization problem. Accordingly, we have developed a specially designed algorithm to solve it. First of all, our algorithm takes into account that $R(T_1,...,T_N)$ is an increasing function of the variables $T_1,...,T_N$, that is, the number of rejected patterns cannot decrease for increasing values of the CRTs. In addition, we assume that $A(T_1,...,T_N)$ is an increasing function of $T_1,...,T_N$. This assumption is often verified in the experiments. According to this assumption, the basic idea of our algorithm is to solve equation 13 iteratively, starting from CRT values that provide a reject rate equal to zero (i.e., $T_i \leq 1/N, \; i=1,..,N$), and varying such values in order to increase the function $A(T_1,...,T_N)$. At each step, each threshold T_i is increased according to the equation $T_i + k\Delta t$, where Δt is a positive constant, and k is an integer varying between 1 and k_{MAX}. Then the variations of accuracy ΔA and reject ΔR due to such changes are evaluated. The changes that provide the maximum positive value of $\Delta A/\Delta R$, and do not make to exceed the reject threshold R_{MAX}, are selected to generate the next CRT

values. The algorithm stops when it is not possible to increase $A(T_1,...,T_N)$ while keeping $R(T_1,...,T_N) \leq R_{MAX}$. It is worth noting that the proposed algorithm does not guarantee to find the optimal solution of equation 13. Nevertheless, experimental results reported in the next section show that it affords CRT values that provide a better error-reject trade-off than in Chow's rule.

4   Experimental Results

The data set used for our experiments consists of a set of multisensor remote-sensing images related to an agricultural area near the village of Feltwell (UK). We selected 10944 pixels belonging to five agricultural classes (i.e., sugar beets, stubble, bare soil, potatoes, carrots), and randomly subdivided them into a training set (5124 pixels) and a test set (5820 pixels). Each pixel was characterized by a fifteen-element feature vector containing the brightness values in the six optical bands, and over the nine radar channels considered. More details about the selected data set can be found in [7,8].

Two different classifiers have been used in our experiments: a k-nearest neighbors (k-nn) classifier and a multi-layer perceptron (MLP) neural network. For the k-nn classifier, a value of the "k" parameter of twenty-one was used. The MLP network had fifteen input units and five output units, as the numbers of input features and data classes, respectively. Fifteen hidden neurons were used.
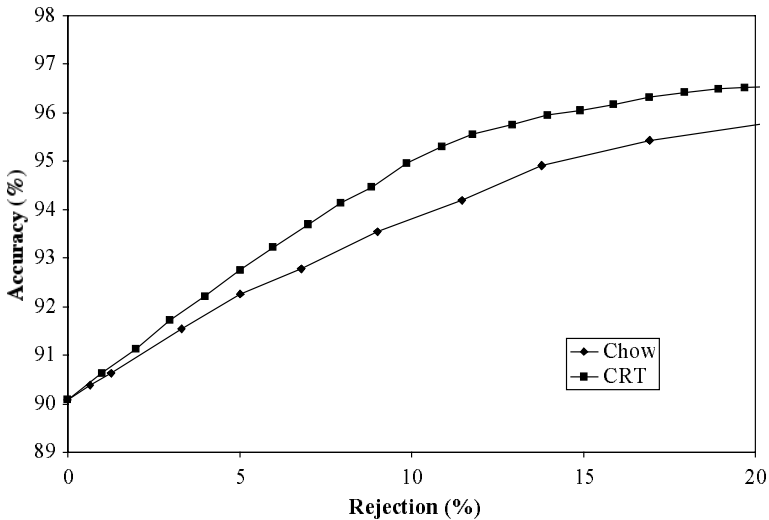

