PERGAMON

Pattern Recognition 33 (2000) 2099–2101

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

Reject option with multiple thresholds

Giorgio Fumera, Fabio Roli*, Giorgio Giacinto

Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09123, Cagliari, Italy

Received 24 February 2000; accepted 22 March 2000

1. Introduction

An N-class classifier 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 statistical pattern recognition theory, such decision regions are defined to maximise the probability of correct recognition, commonly named classifier's accuracy

$$Accuracy = P(correct) = \sum_{i=1}^{N} \int_{D_i} p(x \mid \omega_i) P(\omega_i)\, \mathrm{d}x \tag{1}$$

and, consequently, to minimise the classifier error probability

$$P(error) = \sum_{i=1}^{N} \int_{D_i} \sum_{\substack{j=1 \\ j \neq 1}}^{N} p(x \mid \omega_j) P(\omega_i)\, \mathrm{d}x. \tag{2}$$

To this end, the so-called Bayes decision rule assigns each pattern x to the class for which the a posteriori probability $P(\omega_i \mid x)$ is maximum. An error probability lower than the one provided by the above Bayes rule can be obtained using the so-called "reject" option. Namely, the patterns that are the most likely to be misclassified are rejected (i.e., they are not classified); they are then handled by more sophisticated procedures (e.g., a manual classification is performed). However, handling high reject rates is usually too time-consuming for application purposes. 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

$$\max_{k=1,\ldots,N} P(\omega_k \mid x) = P(\omega_i \mid x) < T, \tag{3}$$

where $T \in [0, 1]$. On the other hand, the pattern x is accepted and assigned to the class ω_i, if

$$\max_{k=1,\ \ldots,\ N} P(\omega_k \mid x) = P(\omega_i \mid x) \geqslant T. \tag{4}$$

The feature space is therefore subdivided into $N + 1$ regions. The reject region D_0 is defined according to Eq. (3), while the decision regions $D_1, \ \ldots, D_N$ are defined according to Eq. (4). It is easy to see that the probability of a pattern being rejected can be computed as follows:

$$P(reject) = \int_{D_0} p(x)\, \mathrm{d}x. \tag{5}$$

On the other hand, the classifier's accuracy is defined as the conditional probability that a pattern is correctly classified, given that it has been accepted:

$$Accuracy = P(correct \mid acceptd) = \frac{P(correct)}{P(correct) + P(err)}. \tag{6}$$

A careful analysis of Chow's work allows to point out that his reject rule provides the optimal error–reject trade-off, only if the a posteriori probabilities are exactly known. Unfortunately, in real applications, such probabilities are affected by significant estimate errors. However, to the best of our knowledge, no previous work has clearly investigated the effects of estimate errors on the optimality of Chow's rule. In addition, alternative reject rules described in the literature were not specifically designed to handle estimate errors [2]. In this paper, we investigate the effects of estimate errors on Chow's rule, and propose the use of multiple reject thresholds related to the data classes. The reported experimental results show that such class-related reject thresholds provide an error-reject trade-off better than the one in Chow's rule.

2. Reject option with class-related thresholds

As previously mentioned, Chow's reject rule provides the optimal error-reject trade-off, only if the posterior probabilities of the data classes are exactly known. This fact can be illustrated by an example. Fig. 1 shows a simple one-dimensional classification task with two

* Corresponding author. Tel.: + 39-070-675-5874; fax: + 39-070-675-5900.

E-mail addresses: fumera@diee.unica.it (G. Fumera), roli@diee.unica.it (F. Roli), giacinto@diee.unica.it (G. Giacinto).

PII: S0031-3203(00)00059-5

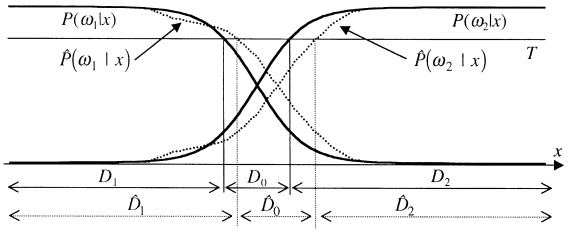

Fig. 1. Application of Chow's rule to the "true" and "estimated" a posteriori probabilities.

