# Separability of Ternary Error-Correcting Output Codes

Sergio Escalera, Oriol Pujol, and Petia Radeva
*Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007, Barcelona.*
{*sergio,oriol,petia*}*@maia.ub.es*

## Abstract

*Error Correcting Output Codes (ECOC) represent a successful framework to deal with multi-class categorization problems based on combining binary classifiers. In this paper, we present a new formulation of the ternary ECOC distance and the error-correcting capabilities in the ternary ECOC framework. Based on the new measure, we stress on how to design coding matrices preventing codification ambiguity and propose a new Sparse Random coding matrix with ternary distance maximization. The results on the UCI Repository and in a real speed traffic categorization problem show that when the coding design satisfies the new ternary measures, significant performance improvement is obtained independently of the decoding strategy applied.*

## 1 Introduction

Error Correcting Output Codes were born as a general framework to combine binary problems to address the multi-class problem [4]. The ECOC technique can be broken down into two general stages: encoding and decoding. At the coding step, given a set of $N$ classes to be learnt, $n$ different bi-partitions (groups of classes) are formed, and $n$ binary problems (dichotomizers) are trained. As a result, a codeword of length $n$ is obtained for each class, where each bit of the code corresponds to the response of a given dichotomizer (coded by +1, -1, according to its class set membership). Arranging the codewords as rows of a matrix, we define a *coding matrix* $M$, where $M \in \{-1, 1\}^{N \times n}$ in the binary case. It was when Allwein et al. [1] introduced a third symbol (the zero symbol) in the coding process that the coding step received special attention. This symbol increases the number of partitions of classes to be considered in a ternary ECOC framework by allowing some classes to be ignored. Then, the ternary coding matrix becomes $M \in \{-1, 0, 1\}^{N \times n}$. In this case, the symbol zero means that a particular class is not considered by a certain binary classifier. Thanks to this, strategies such as the Sparse Random coding [1] have been formulated in the ECOC framework. The decoding step was originally based on error-correcting principles under the assumption that the learning task can be modeled as a communication problem, in which class information is transmitted over a channel [4]. During the decoding process, applying the $n$ binary classifiers, a code is obtained for each data point in the test set. This code is compared to the base codewords of each class defined in the matrix $M$, and the data point is assigned to the class with the *closest* codeword. The most frequently applied decoding strategies are the Hamming ($HD$) and the Euclidean ($ED$) decoding distances [6].

To deal with multi-class categorization problems in the ternary ECOC framework, recent works redefined decoding strategies that were formulated to deal with just two symbols [1]. However, the influence of the zero symbol to the error-correction capabilities and the design of the coding strategies have not been taken into account. In this paper, we formulate the ternary distance in the ECOC framework. Based on the new measure and the ternary error-correcting capabilities, we propose a new sparse coding design. We evaluate the methodology on a wide set of UCI data sets and in a real speed traffic sign categorization problem. The results show that when the new ternary distance is considered on sparse designs, significant performance improvement is obtained.

The paper is organized as follows: Section 2 presents a new sparse coding design based on ternary distance maximization, section 3 presents the experimental results, and finally, section 4 concludes the paper.

## 2 Random ECOC Designs

In this section, we overview both Dense and Sparse Random ECOC designs [1]. We show the inconsistence of the classical Sparse Random design and introduce a new measure for sparse coding designs.

### 2.1 Dense Random Design

Given a binary ECOC matrix $M \in \{-1, 1\}^{N \times n}$, where $N$ is the number of classes and $n$ the

codeword length, the minimum Hamming distance $d_r$ among all pairs of rows is defined as $d_r = min\left\{\sum_{j=1}^{n}(1 - sign(y_{i_1}^j \cdot y_{i_2}^j))/2\right\}$, for $i_1, i_2 \in \{1, ..., N\}$, $i_1 \neq i_2$, being $y_{i_1}^j$ the $j^{th}$ position of the codeword for class $c_{i_1}$. Suppose that two codewords coded using $\{-1, +1\}$ values have a Hamming distance of three. Then, it means that even if we fail in a bit, we still are able to obtain the correct classification. It suggests that a distance $d_r$ in a binary ECOC matrix $M$ can correct $[d_r - 1]/2$ codeword errors at the decoding step [4]. Because of these binary error-correction capabilities, many ECOC designs, such as random ECOC strategies, base the design of the ECOC coding matrix on maximizing the value $d_r$ [1]. Let us now consider the distance $d_c$ between all pairs of columns and their opposites $d_c = min_{j_1,j_2}\{min\,(A(j_1, j_2), B(j_1, j_2))\}$, being:

