Dependence among Codeword Bits Errors in ECOC Learning Machines: An Experimental Analysis

Francesco Masulli[1,2] and Giorgio Valentini[1,2]

[1] DISI - Dipartimento di Informatica e Scienze dell'Informazione
Università di Genova, via Dodecaneso 35, 16146 Genova, Italia
[2] Istituto Nazionale per la Fisica della Materia
via Dodecaneso 33, 16146 Genova, Italia
{masulli, valenti}@disi.unige.it

Abstract. One of the main factors affecting the effectiveness of ECOC methods for classification is the dependence among the errors of the computed codeword bits. We present an extensive experimental work for evaluating the dependence among output errors of the decomposition unit of ECOC learning machines. In particular, we compare the dependence between ECOC Multi Layer Perceptrons (ECOC *monolithic*), made up by a single *MLP*, and ECOC ensembles made up by a set of independent and parallel dichotomizers (ECOC *PND*), using measures based on mutual information. In this way we can analyze the relations between performances, design and dependence among output errors in ECOC learning machines. Results quantitatively show that the dependence among computed codeword bits is significantly smaller for ECOC *PND*, pointing out that ensembles of independent dichotomizers are better suited for implementing ECOC classification methods.

1 Introduction

Error Correcting Output Coding (ECOC) [4] is a two-stage Output Coding (OC) decomposition method [10,8] that has been successfully applied to several classification problem [2,5]. In its first stage it decomposes a multiclass classification problem in a set of two-class subproblems, and in a second stage recomposes the original problem combining them to achieve the class label.

ECOC methods present several open problems such us the tradeoff between error recovering capabilities and learnability of the dichotomies induced by the decomposition scheme [1]. A connected problem is the analysis of the relation between codeword length and performances [5], while the selection of optimal dichotomic learning machines and the design of optimal codes for a given multiclass problem are other open questions subject to active research [3].

Another problem tackled by different works [7,6] is the relation between performances of ECOC and dependence among output errors. In the framework of coding theory Peterson [12] has shown that the error recovering capabilities of

J. Kittler and F. Roli (Eds.): MCS 2001, LNCS 2096, pp. 158–167, 2001.

ECOC codes hold if there is a low dependence among codeword bits. In particular, in a previous work [8] we qualitatively identify the dependence among output errors as one of the factors affecting the effectiveness of ECOC decomposition methods. In that work we outlined that we would expect an higher dependence among codeword bits in *monolithic Error Correcting Output Coding* [4,8] (ECOC *monolithic* for short) compared with *ECOC Parallel Non linear Dichotomizers (PND)* [8] (ECOC *PND* for short) learning machines, considering that ECOC *monolithic* share the same hidden layer of a single *MLP*, while *PND* dichotomizers, implemented by a separate *MLP* for each codeword bit, have their own layer of hidden units, specialized for a specific dichotomic task.

The aim of this work is to *quantitatively* test if the dependence among output errors between ECOC *monolithic* and ECOC *PND* is significantly different. In particular, we perform an extensive experimentation for comparing the dependence among output errors of the decomposition unit of ECOC *monolithic* and ECOC *PND* using measures based on mutual information [9], in order to evaluate if a low dependence among output errors is related to better classification performances.

The paper is structured as follows. In the next section we summarize the main characteristics of the measures based on mutual information we propose for evaluating the dependence among output errors in learning machines. Sect. 3 presents the experimental setup, the results and the discussion about the quantitative comparison of dependence among output errors between ECOC *monolithic* and ECOC *PND* learning machines. The conclusions summarize the main results and the incoming developments of this work.

2 Mutual Information Based Measures of Dependence among Output Errors

In this section we present a brief overview of the mutual information based measures for evaluating the dependence among output errors in learning machines. A more detailed discussion can be found in [9].

