An Optimum Class-Rejective Decision Rule and Its Evaluation

Hoel Le Capitaine, Carl Frelicot

An optimum class-rejective decision rule and its evaluation

Hoel Le Capitaine and Carl Frélicot
Mathematics, Image and Applications Lab.,
University of La Rochelle, France
{*hoel.le_capitaine*},{*carl.frelicot*}@*univ-lr.fr*

***Abstract*—Decision-making systems intend to copy human reasoning which often consists in eliminating highly non probable situations (e.g. diseases, suspects) rather than selecting the most reliable ones. In this paper, we present the concept of class-rejective rules for pattern recognition. Contrary to usual reject option schemes where classes are selected when they may correspond to the true class of the input pattern, it allows to discard classes that can not be the true one. Optimality of the rule is proven and an upper-bound for the error probability is given. We also propose a criterion to evaluate such class-rejective rules. Classification results on artificial and real datasets are provided.**

***Keywords*-bayesian classification; decision rules; loss structure; reject option; risk minimization;**

I. Introduction and motivation

Let $\Omega = \{\omega_1, ..., \omega_c\}$ be a set of classes with prior probabilities $p(\omega_i)$ and assume that, given a pattern $\mathbf{x}$ in a feature space X, the conditional densities $f(\mathbf{x}|\omega_i)$ and the mixture density $f(\mathbf{x}) = \sum_{i=1}^{c} p(\omega_i)\, f(\mathbf{x}|\omega_i)$ can be calculated (or estimated). The probability that $\mathbf{x}$ belongs to class ω_i is given by

$$p(\omega_i|\mathbf{x}) = \frac{p(\omega_i)\, f(\mathbf{x}|\omega_i)}{f(\mathbf{x})}. \qquad (1)$$

In the sequel, we assume that we use a Bayesian classifier, and focus our study on posterior probabilities.

If $risk(\mathbf{x})$ denotes the risk of taking a wrong decision for $\mathbf{x}$, the so-called *Bayes decision rule*, which assigns $\mathbf{x}$ to the class of maximum posterior probability (1), minimizes the error probability defined by

$$e = \int_X risk(\mathbf{x})\; f(\mathbf{x})\; \mathrm{d}\mathbf{x}. \qquad (2)$$

However, when some classes are not known or when classes overlap in X, this rule may not be always efficient. Chow modified it so that one can reject a pattern if its maximum posterior probability is less than a user-defined threshold t [1]. Chow's rule minimizes (2) as a function of t which specifies the reject probability and allows to assign $\mathbf{x}$ to $n^\star \in \{1, c\}$ classes. Ha proposed a *class-selective* rule which minimizes (2) for an average number of selected classes $\overline{n}$ [2]. It allows to assign $\mathbf{x}$ to $n^\star \in [1, c]$ classes. In recent years, research on reject option has focused on binary classification, see *e.g.* [3], [4], neglecting the multi-class case, with the notable exception of [5], where ad-hoc transformations are operated on both distance and density base classifiers. In this paper, the multi-class case is considered without requiring a specific transformation on output models.

We are interested in applications, *e.g.* diagnosis of diseases or police investigation, where it can be more efficient to first discard less reliable cases, *e.g.* diseases or suspects, instead of selecting most reliable ones. To this aim, we introduce the concept of *class-rejective* rules: classes are eliminated and $\mathbf{x}$ is assigned to the remaining ones. In order to obtain good performances with respect to the costs associated with specific actions, one may reject as much classes as possible, while preserving a low error rate. Moreover, such a rule is able to reject all the c classes if none of them corresponds to $\mathbf{x}$, resulting to an usual *distance reject* scheme [6].

II. Decision rule and optimality

Let $\alpha_j = \alpha_j(\mathbf{x})$ be the action of selecting, for $\mathbf{x}$, a subset S_j of classes among the 2^c possible ones, and $L(\alpha_j|\omega_i)$ the loss incurred by action α_j when the true class of $\mathbf{x}$ is ω_i. The overall risk R to be minimized with respect to actions α_j is the average expected loss

$$R = \int_X R(\alpha_j|\mathbf{x})\; f(\mathbf{x})\; \mathrm{d}\mathbf{x} \qquad (3)$$

where $R(\alpha_j|\mathbf{x})$ is the conditional risk defined by

$$R(\alpha_j|\mathbf{x}) = \sum_{i=1}^{c} L\big(\alpha_j(\mathbf{x})|\omega_i\big)\; p(\omega_i|\mathbf{x}). \qquad (4)$$

