Evidential multi-class classification from binary classifiers: application to waste sorting quality control from hyperspectral data

Marie Lachaize, Sylvie Le Hégarat-Mascle, Emanuel Aldea, Aude Maitrot, Roger Reynaud

▶ **To cite this version:**

Marie Lachaize, Sylvie Le Hégarat-Mascle, Emanuel Aldea, Aude Maitrot, Roger Reynaud. Evidential multi-class classification from binary classifiers: application to waste sorting quality control from hyperspectral data. The International Conference on Quality Control by Artificial Vision 2017, May 2017, Tokyo, Japan. 10.1117/12.2266961 . hal-01691773

HAL Id: hal-01691773
https://hal.science/hal-01691773

Submitted on 24 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evidential multi-class classification from binary classifiers: application to waste sorting quality control from hyperspectral data

Marie Lachaize[1,2], Sylvie Le Hégarat-Mascle[1], Emanuel Aldea[1], Aude Maitrot[2], Roger Reynaud[1]
[1]SATIE – CNRS UMR 8029, Université Paris-Sud, Université Paris-Saclay, France
[2]VEOLIA RECHERCHE & INNOVATION, 291 av. Dreyfous Ducas, Limay, France

ABSTRACT

Our application deals with waste sorting using an automatic system involving a hyperspectral camera. This latter provides the data for classification of the different kinds of waste allowing the evaluation of mechanical pre-sorting and its refinement. Hyperspectral data are processed using Support Vector Machine (SVM) binary classifiers that we propose to combine in the belief function theory (BFT) framework to take into account not only the performance of each binary classifier, but also its imprecision related for instance to the number of samples during the learning step. Having underlined the interest of BFT framework to deal with sparse classifiers, we study the performance of different combinations of classifiers.

Keywords: Multi-class classification, waste sorting, hyperspectral data, belief function theory, Support Vector Machine

1. INTRODUCTION

Hyperspectral imaging is a powerful source of information that has gained popularity over the last decade. For each pixel of a scene, hyperspectral sensors collect an almost continuous spectrum of reflectance values in a chosen waveband. This detailed information allows detecting even minor variations within and between spectra which may be most helpful for classification. This type of imaging is used in numerous fields or domains: from military applications to atmosphere analysis, ecosystems monitoring and industrial applications. Among these latter, the waste sorting field has an increasing need for new automated systems. They are required to improve both the quality of recycled materials and the working conditions, in particular in the quality control step that is traditionally carried out by manual operators. Clearly, such an application can benefit from hyperspectral sensors since they are contact-free, and the data acquired in the near infra-red range may allow the distinction of materials that are close in terms of spectral response, such as different kinds of polymers or fibrous materials. However, if hyperspectral sensors are very informative, they also have an obvious drawback related to the significant amount of data they generate. This is all the more a drawback that in an industrial context, processing time is a strong constraint and the time and memory resources are a significant issue.

Support Vector Machines (SVMs) introduced by Vapnik and Cortes[1,2] are commonly used for hyperspectral classification[3,4,5] due to their high classification accuracy and the relative simplicity of their architecture design. These classifiers are well adapted for binary classification. Then, faced to a multi-class problem, classically, one splits it in several binary problems whose outputs should be combined. However, the way to choose (to define) and then to combine these simple classifiers is still an open question. Specifically, in terms of decomposition of the multi-class problem the classic "one-versus-one" (OVO) and "one-versus-all" (OVA) strategies consist respectively in training either a classifier to discriminate between each pair of classes or in training a classifier for each class against all the others. A third approach[6] consists in the use of error-correcting codes (ECOC). In the following we will omit the word "multiclass classification" when referring to OVO, OVA and ECOC multiclass classification. Considering n classifiers and c classes, ECOC approach involves building a matrix M of size $c \times n$ and values in $\{-1,1\}$ such that the j^{th} column of M represents the two subsets of classes considered in the j^{th} binary classifier. Then, each line corresponds to a different code-word that represents a given class (among the c classes) in terms of expected answers of the binary classifiers. A unified framework has been proposed[7] for these three methods using a ternary coding matrix (values in $\{-1,0,1\}$) so that the two subsets of classes handled by a binary classifier do not have to form a partition of the whole class set. The OVO can then also be represented by a coding matrix. The simplest way to combine the classifiers decision is either a vote system or a decoding based on the Hamming distance[6], such that each classifier "casts a vote" for its winning class(es). However, such decision rules do not take into account the confidence level of each classifier in its own decision. Then, more elaborate methods have been proposed[7] to combine classifiers also considering the confidences associated with binary