The main idea behind the evaluation of dependence among output errors of learning machines through mutual information based measures consists in interpreting the dependence among the outputs as the common information shared among them. Mutual information takes into account the marginal and joint probability distributions of the output errors, measuring in a sense the information shared among them. Using standard statistical measures such as the covariance or the coefficient of correlation we estimate only the linear relation between output errors. Conversely, a suitable measure of dependence must evaluate directly the probability distribution of the output errors in order to properly evaluate the stochastic independence between random variables. Mutual information, being a special case of the Kullback-Leibler divergence between two distributions, measures the matching between the joint density distribution and the product of the marginal density distribution of the output errors. If we a have a complete matching, the mutual information is 0 and the output errors are independent,

otherwise higher is the value of the mutual information between output errors, higher will be the dependence between them.

The first measure based on mutual information we define is the *mutual information error* I_E:

$$I_E(e_1, \ldots, e_l) = \sum_{j_1=1}^{b} \ldots \sum_{j_l=1}^{b} p(e_{1j_1}, \ldots, e_{lj_l}) \log \left(\frac{p(e_{1j_1}, \ldots, e_{lj_l})}{p(e_{1j_1}) \ldots p(e_{lj_l})} \right) \quad (1)$$

where $p(e_{1j_1}, \ldots, e_{lj_l})$ is the *discrete joint probability distribution* among all the l output errors and $p(e_{ij_i})$ is the *discrete probability distribution* of the i^{th} output error, with $i \in \{1, \ldots, l\}$ and with the $j_i \in \{1, \ldots, b\}$ corresponding to the discretization of the output errors in b intervals. The mutual information error (eq. 1) expresses the dependence among all output errors of a learning machine. If it is equal to 0 then the distributions of the output errors are statistically independent. It expresses also how are similar the probability distribution of the output errors.

Considering the outputs of a learning machine correct if their errors are below a certain threshold, i.e if $\forall i,\ e_i < \delta,\ \delta > 0$, we define the *mutual information specific error* I_{SE}:

$$I_{SE}(e_1, \ldots, e_l) = \sum_{\mathcal{J}} p(e_{1j_1}, \ldots, e_{lj_l}) \log \left(\frac{p(e_{1j_1}, \ldots, e_{lj_l})}{p(e_{1j_1}) \ldots p(e_{lj_l})} \right) \quad (2)$$

where

$$\mathcal{J} = \left\{ [j_1, \ldots j_l] | \exists (j_v, j_w) | (j_v \neq 1) \wedge (j_w \neq 1) \right\}$$

with $v, w \in \{1 \ldots l\}$. This measure takes into account the output errors only when two or more errors spring from the output, disregarding all cases with no errors or with only one error. For evaluating the dependence among specific pairs of output errors, we introduce the *pairwise mutual information error matrix* R composed by the elements $I_E(e_i, e_j) = [R_{ij}]$ and the *pairwise mutual information specific error matrix* S, composed by the elements $I_{SE}(e_i, e_j) = [S_{ij}]$. We then define also two other global indices: the *pairwise mutual information error matrix index* Φ_R:

$$\Phi_R = \sum_{i=1}^{l} \sum_{j=1}^{l} I_E(e_i, e_j) \quad (3)$$

and the *pairwise mutual information specific error matrix index* Φ_S:

$$\Phi_S = \sum_{i=1}^{l} \sum_{j=1}^{l} I_{SE}(e_i, e_j) \quad (4)$$

These indices measure the sum of the the mutual information error and the mutual information specific error between all the output pairs of the learning machines, and in this sense can be regarded as global measures of dependence

Table 1. Main features of the data sets.

Data set	Number of attributes	Number of classes	Number of training samples	Number of testing samples
d5	3	5	30000	30000
glass	9	6	214	10-fold cross-val
letter	16	26	16000	4000
optdigits	64	10	3823	1797

between output errors. Note that these indices (Eq. 3 and 4) are not equivalent to the corresponding Eq. 1 and 2 of the mutual information among all output errors: Eq. 3 and 4 consider only the mutual information between pairs of output errors, while Eq. 1 and 2 consider the overall mutual information among all output errors.

These mutual information related quantities can be used to compare the dependence of the output errors among different learning machines on the same learning problem, using, of course, the same data sets.

