Shaping the Error-Reject Curve of Error Correcting Output Coding Systems

Paolo Simeone, Claudio Marrocco, and Francesco Tortorella

DAEIMI, Università degli Studi di Cassino,
Via G. Di Biasio 43, 03043 Cassino, Italy
{paolo.simeone,c.marrocco,tortorella}@unicas.it

Abstract. A common approach in many classification tasks consists in reducing the costs by turning as many errors as possible into rejects. This can be accomplished by introducing a reject rule which, working on the reliability of the decision, aims at increasing the performance of the classification system. When facing multiclass classification, Error Correcting Output Coding is a diffused and successful technique to implement a system by decomposing the original problem into a set of two class problems. The novelty in this paper is to consider different levels where the reject can be applied in the ECOC systems. A study for the behavior of such rules in terms of Error-Reject curves is also proposed and tested on several benchmark datasets.

Keywords: Error-Reject Curve, reject option, multiclass problem, Error Correcting Output Coding.

1 Introduction

The reduction of misclassification errors is a key point in Pattern Recognition. Such errors, in fact, can have a heavy impact on the applications accomplished by a classification system and can lead to serious consequences. Typically, error costs are defined and are helpful in defining which kind of error is convenient to avoid. However, those costs can be so high that the best choice could be an abstention from the decision so as to demand the last decision to a further and more efficient test. Since even the decision to abstain brings along some costs (e.g. the intervention of a human expert), the best approach is to find the optimal trade off between the numbers of errors and rejects.

The reject option in a classification system was introduced by Chow in [4] which demonstrated how the optimality could be reached when the prior probabilities and the conditional densities for each class were known. Since then, many approaches have been tried to introduce a reject rule for tuning the performance of a classification scheme. For neural networks a criteria to evaluate the reliability of the decision can be fixed as shown in [5,11]. Similar approaches were followed for Support Vector Machines as proposed in [12,3]. Thus, depending on the implementation of the system a criterion has to be found to evaluate the reliability of the decision and to fix a threshold for the application of the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 118–127, 2011.

reject rule. A profitable instrument to assess the performance of the reject rule independently from the classification costs is the Error-Reject curve that plots the percentage of errors versus rejects for each decision threshold.

Our work focuses on the application of a rejection rule in a system which face a multiple class problem through a pool of dichotomizers arranged according to an Error Correcting Output Coding (ECOC). Such technique was introduced by [6] to split a multiclass problem in many binary subproblems and has proved to be an efficient way to increase the performance attainable by a single monolithic classifier able to produce multiple outputs. The rationale lies on the capability of the code to correct errors and on the stronger theoretical roots and the better comprehension which characterize popular dichotomizers like Decision Trees or Support Vector Machines. Moreover, ECOC systems have been currently used as a starting point to extend boosting techniques to multiclass problems [7].

An analysis of reject for ECOC systems has been analyzed in our previous paper [10] where a reject rule evaluating the reliability of the final decision was proposed. In this paper to improve the performance of the classification system two consecutive thresholds are applied to focus more on the reliability of each level of the system. Each classifier, in fact, has an *internal* decision level where a first reject can be applied; then, in the decoding stage, an *external* reject threshold can be fixed uniquely based on the observation of the output of the ensemble of classifiers. Meanwhile, since each internal threshold induces an Error-Reject curve plotted according to an external threshold, we also show how to obtain a proper description of the system from this range of curves by using the Error-Reject curve given by their convex hull.

The paper is organized as follows: in section 2 we briefly analyze ECOC framework while in section 3 we introduce the two proposed reject rules. An extended analysis on how such rule modifies the Error-Reject curve is done in 4. Experimental results on many benchmarks data are reported in section 5 while the final section 6 presents some conclusions and some possible future developments.

2 The ECOC Classification System

Several multiclass classification systems use a decomposition of the original problem in many binary subproblems. Among them ECOC has been proved to be one of the more efficient and flexible to the application needs. Each original class ω_i with $i = 1, \ldots, n$ is associated to a *codeword* of length L. The collection of these codewords in a matrix, as shown in table 1, represents a coding matrix $\mathbf{C} = \{c_{hk}\}$ where $c_{hk} \in \{-1, +1\}$. Such matrix maps the original multiple class classification task in n different binary tasks defined by the matrix columns. Binary classifiers can be trained on each of these new binary data sets.