$$A(j_1, j_2) = \sum_{i=1}^{N}(1 - sign(y_i^{j_1} \cdot y_i^{j_2}))/2 \quad (1)$$

$$B(j_1, j_2) = \sum_{i=1}^{N}(1 - sign(-1 \cdot (y_i^{j_1}) \cdot y_i^{j_2}))/2 \quad (2)$$

where $j_1, j_2 \in \{1, ..., n\}$, $j_1 \neq j_2$. High value of $d_c$ contributes to consider different sub-partitions of classes and to increase the variability of the knowledge of the classifiers. Note that in eq.(2) the factor (-1) is used to take into account the independence of the class ordering, i.e. the base classifier learns the same problem from the partition $C_1$ versus $C_2$ and from $C_2$ versus $C_1$.

The Dense Random ECOC strategy [1] tries to maximize simultaneously both previous $d_r$ and $d_c$ distances to design matrices where the decoding strategies are able to obtain a correct classification still when there exist failures in some bits of the tested codewords. The Dense random strategy generates a high number of random coding matrices $M$ of length $n$, where the values $\{+1, -1\}$ have a certain probability to appear (usually $P(1) = P(-1) = 0.5$). Studies on the performance of the dense random strategy suggests a length of $n = 10 \log N$ [1]. In order to assure optimal performance of ECOC classification, for the set of generated dense random matrices, the optimal one should maximize the $HD$ between rows $d_r$ and columns $d_c$, taking into account that each column of the matrix $M$ must contain both different symbols $\{-1, +1\}$.

## 2.2 Classical Sparse Random Design

One of the main limitations of the binary ECOC framework is the need of considering all classes for each binary classifier. Although a high distance $d_r$ and $d_c$ can be computed, the selection of the most relevant sub-partition of classes for different multi-classification problems is not assured in the coding design. This fact implies the need of designing large codes to increase the discriminating ability of the combined set of binary problems. Moreover, taking into account the whole set of classes for each classifier significantly reduces the number of possible sub-partitions of classes to consider.

To take into account a higher number of possible classifiers, a third symbol was introduced in the ECOC framework [1]. In this sense, the Sparse Random strategy is designed in the same way than the Dense design, but it includes the third symbol zero with another probability to appear, given by $P(0) = 1 - P(-1) - P(1)$. Studies suggest a sparse code length of $15 \log N$ [1].

