

Error-Correcting Output Coding for Chagasic Patients Characterization

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

The Chagas' disease is endemic in all Latin America, affecting millions of people in the continent. In order to diagnose and treat the chagas' disease, it is important to detect and measure the coronary damage of the patient. In this paper, we analyze and categorize patients into different groups based on the coronary damage produced by the disease. Based on the features of the heart cycle extracted using high resolution ECG, a multi-class scheme of Error-Correcting Output Codes (ECOC) is formulated and successfully applied. The results show that the proposed scheme obtains significant performance improvements compared to previous works and state-of-the-art ECOC designs.

1 Introduction

Chagas' disease is an infectious illness caused by the parasite *Tripanosoma Cruzi*, which is transmitted to humans through the feces of a bug called *Triatoma infestans*. The World Health Organization (WHO) estimates that 16 to 18 million people in Latin American countries are already infected by the disease and other 100 million people are at risk of being infected [5]. In areas where the illness is endemic, Chagas' cardiomyopathy represents the first cause of cardiovascular death [4].

In order to optimize treatment for chronic chagasic patients, it is essential to make use of an effective diagnosis tool able to determine the existence of cardiac injury and, if positive, its magnitude. Since Chagas' cardiomyopathy frequently leads to alterations in the heart's electrical conduction, recently it has been proposed the slopes of QRS complex in order to determine the myocardial damage associated with the disease [7]. Based on the temporal indices and slopes of QRS complex as extracted features, an automatic system that categorized patients into different groups is presented. To perform a multi-classification system able to learn the level of damage produced by the disease, we focus on a new design of Error-Correcting Output Codes [2].

The ECOC technique can be broken down into two

distinct stages: encoding and decoding. Given a set of classes, the coding stage designs a codeword¹ for each class based on different binary problems. The decoding stage makes a classification decision for a given test sample based on the value of the output code. Though many coding designs have been proposed to codify an ECOC coding matrix, obtaining successful results [3], the use of a proper decoding strategy is still an open issue. In this paper, we propose the Loss-Weighted decoding strategy, which exploits the information provided at the coding stage to perform a successful classification. As a result, our system automatically diagnoses the level of coronary damage of patients with the Chagas' disease, outperforming state-of-the-art ECOC designs and related works.

The paper is organized as follows: Section 2 overviews the ECOC framework, section 3 presents the Loss-Weighted decoding strategy, section 4 shows the design of the data set and the experimental results, and finally, section 5 concludes the paper.

2 Error-Correcting Output Codes

Given a set of N_c classes to be learnt, at the coding step of the ECOC framework, n different bi-partitions (groups of classes) are formed, and n binary problems (dichotomies) are trained. As a result, a codeword of length n is obtained for each class, where each bin of the code corresponds to a response of a given dichotomy. Arranging the codewords as rows of a matrix, we define a "coding matrix" M , where $M \in \{-1, 0, 1\}^{N_c \times n}$ in the ternary case. Joining classes in sets, each dichotomy, that defined a partition of classes, codes by $\{+1, -1\}$ according to their class set membership, or 0 if the class is not considered by the dichotomy. In fig.1 we show an example of a ternary matrix M . The matrix is coded using 7 dichotomies $\{h_1, \dots, h_7\}$ for a four class problem (c_1, c_2, c_3 , and c_4). The white regions are coded by 1 (considered as positive for its respective di-

¹The codeword is a sequence of bits of a code representing each class, where each bit identifies the membership of the class for a given binary classifier.

chotomy, h_i), the dark regions by -1 (considered as negative), and the grey regions correspond to the zero symbol (not considered classes by the current dichotomy). For example, the first classifier (h_1) is trained to discriminate c_3 versus c_1 and c_2 ignoring c_1 , and so on.

During the decoding process, applying the n trained binary classifiers, a code x is obtained for each data point in the test set. This code is compared to the base codewords of each class $\{y_1, \dots, y_4\}$ defined in the matrix M , and the data point is assigned to the class with the "closest" codeword [1].

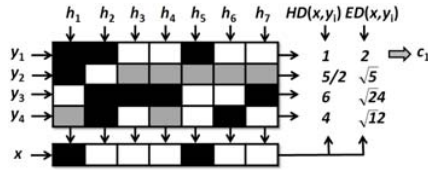


Figure 1. Example of ternary matrix M for a 4-class problem. A new test codeword is classified by class c_1 when using the traditional Hamming and Euclidean decoding strategies.