The classification is then performed by feeding each sample $\mathbf{x}$ to all the dichotomizers and collecting their outputs in a vector $\mathbf{o}$ (*output vector*) that is compared with the original coding matrix words. Several decoding rules have been proposed in literature and it has been largely proved that a *loss decoding rule* is the most sensitive and outperforming one [1]. Such rule takes into account

Table 1. Example of a coding matrix of length 15 for a 4 classes problem

classes	codewords						
ω_1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$
ω_2	-1	-1	-1	-1	$+1$	$+1$	$+1$
ω_3	-1	-1	$+1$	$+1$	-1	-1	$+1$
ω_4	-1	$+1$	-1	$+1$	-1	$+1$	-1

the reliability of the decision evaluating the loss function on the margin of the classifier. For ECOC the margins related to a particular codeword $\mathbf{c_i}$ are given by $c_{ih} f_h(\mathbf{x})$ with $h = 1, \ldots, L$. If we know the original loss function $\mathcal{L}(\cdot)$ of the employed dichotomizers, a global *loss-based distance* can be evaluated as:

$$D_{\mathcal{L}}(\mathbf{c}_i, \mathbf{f}) = \sum_{h=1}^{L} \mathcal{L}(c_{ih} f_h(\mathbf{x})). \tag{1}$$

and thus, the following rule can be defined to predict the k-th class:

$$\omega_k = \arg\min_i D_{\mathcal{L}}(\mathbf{c}_i, \mathbf{f}). \tag{2}$$

3 Two Levels of Rejection for ECOC System

In the description of ECOC technique, it is possible to observe that there are two different levels where a reliability parameter can be evaluated and thus a reject applied. The first simple option is at the output of the classification system where a threshold can be externally set on the output value without any assumption neither on the dichotomizers nor on the coding matrix. We already analyzed this approach [10] by applying a reject option for a loss decoding technique which proved to work sensibly better than other traditional decoding techniques, like ones based on Hamming distance. If we assume a loss value normalized in the range $[0, 1]$, such a criterion (indicated as *Loss Decoding*) can be formalized as:

$$r(\mathbf{f}, t_l) = \begin{cases} \omega_k & \text{if } D_{\mathcal{L}}(\mathbf{c}_k, \mathbf{f}) < t_l, \\ reject & \text{if } D_{\mathcal{L}}(\mathbf{c}_k, \mathbf{f}) \geq t_l. \end{cases} \tag{3}$$

where ω_k is the class chosen according to eq. (2) and $t_l \in [0, 1]$.

A second level of decision can be, instead, localized for each single dichotomizer before grouping the outcomes in the output vector. In fact, each dichotomizer outcome $f_h(\mathbf{x})$ is tipically compared with a threshold τ_h to decide to which of the two classes the sample belongs. This means that $\mathbf{x}$ is assigned to class $+1$ if $f_h(\mathbf{x}) \geq \tau_h$ and to class -1 otherwise.

Independently from the choice of each threshold the most unreliable outcomes will be on its proximity and thus, it could be convenient to reject those samples on each dichotomizer. This can be accomplished by choosing two different thresholds

τ_{h1} and τ_{h2} (with $\tau_{h1} \leq \tau_{h2}$) and defining a reject rule on each binary classifier as:

$$r(f_h, \tau_{h1}, \tau_{h2}) = \begin{cases} +1 & \text{if } f_h(\mathbf{x}) > \tau_{h2}, \\ -1 & \text{if } f_h(\mathbf{x}) < \tau_{h1}, \\ reject & \text{if } f_h(\mathbf{x}) \in [\tau_{h1}, \tau_{h2}]. \end{cases} \tag{4}$$

It is worth remarking that the choice of the thresholds should be made so as to encapsulate the class overlap region into the *reject interval* $[\tau_{h1}, \tau_{h2}]$ and turn most of the errors into rejects. However, to avoid the bias that can occur because of the high differences between each classifier outcomes, instead of choosing the same pair of thresholds for all the dichotomizers, we chose to let all dichotomizers work at the same level of reliability by fixing a common rejection rate ρ. Accordingly, the ROC curve of each dichotomizer has been used to evaluate the pair of thresholds (τ_{h1}, τ_{h2}) such that f_h abstains for no more than ρ samples at the lowest possible error rate [9].