3 Experimental Results

In this section we present a *quantitative* comparison of the dependence among output errors of the decomposition unit of ECOC *monolithic* and ECOC *PND* learning machines, and we analyze the relations between performances, design and dependence among output errors. For this purpose we experimentally compare the mutual information error I_E, the mutual information specific error I_{SE} and the pairwise indices Φ_R and Φ_R (Sect. 2) of the ECOC *monolithic* and *PND* learning machines using different data sets.

3.1 Experimental Setup

We have used four different data sets: the first one, *d5* [1] is generated by NEUROObjects [13], a set of C++ library classes for neural networks development, and the other three, *glass, letter* and *optdigits* are from the UCI machine learning repository of Irvine [11]. The synthetic data set *d5* is made up by five three-dimensional classes, each composed by two normal distributed disjoint clusters of data. The main characteristics of the data sets are shown in Tab. 1.

In order to perform training and testing of the considered learning machines, we have applied multiple runs of different random initializations of weights using a single pair of training and testing data sets and *k-fold cross validation* methods. The results are summarized in Tab.2: errors on the test set are expressed as percent rates, and for each data set the minimum (min), average (mean), and standard deviation (stdev) of the error is given. We have used, both for training

[1] *d5* is on line available at ftp://ftp.disi.unige.it/person/ValentiniG/Data.

the learning machines and for evaluating the dependence among the output errors the software library *NEURObjects* [13].

We have compared the dependence among output errors of ECOC *monolithic* and ECOC *PND* learning machines varying the structure (number of hidden units), the number of discretization intervals of the output errors, and the values of δ (Sect. 2) that define the notion of "correctness" of the outputs.

3.2 Results and Discussion

In this section we present the results of the comparison of I_E and I_{SE} among all outputs, of the Φ_R and Φ_S pairwise indices and the comparison of R and S matrices.

In Fig. 1 we compare I_E and I_{SE} among all output errors of the *monolithic* and ECOC *PND* learning machines on the data sets *d5* and *glass*. On the axes are represented the computed I_E (Fig. 1 a and b) and I_{SE} (Fig. 1 c and d) values. Each point corresponds to a different triplet number of hidden units, number of intervals and values of δ. We point out that all points are above the dotted line, showing that both I_E (Fig. 1 a and b) and I_{SE} (Fig. 1 c and d) are greater for ECOC *monolithic* respect to ECOC *PND*, no matter the structure, the number of intervals and the δ values used. Fig. 2 shows that on all the data sets about all the points are above the dotted line, i.e. all the values of Φ_R are greater for ECOC *monolithic* compared with ECOC *PND*. Similar results hold also considering the Φ_S index. The examination of the pairwise mutual information error matrices can provide us with information about the dependence of specific pairs of output errors. The S and R matrices are represented as triangular matrices, without the diagonal, because they are symmetric and the elements on the diagonal are the entropy of output errors.

Comparing the mutual information matrices of ECOC *monolithic* and *PND* learning machines, we find that about all the pairwise mutual information errors are higher in ECOC *monolithic*: on the d5 data set no element of the R matrix is higher for *PND* and only 1 of 21 is higher considering the S matrix; on optdigits only 3 of 91 both for R and S matrices are higher, and no element of the 435 composing the triangular matrices R and S is higher for *PND* on letter data set.

Table 2. Performance of ECOC monolithic and ECOC *PND* ensemble on four data sets (percent error rates).