decisions, e.g. via loss functions. For these methods to be relevant, the classifiers should be calibrated in order to provide comparable outputs. The SVMs are widely used in a probabilistic framework (e.g. through logistic regression). However, if they model successfully the uncertainty of decisions, probabilities miss modeling the imprecision due to the training and calibration steps of the classifiers.

To handle both uncertainty and imprecision, the evidential framework has been initially proposed[8] by A. Dempster and G. Shafer, while Ph. Smets proposed[9] his interpretation in terms of belief transfer. The belief functions are a useful and intuitive tool to model simple classifiers that will be combined into a multi-class classifier. Previous works[10] proposed a robust calibration of binary SVMs to derive belief functions from scores taking into account the number of samples of each score value during the calibration step. , Xu's approach[10] was applied[11] to classification of waste categories. The multiclass problem was mainly handled using OVA that was shown to fit the industrial constraints in terms of computation load and compared to OVO strategy results. In this study, we investigate more sophisticated binary classifier strategies, e.g. data driven strategies[12], random selection[7], and expert knowledge-based strategies. If these approaches have been evaluated in the framework of classic loss-functions, they have never been tested in the framework of belief function theory (BFT). In this study, we show that BFT allows the combination of imprecise and non-independent classifiers. BFT also offers a new measure to evaluate the results based on the belief function itself. Finally, using belief functions, we also combine classifier outputs provided by different input data (derivation orders of the hyperspectral spectrum in our case).

In Section 2, we specify the constraints and the different strategies that will be compared in this study. We also recall some of the basics of the BFT and we explain briefly the evidential calibration we use and the way we derive the basic belief assignments (bbas) in the common multi-class frame of discernment. Section 3 presents the industrial context of our waste sorting application and main experimental results using hyperspectral data. Finally, Section 4 gives some conclusions and perspectives.

2. USE OF THE EVIDENTIAL FRAMEWORK FOR COMBINATION OF BINARY CLASSIFIERS

The use of the evidential framework to combine binary classifiers handling different subsets of classes against another is motivated by the ability of belief function theory to model partial ignorance or imprecision.

In the previous section, we recalled that Allwein et al. have introduced[7] a ternary code matrix to handle any binary classifier, said "sparse", that does not consider a partition of the whole set of classes. (This rises the question of the classifier interpretation or even relevance. Specifically, for a classifier not involving a given hypothesis in the considered subsets of classes, Allwein et al. propose to entirely ignore its output when computing the loss function (on which multi-class decision is performed). For instance, this means that, considering a 4-class problem (classes in $\Omega=\{\omega_1,\omega_2,\omega_3,\omega_4\}$), the outputs of the classifiers "ω_2 versus ω_3", "ω_2 versus ω_4" and "ω_3 versus ω_4" would not be involved in computation of the loss associated to ω_1 hypothesis. More generally, using the OVO strategy in a c -class problem, $c(c-1)/2$ classifiers are run but, for a given class i, only $(c-1)$ classifiers are used in the multi-class decision. Such a strategy is based on the assumption that only (c-1) classifiers are really reliable in the sense that they were trained on actual samples from the class i. In this study we affirm that the outputs of such classifiers could also be used to quantify the confidence in an imprecise decision. Following the 4-class example, the classifier "ω_2 versus ω_4" actually provides binary decision on "$\Omega\backslash\omega_4$ versus $\Omega\backslash\omega_3$" where $\Omega\backslash\omega_i$ is the whole set of classes but ω_i. However, to model such information in an unbiased way, one should use a framework handling imprecise hypotheses such as the Belief Function Theory (BFT). In the second part of this section, we will explain how BFT allows us to handle partial knowledge/ partial ignorance but let us first specify the kind of partial knowledge/ partial ignorance ECOC strategies introduce.