2.1 Decoding designs

The decoding step decides the final category of an input test by comparing the codewords. In this way, a robust decoding strategy is required to obtain accurate results. Several techniques for the binary decoding step have been proposed in the literature, the most common ones are the Hamming (HD) and the Euclidean (ED) approaches [1]. In fig.1, a new test input x is evaluated by all the classifiers and the method assigns label c_1 with the closest decoding distances. Note that in the particular example of fig. 1 both distances agree.

In [1], the authors propose a Loss-based technique when a confidence on the classifier output is available. For each row of M and each data sample φ , the authors compute the similarity between $f^j(\varphi)$ and $M(i, j)$, where f^j is the j^{th} dichotomy of the set of hypothesis F , considering a loss estimation on their scalar product, as follows: $D(\varphi, y_i) = \sum_{j=1}^n L(M(i, j) \cdot f^j(\varphi))$, where L is a loss function that depends on the nature of the binary classifier. The most common loss functions are the linear and the exponential one. The final decision is achieved by assigning a label to example φ according to the class c_i with the minimal distance.

Recently, the authors of [6] proposed a probabilistic decoding strategy based on the margin of the output of the classifier to deal with the ternary decoding. The decoding measure is given by $D(y_i, F) = -\log \left(\prod_{j \in [1, \dots, n]: M(i, j) \neq 0} P(x^j = M(i, j) | f^j) + \alpha \right)$, where α is a constant factor that collects the probability mass dispersed on the invalid codes, and the probability $P(x^j = M(i, j) | f^j)$ is estimated by means of

$P(x^j = y_i^j | f^j) = \frac{1}{1 + \exp(y_i^j (A^j f^j + B^j))}$, where vectors A and B are obtained by solving an optimization problem [6].

3 Loss-Weighted decoding (LW)

The ternary symbol-base ECOC allows to increase the number of bi-partitions of classes (thus, the number of possible binary classifiers) to be considered, resulting in a higher number of binary problems to be learnt. However, the effect of the ternary symbol is still an open issue. Since a zero symbol means that the corresponding classifier is not trained on a certain class, to consider the "decision" of this classifier on those zero coded position does not make sense. Moreover, the response of the classifier on a test sample will always be different to 0, so it will register an error. Let return to fig. 1, where an example about the effect of the 0 symbol is shown. The classification result using the Hamming distance as well as the Euclidean distance is class c_1 . On the other hand, class c_2 has only coded first both positions, thus it is the only information provided about class c_2 . The first two coded locations of the test codeword x correspond exactly to these positions. Note that each position of the codeword coded by 0 means that both -1 and +1 values are possible. Hence the correct classification should be class c_2 instead of c_1 . The use of standard decoding techniques that do not consider the effect of the third symbol (zero) frequently fails. In the figure, the HD and ED strategies accumulate an error value proportional to the number of zero symbols by row, and finally miss-classify the sample x .

<p>Given a coding matrix M,</p> <p>1) Calculate the matrix of hypothesis H:</p> $H(i, j) = \frac{1}{m_i} \sum_{k=1}^{m_i} \gamma(h_j(\varphi_k^i), i, j) \quad (1)$ <p>based on $\gamma(x_j, i, j) = \begin{cases} 1, & \text{if } x_j = M(i, j) \\ 0, & \text{otherwise.} \end{cases} \quad (2)$</p> <p>2) Normalize H so that $\sum_{j=1}^n M_W(i, j) = 1, \forall i = 1, \dots, N_c$:</p> $M_W(i, j) = \frac{H(i, j)}{\sum_{j=1}^n H(i, j)},$ <p>$\forall i \in [1, \dots, N_c], \forall j \in [1, \dots, n]$</p> <p>Given a test input φ, decode based on:</p> $d(\varphi, i) = \sum_{j=1}^n M_W(i, j) L(M(i, j) \cdot f(\varphi, j)) \quad (3)$

Table 1. Loss-Weighted algorithm.

To solve the commented problems, we propose a Loss-Weighted decoding. The main objective is to find

a weighting matrix M_W that weights a loss function to adjust the decisions of the classifiers, either in the binary and in the ternary ECOC frameworks. To obtain the weighting matrix M_W , we assign to each position (i, j) of the matrix of hypothesis H a continuous value that corresponds to the accuracy of the dichotomy h_j classifying the samples of class i (1). We make H to have zero probability at those positions corresponding to unconsidered classes (2), since these positions do not have representative information. The next step is to normalize each row of the matrix H so that M_W can be considered as a discrete probability density function (3). In fig. 2 a weighting matrix M_W for a 3-class problem with four hypothesis is estimated. Figure 2(a) shows the coding matrix M . The matrix H of fig. 2(b) represents the accuracy of the hypothesis classifying the instances of the training set. The normalization of H results in the weighting matrix M_W of fig. 2(c)².