	ECOC monolithic			**ECOC *PND* ensemble**		
Data set	**min**	**mean**	**stdev**	**min**	**mean**	**stdev**
d5	13.27	18.31	6.44	11.91	12.34	0.74
glass	33.18	36.17	4.54	30.37	32.05	1.77
letter	4.95	6.55	1.91	3.05	3.24	0.24
optdigits	2.61	3.08	0.47	1.89	1.95	0.10

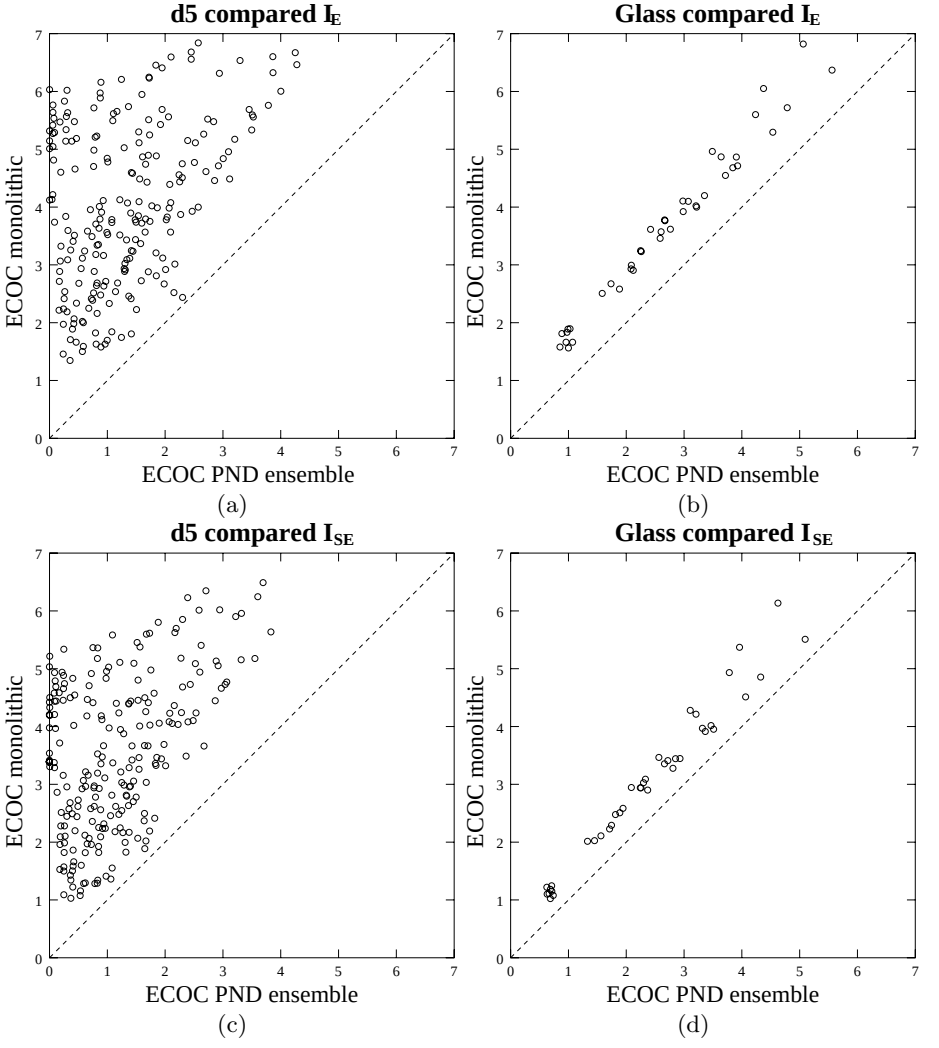

Fig. 1. Compared mutual information error I_E and mutual information specific error I_{SE} among all outputs between ECOC *monolithic* and *PND* learning machines on d5 (a)(c) and glass (b)(d) data sets.

