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*

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.

Keywords: Classification, Error Coding Output Codes, Belief Function Theory, hyperspectral data

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); Krizhevsky et al. (2012). With regard to the complexity of some types of data (e.g. hyperspectral data images) and the increasing number

*Sylvie Le Hégarat-Mascle
Email addresses: marie.lachaize@u-psud.fr (Marie Lachaize), sylvie.le-hegarat@u-psud.fr (Sylvie Le Hégarat-Mascle), emanuel.aldea@u-psud.fr (Emanuel Aldea), aude.maitrot@veolia.fr (Aude Maitrot), roger.reynaud@u-psud.fr (Roger Reynaud)

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); Bautista et al. (2017).

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 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 analyzed to understand which classes are difficult to separate and to infer the ECOC matrix optimizing their separation. Among the criteria to analyze the data, the use of a pre-computed confusion matrix is rather popular, e.g. Escalera et al. (2008); Gao and Koller (2011); 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. (2017), 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. Le 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 modeled 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 $\left|2^\Omega\right|$ 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 $\rho : \Theta \to 2^\Omega$ 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 m the discernment frame, a bba initially defined on 2^Θ may be refined on 2^Ω using

$$\forall A \in 2^\Omega, m^\Omega(A) = \begin{cases} m^\Theta(\cup_{i=1}^n B_i) & \text{if } \exists \{B_1, \ldots, B_n\} \in \Theta^n \mid A = \cup_{i=1}^n \rho(B_i), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Essentially, if Θ is a set of superclasses of Ω classes, refinement boils down to mapping the elements of $\mathcal{F}_{m^\Theta}$ to disjunctions of Ω classes.

The **discounting** operator allows us to take into account prior knowledge about the bba reliability. Since $m^\Omega(\Omega)$ represents the fraction of ignorance carried by bba m^Ω, classical discounting Appriou (1997) modelling the global degree of reliability α of a bba boils down to increasing $m^\Omega(\Omega)$:

$$\begin{cases} m_\alpha^\Omega(A) &= \alpha \times m^\Omega(A), \forall A \in 2^\Omega \backslash \Omega, \\ m_\alpha^\Omega(\Omega) &= \alpha \times m^\Omega(\Omega) + 1 - \alpha. \end{cases} \tag{2}$$

The less reliable is the initial bba, the lower α, the more belief is transferred to Ω.

The **conjunctive combination rule** is the most widely used combination rule because of its simplicity, ability to specify the information and convenient mathematical properties (in particular commutativity and associativity). In the case of two independent bbas m_1^Ω and m_2^Ω defined on the same discernment frame, it can be written as follows:

$$\forall A \in 2^\Omega, m_{1\textcircled{\cap}2}^\Omega(A) = m_1^\Omega \textcircled{\cap} m_2^\Omega(A) = \sum_{\substack{(B,C) \in \mathcal{F}_{m_1} \times \mathcal{F}_{m_2}, \\ B \cap C = A}} m_1^\Omega(B) m_2^\Omega(C). \tag{3}$$

The **conditioning** operator allows us to take into account the certainty we have that the solution is in a given subset B of 2^Ω. It can be expressed as the combination of the initial bba m^Ω with the categorical bba $m_B^\Omega \mid m_B^\Omega(B) = 1$:

$$\forall A \in 2^\Omega, m^\Omega[B](A \cap B) = m^\Omega(A). \tag{4}$$

The **ballooning extension** is the inverse operator to conditioning. It is defined according to the minimal commitment principle Smets and Kennes (1994). If $\bar{B} = \Omega \setminus B$ is the complementary of B in Ω:

$$\begin{cases} \forall A \subseteq B \in 2^\Omega, & m^\Omega\left(A \cup \bar{B}\right) &= m^\Omega[B](A), \\ \forall A \subseteq \bar{B} \in 2^\Omega, & m^\Omega(A) &= 0. \end{cases} \tag{5}$$

The **decision** is generally taken based on a function that carries a probabilistic interpretation. The three most used functions are the plausibility, Pl, the credibility, Bel, and the pignistic probability, $BetP$. The first two are in one-to-one relationship with m^{Ω} and may be interpreted as upper and lower bounds of an imprecise probability function, whereas $BetP$ was defined in Smets and Kennes (1994) to have the same mathematical properties as a probability defined on Ω (provided that $m(\emptyset) < 1$).

$$\forall A \in 2^{\Omega}, Pl(A) = \sum_{B \in \mathcal{F}_{m^{\Omega}} | A \cup B \neq \emptyset} m^{\Omega}(B), \tag{6}$$

$$\forall A \in 2^{\Omega}, Bel(A) = \sum_{B \in \mathcal{F}_{m^{\Omega}} | B \subseteq A} m^{\Omega}(B), \tag{7}$$

$$\forall A \in \Omega, BetP(A) = \frac{1}{1 - m^{\Omega}(\emptyset)} \sum_{B \in \mathcal{F}_{m^{\Omega}} | A \in B} \frac{m^{\Omega}(B)}{|B|}. \tag{8}$$

3. Proposed approach

3.1. General scheme

The proposed approach arises from the following key points:

- Firstly, following data-driven approaches, we assume that some classes are more difficult to separate than others and that, the decisions in favour of one of them shall be robustifed. This can be achieved using well-chosen additional dichotomizers aimed to address these class ambiguities, either directly or indirectly.

- Secondly, sparse dichotomizers allow us to focus on subsets of classes (typically subsets involving the ambiguous classes) in a more flexible way than dense dichotomizers. Indeed, even if authors, e.g. in Rifkin and Klautau (2004); Pujol et al. (2008), argued in favour of dense dichotomizers because their frontiers may be more complex and interesting than those of sparse ones, those works also underlined the fact that establishing the right parameter tuning for dense classifiers is not easy, notably because of unbalanced class subsets in the training step.

- Thirdly, sparse classifier outputs shall be correctly interpreted. In particular, the *missing* classes (i.e. the classes not included in any of the two subsets of classes considered by a sparse dichotomizer) should not introduce a bias in the decoding process, Escalera et al. (2010).

- Fourthly, except in the case of performance-driven approaches, coding and decoding are not independent. For instance, the 'best' ECOC matrix $\boldsymbol{M}$ is estimated assuming a decoding criterion (generally the Hamming distance) and/or assuming the class ambiguities (generally from a confusion matrix that depends not only on the data but also on the classification algorithm).

The proposed solution was built to address these points. It is based on the interpretation of each dichotomizer as an individual source of information so that the multiclass classification problem is viewed as a fusion problem between imprecise and uncertain pieces of information (dichotomizer outputs). Focusing on belief function framework to handle both imprecision and uncertainty, ECOC decoding is carried out in three stages: bba allocation, bba combination and decision on the final bba. The bba allocation explicitly models the ambiguities within a subset of classes not distinguished by a dichotomizer as well as the ambiguities with sparse dichotomizer missing classes. Such modelling can be readily achieved by handling compound hypotheses representing the two subsets of classes corresponding to the dichotomizer's hypotheses. Then, the bba combination allows us to derive a single bba gathering all the dichotomizer's soft outputs. This bba brings us a wealth of information: the most likely class(es) and also the evidential measures of imprecision and conflict outcoming from the combination. In this work, we propose to exploit these measures for ECOC coding by determining it from the current ECOC decoding. Indeed, these measures provide us with valuable information on the remaining ambiguities between classes and with hints to choose the more useful dichotomizers to raise these ambiguities.

Figure 1 presents the general scheme of the proposed approach for a sample s. The first horizontal block represents the main steps of the decoding module, whereas the coding block (vertical) shows the extension and concatenation of the initial ECOC matrix with additional dichotomizers. For the first iteration, the initial ECOC M_1 is provided *a priori* (e.g., OVA) and the evidential decoding (cf. Section 3.2) is performed to get the bba m_1 that gathers the information of the soft outputs of the dichotomizers of M_1. From m_1, the first decision to take is whether to continue the iterative process or not. If the decision is to stop, then m_1 also allows us to decide the most likely class and to assign its label to the considered sample. Otherwise, as explained in Section 3.3, the complementary dichotomizer(s) able to remove the ambiguities are estimated and a new bba is derived from their soft outputs (by evidential decoding). This bba, called m_i at iteration i, is combined with the bba(s) obtained at previous iteration(s). Note that thanks to the associativity of the conjunctive combination rule (Eq. (3)), the iterative nature of the process (iterative combination of partial ECOC) has no impact on the obtained bba $m_{1 \textcircled{\cap} \ldots i}$.