$$\begin{array}{cc}
 M = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 1 & -1 \end{bmatrix} & H = \begin{bmatrix} 0.955 & 0.955 & 1.000 & 0.000 \\ 0.900 & 0.800 & 0.000 & 0.000 \\ 1.000 & 0.905 & 0.805 & 0.805 \end{bmatrix} \\
 \text{(a)} & \text{(b)} \\
 M_W = \begin{bmatrix} 0.328 & 0.328 & 0.344 & 0.000 \\ 0.529 & 0.471 & 0.000 & 0.000 \\ 0.285 & 0.257 & 0.229 & 0.229 \end{bmatrix} & \\
 & \text{(c)}
 \end{array}$$

Figure 2. (a) Coding matrix M of four hypotheses for a 3-class problem. (b) Matrix H of hypothesis accuracy. (c) Weighting matrix.

The Loss-weighted algorithm is shown in table 1. As commented before, possible loss functions applied in equation (3) are be the linear or the exponential ones. The linear function is defined by $L(\theta) = \theta$, and the exponential loss function by $L(\theta) = e^{-\theta}$, where in our case θ corresponds to $M(i, j) \cdot f^j(\varphi)$. Function $f^j(\varphi)$ may return either the binary label or the confidence value of applying the j^{th} classifier to sample φ .

4 Results

Before the experimental results are presented, we comment the data, methods, and evaluation measurements.

- **Data:** In this work, we analyzed a population composed of 107 individuals from the Chagas data set recorded at Simón Bolívar University (Venezuela). For each individual, a continuous 10-minute HRECG was recorded using orthogonal XYZ lead configuration. All the recordings were digitalized with a sampling frequency of 1 kHz and amplitude resolution of 16 bits.

²Note that the presented Weighting Matrix M_W can also be applied over any existing decoding strategy.

Out of the total 107 individuals of the study population, 96 are chagasic patients with positive serology for *Trypanosoma Cruzy*, clinically classified into three different groups according on their degree of cardiac damage (Groups I, II, and III). This grouping is based on the clinical history, Machado-Guerreiro test, conventional ECG of twelve derivations, Holter ECG of 24 hours, and myocardiograph study for each patient. The other 11 individuals are healthy subjects with negative serology taken as a control group (Group 0). All individuals of the data set are described with a features vector of 16 features based on [7]. The four analyzed groups are described in detail next:

- Group 0: 11 healthy subjects in the age 33.6 ± 10.9 years, 9 men and 2 women.
- Group I: 41 total patients with the Chagas' disease in the age of 41.4 ± 8.1 years, 21 men and 20 women, but without evidences of cardiac damage in cardiographic study.
- Group II: 39 total patients with the Chagas' disease in the age of 45.8 ± 8.8 years, 19 men and 20 women, with normal cardiographic study and some evidences of weak or moderate cardiac damage registered in the conventional ECG or in the Holter ECG of 24 hours.
- Group III: 16 total patients with the Chagas' disease in the age of 53.6 ± 9.3 years, 9 men and 7 women, with significant evidences of cardiac damage detected in the conventional ECG, premature ventricular contractions and/or cases of ventricular tachycardiac registered in the Holter ECG and reduced fraction of ejection estimated in the cardiographic study.

- **Methods:** We compare our results with the performances reported in [7] for the same data. Moreover, we compare different ECOC designs: the one-versus-one ECOC coding strategy [8] applied with the Hamming [1], Euclidean [1], Probabilistic [6], and the presented Loss-Weighted decoding strategies. We selected the one-versus-one ECOC coding strategy because the individual classifiers are usually smaller in size than they would be in the rest of ECOC approaches, and the problems to be learnt are usually easier, since the classes have less overlap. Each ECOC configuration is evaluated for three different base classifiers: Fisher Linear Discriminant Analysis (*FLDA*) with a previous 99.9% of Principal Components, Discrete Adaboost with 50 runs of Decision Stumps, and Linear Support Vector Machines with the regularization parameter C set to 1.

- **Evaluation measurements:** To evaluate the methodology we apply leave-one-patient-out classification on the Chagas data set.