We define a loss structure modeling the specifical needs of the class-rejection problem. Let R_j be the subset of r_j rejected classes and set $L(\alpha_j|\omega_i) = L_e(\alpha_j|\omega_i) + L_r(\alpha_j)$ where $L_e(\alpha_j|\omega_i)$ and $L_r(\alpha_j)$ denote respectively the loss of eliminating the true class ω_i and the loss of having eliminated a few number of classes, when taking action α_j. We propose to define both losses as:

$$L_e(\alpha_j|\omega_i) = \begin{cases} C_e & \text{if } \omega_i \in R_j \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

$$L_r(\alpha_j) = C_r(c - r_j) \qquad (6)$$

where C_e and C_r are user-defined positive costs depending on the application but not on the decision rule. Thus, (4) can be rewritten as

$$R(\alpha_j|\mathbf{x}) = C_e \sum_{\omega_i \in R_j} p(\omega_i|\mathbf{x}) + C_r(c - r_j). \qquad (7)$$

Since $p(\mathbf{x}) > 0$ for all $\mathbf{x}$, minimizing (3) is equivalent to minimizing (4) or (7) over all possible numbers of rejected classes $r \in [0, c]$ and we have to solve

$$\min_{r\in[0,c]} \left\{ \min_{r_j=r} R(\alpha_j|\mathbf{x}) \right\}. \qquad (8)$$

For a fixed r, it is obvious that the solution of (8) is such that R_j contains the r classes of lowest posterior probabilities. Let us use the decreasing ordered sequence of posterior probabilities such as $p(\omega_{(i)}|\mathbf{x}) \geq p(\omega_{(i+1)}|\mathbf{x})$, $\forall i = 1, c$. Then, the conditional risk (7) can be defined as a function of the r worst classes by

$$R(r|\mathbf{x}) = C_e \sum_{i=c-r+1}^{c} p(\omega_{(i)}|\mathbf{x}) + C_r\big(c - r\big). \qquad (9)$$

Setting $t = C_r/C_e$ and simplifying by Ce, (8) becomes

$$\min_{r\in[0,c]} \left\{ R(r|\mathbf{x}) \right\} = \min_{r\in[0,c]} \left\{ \sum_{i=c-r+1}^{c} p(\omega_{(i)}|\mathbf{x}) + t\,(c-r) \right\}. \qquad (10)$$

The first term in the right-hand side of (10) is (not strictly) convex because it sums increasing values and the second term $t(c - r)$ is a (not strictly) convex function, so $R(r|\mathbf{x})$ is convex. Therefore, if there is a unique solution $r^\star$, it satisfies:

$$\begin{aligned} R(r^\star - 1|\mathbf{x}) &> R(r^\star|\mathbf{x}) & (11)\\ R(r^\star + 1|\mathbf{x}) &> R(r^\star|\mathbf{x}). & (12) \end{aligned}$$

Using (10), it results in $p(\omega_{(c-r^\star+1)}|\mathbf{x}) < t$ and $p(\omega_{(c-r^\star)}|\mathbf{x}) > t$. Since $p(\omega_{(i)}|\mathbf{x})$'s are ordered, it ensues the optimal number of rejected classes:

$$r^\star(\mathbf{x}, t) = \max_{r\in[0,c]} \left\{ r : p(\omega_{(c-r+1)}|\mathbf{x}) < t \right\} \qquad (13)$$

or, equivalently,

$$r^\star(\mathbf{x}, t) = \min_{r\in[0,c]} \left\{ r : p(\omega_{(c-r)}|\mathbf{x}) > t \right\} \qquad (14)$$

with $p(\omega_{(c+1)}|\mathbf{x}) = 0$ and $p(\omega_{(0)}|\mathbf{x}) = 1$ by convention.

What about several solutions for (10)? Since $R(r|\mathbf{x})$ is convex, multiple solutions are necessary consecutive and taking the highest (respectively the lowest) only change '$>$' to '$\geq$' in (11) (respectively (12)) and '$<$' to '$\leq$' in (13) (respectively '$>$' to '$\geq$' in (14)).