Therefore, considering that the value produced by the classifier in the case of a reject is assumed to be 0, we can apply a reject rule defined as:

$$f_h^{(\rho)}(\mathbf{x}, \tau_{h1}, \tau_{h2}) = \begin{cases} 0 & \text{if } f_h(\mathbf{x}) \in [\tau_{h1}, \tau_{h2}] \\ f_h(\mathbf{x}) & \text{otherwise} \end{cases}. \tag{5}$$

It is worth noting that the null value is a possible outcome also for the dichotomizer without the reject option. It corresponds to the particular case when the sample falls on the decision boundary and thus it is assigned neither to the positive nor to the negative label. In this case, the loss calculated on the margin is $\mathcal{L}(c_{ih} f(\mathbf{x})) = \mathcal{L}(0)$, whichever the value of c_{ih}, while it is higher or lower than the "don't care" loss value $\mathcal{L}(0)$ if $f_h(\mathbf{x}) \neq 0$ (depending on whether $c_{ih} f(\mathbf{x})$ is positive or negative). The reject option actually extends such behavior to all the samples whose outcome $f_h(\mathbf{x})$ falls within the reject interval $[\tau_{h1}, \tau_{h2}]$. In this way, a part of the values assumed by the loss is not considered in the final decision procedure, which we indicate as *Trimmed Loss Decoding*.

Loss distance is now modified by the presence of the zero values in the output word $\mathbf{f}^{(\rho)}$ and it is given by:

$$D_{\mathcal{L}}(\mathbf{c}_i, \mathbf{f}^{(\rho)}) = \sum_{h \in I_{nz}} \mathcal{L}(c_{ih} f_h(\mathbf{x})) + |I_z| \cdot \mathcal{L}(0) \tag{6}$$

where I_{nz} and I_z are the sets of indexes of the nonzero values and zero values in the output word, respectively. In practice the loss is given by two contributions, where the second one is independent from the codeword that is compared to the output word. Through this new loss distance, a rejection rule can be immediately applied at the output of the ECOC system by choosing a threshold value t_l:

$$r(\mathbf{f}^{(\rho)}, t_l) = \begin{cases} \omega_k & \text{if } D_{\mathcal{L}}(\mathbf{c}_k, \mathbf{f}^{(\rho)}) < t_l, \\ reject & \text{if } D_{\mathcal{L}}(\mathbf{c}_k, \mathbf{f}^{(\rho)}) \geq t_l. \end{cases} \tag{7}$$

where ω_k is the class chosen according to eq. (2).

4 How to Evaluate the Reject Rule

Generally speaking, a reject option is accomplished by evaluating the reliability of the decision taken by the classifier and rejecting such decision if it is lower than some given threshold. A complete description of the classification system with the reject option is given by the *Error-Reject (ER) curve* which plots the error rate $E(t)$ against the reject rate $R(t)$ when varying the threshold t on the reliability estimate. In the ECOC approach when the threshold is applied at the output of the classification system as in eq. (3), it is very simple to build the error-reject curve by varying the threshold t_l and observing the errors and rejects obtained (see fig. 1).

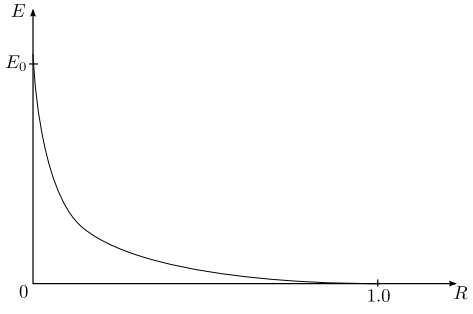

Fig. 1. A typical Error-Reject curve for an external reject rule. E_0 is the error at 0-reject while the error rate becomes null for a reject rate equal to 1.0.

Some remarks have, instead, to be done on the resulting decision rule which depends on two different thresholds (ρ and t_l) while the previously described rules depend only on one parameter. In the previous cases the reject option generates a unique curve where each point is function of only one threshold, while now we have a family of ER curves, each produced by a particular value of ρ. There is thus some ambiguity in defining the ER-curve representative of the performance of the whole system. To solve this problem, let us consider two ER-curves corresponding to two different values ρ_1 and ρ_2 of the internal reject threshold. They can be arranged into two different ways: one of the curve can be completely below the other one (see fig. 2.a) or they can intersect (see fig. 2.b). In the first case, the lower curve (and the corresponding ρ) must be preferred because it achieves a better error rate at the same reject rate (and vice versa). The second case shows different regions in which one of the curves is better than the other one and thus, there is not a curve (and an internal reject threshold value) definitely optimal. Therefore, to obtain an optimal ECOC system under all circumstances, the ER-curve should include the locally optimal parts of the two curves. This is obtained if we assume as ER-curve of the ECOC system the convex hull of the two curves (see fig. 2.c).