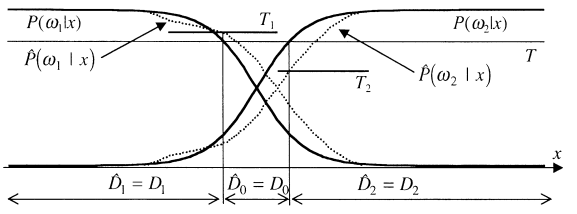

Fig. 2. Two different reject thresholds T_1 and T_2 applied to the estimated a posteriori probabilities of the classification task in Fig. 1.

data classes ω_1 and ω_2 characterised by Gaussian distributions. The terms $P(\omega_i \mid x)$ and $P(\omega_i \mid x)$, $i = 1, 2$, indicate the "true" and "estimated" a posteriori probabilities, respectively. We hypothesised that significant errors affect the estimated probabilities in the range of feature values in which the two classes are "overlapped". Other researchers share this assumption, which is in agreement with real experiments [3]. 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 Chow's reject threshold. 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 no threshold T value applied to the estimated probabilities can provide both the optimal decision regions and the optimal reject region. Therefore, this example points out that Chow's rule cannot provide the optimal error–reject trade-off, if the posterior probabilities are affected by errors. The authors have proved the general validity of this conclusion [4].

A careful analysis of Fig. 1 suggests a different approach from Chow's rule for obtaining the optimal error–reject trade-off, when the a posteriori probabilities are affected by errors. Fig. 1 shows that the estimated regions differ from the optimal ones in the ranges $(\hat{D}_1 - D_1)$ and $(D_2 - \hat{D}_2)$. In particular, Chow's rule erroneously accepts the patterns belonging to the range $(\hat{D}_1 - D_1)$, since the posterior probability $\hat{P}(\omega_1, \mid x)$ takes on 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 $T_1 > T$. Analogously, the patterns belonging to $(D_2 - \hat{D}_2)$ are erroneously rejected, since the posterior probability $\hat{P}(\omega_2 \mid x)$ takes on lower values than the true ones within this range. Such patterns would be correctly accepted using a threshold value $T_2 < T$.

The above analysis suggests the use of multiple reject thresholds for the different data classes to obtain the optimal decision and reject regions, even if the a posteriori probabilities are affected by errors. Fig. 2 shows the use of two different reject thresholds T_1 and T_2 for the classification task of Fig. 1. It is easy to see that such thresholds applied to the estimated probabilities allow to obtain both the optimal decision regions and reject region. Therefore, this example suggests that the use of N class-related reject thresholds (CRTs) can provide a better error–reject trade-off than Chow's. The general validity of this conclusion has been proved by the authors [4]. 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 CRTs $T_1, \ldots, T_N$ exist that the corresponding classifier's accuracy $A(T_1, \ldots, T_N)$ is equal or higher than the accuracy $A(T)$ provided by Chow's rule

$$\forall R \; \exists T_1, T_2, \ldots, T_N: \quad A(T_1, T_2, \ldots, T_N) \geqslant A(T). \tag{7}$$

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

$$\max_{k=1,\ldots,N} \hat{P}(\omega_k \mid x) = \hat{P}(\omega_i \mid x) < T_i, \tag{8}$$

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

$$\max_{k=1,\ldots,N} \hat{P}(\omega_k \mid x) = \hat{P}(\omega_i \mid x) \geqslant T_i. \tag{9}$$

The CRTs take on values in the range [0,1]. It is worth noting that, analogously to Chow's rule, in real applications, the values of the CRTs must be estimated according to the classification task at hand.

3. 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 and randomly subdivided them into a training set (5124 pixels) and a test set (5820 pixels). Each pixel was characterised by a feature vector containing the brightness values in the six optical bands, and over the nine radar channels considered. Further details on the selected data set can be found in Ref. [5]. In our experiments, we considered the usual error–reject requirement of real

pattern recognition applications, that of obtaining the highest accuracy and a reject rate below a given value R_{MAX}. Accordingly, the CRT values were estimated by solving the following constrained maximisation problem [4]:

$$\begin{cases} \max\limits_{T_1,\ \ldots,\ T_N} A(T_1, \ldots, T_N) \\ R(T_1,\ \ldots,\ T_N) \leqslant R_{MAX}. \end{cases} \quad (10)$$

It is worth noting that, according to Eq. (7), for any given R_{MAX}, the CRT values obtained as solutions of the above maximisation problem provide an accuracy equal or higher than in Chow's rule.

