Engineering Applications of Artificial Intelligence 73 (2018) 10–21

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Evidential framework for Error Correcting Output Code classification

Marie Lachaize [a,b], Sylvie Le Hégarat-Mascle [a,*], Emanuel Aldea [a], Aude Maitrot [b], Roger Reynaud [a]

[a] SATIE laboratory, Université Paris-Sud, Université Paris-Saclay, 91405 Orsay, France
[b] VEOLIA RECHERCHE & INNOVATION, 291 av. Dreyfous Ducas, Limay, France

ARTICLE INFO

Keywords:
Classification
Error Coding Output Codes
Belief Function Theory
Hyperspectral data

ABSTRACT

The Error Correcting Output Codes offer a proper matrix framework to model the decomposition of a multiclass classification problem into simpler subproblems. How to perform the decomposition to best fit the data while using a small number of classifiers has been a research hotspot, as well as the decoding part, which deals with the subproblem combination. In this work, we propose an evidential unified framework that handles both the coding and decoding steps. Using the Belief Function Theory, we propose an efficient modelling, where each dichotomizer in the ECOC strategy is considered as an independent information source. This framework allows us to easily model the refutation information provided by sparse dichotomizers and also to derive measures to detect tricky samples for which additional dichotomizers could be needed to ensure decisions. Our approach was tested on hyperspectral data used to classify nine different types of material. According to the results obtained, our approach allows us to achieve top performance using compact ECOC while presenting a high level of modularity.

1. Introduction

Automatic multiclass image classification is a major topic in pattern recognition in computer vision and numerous methods have already been proposed, e.g. Geman and Geman (1987), Boser et al. (1992), Crammer and Singer (2002), Wang et al. (2010) and Krizhevsky et al. (2012). With regard to the complexity of some types of data (e.g. hyperspectral data images) and the increasing number of classes (e.g. for applications requiring finer and finer classes), the 'Divide and Conquer' strategy has been proposed (Brassard and Bratley, 1996). This strategy consists of splitting the multiclass problem in a set of binary classification problems simpler to solve. Following such a strategy, the Error Correcting Output Codes (Dietterich and Bakiri, 1995; Allwein et al., 2000) have been designed to address both involved problems of decomposition of the multiclass problem and interpretation of binary classification outputs. For instance, the one-versus-one (OVO) and one-versus-all (OVA) strategies (Hastie and Tibshirani, 1998; Rifkin and Klautau, 2004) are specific ECOC. More generally, given a set of classes Ω of cardinality N, an ECOC matrix $\boldsymbol{M}$ of size $N \times l$ with values in $\{-1, 0, 1\}$ corresponds to a decomposition of the multiclass problem in l binary problems called dichotomizers. Each dichotomizer, coded by one $\boldsymbol{M}$ column, aims at classifying any given sample between two non overlapping subsets of classes. If the two class subsets form a partition of Ω, the dichotomizer is said to be dense and $\boldsymbol{M}_{ij} \in \{-1, 1\}$, where 1 and -1 designate the opposing classes. Otherwise, it is said to be sparse and $\boldsymbol{M}_{ij} \in \{-1, 0, 1\}$, where 0 designates the classes that are not involved in the classifier training. Now, ECOC research still includes open-ended questions either for coding (i.e. defining $\boldsymbol{M}$) or for decoding (i.e. assigning class label according to $\boldsymbol{M}$ answers), e.g. Bai et al. (2016), Santhanam et al. (2016), Xu et al. (2016) and Bautista et al. (in press).

1.1. ECOC coding related work

Concerning coding, initial methods such as Allwein et al. (2000) only consider constraints on $\boldsymbol{M}$: size, type of dichotomizers and distance between $\boldsymbol{M}$ rows, i.e. class codeword. However, using this approach, the number of dichotomizers remains an *a priori* parameter difficult to set and this predetermined behaviour does not allow us to take into account the dichotomizer's specific performance.