Fig. 3. The accuracy-rejection trade-offs of the k-nn classifier using the CRT and Chow's rules are represented on the A-R plane for values of the rejection rate ranging from 0% to 20%.
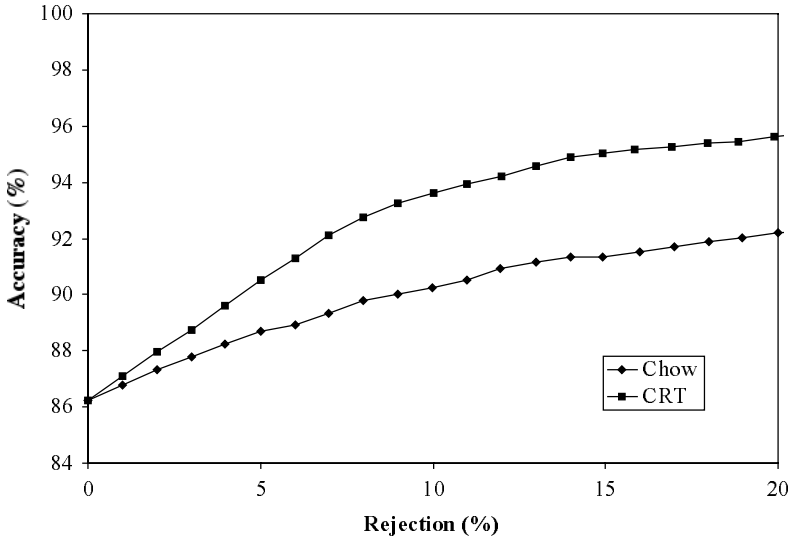
Fig. 4. The accuracy-rejection trade-offs of the MLP neural network using the CRT and Chow's rules are represented on the *A-R* plane for values of the rejection rate ranging from 0% to 20%.

According to the algorithm described in Section 2, test data were used to estimate the values of the CRTs and of Chow's reject threshold. Values of the Δt and k_{MAX} parameters equal to 0.001 and 200, respectively, were adopted. The CRT and Chow's rules were compared in the so-called accuracy-reject plane (*A-R* plane), introduced in [4]. In the *A-R* plane, the accuracy-reject trade-offs provided by a given reject rule are described by the curve $A(R)$ connecting the points that represent the accuracy values for different rejection rates. A range of reject rates from 0% to 20% was considered. This range is usually the most significant for application purposes.

Figure 3 shows the accuracy-reject trade-offs provided by the *k*-nn classifier using the CRT and Chow's rules. The results are related to the test set and they are shown in the *A-R* plane. It is worth noting that, for any value of reject rate, the accuracy provided by the CRT rule is higher that in Chow's rule. Accordingly, we can say that the CRT reject rule provides a better error-reject trade-off than in Chow's rule. Figure 4 shows the results related to the MLP neural network. It is easy to see that conclusions similar to the ones of the experiment with the *k*-nn classifier can be drawn.

5   Conclusions

In this paper, we addressed the problem of the optimality of Chow's reject rule when the a posteriori probabilities are affected by estimate errors. We showed that Chow's rule cannot provide the optimal error-reject trade-off if significant estimate errors are present. We then proposed the use of class-related reject thresholds. The authors have

proved in [6] that the related reject rule provides a better error-reject trade-off than in Chow's rule. Reported experimental results confirmed the proposed theory. Finally, it is worth noting that the use of class-related reject thresholds was previously proposed for different purposes by Yau and Manry [3]. They have shown that such multiple thresholds allow to equalize the error and reject probabilities for different data classes.

Acknowledgements

References

1. Chow, C.K.: On Optimum Error and Reject Trade-off. IEEE Transactions on Information Theory, Vol.It-16, 1 (1970) 41-46
2. Cordella, L.P., De Stefano, C., Tortorella, F., Vento, M.: A Method for Improving Classification Reliability of Multilayer Perceptrons. IEEE Transactions on Neural Networks, Vol. 6, 5 (1995) 1140-1147
3. Yau, H.C., Manry, M.T.: Automatic Determination of Reject Thresholds in Classifiers Employing Discriminant Functions. IEEE Transactions on Signal Processing, Vol. 40, 3 (1992) 711-713
4. Battiti R. and Colla A.M. Democracy in neural Nets: Voting Schemes for Classification. Neural Networks, 7, (1994) 691-707
5. Tumer, K. and J. Ghosh. Error correlation and error reduction in ensemble classifiers, Connection Science, 8, (1996) 385-404.
6. G.Fumera and F.Roli: Multiple Reject Thresholds for Improving Classification Reliability, Internal Report, Univ. of Cagliari, 1999
7. Roli F.: Multisensor image recognition by neural networks with understandable behaviour. International Journal of Pattern Recognition and Artificial Intelligence 10 (1996) 887-917.
8. Serpico, S.B., and Roli F.: Classification of multi-sensor remote-sensing images by structured neural networks, IEEE Trans. On Geoscience and Remote Sensing 33 (1995) 562-578.