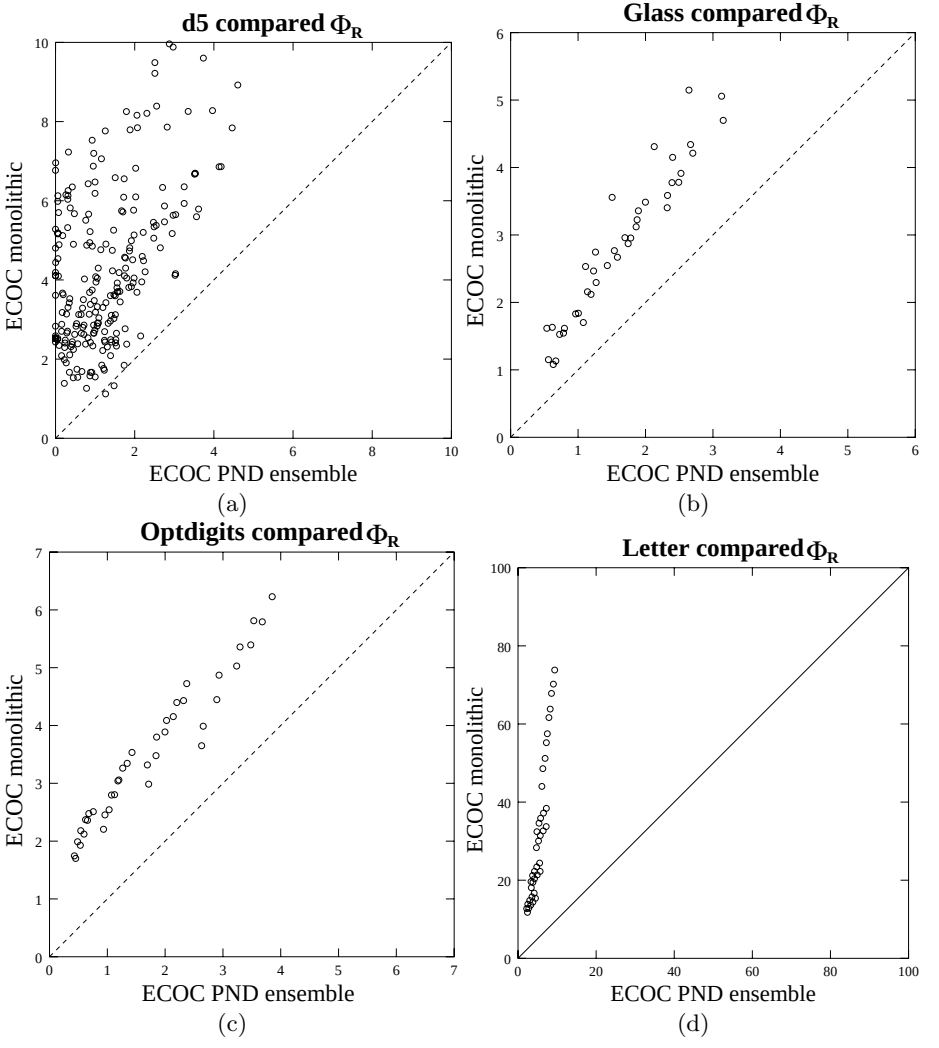

Fig. 2. Compared mutual information error matrix indices Φ_R between ECOC *monolithic* and *PND* learning machines on d5 (a), glass (b), optdigits (c) and letter (d) data sets.

Fig. 3 shows the relations between error rates and mutual information based measures I_E and I_{SE} considering the *d5* data set. Both I_E and I_{SE} curves of ECOC *PND* ensemble lie below the corresponding curves of ECOC *monolithic* learning machines: These figures confirm that the dependence among output errors is smaller for ECOC *PND*. It is worth noting that, as expected, I_E and I_{SE} grow with error rates, but their values are mostly related to a specific learning machine architecture.

We have seen that all the results relative to the mutual information error I_E and the mutual information specific error I_{SE} among all the outputs on the data sets *d5* and *glass* show greater values for ECOC *monolithic* respect to ECOC *PND* (Fig. 1). These results are confirmed by the evaluation of the mutual information error matrix indices Φ_R and Φ_S (Fig. 2), concerning also the *optdigits* and *letter* data sets. The analysis of the pairwise mutual information matrices R and S converges on showing that also about all the I_E and I_{SE} values between each pair of output errors are greater for ECOC *monolithic* learning machines. Moreover, applying the *mutual information error t-test* [9] for evaluating the significance of the differences between the I_E and I_{SE} values of the two ECOC learning machines, we have verified that in almost all the comparisons we have registered a significant difference with a degree of confidence of 95%.