The results of categorization for the four groups of patients reported by [7] are shown in fig. 4. Considering the number of patients from each group, the mean classification accuracy of [7] is of 57%. The results us-

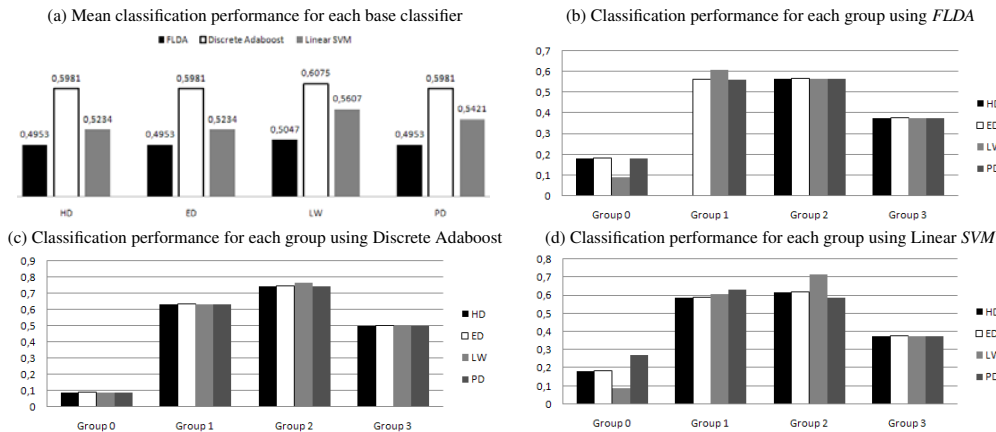


Figure 3. Leave-one-patient-out classification using one-versus-one ECOC design (HD: Hamming decoding, ED: Euclidean decoding, LW: Loss-Weighted decoding, PD: Probabilistic decoding) for the four groups with and without Chagas' disease.

ing the different ECOC configurations for the same four groups are shown in fig. 3. In fig. 3(a), the mean accuracy for each base classifier and decoding strategy is shown. The individual performances of each group of patients for each base classifier are shown in fig. 3(b), fig. 3(c), and fig. 3(d), respectively. Observing the mean results of fig. 3(a), one can see that any ECOC configuration outperforms the results reported by [7]. Moreover, even if we use *FLDA*, Discrete Adaboost, or Linear *SVM* in the one-vs-one ECOC design, the best performance is always obtained with the proposed Loss-Weighted decoding strategy. In particular, the one-versus-one ECOC coding with Discrete Adaboost as the base classifier and Loss-Weighted decoding attains the best performance, with a classification accuracy upon 65% considering the four groups of patients.

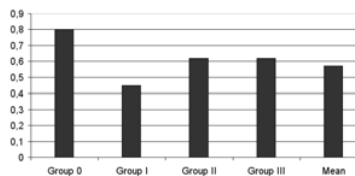


Figure 4. Classification performance reported by [7] for the four groups of patients.

5 Conclusions

In this paper, we characterized patients with the Chagas' disease based on the coronary damage produced by the disease. We used the features extracted using the ECG of high resolution from the heart cycle of 107 patients, and presented a decoding strategy of Error-Correcting Output Codes to learn a multi-class system. The results show that the proposed scheme outperforms previous works characterizing patients with different coronary damage produced by the Chagas'

disease (upon 10% performance improvements), at the same time that it achieves better results compared with the state-of-the-art ECOC designs for different base classifiers.

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, 2002.
- [2] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–282, 1995.
- [3] S. Escalera, O. Pujol, and P. Radeva. Boosted landmarks of contextual descriptors and forest-ecoc: A novel framework to detect and classify objects in clutter scenes. *Pattern Recognition Letters*, 28(13):1759–1768, 2007.
- [4] J. H. Maguire, R. Hoff, I. Sherlock, A. C. Guimaraes, A. C. Sleight, N. B. Ramos, K. E. Mott, and T. H. Séller. Cardiac morbidity and mortality due to chagas disease: prospective electrocardiographic study of a brazilian community circulation. 75:1140–1145, 1987.
- [5] W. D. of Control of Tropical Diseases. Chagas disease elimination. burden and trends. *WHO web site www.who.int/ctd/html/chagburtre.html*.
- [6] A. Passerini, M. Pontil, and P. Frasconi. New results on error correcting output codes of kernel machines. *IEEE Transactions on Neural Networks*, 15(1):45–54, 2004.
- [7] E. Pueyo, E. Anzuola, E. Laciari, P. Laguna, and R. Jané. Evaluation of QRS slopes for determination of myocardial damage in chronic chagasic patients. *Computers in Cardiology*, 2007.
- [8] T. Hastie and R. Tibshirani. Classification by pairwise grouping. *NIPS*, 26:451–471, 1998.