In real applications, the functions $R(T_1, \ldots, T_N)$ and $A(T_1, \ldots, T_N)$ can be estimated according to Eqs. (5) and (6) using a finite validation set. Therefore, they take on a finite number of values in the range [0,1], and Eq. (10) represents a constrained maximisation problem, whose "target" and "constraint" functions are discrete valued functions of continuous variables. To the best of our knowledge, no algorithm reported in the literature fits well the characteristics of this problem. Therefore, we have developed a specially designed algorithm to solve it [4]. Our algorithm takes into account that $R(T_1, \ldots, T_N)$ is an increasing function of the variables $T_1, \ldots, T_N$ (i.e., the number of rejected patterns cannot decrease for increasing values of the CRTs) and also assumes that $A(T_1, \ldots, T_N)$ is an increasing function of $T_1, \ldots, T_N$. (This assumption is often verified in the experiments). The basic idea is to solve Eq. (10) iteratively. We start from CRT values that provide a reject rate equal to zero, and at each step increase the value of one of the CRTs in order to increase accuracy until the reject rate exceeds the value R_{MAX}. It is worth noting that our algorithm does not guarantee an optimal solution to Eq. (10). Nevertheless, experimental results reported in the following show that it affords CRT values that provide a better error–reject trade-off than in Chow's rule.

In our experiments, we used two different classifiers: a K-nearest neighbours (K-NN) classifier, with a "K" parameter value of 21, and a multi-layer perceptron (MLP) neural network. We used a net architecture with 15 input units and five output units as the number of input features and data classes, respectively. The architecture also included fifteen hidden neurons. Test data were used to estimate the value of Chow's reject threshold and the values of the CRTs. We considered a range of reject rates from 0 to 20%, since this range is usually the most relevant for application purposes. Figs. 3 and 4 show a comparison of the two reject rules in the accuracy-reject plane. For any value of the reject rate, the CRT rule provides an accuracy higher than Chow's rule. Accordingly, we can say that with the proposed reject rule a better error–reject trade-off can be obtained.

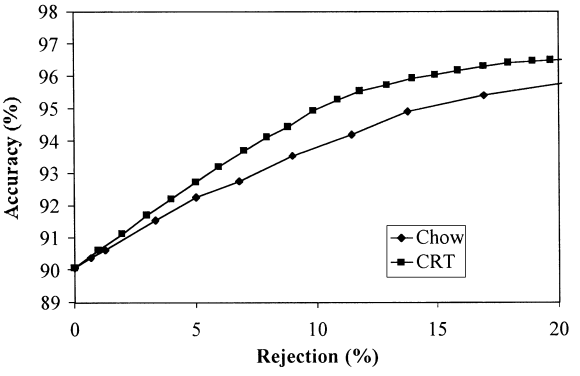

Fig. 3. Accuracy–rejection trade-offs of the k-nn classifier using the CRT and Chow's rules.

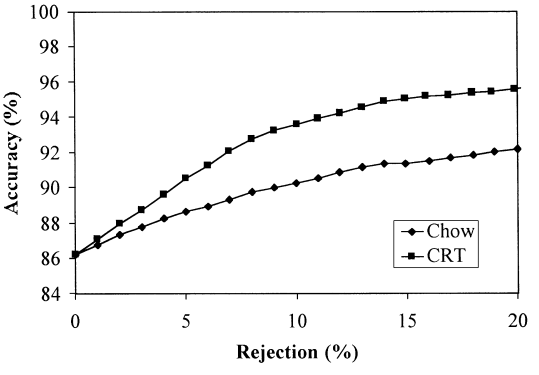

Fig. 4. Accuracy–rejection trade-offs of the MLP neural network using the CRT and Chow's rules.

References

[1] C.K. Chow, On optimum error and reject tradeoff, IEEE Trans. Inform. Theory IT-16 1 (1970) 41–46.

[2] L.P. Cordella, C. De Stefano, F. Tortorella, M. Vento, A method for improving classification reliability of multilayer perceptrons, IEEE Trans. Neutral Networks 5 (6) (1995) 1140–1147.

[3] K. Tumer, J. Ghosh, Error correlation and error reduction in ensemble classifiers, Connection Sci. 8 (1996) 385–404.

[4] G. Fumera, F. Roli, Multiple reject thresholds for improving classification reliability, Internal Report, University of Cagliari, 1999, pp. 1–13.

[5] F. Roli, Multisensor image recognition by neural networks with understandable behaviour, Int. J. Pattern Recognition Artif. Intell. 10 (1996) 887–917.