Consequently the experimental results on the selected data sets confirm that ECOC Parallel Non linear Dichotomizers show a lower dependence among the output errors of their decomposition unit compared with the output errors of the corresponding ECOC *monolithic* multi layer perceptron.

4 Conclusions

In this paper, we have compared the dependence among output errors between ECOC *monolithic MLP* and ECOC *PND* learning machines using measures based on mutual information.

The measurements of the mutual information error I_E, the mutual information specific error I_{SE} and the mutual information error matrix indices Φ_R and Φ_S show that ECOC *PND* have a lower dependence among the output errors of their decomposition unit compared with the output errors of the corresponding ECOC *monolithic MLP*. Hence ECOC *PND* ensembles appear more suited to exploit the error recovering capabilities of ECOC methods, whose effectiveness depends on the independence among codeword bits errors [12,8].

The observed difference in the dependence among output errors is related to the different design of the two learning machines and in particular to the design of the decomposition unit. Our experimentation suggests that a low dependence can be achieved implementing the decomposition unit through an ensemble of parallel and independent dichotomizers, such as the dichotomic *MLPs* proposed in our experimentation, or other suitable dichotomizers such as decision trees or support vector machines.

An ongoing development of this work consists in quantitatively studying how boosting methods can increase the diversity among the dichotomizers and