2.1 ECOC strategies

Let Ω denote the class set and A and B two distinct subsets of Ω: $\varnothing\subset A\subseteq\Omega$, $\varnothing\subset B\subseteq\Omega$, $A\cap B=\varnothing$ (note that A and B do not have to form a partition of Ω so that possibly $A\cup B\subset\Omega$). In the following A and B are also called superclasses. The base classifiers are binary A versus B classifiers that are binary SVMs in this study. Denoting $|X|$ the cardinality of a superclass X, when $|A|=|B|=1$, the base classifiers correspond to the OVO case; when $|A|=1$ and $|B|=c$-1, they correspond to the OVA case, that is a special case of $|A|=k$ and $|B|=c-k$, $k\in\{1,\ldots,c/2\}$ called KVR ("k versus rest"). Many works have been done on the optimal choice of these base classifiers. As said in the Introduction, these works usually involve the construction of an ECOC matrix, noted $\mathbf{M}$ with dimensions $c\times n$ with c the number of classes ($c=|\Omega|$) and n the number

of binary classifiers and values in {-1,0,1}. Constraints and criteria may be formulated on **M** rows to increase the separability of the classes. For instance, Allwein et al. measure the minimum distance[7], noted ρ, between two rows of **M**: $\rho = \min_{\substack{(i,l)\in\{1,..,k\}^2 \\ i\neq l}} \rho_{il}$ with $\rho_{il} = \sum_{j=1}^{n} \frac{1-m_{i,j}m_{l,j}}{2}$ and $m_{i,j}$ the term (i,j) of **M**.

Measure ρ is a very simple measure to evaluate the complementarity of the simple classifiers. Then, a simple way to choose **M** is to generate random codes and keep the matrix **M** with the highest ρ value (or the first **M** with a ρ value greater than a given threshold). The parameters of such a pseudo-random **M** construction are: the kind of simple classifiers, namely only KVR corresponding to dense codes (i.e. the matrix contains only $\{-1,1\}$ values) or any simple classifiers corresponding to sparse codes (i.e. value 0 is allowed in **M**), the minimum ρ value and the number of classifiers n. As said, such a construction of **M** is only based on the theoretical complementarity of the simple classifiers. Then, other approaches have been proposed to consider more problem-specific information. They are either based on the observed performance of the base classifiers or on the observed separability of the classes given the data.

The performance-driven approach is based on the performance of individual simple classifiers. For instance, Bai et al. investigate[13] each of the possible partitions of the class set by choosing the optimized input data for each of the corresponding binary classifiers (called "dichotomizers") and evaluating their performances in terms of correct classification rate (CCR). Then, the performance of each classifier is taken into account to define the coding matrix: The selection avoids selecting some dichotomizers that perform badly even if they provide a good separability of the codewords (e.g. in terms of ρ_{il}). However, such approach requires to train and evaluate each possible dichotomizer that is tedious when the number of classes grows (for a nine-class problem, 9330 classifiers have to be evaluated). Moreover, the performance of individual classifiers does not allow to predict the performance of their combination since in such a case the independence of the errors is a key point for fusion success.