With this formulation, the risk associated to the optimal number of rejected classes $r^\star$ can be written as

$$risk(\mathbf{x}, r^\star) = \sum_{i=c-r^\star+1}^{c} p(\omega_{(i)}|\mathbf{x}) \qquad (15)$$

Special values of the solution can be emphasized:

- $r^\star = 0$: $\mathbf{x}$ is associated to all classes as it would be using Chow's rule [1],
- $r^\star = c - 1$: $\mathbf{x}$ is rejected from all classes except one, so the rule reduces to the Bayes decision rule and the risk (15) is the Bayes one $risk(\mathbf{x}) = 1 - p(\omega_{(1)}|\mathbf{x})$,
- $r^\star = c$: $\mathbf{x}$ is rejected from all (known) classes, *i.e.* it is distance rejected in the sense of [6].

As a final proposition, let us give an upper-bound of the induced probability error $e(t)$. From (13) and the ordering of $p(\omega_{(i)}|\mathbf{x})$'s, we have

$$t \geq p(\omega_{(c-r^\star+1)}|\mathbf{x}) \geq ... \geq p(\omega_{(c)}|\mathbf{x}). \qquad (16)$$

Therefore,

$$\begin{aligned} \sum_{i=c-r^\star+1}^{c} t &\geq p(\omega_{(c-r^\star+1)}|\mathbf{x}) + ... + p(\omega_{(c)}|\mathbf{x}) \\ \Leftrightarrow \quad r^\star \times t &\geq risk(\mathbf{x}, r^\star). & (17) \end{aligned}$$

Integrating both sides, we finally get

$$\begin{aligned} t \int_X r^\star(\mathbf{x}, t)\; f(\mathbf{x})\;\mathrm{d}\mathbf{x} &\geq \int_X risk(\mathbf{x}, r^\star)\; f(\mathbf{x})\;\mathrm{d}\mathbf{x} \\ \Leftrightarrow \quad t \times \overline{r^\star}(t) &\geq e(t) & (18) \end{aligned}$$

where $\overline{r^\star}$ is the average number of rejected classes.

III. Evaluation of class-rejecting rules

In order to evaluate a decision rule whose parameters are Θ, one generally use: the correct classification $C(\Theta)$, the misclassification (or error) $E(\Theta)$ and the reject $R(\Theta)$ probabilities or rates. Chow introduced the error-reject (ER) trade-off (both should be minimized, see also [7]) and proposed to analyze the so-called $(E, R)-$curve ($E(\Theta)$ $vs.$ $R(\Theta)$) for all possible values of $\Theta = t$ to find the optimal or an operational value for t [1]. When comparing rules, the least the AUC (*Area Under Curve*) is, the better the rule, see [8] for its description in ROC analysis. For class-selective rules, *e.g.* Ha's rule [2], an object is considered as misclassified if the class it is issued from does not belong to the subset of selected ones. Then, the (E, R) trade-off is replaced by the $(E, \overline{n})$ trade-off, where $\overline{n}$ is the average number of selected classes, and the $AUC(E, \overline{n})$ can be used to compare rules [9].

For the new kind of class-rejective rules we introduce, we set that an object is misclassified if the class it is issued from belongs to the subset of rejected classes, and we introduce the $(E, \overline{r})$ trade-off which consists in jointly minimizing $E(\Theta)$ and maximizing the average number of rejected classes $\overline{r}(\Theta)$. Given such rules, the best one is the rule which has the least $AUC(E,\overline{r})$:

$$\min_{\substack{\text{available}\\ \text{rules}}} AUC(E,\overline{r}) = \min_{\substack{\text{available}\\ \text{rules}}} \int_D E\big(\overline{r^{\star}(t)}\big)\,\mathrm{d}t \tag{19}$$

where D is the definition domain for t. For a single Bayesian classifier, we compare the $(E,\overline{r})$ curves, which are the counterparts of Chow's error-reject trade off curves, obtained using the available rules.

IV. Experimental results

With respect to $AUC(E,\overline{r})$, we compare the new *Optimal Class-Rejective* (OCR) rule to what we call the *Bottom-r Ranking* (BrR) rule. By analogy to the *top-n ranking* rule used by Ha [2], it simply consists in setting a constant number of rejected classes r for the whole data set, *i.e.* posterior probabilities are not considered. The error rate is then computed by considering that the r lowest probabilities correspond to the r rejected classes.