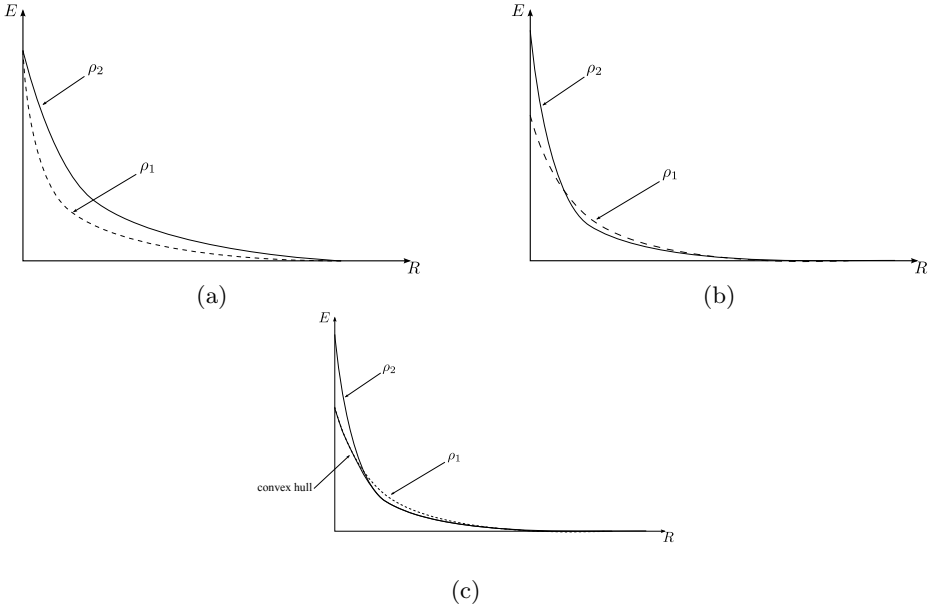

Fig. 2. Different cases for two ER-curves produced by two rejection rates ρ_1 and ρ_2 when varying the external threshold t_l. (a) The ER-curve produced by ρ_1 dominates the curve produced by ρ_2. (b) There is no dominating ER-curve. (c) The convex hull of the ER-curves shown in (b) including the locally optimal parts of the two curves.

This can be easily extended to the curves related to all the values considered for ρ so as to assume as the ER-curve of the ECOC system the convex hull of all the curves (see fig. 3).

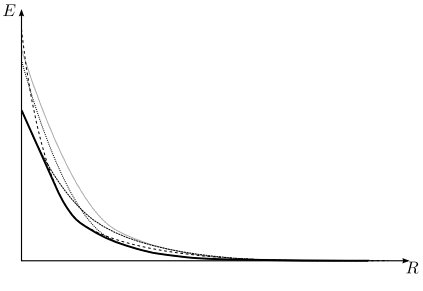

Fig. 3. The ER-curves generated for different ρ and their convex hull assumed as the ER-curve of the ECOC system

5 Experiments

To test the performance of the proposed reject rules, six multiclass data sets publicly available at the UCI machine learning repository [2] have been used. To avoid any bias in the comparison, 12 runs of a multiple hold out procedure have been performed on all the data sets. In each run, the data set has been split into three subsets: a training set (containing the 70% of the samples of each class), a validation set and a test set (each containing the 15% of the samples of each class). The training set is used to train the base classifiers, the validation set to normalize the outputs into the range $[-1, 1]$ and to calculate the thresholds (τ_{h1}, τ_{h2}) and the test set to evaluate the performance of the classification system. A short description of the data sets is given in table 2. In the same table we also report the number of columns of the coding matrix chosen for each data set according to [6]. As base dichotomizer Support Vector Machines (SVM) have been implemented through the SVM^{Light} [8] software library using a linear kernel and an RBF kernel with $\sigma = 1$. In both cases the C parameter has been set to the default value calculated by the learning procedure and the "hinge" loss $\mathcal{L}(z) = \max\{1 - z, 0\}$ has been adopted.