Let us underline that the proposed approach differs from classic data-driven ones in the sense that it automatically adapts to the basis multiclass classifier (M_1) whereas, except Pujol et al. (2008) that extends an initial multiclass classifier (ECOC matrix), classic data-driven approaches assume that main class ambiguities subsist irrespective of the classifier and can be estimated independently.

To illustrate our algorithm, we will consider a classification problem from image data (specifically classification of kinds of material based on hyperspectral sensors). For this problem, the pixels of the image are the samples to be classified. Now, when the samples of two (or more) classes are close in the feature space, the classes are said *ambiguous* since the labels of their samples are

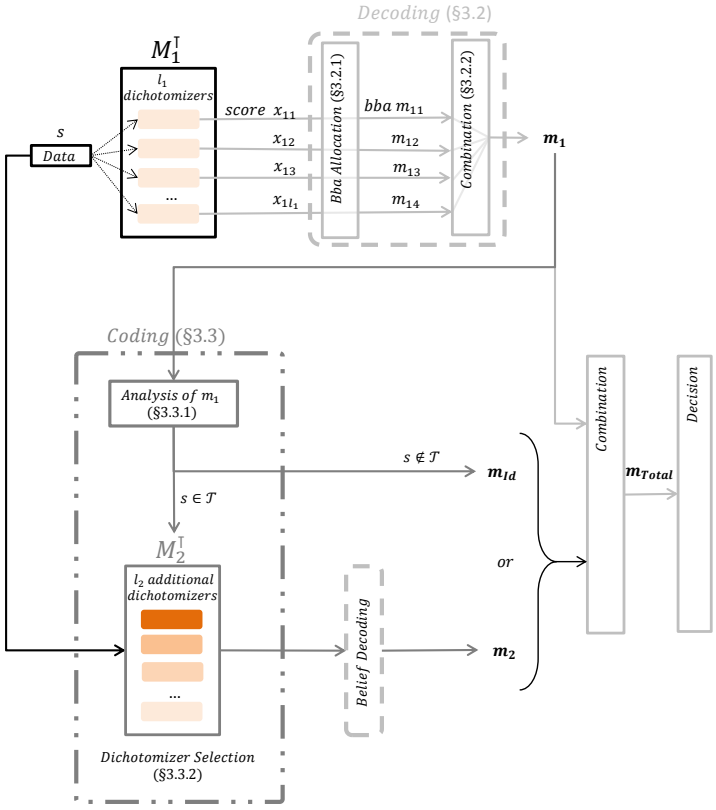


Figure 1: Two iterations of the general scheme of the proposed approach (Note that M_1 and M_2 are represented as transposed).

easily confused one for the other. These *ambiguous* classes generate samples for which the classification is *tricky*, typically involving unclear decisions between the ambiguous classes. In the following, by language extension, we call *tricky* such pixels.

3.2. Evidential decoding

Let us now specify the three announced stages of evidential decoding.

3.2.1. Bba allocation for dichotomizers

Bba allocation is often presented as a crucial issue since it contains the modelling of the source uncertainty, imprecision (or partial ambiguities/ignorance) and reliability. Among the numerous works already proposed, Xu et al. (2016) seems the closest to our purpose. It deals with the derivation of a bba from the soft output of a dichotomizer, i.e. its score for a given pixel. The relationship between a score value and the corresponding bba is obtained through a

calibration step that extends the logistic regression used to derive a probability from the score value to the derivation of a bba. Xu's calibration step takes into account the imprecision related to the number of samples observed for the considered score value so that it is all the more interesting as the number of samples is rather low and variable versus the score values, Lachaize et al. (2016).

Indeed, when the number of samples achieving a given score is important, the calibrated bba tends towards a Bayesian bba (null mass on the compound hypothesis representing the disjunction between the dichotomizer classes).

Now, if there is an important overlap between the scores of the two classes, the samples in the overlapping area are numerous so that the calibrated bba tends towards the dogmatic bba $m(\{-1\}) = m(\{1\}) = 0.5$. Such a committed bba is not in line with the idea that in such a case a score observation has a low reliability and shall be discounted to defer the decision to other sources.

In this study, we have investigated how relevant the approach proposed in Bloch (2008) is for dichotomizer evidential calibration problem. The idea is to modify (make it less committed) an initial bba depending on the context. This latter is perceived via dilation or erosion operators defined in fuzzy mathematical morphology according to Bloch and Maitre (1995). For our problem, we implement Bloch (2008) as follows.

Let $h_\theta(s) = \frac{1}{1+exp(\theta_0+\theta_1 x)}, \forall x \in \mathbb{R}$, be the sigmoid function fitted by the logistic regression. Denoting by $\top$ the used t-norm, $\perp$ the t-conorm dual of $\top$ with respect to a complementation c, the fuzzy dilation and the fuzzy erosion of μ function of $\mathbb{R}$ are defined by:

$$\forall x \in \mathbb{R}, \delta_\nu(\mu)(x) = \sup_{y \in \mathbb{R}} \top [\mu(y), \nu(x-y)], \tag{9}$$

$$\forall x \in \mathbb{R}, \epsilon_\nu(\mu)(x) = \inf_{y \in \mathbb{R}} \perp [\mu(y), c(\nu(x-y))], \tag{10}$$

with ν the structuring element. In our case, we use Lukasiewicz' t-norm: $\top_{\text{Luk}}(x,y) = \max\{0, x+y-1\}$, $c(x) = 1-x$, $\perp_{\text{Luk}}(x,y) = \min\{1, x+y\}$, and a Gaussian structuring element parametrized by α: $\nu(x) = e^{(-x^2/\alpha)}$, the bigger α, the wider ν. Fuzzy erosion and dilation are applied to h_θ or to h_θ erosion/dilation result in order to achieve h_θ opening: $\epsilon_\nu \circ \delta_\nu(h_\theta)$, or closing: $\delta_\nu \circ \epsilon_\nu(h_\theta)$.

Then, let $\{-1, 1\}$ be the dichotomizer discernment frame and m_0 the Bayesian bba associated to the score x such that $m_0[x](1) = h_\theta(x)$ and $m_0[x](-1) = 1 - m_0[x](1)$ (for notation shortness, the discernment frame superscript has been omitted). Bba assignment Bloch (2008) is valid only if, for m_0, only pairs of focal elements overlap. This is our case since $m_0[x]$ has only two focal elements. Therefore, the duality property of dilation/erosion operators or closing/opening allows for the derivation of credibility and plausibility functions of a well-defined bba m_i^b (where superscript recalls that the discernment frame is binary and subscript refers to the index of the dichotomizer). In our case, we focus on dilation/erosion so that:

$$\begin{cases} m_i^b(A) &= \epsilon_\nu(m_0[x])(A), & \forall A \in \{-1,1\}, \\ m_i^b(\{-1,1\}) &= 1 - m_i^b(\{-1\}) - m_i^b(\{1\}). \end{cases} \tag{11}$$

In the absence of overlapping between scores of the dichotomizer classes, the closer the bounds of the scores intervals, the steeper the slope of the sigmoid, and the higher the mass transferred to $\{-1, 1\}$ (for the scores located in the steep part of the sigmoid). Such a modelling allows a greater robustness versus class border estimation. In the case of overlapping, we assume that the reliability of the dichotomizer decreases with the size of the overlapping (indeed, any score included within the overlapping does not allow us to derive the sample class with certainty). Then, by increasing the width of the structuring element in relation with the length of the overlapping, the bba is automatically discounted ($m(\{-1, 1\})$ increases).

Figure 2 shows, for different distributions of the scores, the mass value $m(\{-1, 1\})$ versus score values, derived either by Xu's calibration, Xu et al. (2016), or by Bloch's allocation, Bloch (2008), with variable structuring element width. On the same plot, we have also represented the sigmoid obtained by the logistic regression and the class samples labelled 0 or 1 on the y-axis. We note that in the absence of overlapping (left sub-figure), Xu's bba is less committed than Bloch's one, but the modelled ignorance ($m(\{-1, 1\})$) decreases with the appearance of the overlapping conversely to the Bloch's allocation. Indeed, in this latter, $m(\{-1, 1\})$ models the ambiguity between the two classes (by allocating the mass to the disjunction rather than by equidistributing it between the two classes) either due to their overlapping or to the imprecision of the border between classes.