Both rules are tested on fourteen datasets whose characteristics (number n of patterns, number p of features, number c of classes, degree of overlap) are reported in Table I. Eleven datasets are from the UCI ML-Repository [10], three are the following synthetic ones:

- DH consisting of two overlapping Gaussian classes with different covariance matrices according to the Highleyman distribution, each composed of 800 observations in $\mathbb{R}^2$, see [11].
- $D1$ containing 2000 points drawn from two Gaussian seven-dimensional distributions of 1000 points each with means $\mathbf{v}_1 = {}^t(1,0,...,0)$ and $\mathbf{v}_2 = {}^t(-1,0,...,0)$, and equal covariance matrices $\Sigma_1 = \Sigma_2 = I$.
- $D2$ which contains 4000 points drawn from four Gaussian bi-dimensional distributions of 1000 points each. Their respective means are $\mathbf{v}_1 = {}^t(1,1)$, $\mathbf{v}_2 = {}^t(1,-1)$, $\mathbf{v}_3 = {}^t(-1,1)$ and $\mathbf{v}_4 = {}^t(-1,-1)$, and their covariance matrices are all equal $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = I$.

Classes composing each real dataset are also assumed to be Gaussian. Note that this does not limitate the OCR rule which only need posterior probabilities to be computed using (1) whatever the class-conditional density model.

The threshold t is varied within the range $[0,1]$ for the OCR rule and the number of rejected classes is discretely increased from 0 to $c-1$ (to skip the distance reject case) for the BrR rule. For both rules, error rates are computed through a resubstitution procedure, and the performance is given in Table II where best scores are indicated in bold.

Table I
Description of the tested datasets.

Dataset	n	p	c	Overlap
DH	1600	2	2	slight
$D1$	2000	7	2	moderate
$D2$	4000	2	4	strong
Ionosphere	351	34	2	strong
Forest	495411	10	2	strong
Vowel	528	10	11	moderate
Digits	10992	16	10	slight
Thyroid	215	5	3	slight
Iris	150	4	3	slight
Pima	768	9	2	strong
Statlog	6435	36	6	moderate
Glass	214	9	6	strong
Yeast	1484	8	10	strong
Page-blocks	5473	10	5	moderate

Table II
$AUC(E,\overline{r})$ for the Optimal Class-Rejective and the Bottom-r Ranking rules.

Dataset	OCR Rule	B-rR Rule
DH	**0.0053**	0.0262
$D1$	**0.0406**	0.0727
$D2$	**0.1794**	0.2373
Ionosphere	**0.1644**	0.3205
Forest	**0.1309**	0.1577
Vowel	**0.0374**	0.0473
Digits	**0.0098**	0.0319
Thyroid	**0.0064**	0.0302
Iris	**0.0092**	0.0200
Pima	**0.1146**	0.1628
Statlog	**0.0959**	0.1490
Glass	**0.1133**	0.1472
Yeast	**0.7572**	0.7716
Page-blocks	**0.0906**	0.0987

These results show that the proposed OCR rule always outperforms the BrR rule. Looking at the amount of overlap between classes, one can go deeper into the analysis of the relative scores considering their ratio $AUC(BrR)/AUC(OCR)$ and their difference $AUC(BrR) - AUC(OCR)$. The ratios are much more important for datasets presenting a slight overlap of classes $\{DH$,Digits,Thyroid,Iris$\}$ than for datasets presenting a strong overlap $\{D2$,Ionosphere,Forest,Pima,Glass,Yeast$\}$. The reason is that for better separated classes, there are more patterns for which the maximum posterior probability is much higher than the others, so the associated risk (and implicitly the error) is lower for the OCR rule whereas this is not taken into account for the BrR rule.

While the ratios are lower for datasets presenting a strong overlap, the differences are much more important (except for Yeast) than for datasets presenting better separated classes. This means that the less the classes overlap, the less the benefit is, as it is for any decision rule including a reject option, *e.g.* [1], [6], [2].

Figure 1 shows the $(E,\overline{r})$ curves obtained on the three artificial datasets. One can see that for every average number of rejected classes, the error rate obtained with the proposed OCR rule is lower than the one obtained with the BrR rule. Obviously, when the number of rejected classes is equal to $c-1$ (respectively 1,1 and 3 for $D1$, $D2$ and DH), the error rate is the same for both rules as one could expect. As pointed out in the previous section, this case corresponds to the Bayes decision rule, so that both rules are the same one.