The data-driven approach consists in building the ECOC using information derived from the data themselves, for example confusion matrices. This is a research hotspot of ECOC applications[14,15]. Zhou et al. proposed[12] a method derived from an idea already hinted[16]. In Zhou's approach, called the CMSECOC (Confusion Matrix Superclass ECOC), superclasses are formed so that two superclasses are easily separated. Then, the class-elements within each superclass are more difficult to separate and need more redundancy to be distinguished. The superclasses are built by analyzing the confusion matrix obtained by an *a priori* multi-class classifier (discussed just afterwards). Then, the OVA (or KVR) is used to separate the superclasses from one another and the OVO is used to classify within each superclass. This construction ensures to add redundancy when it is needed and the two degenerated cases of this method are the OVO and the OVA methods: the first one occurs when all classes are considered too similar, and the second one when all classes are well separate. Now, among the drawbacks or instability sources of such approach, we can mention: the classifier used to derive the confusion matrix and the thresholds that decide if two classes are considered similar or different. Then, an alternative may be to use meta-knowledge on the data to build the multi-class classifier. For example, in our application, since the material "PLA" is well separated from all the others, we can use an OVA classifier to define this class. On the other hand, the "Paper" and "Cardboard" materials have a tendency to be confused, thus an expert could propose to separate them by combining a "2VR" ("Paper or Cardboard" versus the other classes) and a "OVO" ("Paper" vs "Cardboard") classifiers.

Previous cited works mainly focused on the construction of the **M** matrix from modeling the interactions between the simple classifiers. However, very few works deal with the modeling of the simple classifiers themselves, whereas we can state that the base classifiers involve three kinds of imprecision:

- First, as they are learnt on a learning sample set, the precision of a score (as the output to a new sample to classify) depends on the number of samples achieving this score in the learning set;
- Second, playing A versus B subsets, a decision or a score in favor of A (for instance) should not prejudge the probabilities of the hypotheses ω_i in A, not even their equiprobability;
- Third, when $A \cup B \subset \Omega$, any singleton hypothesis in $\Omega \backslash A \cup B$ is also possible.

Let us now present how these three kinds of imprecision can be modeled by the BFT.

2.2 Belief function model

Let us introduce some necessary notations. Let C_j be a binary classifier that is represented by a column j in the **M** matrix. Its output in response to an observation value x is the score (case of SVM classifiers) $s_j(x)$. The basic belief assignment (bba) that is derived from $s_j(x)$ is simply denoted m_j for the sake of brevity even if it obviously depends on $s_j(x)$ and calibration parameters specific to C_j. The bba m_j is a function of the set of the subsets of $\{0,1\}$, i.e. $\{\emptyset,\{0\},\{1\},\{0,1\}\}$. Then, the multi-class discernment frame is Ω (already introduced) and we denote by 2^{Ω} the set of its subsets that contains $2^{|\Omega|}$ elements. A bba derived from output of C_j and defined on 2^{Ω} is noted m_j^{Ω}.

Then, for any observation x, the evidential multi-class classification process involves three different steps: (i) the derivation of the bbas m_j, $j\in\{1,\ldots n\}$, corresponding to each output $s_j(x)$; (ii) the projection to 2^{Ω} of each m_j, defined on $2^{\{0,1\}}$ with binary classes specific to classifier C_j, leading to m_j^{Ω}, $j\in\{1,\ldots n\}$; (iii) the combination of the m_j^{Ω}, $j\in\{1,\ldots n\}$.

Let us now briefly present these three steps.

The first step is called calibration. It follows the approach[10,11] and should be done for each trained classifier C_j. For this step, we use a set of samples for which both the observation x and the actual class y are known. Having x as input, the SVM classifier C_j provides a class and a score, in $(-\infty,+\infty)$, that depends on the distance between the sample and the frontier that the SVM created between the two classes: the higher the absolute value of the score, the wider the distance. Each SVM has its specific scale for these scores so that they should be calibrated in prevision of the combination step. Several methods exist to turn scores into probabilities (for example the logistic regression) or into bba[10]. In the case of the evidential framework, for each score value s_j the derived bba is defined from values $m_j(\{0\})$, $m_j(\{1\})$ and $m_j(\{0,1\})$. Roughly, $m_j(\{0\})$ and $m_j(\{1\})$ represent the uncertainty that the sample belongs to class $\{0\}$ or to class $\{1\}$ according to binary classifier C_j (also involving the two binary class definition) and $m_j(\{0,1\})$ represents the imprecision on previous uncertainty values, imprecision that is strongly correlated to the number of test samples achieving the considered score value (typically, if some scores never appear in the calibration set then the mass on $\{0,1\}$ on these scores will be high). It is also possible to discount the bba by increasing all the more $m_j(\{0,1\})$ to take into account the classifier reliability.