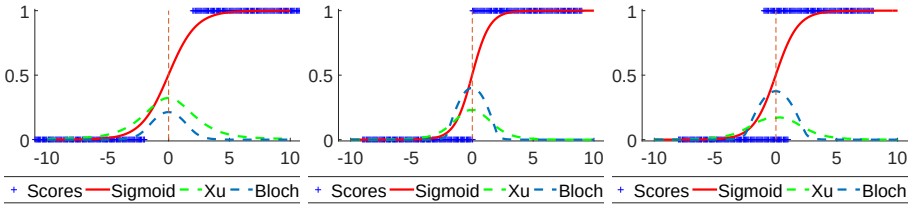


Figure 2: Comparison of $m(\{-1, 1\})$ value (y-axis) versus score (x-axis) in case of Xu's calibration or Bloch's allocation; 3 different cases of score overlapping.

In our application (classification based on ECOC dichotomizers) the number of sources is important (equal to the length of the ECOC, i.e. typically between one and few tens) so that, using the conjunctive combination rule, sources should be as little committed as possible. Then, from m_i^b we derive a consonant bba by transferring the mass of the weaker focal element to the disjunction $\{-1, 1\}$.

3.2.2. Bba combination

Decoding the ECOC involves combining the different outputs of the dichotomizers. The bba allocation step (Section 3.2.1) allows us to derive an elementary bba for each of the dichotomizer's soft output. However, each of these bbas has its own discernment frame determined by the considered dichotomizer classes. Since combination can only be performed in the same discernment

frame, the first step is thus to *project* the elementary bbas from their own dichotomizer discernment frame to the common multiclass discernment frame Ω. This step depends on whether the dichotomizer is sparse or not, i.e. whether its classes form a partition of Ω or do not.

If the dichotomizer is dense, its classes are simply interpreted as compound classes, Lachaize et al. (2016). If ρ_i is the function that provides the subset of Ω classes versus each class of the i^{th} dichotomizer ($\rho_i : \{-1, 1\} \rightarrow 2^{\Omega}$), then bba refinement may be written:

$$\left\{ \begin{array}{rcll} m_i^{\Omega}\left(\rho_i\left(A\right)\right) & = & m_i^b\left(A\right), & \forall A \in \{-1, 1\}, \\ m_i^{\Omega}\left(\Omega\right) & = & m_i^b\left(\{-1, 1\}\right). & \end{array} \right. \tag{12}$$

If the dichotomizer is sparse, in addition to the refinement on $\rho_i\left(-1\right) \cup \rho_i\left(1\right)$, a ballooning extension from $\{\rho_i\left(-1\right), \rho_i\left(1\right)\}$ to Ω shall be performed. The combination of these two operations leads to the following equations:

$$\left\{ \begin{array}{rcl} m_i^{\Omega}\left(\Omega \setminus \rho_i\left(1\right)\right) & = & m_i^b\left(\{-1\}\right), \\ m_i^{\Omega}\left(\Omega \setminus \rho_i\left(-1\right)\right) & = & m_i^b\left(\{1\}\right), \\ m_i^{\Omega}\left(\Omega\right) & = & m_i^b\left(\{-1, 1\}\right). \end{array} \right. \tag{13}$$

Note that handling compound hypotheses allows us to model effectively partial ignorance conversely to usual decoding of the ECOC matrix. Firstly, considering a dense dichotomizer addressing subsets $\rho_i\left(-1\right)$ versus $\rho_i\left(1\right)$, a decision or a score in favour of $\rho_i\left(1\right)$ (for instance) should not prejudge the probabilities of the hypotheses within $\rho_i\left(1\right)$, not even their equiprobability. Secondly, using the ballooning extension allows an appropriate modelling of the information provided by sparse dichotomizers. Indeed, regardless of the output of a sparse dichotomizer, every hypothesis in the complementary of $\rho_i\left(-1\right) \cup \rho_i\left(1\right)$ (in Ω) is possible. Thus, a high score in favour of $\rho_i\left(1\right)$ (for instance) should be interpreted as a refutation, i.e. as a strong disbelief in $\rho_i\left(-1\right)$ (rather than a belief in $\rho_i\left(1\right)$).

Finally, in the proposed model, each dichotomizer is considered as a source providing partial information about the actual class in Ω. These partial pieces of information are formalized by means of the m_i^b bbas that are then combined using the conjunctive combination rule (Eq. (3)) to infer a global belief on Ω. Such an approach is much more flexible and simpler than the one proposed in Quost et al. (2007) in which the dichotomizer answers are considered as conditioned information which implies solving an optimization problem to derive the multiclass bba, while providing equivalent or better results Lachaize et al. (2016).

3.2.3. Results on simulated data

To illustrate this decoding part, we present some first results showing the interest of the belief function framework for ECOC decoding. We consider the one-versus-one (OVO) and one-versus-all (OVA) cases varying the number of classes ($N \in [5, 7]$) and we draw random scores with different error probability rates ($\epsilon \in [0.1, 0.3]$), assuming dichotomizers have similar performance. Table 1

Table 1: Comparison of CCR performance achieved by different decoding strategies including Hamming, loss-based (LB1 and LB2) and Evidential ones (Xu and Bloch), varying the dichotomizer error probability rates (ϵ) and the class number (N); dense (OVO) and sparse (OVA) ECOC cases.

	ϵ	.1	.2	.3
	N	5 / 6 / 7	5 / 6 / 7	5 / 6 / 7
OVA	Hamming	75.6 / 71.1 / 67.4	55.8 / 50.2 / 46.1	40.5 / 35.1 / 30.9
	LB 1	89.5 / 87.5 / **86.0**	70.8 / 67.1 / 64.2	51.3 / 46.8 / 43.2
	LB 2	**89.5** / 87.6 / 85.9	70.8 / 67.1 / 64.2	51.2 / 46.8 / 43.2
	Xu	89.5 / **87.6** / 85.9	70.9 / 67.1 / 64.2	51.3 / 46.8 / 43.2
	Bloch	89.5 / 87.6 / 85.9	**70.9** / **67.2** / **64.8**	**51.3** / **46.8** / **43.2**
OVO	Hamming	78.7 / 77.0 / 75.7	59.2 / 55.5 / 52.5	41.4 / 37.8 / 33.2
	LB 1	85.4 / 84.7 / 82.8	65.0 / 60.6 / 57.1	42.2 / 36.9 / 32.8
	LB 2	85.9 / 85.5 / 83.9	66.6 / 62.5 / 59.3	43.9 / 38.6 / 34.5
	Xu	87.9 / 88.2 / 88.3	70.9 / 69.3 / 67.7	50.7 / 46.8 / 43.2
	Bloch	**88.1** / **88.3** / **88.5**	**71.5** / **70.2** / **69.1**	**51.4** / **47.7** / **44.4**

shows the obtained results in terms of percentages of Correct Classification Rates (CCR). Shown CCR are average values considering different numbers of samples for bba calibration (sigmoid function estimation) and different drawing laws for score (uniform, Gaussian, heavy queues). The bold numbers underline the best performance (considering exact values). We note that highest CCR values are almost always achieved by evidential decoding with a slight advantage for Bloch's allocation in particular in the OVO case. Besides, the dense case is much better handled by the loss-based decoding than the sparse one, so that for the OVA case all considered approaches are equivalent with the exception of Hamming decoding which is significantly worse. In the sparse case, the interest of the belief decoding versus the probabilistic (loss-based) one increases with the error probability rate and also (to a much lesser extent) with the class number, which means that it is all the more interesting as the classification is difficult.

3.3. Evidential coding

Let us now discuss the coding step that deals with the definition of the ECOC matrix. As explained in Section 3.1, we propose that the ECOC matrix is not defined at once but in a dynamic way from the preliminary results obtained from 'partial ECOC(s)' (cf. Figure 1). According to this scheme, coding and decoding are performed altogether by alternate estimation. Then, at each iteration, three decisions have to be taken: (i) Is the information gathered by the current ECOC matrix sufficient to make a reliable decision; (ii) If not, how to extend the current ECOC; and (iii) Should the extended ECOC process all the samples. Besides being connected, the answers of all these issues shall be contained in the current bba $m_{1\ldots i}$.