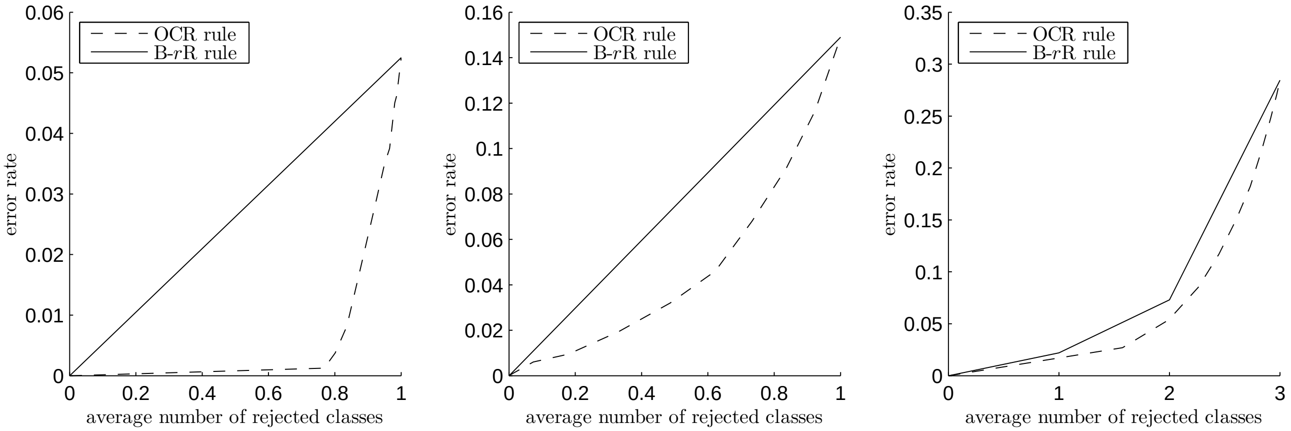

Figure 1. $(E, \overline{r})$ curves of both OCR and BrR rules curve obtained on the three artificial datasets (from left to right: DH, $D1$ and $D2$).

V. Conclusion

In this paper, we propose the new concept of *class-rejective decision rule* for pattern recognition. Instead of selecting the most probable classes, it allows to discard (reject) the less probable ones. Its optimality with respect to the error probability for an average number of rejected classes is proven as well as an upper-bound to the error probability. A measure to evaluate such rules is introduced, similarly to the one used to evaluate class-selective decision rules. Classification performance obtained on artificial and real datasets show that it outperforms the rule which consists in rejecting a constant number of classes. The proposed rule can be used with any classifier, provided posterior probabilities or equivalent values (*e.g.* membership degrees from a fuzzy classifier) are available.

Future research will concern the study of a new mixed *class-selective-rejective* decision rule which should jointly optimize the number of selected and rejected classes over the user-defined threshold definition domain. We also plan to use this optimum decision rule for outliers detection, *i.e.* patterns that do not match any of the known classes so that they must be distance rejected. This problem, as well as its variant for support vector machines, will be addressed using hinge loss minimization [12]. It will be studied by taking into account new results on AUC variants [13].

Acknowledgments

Support for part of this work was provided by the French ministry of research through a PhD grant.

References

[1] C. Chow, “An optimum character recognition system using decision functions,” *IRE Transactions on Electronic Computers*, vol. 6, no. 4, pp. 247–254, 1957.

[2] T. Ha, “The optimum class-selective rejection rule,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[3] M. Yuan and B. Wegkamp, “Classification methods with reject option based on convex risk minimization,” *Journal of Machine Learning Research*, vol. 11, pp. 111–130, 2010.

[4] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu, “Support vector machines with a reject option,” in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. NIPS, 2008, pp. 537–544.

[5] D. Tax and R. Duin, “Growing a multi-class classifier with a reject option,” *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1565–1570, 2008.

[6] B. Dubuisson and M. Masson, “A statistical decision rule with incomplete knowledge about classes,” *Pattern Recognition*, vol. 26, pp. 155–165, 1993.

[7] L. K. Hansen, C. Liisberg, and P. Salamon, “The error-reject tradeoff,” *Open Systems & Information Dynamics*, vol. 4, no. 2, pp. 159–184, 1997.

[8] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[9] H. Le Capitaine, “Aggregation operators for similarity measures. application to ambiguity in pattern recognition,” Ph.D. dissertation, Univ. of La Rochelle, 2009.

[10] A. Asuncion and D. Newman, “UCI machine learning repository,” http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

[11] W. Highleyman, “Linear decision functions, with application to pattern recognition,” *Proceedings of the IRE*, vol. 50, no. 6, pp. 1501–1514, 1962.

[12] P. Bartlett and M. Wegkamp, “Classification with a reject option using a hinge loss,” *Journal of Machine Learning Research*, vol. 9, pp. 1823–1840, 2008.

[13] S. Vanderlooy and E. Hüllermeier, “A critical analysis of variants of the auc,” *Machine Learning*, vol. 72, no. 3, pp. 247–262, 2008.