Alternatively, performance-driven methods have been proposed. For example, Bai et al. (2016) assesses the performance of every dichotomizer (among the whole set of potential dichotomizers given the set of classes) and builds $\boldsymbol{M}$ by favouring dichotomizers exhibiting the highest performance. However, besides being computationally very expensive, such an approach fails to provide some redundancy where

* Corresponding author.
E-mail addresses: marie.lachaize@veolia.com (M. Lachaize), sylvie.le-hegarat@u-psud.fr (S.L. Hégarat-Mascle), emanuel.aldea@u-psud.fr (E. Aldea), aude.maitrot@veolia.com (A. Maitrot), roger.reynaud@u-psud.fr (R. Reynaud).

https://doi.org/10.1016/j.engappai.2018.04.019
Received 18 October 2017; Received in revised form 19 March 2018; Accepted 16 April 2018
Available online 7 May 2018

it is the most needed, namely in order to separate close or ambiguous classes.

Then, to address this point, data-driven approaches have been proposed. The data are analysed to understand which classes are difficult to separate and to infer the ECOC matrix optimizing their separation. Among the criteria to analyse the data, the use of a pre-computed confusion matrix is rather popular, e.g. Escalera et al. (2008), Gao and Koller (2011) and Zhou et al. (2016), whereas Pujol et al. (2006) considers the mutual information within the dichotomizer sets. For the construction of ECOC, some hierarchical constraints are often introduced, i.e. starting from easily distinguishable superclasses and adding dichotomizers to distinguish classes within these superclasses (Pujol et al., 2006; Zhou et al., 2016), or conversely (Escalera et al., 2008). For instance, Gao and Koller (2011) proposes a joint optimization process to learn a hierarchy of classifiers in which each node corresponds to a binary subproblem. Nonetheless, although the hierarchical configuration speeds up the testing step, it is highly prone to error propagation. Some other data-driven approaches explicitly focus on removing ambiguities between similar classes. In Pujol et al. (2008), the ECOC matrix is iteratively constructed as follows: at each iteration, the pair of the most confused classes is derived from the current confusion matrix and the ECOC matrix is extended with new dichotomizers that both separate the ambiguous classes and that show good performance. In Bautista et al. (in press), by factorizing the confusion matrix, a dense ECOC matrix is generated so that the ambiguous classes have distant codewords. Finally, note that all cited data-driven ECOC matrix design solutions rely on learning data, which may make their results prone to errors when faced with unexpected class ambiguities.

1.2. ECOC decoding related work

The simplest decoding is the minimization of the Hamming distance, (Nilsson, 1965), based on the binary decisions of the dichotomizers. Then, the loss-based decoding, (Allwein et al., 2000), has been proposed to take into account the confidence levels associated with binary decisions, according to the considered loss function and a calibration process of the dichotomizer outputs or scores. If these approaches have shown to be efficient for dense ECOC, they come up against modelling the ambiguity introduced by the absent classes in the sparse classifiers (0 values in $\boldsymbol{M}$). In the Hamming and classic loss-based decoding, any answer of a 0-valued class is considered irrelevant and a fixed weight is assigned. However, as underlined by Pujol et al. (2008), this fixed weight creates a bias when there is an imbalance among the classes involved in sparse classifiers. Therefore Escalera et al. (2010) proposed a new ternary decoding method that is robust to this bias. However, Escalera et al. (2010) still misses the opportunity to exploit additional information from the 0-valued class answers, e.g. in terms of refutation of some classes, as we propose in this work using belief functions.