3.3.1. Decision criteria

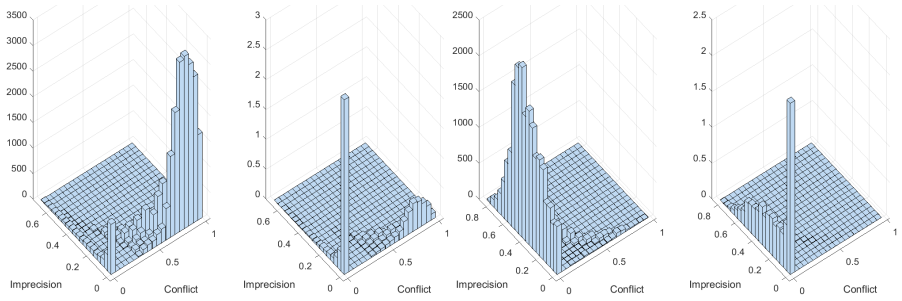
From the current bba $m_{1\ldots i}$, two main pieces of information may be derived:

- How reliable would an immediate decision be?
- Which are the most likely classes?

In relation to decision reliability, several indices intrinsic to a bba allow for the detection of a questionable labelling decision. In this work we point out two reasons that prevent a reliable decision. The first one is when we lack information. That would be, for instance, if the dichotomizers considered in the ECOC do not allow for distinguishing some classes, or if the obtained dichotomizer scores are very close to 0 (i.e., based on the sigmoid calibration, dichotomizer hypotheses would be roughly equiprobable). In this work, we assume that the lack of information may be assessed in terms of the imprecision of the decision. Using belief functions, this latter is measured by the imprecision interval defined as the difference between functions Pl and Bel (Eq. (6-7)) for the decided class may:

$$\iota = Pl\left(\arg\max_{A\in\Omega} BetP\left(A\right)\right) - Bel\left(\arg\max_{A\in\Omega} BetP\left(A\right)\right). \tag{14}$$

The second reason to suspect a decision is when the considered dichotomizers produce conflictual outputs. In this case, probably at least one of them provides an erroneous output involving the observed conflict when it is combined with the other outputs. Using consonant bbas for elementary bbas (cf. Section 3.2.1), the mass on $\emptyset$ can only come from their conjunctive combination (belief in incompatible hypotheses, i.e. having an empty intersection) so that the conflict index seems to be a good indicator to measure if the dichotomizers do not agree and therefore denotes a potential misclassification.



(a) OVA (b) OVO

Figure 3: 2D histograms of decision reliability indices computed on subsets of pixels either incorrectly classified (left) or correctly classified (right); (a) case of dense ECOC, (b) case of sparse ECOC.

Thus, to each of the two reasons to make a decision questionable, we associated an index (ι or $m(\emptyset)$) derived from the bba supporting the decision. Anticipating Section 4, we present some results obtained from our hyperspectral data involving nine classes. Figure 3 shows the histograms of conflict and imprecision values. We distinguish the case of the OVA and the OVO ECOC. For the OVA, bbas obtained from dichotomizer outputs are much more committed than in the case of the OVO because of the ballooning extension (Eq. (13)). Then, the OVA bba is characterized by a higher level of conflict than the OVO one. We also distinguish the cases of the ill-labelled pixel set and the correct-labelled pixel set. For the correct-label set, there is a very sharp peak at $(0, 0)$ meaning that both conflict and imprecision are low in the majority of cases. For the ill-label set, histograms present high values either for the conflict index for the OVA case or for the imprecision index for the OVO case. In the following, the set of samples flagged by one or two of these indices, i.e. *tricky* pixels, is denoted $\mathcal{T}$. Even if $\mathcal{T}$ do not coincide exactly with the set of wrongly-classified samples (as shown by the small peaks beside the main ones in Figure 3), the proposed indices seem to be good indicators of potentially erroneous samples, i.e. for which an immediate decision is questionable.

Then, we aim at identifying the most likely classes for each element of $\mathcal{T}$. In this work, we consider the generalisation of the pignistic transform (Eq. (8)):

$$\forall A \subseteq \Omega, BetP^{\Omega}(A) = \frac{1}{1 - m^{\Omega}(\emptyset)} \sum_{B \in \mathcal{F}_m} \frac{|A \cap B|}{|B|} m^{\Omega}(B). \tag{15}$$

Since $BetP^{\Omega}$ increases with the cardinality of the hypothesis (e.g., consonant case), to remove bias, the comparison shall be performed between hypotheses of the same cardinality. Since, while looking for the most likely classes, we are considering tricky pixels, the considered cardinality shall be greater than 1. Assuming that ambiguities can be solved pair by pair, for each selected pixel, we derive the pair of the most likely classes as $(\omega_1, \omega_2) = \underset{A \in 2^{\Omega}, |A|=2}{\arg\max}\ BetP^{\Omega}(A)$.

We make the assumption that the pairs of ambiguous classes maximize the previous $BetP$. The number of times a pair of classes maximizes $BetP$ is then a helpful information to choose the additional binary classifiers (that shall solve the ambiguities). Considering a given sample set, we derive, for each pair of classes, how frequently the respective pair maximizes $BetP$ and order class pairs by decreasing frequency. The most frequent pairs of classes (lower ranks in the ordering) correspond to the more frequent class ambiguities, thus the ones we aim to solve prioritarily. Note that this ordering is relevant only for samples we would like to correct (i.e. the wrongly classified ones). To establish this ordering in the absence of ground truth, we suggest to consider $\mathcal{T}$ as the sample set which represents the set of pixels for which the current label appears questionable. Considering the same dataset as for Figure 3, Figure 4 shows for each pair among 9 classes ($\binom{9}{2} = 36$ pairs in all), its rank in a given ordering that is established considering different sets of pixels, namely: (i) any pixel, (ii) only wrongly-classified pixels, (iii) only tricky pixels i.e. $\mathcal{T}$. The ranks

are displayed in colour gradient (from red to blue) so that the pairs in the warm colours (red to orange) are the ones that should be chosen in priority to be distinguished by additional dichotomizers. The relevance of the proposed conflict index is stressed by the similarity between the two orderings according to conflict thresholded pixels and the ordering according to wrongly-classified pixels.

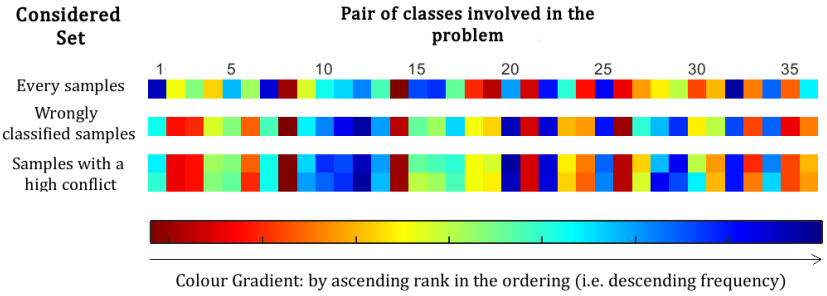


Figure 4: Frequency-ordering (among the 36 considered pairs) derived considering either (i) all pixels (1^{st} line), (ii) only wrongly-classified pixels (2^{nd} line), (iii-iv) conflict thresholded pixels (conflict greater than 0.2 and 0.5, 3^{rd} and 4^{th} lines respectively); order value coded in colour gradient from red to blue.

3.3.2. Supplementary dichotomizer selection

Having derived the ordering of the pairs of classes $\{(\omega_1, \omega_2)\}$, for each of them we have to choose the dichotomizer that will allow us to remove the ambiguity between ω_1 and ω_2. Even if some dichotomizers which handle only one of these two classes may bring some useful information (depending on the ambiguity left by the other dichotomizers), we focus on dichotomizer(s) which explicitly distinguish ω_1 and ω_2: Denoting the dichotomizer classes $\rho_i(-1)$ and $\rho_i(1)$ (like in Section 3.2.2),

$$\begin{cases} \vee & \omega_1 \in \rho_i(-1) \wedge \omega_2 \in \rho_i(1), \\ & \omega_2 \in \rho_i(-1) \wedge \omega_1 \in \rho_i(1). \end{cases} \tag{16}$$

Even with constraint (16), the number of possible dichotomizers remains intractable. In actual applications, it is generally impractical to train every dichotomizer and even if new training might be considered during the ECOC construction, we assume a *pool* of dichotomizers at our disposal. Then, denoting $\mathcal{Q}_{\omega_1|\omega_2}$ the set of dichotomizers of the pool satisfying constraint (16), we select the best performing dichotomizer in $\mathcal{Q}_{\omega_1|\omega_2}$.

Indeed, the fact of considering the partial ECOC result (selection of tricky pixels and estimation of the ambiguous classes for each of these pixels) is likened to a data-driven approach. Therefore, by adopting a performance criterion to order the elements of any given $\mathcal{Q}_{\omega_1|\omega_2}$, our approach combines both criteria

in a hierarchical way. This idea of mixing performance-driven and data-driven strategies in a hierarchical way can also be found in Pujol et al. (2008).