### 2.2.1 Sparse Design with Ternary Separability

Let us show an example to analyze sparse designs. A zero symbol in a class code introduces *one degree of freedom*, that means that both +1 and -1 are possible values during the test classification since the class has not been taken into account to train the corresponding dichotomizer. Any codeword $y_i$ containing the zero symbol defines an extended set of possible codewords that could be obtained by examples of the class $c_i$. In this sense, a possible codeword $y_1 = \{1, 0, 0\}$ can be disambiguated into its extended set of codewords $Y_1^e = \{\{1, 1, 1\}, \{1, 1, -1\}, \{1, -1, 1\}, \{1, -1, -1\}\}$, where each of the four codewords of $y_1$ is a possible representation[1] of the same codeword $y_1$. Now, a possible codeword for a second class $y_2 = \{1, 1, 1\}$ corresponds to one of the four possible representations of $y_1$ ($y_2 \in Y_1^e$). Let us consider another example of codewords of length six. Suppose that we randomly define two codewords $y_1 = \{1, 1, 1, 0, 0, 0\}$ and $y_2 = \{0, 0, 0, 1, 1, 1\}$ in a Sparse Random design. If we use the classical distance $d_r$ between $y_1$ and $y_2$, we obtain a class separability of three. However, based on the previous example, if we disambiguate $y_1$ and $y_2$, we obtain that $Y_1^e \cap Y_2^e = \{1, 1, 1, 1, 1, 1\}$. Thus, an input test codeword $X = \{1, 1, 1, 1, 1, 1\}$ belongs to both previous codewords, which implies a wrong Sparse design. Finally, observe the ternary coding matrix $M$ of fig. 1. Suppose that the matrix $M$ of the figure receives an input test data sample which codeword corresponds to $X = \{-1, 1, 1, 1, 1\}$. This codeword matches with the four positions different of zero from class $c_1$ and the three from class $c_3$. In this case, $X \in Y_1^e$ and $X \in Y_3^e$. Thus, both classes can be a possible solution. However, the $HD$ between codewords $y_1$ and $y_3$ produces a value of 1.5. Note that in the literature [1], a Sparse Random matrix is generated by selecting the matrix from a previous set of matrices that maximizes the distances $d_r$ and $d_c$. The $HD$ between columns containing the third

[1] Possible representation means that any test example of class $c_1$ would give a codeword from $Y_1^e$.

symbol is still useful since the zero positions help to create create a rich set of partitions to be learnt. However, the measure $d_r$ for the row separability in terms of the $HD$ is inconsistent. Instead, to assure that the coding matrix $M$ splits all pairs of classes, each pair of codewords of $M$ should be split by at least one hypothesis:

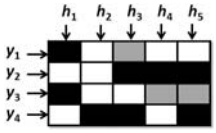


**Figure 1.** Codification error between classes $c_1$ and $c_3$.

**Definition 1.**: The **ternary separability** condition of a matrix $M$ is defined as:

$\forall(y_{i_1}, y_{i_2})|i_1, i_2 \in \{1, ..., N\}, i_1 \neq i_2, \exists h_j|$
$(c_{i_1} \in C_1^j, c_{i_2} \in C_2^j) \vee (c_{i_2} \in C_1^j, c_{i_1} \in C_2^j)$

where $C_1^j$ and $C_2^j$ are the two subsets of classes for hypothesis $h_j$, respectively. Then, we define the distance between two codewords in a ternary ECOC:

**Definition 2.**: The **ternary distance** between two codewords ($y_1$,$y_2$) is defined as:

$$d(y_1, y_2) = \sum_{j=1}^{n} \frac{1}{2}|y_1^j||y_2^j|(1 - y_1 y_2) \qquad (3)$$

It defines the number of different bits between two codewords without taking into account the positions coded by zero. The weighting term $|y_1^j||y_2^j|$ makes the distance to ignore the zero positions which do not give information about the classes separability. Then, the pair of codewords $(y_{i_1}, y_{i_2})$ that are split by the minimum number of hypothesis in a ternary ECOC matrix $M$ defines the new distance $d_t$:

**Definition 3.**: The **distance** $d_t$ of a coding matrix $M$ is defined as follows:

$$d_t = argmin_{i_1,i_2} \sum_{j=1}^{n} \frac{1}{2}|y_{i_1}^j||y_{i_2}^j|(1 - y_{i_1} y_{i_2}) \qquad (4)$$

where the term $d_t$ defines the distance between the pair of codewords that are split by the minimum number of binary problems in a ternary ECOC matrix. Then, as the distance in the ternary case is reformulated, the new measure of error-correction changes. Having a $N$-multi-class classification problem in the binary ECOC framework, a distance $d_r$ between rows of $M$ can correct $[d_r - 1]/2$ bits errors. In the ternary case, the maximum class separability is defined by the measure $d_t$. Thus, on a sparse ECOC matrix, $[d_t - 1]/2$ bits errors can be corrected[2]. Then, as the use of the distance $d_r$ applied to the classical design of the Sparse Random matrix $M$ produces inconsistences, we suggest to redefine the coding stage of the Sparse Random designs. A good codification of a ternary matrix should assure the highest number of codeword bits splitting each pair of rows; that is to maximize the value $d_t$. Therefore, we propose to use the new measure of ternary separability for the Sparse Random design. In this case, the selected random matrix should be that one which maximizes simultaneously $d_c$ and $d_t$.