We show in fig. 4 and fig. 5 the results obtained for the two classifiers in terms of the Error-Reject curves calculated by averaging the Error-Reject curves obtained in the 12 runs of the multiple hold out procedure. The range for the reject rate on the x-axis has been limited to $[0, 0.30]$ since higher reject rates are typically not of interest in real applications. Both figures reports a comparison of the two considered reject rules: Loss Decoding (LD) and TLD (Trimmed Loss Decoding). For the LD the loss output was normalized in the range $[0, 1]$ and consequently the thresholds were varied in this interval with the step 0.01. In the case of TLD we have varied the parameter ρ from 0 to 1 with step 0.05 and, as in the LD case, the loss output has been normalized in the range $[0, 1]$ and the external threshold varied with step 0.01 into the same range.

In the majority of the analyzed cases, the two figures show that the ER-curves generated by the TLD method dominate those of LD. Only in two cases, i.e., Abalone and Vowel with an SVM with RBF kernel, there is a complete equivalence of the two curves. Consequently, we can say that TLD approach

Table 2. Data sets used in the experiments

Data Set	Classes	Features	Length (L)	Samples
Abalone	29	8	30	4177
Ecoli	8	7	62	341
Glass	6	9	31	214
Letter	26	16	63	5003
Vowel	11	10	14	435
Yeast	10	8	31	1484

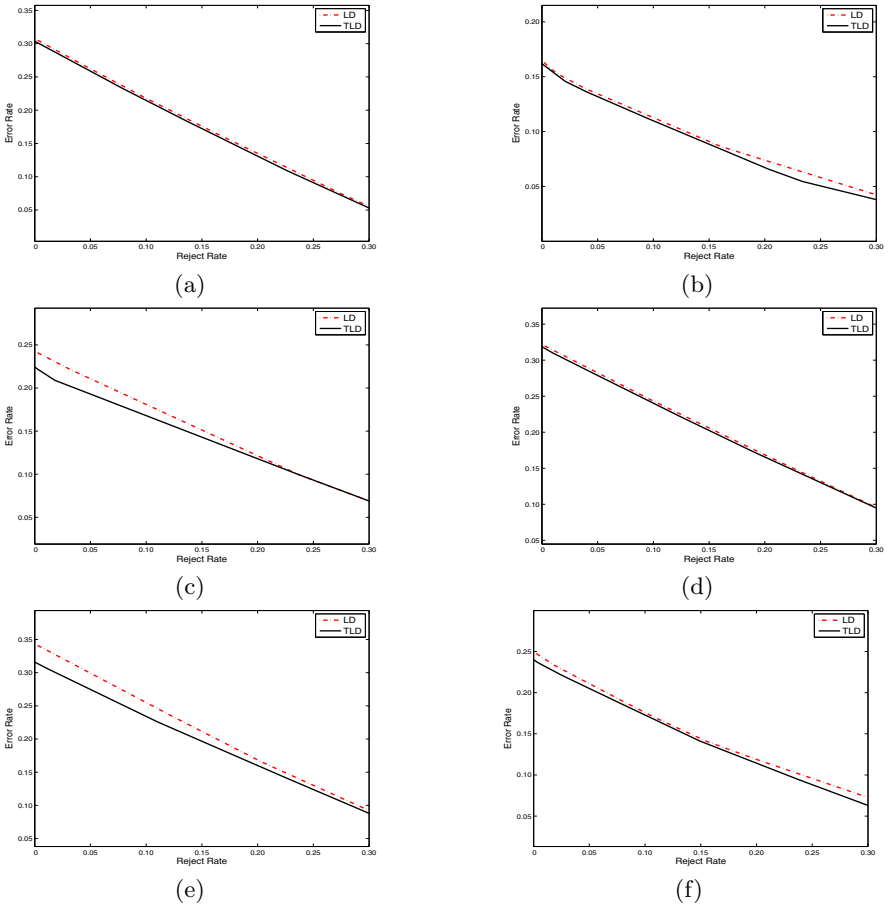

Fig. 4. Error-Reject curves obtained on Abalone (a), Ecoli (b), Glass (c), Letter, (d), Vowel (e) and Yeast (f) with linear SVM

is superior than LD since it allows us to control the individual errors of the base classifiers. Thus, the use of an internal reject rule can be profitably used to improve the performance of the ECOC systems. In conclusion, knowing the architectural details of the base classifiers (i.e. the nature of their outputs and their loss functions), the system has the possibility to face the uncertainty of wrong predictions in a more precise and effective way than a simple external technique.