3.3.3. Global ECOC versus local ECOC

In previous works about ECOC coding, either data-driven approaches (e.g. Pujol et al. (2006); Escalera et al. (2008); Santhanam et al. (2016); Zhou et al. (2016)) or performance-driven ones (e.g. Passerini et al. (2004); Bai et al. (2016)), the same 'optimal' ECOC is applied to every sample (pixel for an image). In this work, having derived the dynamic ECOC (at a given iteration of our general scheme), we may apply it either to the whole set of pixels, as usual, or to subsets.

In the case where our dynamic ECOC is applied to the whole image, we call it 'global'. Then, our approach may seem close to Pujol et al. (2008) where the authors propose to extend any ECOC matrix by adding dichotomizers considering current result performance. However their approach differs from ours on the following points. The difficult classes (that need extra classification effort) are established by analysing the confusion matrix computed via a weighted decoding method. However, estimating this confusion matrix requires several sets of labelled samples (in addition to the training of the dichotomizers), whereas, in our work, pairs of difficult classes are identified by a conflict analysis that do not use ground truth data. Therefore, a strong advantage of the proposed approach is its ability to dynamically cope with new ambiguities.

An alternative to previous global ECOC is to consider the supplementary partial ECOC only for the pixels of $\mathcal{T}$, mainly in order to reduce the processing time. This strategy is called 'semi-global' since, at the end it comes down to having partitioned the image in subsets of pixels (one splitting per iteration) and having applied to each of these pixel subsets an ECOC derived from the restriction of a common global ECOC to a subset of dichotomizers. For instance, in the case of two iterations and initial ECOC $\boldsymbol{M_1}$, the image is partitioned between $\mathcal{T}_1$ computed at the end of the first iteration and its complementary $\overline{\mathcal{T}_1}$ in the whole set of pixels, so that $\overline{\mathcal{T}_1}$ pixels are finally classified using $\boldsymbol{M_1}$ and $\mathcal{T}_1$ pixels are classified using $\widehat{\boldsymbol{M_1}\boldsymbol{M_2}}$ where $\widehat{\ }$ denotes the matrix concatenation.

Another variant, called 'local', consists in customizing the additional dichotomizers for every pixel, by extending the basic ECOC matrix only with dichotomizer(s) which help to remove the ambiguity between the most likely classes at the considered pixel. For instance, from $\boldsymbol{M_1}$ and $\mathcal{T}_1$, $\overline{\mathcal{T}_1}$ pixels will be classified by $\boldsymbol{M_1}$ whereas each pixel s of $\mathcal{T}_1$ is classified using a specific ECOC $\widehat{\boldsymbol{M_1}\boldsymbol{D}}(s)$, where $\boldsymbol{D}(s)$ is chosen in the subset of dichotomizers $\mathcal{Q}_{\omega_1|\omega_2}(s)$ specific to the separation of the ambiguous classes in s.

Finally, it appears that the ending condition depends on too many criteria related to the specific application: complexity of the classification (number of classes and quality of data), minimum performance requirements, processing time, etc. Therefore, like in most ECOC coding works that usually consider an *a priori* fixed number of dichotomizers (Allwein et al. (2000); Crammer and Singer (2002); Pujol et al. (2006)), the stop criterion is left to the user's

choice and our approach focuses on providing an ordering of the most interesting dichotomizers to remove the classification uncertainties.

4. Experimental results

To illustrate our algorithm, we focus on the classification of different kinds of materials, such as different types of plastics, papers, etc. To distinguish these materials that may be very close in terms of appearance, the whole spectral response in the near infrared range represents a major asset. Hyperspectral imaging appears then as the most suitable source of information for their classification.

4.1. Data

For each pixel of a scene, hyperspectral sensors collect an almost continuous spectrum of reflectance values in a chosen waveband. In this study, we use two different hyperspectral cameras (called HSI_1 and HSI_2 sensors, both tested in Veolia laboratories) with spectral resolution of respectively 212 and 275 wavelengths between 900 nm and 2500 nm. Classic preprocessing of the spectra involves the computation of different derivative orders (0 and 1) of the spectrum by applying the Savitzky-Golay filter, Savitzky and Golay (1964). For each of these derivatives, the computation of the Principal Component Analysis (PCA, Hotelling (1933); Chen and Qian (2011)) provides the input data for the classifier. The PCA aims at reducing both the data dimensionality and the correlation between the bands. The number of selected components is set to represent 99% of the information, that corresponds to less than 20 components in most cases.

The data considered in our experiments has been acquired by imaging specimen boards with small material samples. In all, we have four boards called Paper, Plastic1, Plastic2a, Plastic2b involving 9 classes, namely 7 polymers classes: Acrylonitrile butadiene styrene (ABS), Polycarbonate (PC), Polyethylene (PE), Polyethylene terephthalate (PET), Polylactic acid (PLA), Polypropylene Polystyrene and Polyvinyl chloride (PP - PS - PVC), and rubber; and 2 fibrous classes (paper and cardboard). The average board size in pixels is 250×250.

4.2. Dichotomizers

Support Vector Machines (SVMs) introduced by Cortes and Vapnik (1995) are commonly used for hyperspectral classification (e.g., Melgani and Bruzzone (2004); Kuo et al. (2014)) due to their high classification accuracy and the relative simplicity of their architecture design. SVMs being particularly efficient for binary classification, we focus on them as our dichotomizers.

Since SVMs are learning based classifiers, three distinct datasets were extracted from the acquired images:

- The *training* dataset has 1000 samples per class and is used for SVM training. The training set allows for the estimation of each dichotomizer parameters, determined by 5 fold cross validation and grid search, using Gaussian kernels.

- The *calibration* dataset has 200 samples per class and is used for bba calibration. The calibration set allows for determining the logistic regression used in the derivation of the bbas from the SVM scores (Section 3.2.1).

- The *validation* dataset has 1000 samples per class and is used for test and performance estimation. In addition to samples from specimen boards, the test dataset also include a board that presents real objects.

4.3. Experimental results

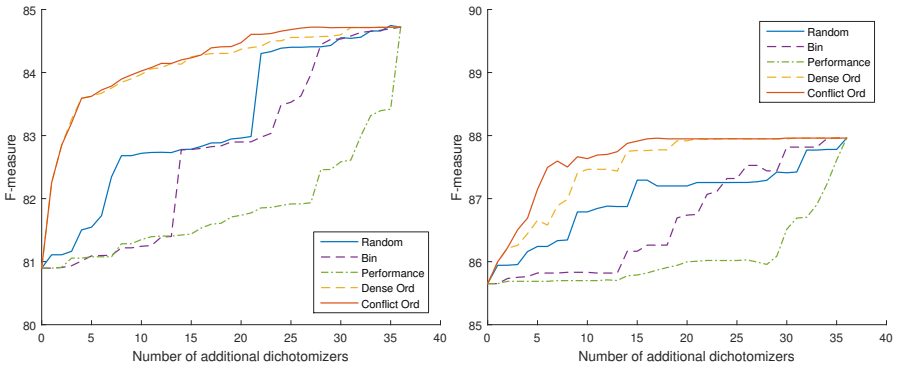
In order to analyze the proposed approach, several experiments have been conducted. Our results are quantitatively evaluated in terms of Correct Classification Rate (CCR) and F-measure criterion. We recall that for two classes, the F-measure is equal to $\frac{2TP}{2TP+FN+FP}$, where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. The higher these two criteria are, the better is the classification result. In case of more than two classes, the multiclass F-measure is the average of the two class F-measure values considering each class and its complementary.

4.3.1. Usefulness of additional dichotomizer ordering

First of all, we check the interest of the proposed selection of additional dichotomizers based on analysis of the bbas of high conflict pixels. Our basic ECOC is the OVA involving 9 dense dichotomizers in our application case with 9 classes. Figure 5 shows the increase of F-measure values (CCR values are not shown since they are highly correlated and comments would be the same) when adding one-versus-one dichotomizers one after another according to different orderings. Among the five considered orderings, *Random* and *Binary* are examples of random orderings of the 36 available dichotomizers and *Performance* is derived from the sorting of the dichotomizers according to the validation set. The two last orderings are provided by the frequency analysis of ambiguous pairs of classes, either on the whole set of pixels (*Dense Ord*) or on $\mathcal{T}$ samples (*Conflict Ord*). On the two different HSI sensor data sets, the ordering according to dichotomizer performance provides the latest (most delayed) result improvement. This may be explained by the fact that the best performing dichotomizers (chosen among the firsts) deal with classes already well-distinguished by the basic ECOC, namely the less ambiguous classes. The examples of random orderings allow for earlier F-measure improvement. Finally, for both data sets, the proposed orderings show the earliest improvement, i.e. the greatest increase achieved by the first chosen dichotomizers. Besides, the advantage of the *Conflict Ord* is more pronounced for HSI_2 data than for HSI_1 data.