[2]We realize that the error-correcting capability also depends on the way that the decoding strategies are applied.

## 3 Results

First, we discuss the data, comparatives, and measurements: • *Data*: We used the 16 multi-class data sets from the UCI Repository [2] described in table 1. We also use the video sequences obtained from a Mobile Mapping System [3] to test a real traffic sign categorization problem. • *Comparatives*: We use the classical Sparse Random design [1] and the new Sparse Random with ternary distance maximization. The sparse matrices are selected from a set of 20000 randomly generated matrices with a length of codewords of $N$, where $P(0) = P(1) = P(-1) = 1/3$. To decode, we use nine state-of-the-art decoding strategies: $HD$ [4], $ED$ [6], Inverse Hamming Decoding ($IHD$) [8], Attenuated Euclidean Decoding ($AED$) [5], Linear ($LLB$) and Exponential ($ELB$) Loss-based [1], Probabilistic Decoding ($PD$) [7], Laplacian Decoding ($LAP$) [5], and Pessimistic $\beta$-Density Distribution Decoding($\beta - DEN$) [5]. • *Measurements*: We apply stratified ten-fold cross-validation and test for confidence interval at 95% with a two-tailed t-test. The base classifiers are Gentle Adaboost with 50 runs of decision stumps and Linear Support Vector Machines ($SVM$) with the regularization parameter $C$ set to one.

### 3.1 UCI classification

In this experiment, from exactly the same set of generated matrices, we selected the classical Sparse Random matrix by the one which maximizes $d_r$ and $d_c$, and the new Sparse Random matrix by selecting the one which maximizes $d_t$ and $d_c$. To show the performance improvements by selecting the new Sparse Random matrix, the absolute and relative improvements using the obtained performances are shown in fig. 2 for Gentle Adaboost and Linear $SVM$, respectively. The light bars correspond to the absolute improvement, and the dark lines to the relative one. Note that simply changing the decision on the selection of the sparse matrix from the same set of generated random matrices, the performance significantly increases independently of the decoding strategy applied.

### 3.2 Real multi-class traffic sign categorization

For this experiment, we use the video sequences obtained from a Mobile Mapping System [3] to test

**Table 1.** UCI repository data sets characteristics.

| Problem | Train | Features | Classes | Problem | Train | Features | Classes |
|---|---|---|---|---|---|---|---|
| Dermat. | 366 | 34 | 6 | OptDigits | 5620 | 64 | 10 |
| Iris | 150 | 4 | 3 | Shuttle | 14500 | 9 | 7 |
| Ecoli | 336 | 8 | 8 | Vehicle | 846 | 18 | 4 |
| Wine | 178 | 13 | 3 | Segment. | 2310 | 19 | 7 |
| Glass | 214 | 9 | 7 | Pendigits | 10992 | 16 | 10 |
| Thyroid | 215 | 5 | 3 | Letter | 20000 | 16 | 26 |
| Vowel | 990 | 10 | 11 | Satimage | 6435 | 36 | 7 |
| Balance | 625 | 4 | 3 | Yeast | 1484 | 8 | 10 |

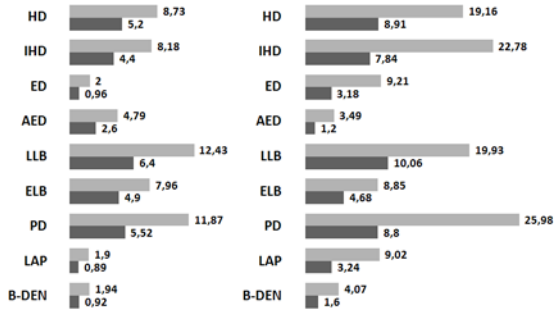


**Figure 2.** Absolute (light lines) and relative (dark lines) improvements for the Sparse Random designs using ternary distance maximization for Gentle Adaboost (left) and Linear $SVM$ (right) on the UCI experiments, respectively.