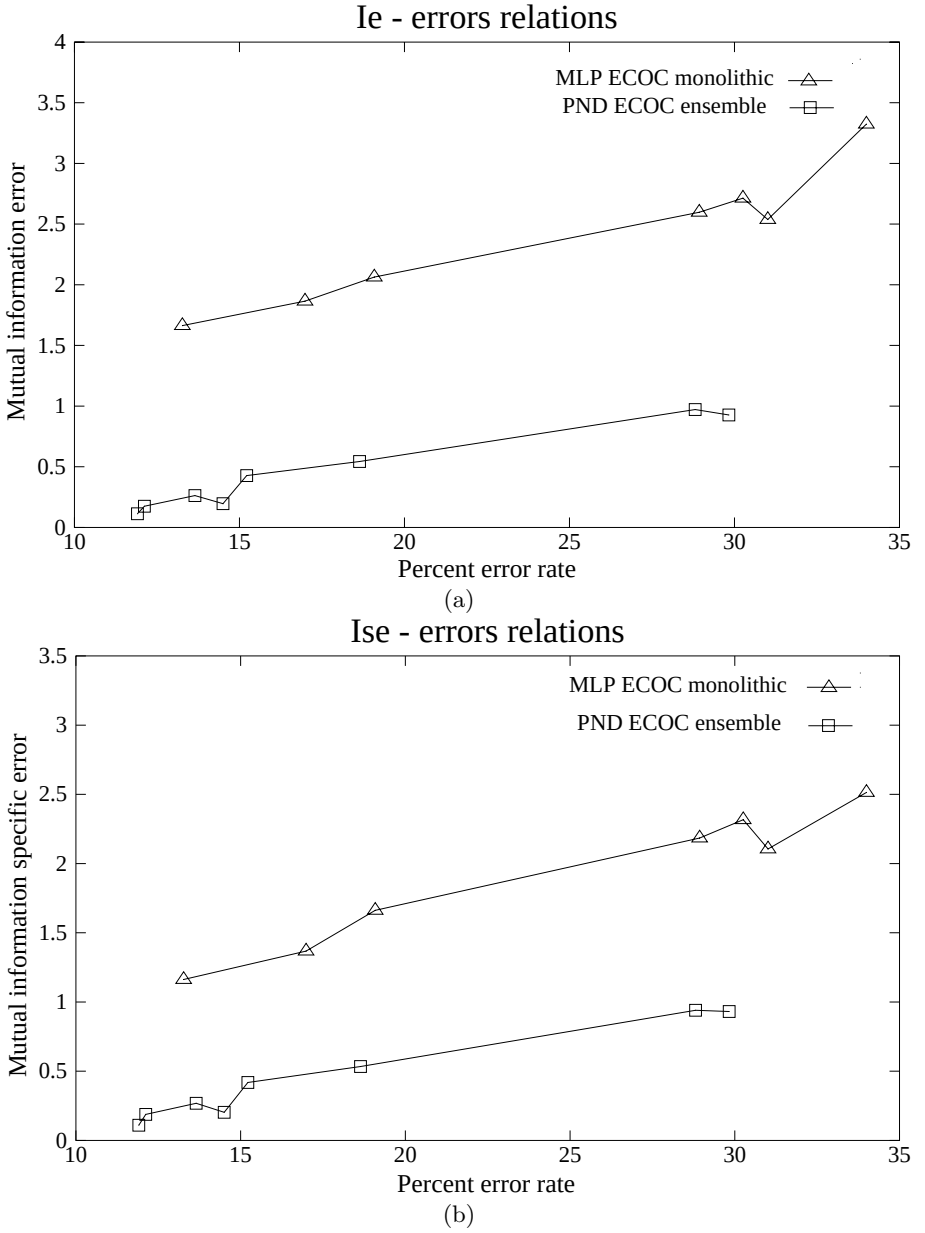

Fig. 3. Relations error rates - mutual information error I_E (a) and error rates - mutual information specific error I_{SE} (b) in ECOC *monolithic* and *PND* learning machines on the d5 data set.

the independence among output errors in ECOC learning machines, using the proposed measures based on mutual information, and extending them to evaluate the diversity between the base learners.

Acknowledgments. We would like to thank the anonymous reviewers for their comments and suggestions. This work has been partially funded by Progetto finalizzato CNR-MADESS II, INFM and University of Genova.

References

1. E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. In *Proc. ICML'2000, The Seventeenth International Conference on Machine Learning*, 2000.
2. A. Berger. Error correcting output coding for text classification. In *IJCAI'99: Workshop on machine learning for information filtering*, 1999.
3. Y. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 35-46, 2000.
4. T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, (2):263–286, 1995.
5. R. Ghani. Using error correcting output codes for text classification. In *ICML 2000: Proceedings of the 17th International Conference on Machine Learning*, pages 303–310, San Francisco, US, 2000. Morgan Kaufmann Publishers.
6. V. Guruswami and A. Sahai. Multiclass learning, boosting, and error-correcting codes. In *Proc. of the Twelfth Annual Conference on Computational Learning Theory*, pages 145–155. ACM Press, 1999.
7. E. Kong and T.G. Dietterich. Error - correcting output coding correct bias and variance. In *The XII International Conference on Machine Learning*, pages 313–321, San Francisco, CA, 1995. Morgan Kauffman.
8. F. Masulli and G. Valentini. Effectiveness of error correcting output codes in multiclass learning problems. In *Lecture Notes in Computer Science*, volume 1857, pages 107–116. Springer-Verlag, Berlin, Heidelberg, 2000.
9. F. Masulli and G. Valentini. Mutual information methods for evaluating dependence among outputs in learning machines. Technical Report TR-01-02, DISI - Dipartimento di Informatica e Scienze dell' Informazione - Università di Genova, 2001. ftp://ftp.disi.unige.it/person/ValentiniG/papers/TR-01-02.ps.gz.
10. E. Mayoraz and M. Moreira. On the decomposition of polychotomies into dichotomies. In *The XIV International Conference on Machine Learning*, pages 219–226, Nashville, TN, July 1997.
11. C.J. Merz and P.M. Murphy. UCI repository of machine learning databases, 1998. www.ics.uci.edu/mlearn/MLRepository.html.
12. W.W. Peterson and E.J.Jr. Weldon. *Error correcting codes*. MIT Press, Cambridge, MA, 1972.
13. G. Valentini and F. Masulli. NEURObjects, a set of library classes for neural networks development. In *Proceedings of IIA'99 and SOCO'99*, pages 184–190, Millet, Canada, 1999. ICSC Academic Press.