Figure 6 shows the increase in the F-measure when adding the dichotomizers in the proposed ordering *Conflict Ord* for different parameters: by varying the threshold of the conflict value defining the subset of pixels $\mathcal{T}$ and by varying the set of pixels affected by a new dichotomizer versus the considered ECOC variant, namely global, semi-global or local (cf. Section 3.3.3). Note that with the local version, only one (different) dichotomizer is added in each pixel. Then it is not surprising that using more additional dichotomizers, the semi-global version slightly outperforms the local version (but at the expense of more computational resource). Comparing global and semi-global results, we note that they are equivalent in terms of performance, whereas the number of processed pixels has been much reduced in the semi-global approach: 40% and 33% of the whole number of pixels, respectively. Furthermore, for the HSI_2 dataset, processing only a subset of pixels ($\mathcal{T}$) allows us to avoid degrading labels of pixels previously correctly classified.

Finally, on Figures 5 and 6, the x-axis varies between 0 and 36 meaning that at the end, the whole pool of possible dichotomizers has been considered. However, in practical applications, the dimension of the ECOC is bounded to lower values, making the choice of the dichotomizer ordering all the more important. For instance for the HSI_1 data set, for 5 supplementary dichotomizers (14 in all), F-measure indices increase by more than 2.5% and for 10 supplementary dichotomizers (19 in total), the F-measure indices increase by more than 3.0%. Such increases are rather satisfying since they represent respectively about 70% and 85% of the whole increase when adding all the dichotomizers of the pool.

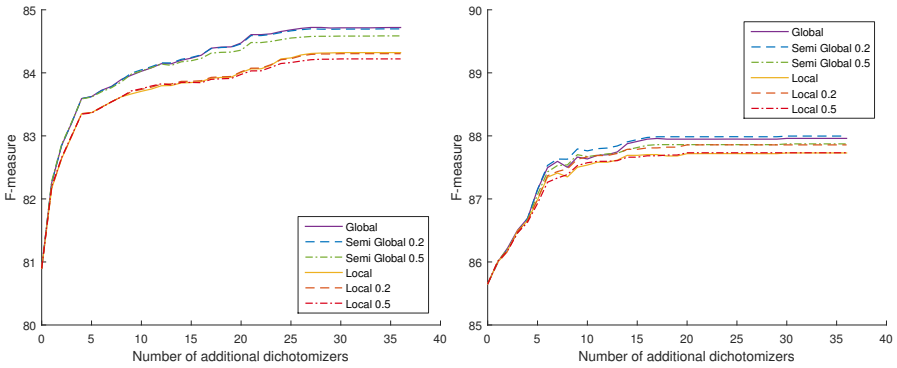


(a) HSI_1 data (b) HSI_2 data

Figure 5: Performance versus number of additional dichotomizers selected according to different orderings. Note that there is a shift between sub-figure y-axis.

4.3.2. Impact of the pool of dichotomizers

Then, we test our approach considering different pools of dichotomizers. Figure 7 shows the increase in performance (still versus the number of additional



(a) HSI_1 data (b) HSI_2 data

Figure 6: Global, Semi Global and Local strategies comparison.

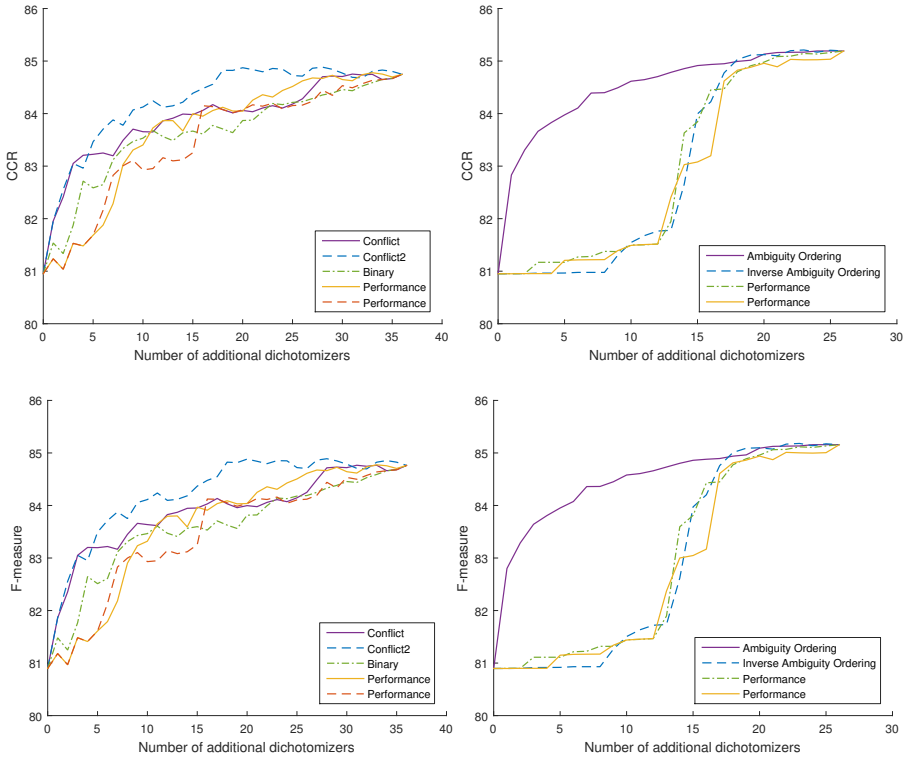
dichotomizers) by considering either the pool of two-versus-all dichotomizers ($\binom{9}{2} = 36$ dichotomizers in all) or considering a subset of the trained two-versus-two dichotomizers (25 dichotomizers among $\binom{9}{4} \times \frac{\binom{2}{4}}{2} = 378$ dichotomizers in all).

We note that, in the case of the two-versus-all pool, the ordering of the dichotomizers matters much less than in the previous case with the one-versus-one pool. Besides, the curves are not monotonic, even if globally increasing. Several points partially explain these observations.

Firstly, the two-versus-all dichotomizers may achieve lower performance than the one-versus-one dichotomizers due to the fact that the binary problems induced by the one-versus-one dichotomizers are simpler. Secondly, two-versus-all dichotomizers separate numerous pairs at a time: each two-versus-all separates $2(N-2)$, whereas one-versus-one dichotomizers separates only 1 and one-versus-all separates $N-1$, where N is the number of classes. Considering a given pair of classes (ω_1, ω_2), information for its class splitting can be provided by only one-versus-one dichotomizer (namely ω_1 versus ω_2), or one among $(N-2)(N-3)$ two-versus-two dichotomizers (namely $\{\omega_1, \omega_i\}$ versus $\{\omega_2, \omega_j\}$, $i, j \notin \{1, 2\}$, $i \neq j$), or one among $2(N-2)$ two-versus-all dichotomizers (namely $\{\omega_1, \omega_i\}$ versus $\Omega \setminus \{\omega_1, \omega_i\}$, or $\{\omega_2, \omega_i\}$ versus $\Omega \setminus \{\omega_2, \omega_i\}$, $i \notin \{1, 2\}$). Thus, even when randomly picking a two-versus-all dichotomizer, there is a high probability that it distinguishes one of the pairs of interest (probability $\frac{2(N-2)}{\binom{N}{2}} = \frac{4(N-2)}{N(N-1)}$ for a given pair). Thus, for this pool, the ordering has much less impact than with the one-versus-one pool. However note that the two-versus-two dichotomizers involve other challenges in terms of learning step (e.g., imbalance in the class representation for high numbers of classes, transfer when adding new classes).