1.3. Belief function related work

The evidential framework was initially defined by A. Dempster and G. Shafer (Shafer, 1976), while Ph. Smets proposed his interpretation in terms of belief transfer, (Smets and Kennes, 1994). This theory has been widely used to model different kinds of uncertainty in classification problems (e.g. Hegarat-Mascle et al., 1997; Tabassian et al., 2012; Liu et al., 2014), detection and recognition (e.g. Xu et al., 1992; Mercier et al., 2009), tracking (e.g. Smets and Ristic, 2007; André et al., 2015), object reconstruction (e.g. Díaz-Más et al., 2010; Rekik et al., 2016) and localization (e.g. Roquel et al., 2014) etc. A major strength of belief function theory is that it avoids introducing bias in cases of partial ignorance (conversely to an equiprobability assumption or the mentioned fixed cost). This makes it all the more important that different sources of information are combined, sources that may correspond to different classifier outputs when dealing with a classification problem. In this case, the basic belief assignment (or bba) allocation step also handles the calibration process of the classifier outputs. Now, numerous bba allocation methods, among the ones already proposed, are actually data-driven approaches. For example, Xu et al. (1992) proposes a method to build bbas for a classifier using the recognition rate, the substitution rate and the rejection rate derived from its confusion matrix; Parikh et al. (2001) considers the classifier's performance values for the different classes; and more recently, Deng et al. (2016) which aims at combining several multiclass classifiers, constructs a bba per multiclass classifier from its crisp outputs (labels) and learned precision–recall rates. Now, conversely to Deng et al. (2016), authors generally consider soft outputs and even Xu et al. (2016) proposes to take into account not only the dichotomizer score value itself but also the number of samples per score value by extending the classic probabilistic calibration methods such as the logistic regression to the belief function framework. Finally note that the final decoding depends on the interpretation of the dichotomizers bbas: either as independent information sources, or as proposed in Quost et al. (2007), as conditioned pieces of information (allowing, at least in the classic OVO and OVA cases, to recover the multiclass bba from an optimization problem).

In our approach, we propose a full ECOC strategy (coding and decoding) that takes advantage of the modelling ability of the belief function theory framework. For the decoding part, each dichotomizer answer will be modelled by a belief function assignment depending on both the confidence score and the parameters of the calibration process. The method we propose extends the work of Lachaize et al. (2016).

For the coding part, we use evidential indices such as conflict to dynamically extend any ECOC matrix in such a way as to identify and remove remaining ambiguities, rendering the proposed coding method auto-adaptive.

The paper is organized as follows: Section 2 introduces the belief function tools and notations used in this work. Section 3 explains the proposed evidential classification including the ECOC coding and decoding processes. Section 4 discusses the results obtained from experiments using hyperspectral data acquired for a material classification application. Section 5 gathers the conclusions and perspectives of this work.

2. Preliminaries on belief function theory (BFT)

In this section, we introduce the tools and notations used in this study. For a reader not familiar with BFT, we refer to the founding book, (Shafer, 1976).

2.1. Basic concepts

Let Ω denote the **discernment frame**, i.e. the set of mutually exclusive hypotheses representing the solution possibilities and let 2^Ω denote the power set of Ω, i.e. the set of subsets of Ω elements. 2^Ω cardinality is denoted $|2^\Omega|$ and it is equal to $2^{|\Omega|}$. A bba (basic belief assignment) is defined through its **mass function** m such that: $m : 2^\Omega \rightarrow [0,1]$, $\sum_{A\in 2^\Omega} m(A) = 1$. If $m(A) > 0$, A is said to be a *focal element* and $m(A)$ represents the belief that the solution is in A, without having to specify the affiliation of the solution to any subset of A. In the following, we denote by $\mathcal{F}_m$ the set of focal elements of the bba m. Under the open world assumption, Ω may be non exhaustive and $\emptyset$ may be a focal element ($\emptyset \in \mathcal{F}_m$), with its mass representing the belief that the solution is not in Ω.

Refinement and **coarsening** are dual operators that allow some transformations of the discernment frame and its associated bbas. Specifically, let Θ and Ω be two discernment frames such that $|\Theta| < |\Omega|$. A refinement from Θ to Ω is defined by a function $\rho : \Theta \rightarrow 2^\Omega$ such that the set of the ρ images ($\{\rho(B), B \in \Theta\}$) is a partition of Ω, noted $\mathcal{P}_\rho(\Omega)$: $\forall A \in \mathcal{P}_\rho(\Omega), \exists! B \in \Theta \mid A = \rho(B)$. Then, specifying by a superscript on