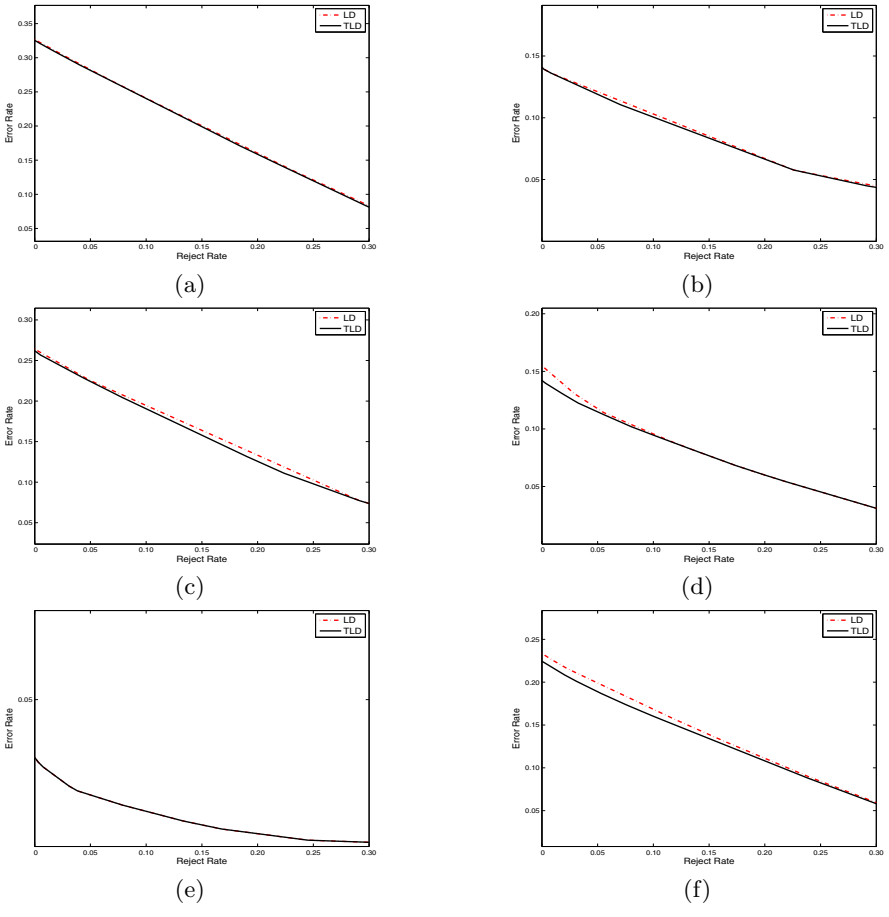

Fig. 5. Error-Reject curves obtained on Abalone (a), Ecoli (b), Glass (c), Letter, (d), Vowel (e) and Yeast (f) with RBF SVM

6 Conclusions

In this paper we have analyzed two different techniques to enrich an ECOC classification system with a reject option. The main difference is in the simplicity of the external approach that only requires an intervention on the decoding stage of the ECOC system while the Trimmed Loss Decoding requires to manage more parameters. However, the geometrical method described for the Error-Reject curve simplifies the use of the internal technique and this can be particularly useful in improving the error/reject trade-off, as shown by the experiments. A possible development of this work can be focused on the investigation of the relation between the rejection rule with the characteristics of the coding matrix.

References

1. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141 (2000)
2. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)
3. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. J. Mach. Learn. Res. 9, 1823–1840 (2008)
4. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1), 41–46 (1970)
5. Cordella, L.P., De Stefano, C., Tortorella, F., Vento, M.: A method for improving classification reliability of multilayer perceptrons. IEEE Transactions on Neural Networks 6(5), 1140–1147 (1995)
6. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)
7. Guruswami, V., Sahai, A.: Multiclass learning, boosting, and error-correcting codes. In: Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, pp. 145–155. ACM, New York (1999)
8. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, ch. 11. MIT Press, Cambridge (1999)
9. Pietraszek, T.: On the use of ROC analysis for the optimization of abstaining classifiers. Machine Learning 68(2), 137–169 (2007)
10. Simeone, P., Marrocco, C., Tortorella, F.: Exploiting system knowledge to improve ecoc reject rules. In: International Conference on Pattern Recognition, pp. 4340–4343. IEEE Computer Society, Los Alamitos (2010)
11. De Stefano, C., Sansone, C., Vento, M.: To reject or not to reject: that is the question-an answer in case of neural classifiers. IEEE Transactions on Systems, Man, and Cybernetics, Part C 30(1), 84–94 (2000)
12. Tortorella, F.: Reducing the classification cost of support vector classifiers through an roc-based reject rule. Pattern Anal. Appl. 7(2), 128–143 (2004)