Much more relevant to illustrate the benefits brought by our approach are the results provided by the two-versus-two pool. For this test, the subset of dichotomizers is composed of 13 'interesting' dichotomizers (that separate the

first ambiguous pairs designated by the conflict analysis), and 13 *a priori* not useful dichotomizers (that separate the last ambiguous pairs of classes according to the conflict analysis or that keep the ambiguous classes in the same side of their separation border). Like in the case of the one-versus-one pool, the choice of the ordering of the dichotomizers has a strong impact on the performance when adding a small number of supplementary dichotomizers. Results derived from HSI_2 data lead to similar conclusions, namely: the ordering matters less when choosing a candidate among a pool of 'dense' dichotomizers that separate several classes at once. On the other hand, more 'dedicated' dichotomizers like two-versus-two or one-versus-one induce simpler binary problems and the complementarity of the resulting errors is easier to understand and manipulate. Thanks to the use of the belief functions decoding, adding sparse dichotomizers involving only a few classes does not create the same bias as Hamming or loss-based decodings. In our further experiments we use the *one-versus-one* pool that is the simplest possible pool and need no additional criterion to decide which dichotomizer to use for a given ambiguous pair (e.g. performance).



(a) *two-vs-all* pool (b) *two-vs-two* pool

Figure 7: Performance versus number of additional dichotomizers selected according to different orderings for the *two-vs-all* and the *two-vs-two* pools of dichotomizers.

4.3.3. Benefit of multiorder data combination

According to our initial experiments and in the spirit of some authors who customize the data or features used by each dichotomizer, Bai et al. (2016), we propose to combine complementary information sources. In our case, highly complementary sources are provided by the different derivation orders of the hyperspectral spectra, Tachwali et al. (2007). Figure 8 illustrates, on two different sample boards, the complementarity of the conflict maps (assessing subsets $\mathcal{T}_{D0}$ and $\mathcal{T}_{D1}$ of tricky pixels) according to OVA applied to $D0$ or $D1$ data respectively.

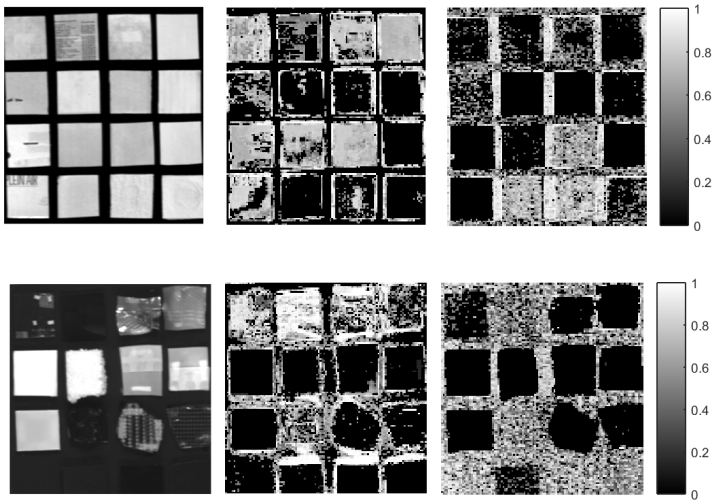


Figure 8: Conflict maps from OVA classifier applied to $D0$ or $D1$ data, respectively; case of two sample boards - HSI_1 Data

We propose to consider $D0$ and $D1$ in the following way. From a given basic ECOC used for processing the data in a given order of derivation ($D0$ or $D1$), the added dichotomizers are chosen in the other order of derivation ($D1$ or $D0$ respectively). We consider different basic ECOCs, including the OVA, because it is both classic and involves a small number of dichotomizers, and two data-driven ECOCs, namely CMSECOC, Zhou et al. (2016). The CMSECOC idea is to gather ambiguous classes in order to build *superclasses*, which are expected to be separable by OVA strategy. In Zhou et al. (2016), the ambiguous classes are selected according to a similarity matrix estimated using ground truth data. Thereafter, each superclass is split using the OVO strategy within it. We focus on these CMSECOC strategies since, using the proposed belief-based decoding (cf. Section 3.2), they allow for high performance of the basic ECOC results. However, since CMSECOC already involves one-versus-one dichotomizers, the additional dichotomizers (also taken in one-versus-one pool) process data corresponding to the complementary order of derivation. As

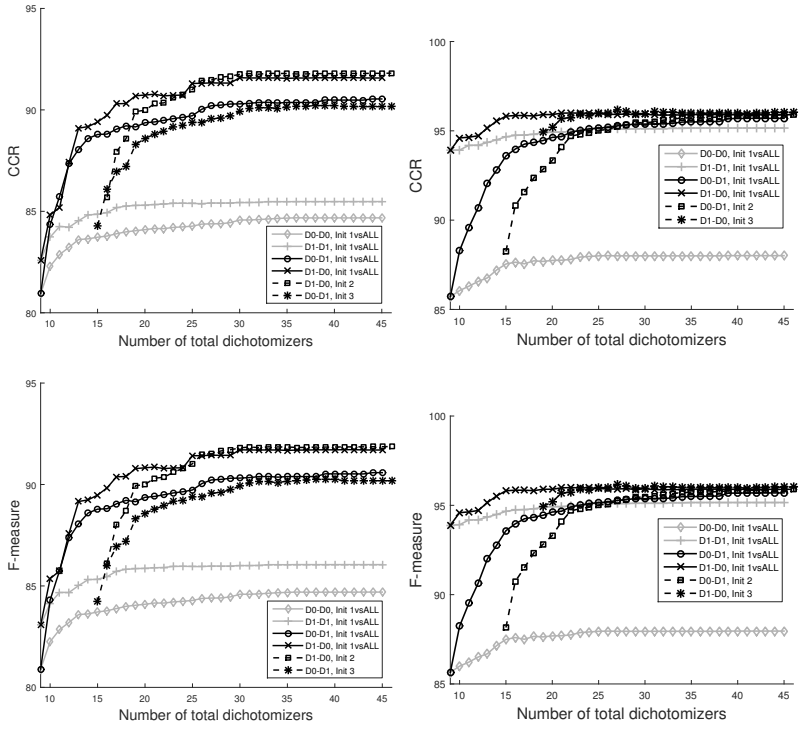
previously, to determine $\mathcal{T}$, both the conflict and the imprecision measures have been used with a threshold of 0.2.

As displayed on Figure 9, the performance is improved (on CCR and F-measure) when combining results on the two orders of derivation. On the HSI_1 dataset, $D0$ and $D1$ have a high complementarity of errors (as seen on the conflict maps of Figure 8). Using the OVA as basic ECOC, the combination with $N = 5$ dichotomizers (for a total of 14 classifiers) from a different order than the OVA improves the CCR by about 5% when starting with $D0$ and by 4% when starting with $D1$ compared to sticking to the same order. The mixed combinations, starting from $D1$ and $D0$, reach about 92% and 90% respectively when adding 20 dichotomizers and up. On the HSI_2 dataset, there is a dissymmetry between the performance achieved by the two orders: the basic $D1$ results achieve almost 8% higher performance than the basic $D0$ results. However, thanks to complementarity, adding the information from the $D0$ dichotomizer improves the results compared to adding information from the same order. Using $D1$ for the basic OVA is the most efficient combination, since it allows for reaching a correct classification rate of 97% with seven dichotomizers added. Now, the most probative of the conflict-based ordering are the results obtained when using the $D0$ for basic OVA since it allows for a +7% improvement on the CCR when adding seven dichotomizers.

Using CMSECOC as M_1, we get the same global tendency for the performance curves on both datasets. Specifically, the results at the output of M_1 decoding are higher than previously. This is due to our decoding, since using Hamming decoding like in Zhou et al. (2016), performance indices are about 10% lower: Hamming decoding is not efficient in the case of numerous sparse classifiers that are not equally distributed among classes. Note also that depending on the data considered ($D0$ or $D1$, HSI_1 or HSI_2), the corresponding basic CMSECOC varies, in particular in terms of code length. A rather interesting point is that achieved performance for about 25 dichotomizers in all is much more robust to the basic ECOC (CMSECOC or OVA) than to the considered data and, for HSI_1, the ordering in which they are considered.

