Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 26 (2005) 83–90

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Using diversity measures for generating error-correcting output codes in classifier ensembles

Ludmila I. Kuncheva *

School of Informatics, University of Wales Bangor, Dean Street, Bangor, Gwynedd LL57 1UT, United Kingdom

Received 5 July 2004
Available online 14 October 2004

Abstract

Error-correcting output codes (ECOC) are used to design diverse classifier ensembles. Diversity within ECOC is traditionally measured by Hamming distance. Here we argue that this measure is insufficient for assessing the quality of code for the purposes of building accurate ensembles. We propose to use diversity measures from the literature on classifier ensembles and suggest an evolutionary algorithm to construct the code.

Keywords: Statistical pattern recognition; Classifier ensembles; Diversity measures; Error-correcting output codes (ECOC); Evolutionary algorithms

1. Introduction

ECOC are developed for pattern recognition problems with multiple classes (Aha and Blankert, 1997; Dietterich and Bakiri, 1991, 1995; Masulli and Valentini, 2000a,b; Shapire, 1997; Windeatt and Ghaderi, 2001, 2003). The idea is to avoid solving the multiclass problem directly and to break it into dichotomies instead. Each classifier in the ensemble discriminates between two (possibly compound) classes. Consider an example where $\Omega = \{\omega_1, \ldots, \omega_{10}\}$ is the set of class labels. We can break Ω into $\Omega = \{\Omega^{(1)}, \Omega^{(0)}\}$ where $\Omega^{(1)} = \{\omega_1, \ldots, \omega_5\}$ and $\Omega^{(0)} = \{\omega_6, \ldots, \omega_{10}\}$, called *a dichotomy*. Discriminating between $\Omega^{(1)}$ and $\Omega^{(0)}$ will be the task of one of the classifiers in the ensemble. Each classifier is assigned a different dichotomy.

Diversity between the classifiers is a highly desirable characteristic of the ensemble. The presumption in using ECOC is that diverse classifiers are obtained from diverse dichotomies. While minimum Hamming distance is the traditional measure for diversity in ECOC, in this paper we propose to use diversity measures originally

* Tel.: +44 1248 383661; fax: +44 1248 361429.
E-mail address: l.i.kuncheva@bangor.ac.uk

0167-8655/$ - see front matter
doi:10.1016/j.patrec.2004.08.019

devised for classifier outputs. Minimum Hamming distance guarantees the error-correcting capability of the code. However, we may wish to compromise on this guarantee in order to get a more diverse ensemble *on the average*, as explained later by an example. This idea brings in diversity measures used in classifier combination (Kuncheva and Whitaker, 2003).

The paper is organized as follows. Section 2 explains ECOC and gives some code generating methods. While minimum Hamming distance is a traditional measure for the error-correcting quality of a code, Section 3 looks into the need for another measure of the quality of the code when used for classifier ensembles. Section 4 suggests an application of diversity measures to evaluating ECOC. Section 5 proposes an evolutionary algorithm for ECOC construction, using a diversity measure as its fitness function. Section 6 gives our comments and conclusions.

2. Error-correcting output codes (ECOC)

Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of class labels. Suppose that each classifier codes the respective compound class $\Omega^{(1)}$ as 1 and compound class $\Omega^{(0)}$ as 0. Then every class $\omega_j, j = 1, \ldots, c$, will have a binary "profile" or a *codeword*. For example, suppose that there are 5 classifiers. A possible class profile (codeword) for ω_1 is $[0, 1, 1, 0, 1]^T$. This means that ω_1 is in the respective $\Omega^{(1)}$ sets for classifiers D_2, D_3, and D_5 and in the respective $\Omega^{(0)}$ sets for classifiers D_1 and D_4.

2.1. The code matrix

We can represent each dichotomy as a binary vector of length c with 1's for the classes in $\Omega^{(1)}$ and 0's for the classes in $\Omega^{(0)}$. The set of all such vectors has 2^c elements. However, not all of them correspond to different splits. Consider $[0, 1, 1, 0, 1]^T$ and $[1, 0, 0, 1, 0]^T$. Even though the Hamming distance between the two binary vectors is equal to the maximum possible value, 5, the two subsets are identical, only with swapped labels. Since there are two copies of each split within the total of 2^c splits, the number of different splits is $2^{(c-1)}$. The splits $\{\Omega, \emptyset\}$ and the corresponding $\{\emptyset, \Omega\}$ are of no use because they do not represent any discrimination task. Therefore the number of possible *different* splits of a set of c class labels into two non-empty disjoint subsets (dichotomies) is $2^{(c-1)} - 1$.

Let L be the chosen number of classifiers in the ensemble. The class assignments for the ensemble (the dichotomies) can be represented as a binary *code matrix* C of size $c \times L$. The (i, j)th entry of C, denoted $C(i, j)$ is 1 if class ω_i is in $\Omega_j^{(1)}$ or 0, if class ω_i is in $\Omega_j^{(0)}$. Thus each row of the code matrix is a codeword and each column is a classifier assignment. An example of a code matrix for $c = 4$ classes with all possible $2^{(4-1)} - 1 = 7$ different dichotomies is shown in Table 1.

Let $[s_1, \ldots, s_L]$, $s_i \in \{0, 1\}$, be the binary output of the L classifiers in the ensemble for a given input $\mathbf{x}$. The Hamming distance between the classifier outputs and the codewords for the classes is calculated as $\sum_{i=1}^{L} |s_i - C(j, i)|$. In the standard set-up the input is labeled in the class with the smallest distance (decoding phase). Ties are broken randomly. A more sophisticated decoding strategies are discussed by Windeatt and Ghaderi (2001, 2003).

To take the most advantage of an ECOC ensemble, the code matrix should be built according to two main criteria.

Row separation. In order to avoid misclassifications, the codewords should be as far apart from one another as possible. We can still recover the correct label for $\mathbf{x}$ even if several classifiers have guessed wrongly. A measure of the quality of an error-correcting code is the minimum Hamming distance, H_c, between any pair of codewords. The number of errors that the code is guaranteed to be able to correct is $\lfloor \frac{H_c - 1}{2} \rfloor$. (Here $\lfloor a \rfloor$ denotes the "floor", i.e., the nearest integer smaller than a.)

Table 1
Exhaustive ECOC for $c = 4$ classes ($L = 7$ classifiers)

	D_1	D_2	D_3	D_4	D_5	D_6	D_7
ω_1	0	0	0	1	0	1	1
ω_2	0	0	1	0	0	0	0
ω_3	0	1	0	0	1	0	1
ω_4	1	0	0	0	1	1	0