the methods in a real categorization problem. Figure 3 shows examples of video sequences and samples of the speed data set used in the experiments. The data set contains a total of 2500 samples divided in nine classes. Each sample is composed by 1200 pixel-based features after smoothing the image and applying histogram equalization. The results of this experiment are shown in fig. 4 using the same criteria. The best performance was obtained by the new Sparse design with $\beta$-Density decoding, with an accuracy upon 80%, meanwhile the traditional Sparse design obtained results inferior to 70%. Moreover, one can see in fig. 4 that the ternary sparse maximization criterion also obtains performance improvements for all decoding strategies.

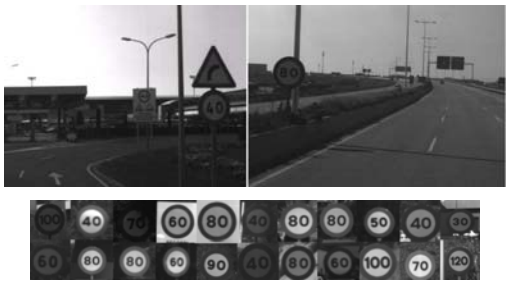

**Figure 3.** Samples from the road video sequences and speed data set samples.

## 4 Conclusions

We introduced a new formulation of the ternary distance that defines the classes separability in the ternary ECOC framework. We showed that the rows separability in terms of the Hamming distance of the binary ECOC framework can not be applied in the ternary case. Based on the new measure, a new Sparse Random construction is presented. The results on a wide set of UCI data sets and in a real speed traffic sign categorization problem show that when the coding designs satisfy the new ternary measures, significant performance improvements are obtained independently of the decoding strategy applied.

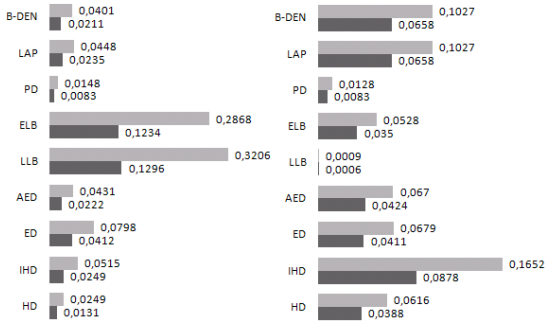


**Figure 4.** Absolute (light lines) and relative (dark lines) improvement for the Sparse Random designs using ternary distance maximization for Gentle Adaboost (left) and Linear $SVM$ (right) on the traffic sign categorization experiment, respectively.


## Acknowledgments

This work has been supported in part by projects TIN2006-15308-C02, FIS PI061290, and CONSOLIDER-INGENIO CSD 2007-00018.


## References


[1] E. Allwein, R. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In *JMLR*, volume 1, pages 113–141, 2000.

[2] A. Asuncion and D. Newman. UCI machine learning repository. In *University of California, Irvine, School of Information and Computer Sciences*, 2007.

[3] J. Casacuberta, J. Miranda, M. Pla, S. Sanchez, A.Serra, and J.Talaya. On the accuracy and performance of the geomobil system. In *International Society for Photogrammetry and Remote Sensing*, 2004.

[4] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. In *Journal of Artificial Intelligence Research*, volume 2, pages 263–282, 1995.

[5] S. Escalera, O. Pujol, and P. Radeva. Decoding of ternary error correcting output codes. In *CIARP*, volume 4225, pages 753–763, 2006.

[6] T. Hastie and R. Tibshirani. Classification by pairwise grouping. In *The annals of statistics*, volume 26, pages 451–471, 1998.

[7] A. Passerini, M. Pontil, and P. Frasconi. New results on error correcting output codes of kernel machines. In *IEEE Transactions on Neural Networks*, volume 15(1), pages 45–54, 2004.

[8] T. Windeatt and R. Ghaderi. Coding and decoding strategies for multi-class learning problems. In *Information Fusion*, volume 4, pages 11–21, 2003.