5. Conclusion and perspectives

This work studied the relevance of the belief function theory (BFT) framework for the ECOC field. This usefulness was clearly shown for both coding and decoding issues. Indeed, firstly the use of BFT allows us to propose a decoding step that models each dichotomizer in the ECOC matrix as an individual source of information. Thanks to the manipulation of compound hypotheses, we were able to model the exact information provided by each dichotomizer, even the sparse ones. Our method therefore uses the belief function framework to elegantly model concepts which are otherwise difficult to formalize. Secondly, BFT provides us with indices particularly relevant for detecting the potentially unreliable decisions, namely the conflict and the imprecision measures. The analysis of these indices and of the basic belief assignment at the output of the decoding step allows us to propose a new method to extend an ECOC matrix in



(a) HSI_1 data (b) HSI_2 data

Figure 9: Performance (CCR and F-measure) versus number of additional dichotomizers for different combination of $D0$ and $D1$ sources.

order to solve the remaining ambiguities. Several variants of this method were proposed depending on the chosen degree of stationarity of the ECOC matrix used. The conflict and imprecision measures are intrinsic indices, so they allow for auto-evaluation to detect outliers or drifts from the training set. Our approach was tested on hyperspectral data, acquired from two different sensors, to classify nine different types of material. We clearly showed the benefit of extending a basic ECOC matrix to derive a compact ECOC with high performance. The semi-global variant that processes only the pixels detected as tricky (about 35% of the image in our case) seems to be a good compromise since it achieves similar performance as the global variant (processing all the pixels). The absence of requirement for the ground truth to build the ECOC extension is obviously a major strength of our approach. Finally, in order to increase the complementarity of the dichotomizer outputs, we propose to combine different features extracted from the data (derivative orders in our case). The obtained results outperform in a significant way any other results (CCR increase between 2% and 8%, depending on the considered sensor and basic ECOC).

Future work will investigate the definition of a measure to assess the redun-

dancy of the discernability of a given class in a given ECOC matrix. Indeed, we saw that some classes were more difficult to separate from others. For these classes, we aim at increasing the number of independent subsets of dichotomizers allowing their separation (i.e., providing an unambiguous codeword). In a deeper analysis, we would also like to estimate the complementarity of errors between the previous independent subsets of dichotomizers. Indeed, the combination of these may remove errors only if they are not correlated (classification performance is strongly related to the complementarity of the errors or ambiguities of subsets of dichotomizers viewed as individual classifiers).

In terms of application, we will investigate how classification quality control may be related to the evolution of the percentage of detected tricky pixels. The basic idea is that an anomalous increase in this percentage should alert the user. The usefulness of the proposed indices is that they do not require a ground truth, even though this may not be the case for the action ensuing a warning, e.g. a new training on updated data.

References

Allwein, E. L., Schapire, R. E., Singer, Y., 2000. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1 (Dec), 113–141.

André, C., Le Hégarat-Mascle, S., Reynaud, R., 2015. Evidential framework for data fusion in a multi-sensor surveillance system. Engineering Applications of Artificial Intelligence 43, 166–180.

Appriou, A., 1997. Multiple signal tracking processes. Aerospace Science and Technology 1 (3), 165–178.

Bai, X., Niwas, S. I., Lin, W., Ju, B.-F., Kwoh, C. K., Wang, L., Sng, C. C., Aquino, M. C., Chew, P. T., 2016. Learning ECOC code matrix for multiclass classification with application to glaucoma diagnosis. Journal of Medical Systems 40 (4), 78.

Bautista, M. A., Pujol, O., De La Torre, F., Escalera, S., 2017. Error-correcting factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence In Press.

Bloch, I., 2008. Defining belief functions using mathematical morphology–application to image fusion under imprecision. International Journal of Approximate Reasoning 48 (2), 437–465.

Bloch, I., Maitre, H., 1995. Fuzzy mathematical morphologies: a comparative study. Pattern Recognition 28 (9), 1341–1387.

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational Learning Theory. ACM, pp. 144–152.

Brassard, G., Bratley, P., 1996. Fundamentals of algorithmics. Vol. 33. Prentice Hall Englewood Cliffs.

Chen, G., Qian, S.-E., 2011. Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage. International Journal of Approximate Reasoning 49 (3), 973–980.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Crammer, K., Singer, Y., 2002. On the learnability and design of output codes for multiclass problems. Machine Learning 47 (2), 201–233.

Deng, X., Liu, Q., Deng, Y., Mahadevan, S., 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences 340, 250–261.

Díaz-Más, L., Muñoz-Salinas, R., Madrid-Cuevas, F. J., Medina-Carnicer, R., 2010. Shape from silhouette using Dempster–Shafer theory. Pattern Recognition 43 (6), 2119–2131.

Dietterich, T. G., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286.

Escalera, S., Pujol, O., Radeva, P., 2010. On the decoding process in ternary error-correcting output codes. IEEE transactions on Pattern Analysis and Machine Intelligence 32 (1), 120–134.

Escalera, S., Tax, D. M., Pujol, O., Radeva, P., Duin, R. P., 2008. Subclass problem-dependent design for error-correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6), 1041–1054.

Gao, T., Koller, D., 2011. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 2072–2079.

Geman, S., Geman, D., 1987. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In: Readings in Computer Vision. Elsevier, pp. 564–584.

Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling. In: Advances in Neural Information Processing Systems. pp. 507–513.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24 (6), 417.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

Kuo, B.-C., Ho, H.-H., Li, C.-H., Hung, C.-C., Taur, J.-S., 2014. A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (1), 317–326.

Lachaize, M., Le Hégarat-Mascle, S., Aldea, E., Maitrot, A., Reynaud, R., 2016. SVM classifier fusion using belief functions: Application to hyperspectral data classification. In: International Conference on Belief Functions. Springer, pp. 113–122.

Le Hegarat-Mascle, S., Bloch, I., Vidal-Madjar, D., 1997. Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. International Journal of Approximate Reasoning 35 (4), 1018–1031.

Liu, Z.-G., Pan, Q., Dezert, J., 2014. A belief classification rule for imprecise data. Applied Intelligence 40 (2), 214–228.

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. International Journal of Approximate Reasoning 42 (8), 1778–1790.

Mercier, D., Cron, G., Denœux, T., Masson, M.-H., 2009. Decision fusion for postal address recognition using belief functions. Expert Systems with Applications 36 (3), 5643–5653.

Nilsson, N. J., 1965. Learning Machines. New York.

Parikh, C. R., Pont, M. J., Jones, N. B., 2001. Application of Dempster–Shafer theory in condition monitoring applications: a case study. Pattern Recognition Letters 22 (6), 777–785.

Passerini, A., Pontil, M., Frasconi, P., 2004. New results on error correcting output codes of kernel machines. IEEE Transactions on Neural Networks 15 (1), 45–54.

Pujol, O., Escalera, S., Radeva, P., 2008. An incremental node embedding technique for error correcting output codes. Pattern Recognition 41 (2), 713–725.

Pujol, O., Radeva, P., Vitria, J., 2006. Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (6), 1007–1012.

Quost, B., Denœux, T., Masson, M.-H., 2007. Pairwise classifier combination using belief functions. Pattern Recognition Letters 28 (5), 644–653.

Rekik, W., Le Hégarat-Mascle, S., Reynaud, R., Kallel, A., Hamida, A. B., 2016. Dynamic object construction using belief function theory. Information Sciences 345, 129–142.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. Journal of Machine Learning Research 5 (Jan), 101–141.

Roquel, A., Le Hégarat-Mascle, S., Bloch, I., Vincke, B., 2014. Decomposition of conflict as a distribution on hypotheses in the framework on belief functions. International Journal of Approximate Reasoning 55 (5), 1129–1146.

Santhanam, V., Morariu, V. I., Harwood, D., Davis, L. S., 2016. A non-parametric approach to extending generic binary classifiers for multi-classification. Pattern Recognition 58, 149–158.

Savitzky, A., Golay, M. J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36 (8), 1627–1639.

Shafer, G., 1976. A Mathematical Theory of Evidence. Vol. 1. Princeton university press Princeton.

Smets, P., Kennes, R., 1994. The transferable belief model. Artificial Intelligence 66 (2), 191–234.

Smets, P., Ristic, B., 2007. Kalman filter and joint tracking and classification based on belief functions in the tbm framework. Information Fusion 8 (1), 16–27.

Tabassian, M., Ghaderi, R., Ebrahimpour, R., 2012. Combining complementary information sources in the Dempster–Shafer framework for solving classification problems with imperfect labels. Knowledge-Based Systems 27, 92–102.

Tachwali, Y., Al-Assaf, Y., Al-Ali, A., 2007. Automatic multistage classification system for plastic bottles recycling. Resources, Conservation and Recycling 52 (2), 266–285.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 3360–3367.

Xu, L., Krzyzak, A., Suen, C. Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE transactions on Systems, Man, and Cybernetics 22 (3), 418–435.

Xu, P., Davoine, F., Zha, H., Denoeux, T., 2016. Evidential calibration of binary SVM classifiers. International Journal of Approximate Reasoning 72, 55–70.

Zhou, J., Yang, Y., Zhang, M., Xing, H., 2016. Constructing ECOC based on confusion matrix for multiclass learning problems. Science China Information Sciences 59 (1), 1–14.