The second step is different depending on whether the two subsets A and B considered by the binary classifier (see Section 2.1) form a partition of Ω ($A\cup B=\Omega$) or not ($A\cup B\subset\Omega$). Following work[11], in the first case, the hypotheses A and B are simply interpreted as compound classes: $m_j^{\Omega}(A)=m_j(\{0\})$, $m_j^{\Omega}(B)=m_j(\{1\})$ and $m_j^{\Omega}(\Omega)=m_j(\{0,1\})$, whereas, in the second case, in addition to the previous interpretation, a deconditioning on $\Omega\backslash A\cup B$ denoted $\overline{A\cup B}$, should be performed: $m_j^{\Omega}(A\cup\overline{A\cup B})=m_j(\{0\})$, $m_j^{\Omega}(B\cup\overline{A\cup B})=m_j(\{1\})$ and $m_j^{\Omega}(\Omega)=m_j(\{0,1\})$. Note that this second step addresses the requirement of modeling and handling the second and third kinds of imprecision mentioned in Section 2.1 (whereas the answer to the first kind of imprecision was provided by the previous calibration step).

The third step deals with the combination of the simple classifiers C_j. However, thanks to the interpretation of their outputs in terms of bba on a same discernment frame Ω, this step is achieved in a trivial way by combining the bbas m_j^{Ω} using a BFT combination rule: either the classic conjunctive rule (noted $\bigcap m$) or the cautious rule (noted $\bigwedge m$) if one aims at taking into account some possible correlations between classifiers[17]. Let then m_x^{Ω} denote the obtained bba after

combination: $m_x^\Omega = \bigcap_{j=1}^n m_j^\Omega$ or $m_x^\Omega = \bigwedge_{j=1}^n m_j^\Omega$. From m_x^Ω, other belief functions can be derived that will be used for class decision: the pignistic probability[9], the belief function and the plausibility function[8]. Classically, a decision by maximizing the belief on singleton hypotheses is interpreted as following a pessimistic criterion whereas maximizing the plausibility (on singleton hypothesis) as following an optimistic criterion and maximizing the pignistic probability is often seen as a compromise by equidistribution of the mass of the compound hypotheses on their singleton hypotheses. In this study, we have considered the optimistic rule.

Finally, let us introduce a performance measure taking into account the plausibility function in a soft way. Indeed, we aim at evaluating the confidence in a correct decision versus the confidence in a wrong decision. Then, for the set of the labeled test samples $(x_i, y_i, \tilde{y}_i)_{i=1\ldots N}$ where x_i is the observation, y_i the actual label and $\tilde{y}_i$ the estimated label, we define:

$$A_{eval} = \frac{\langle Pl_c \rangle}{\langle Pl_w \rangle},$$
$$\text{with} \quad \langle Pl_c \rangle = \left(\frac{\sum_{(\tilde{y}_i = y_i)} Pl_{x_i}(\tilde{y}_i)}{\sum_{(\tilde{y}_i = y_i)} 1} \right), \quad \langle Pl_w \rangle = \left(\frac{\sum_{(\tilde{y}_i \neq y_i)} Pl(\tilde{y}_i)}{\sum_{(\tilde{y}_i \neq y_i)} 1} \right). \tag{1}$$

A_{eval} values are in [0,+∞). The closer to 1 $\langle Pl_c \rangle$ is, and the closer to 0 $\langle Pl_w \rangle$ is, the more reliable the multi-class classifier is. When A_{eval} is close to 1 then the correct and the wrong classifications have the same order of plausibility. When A_{eval} is lower than 1, the wrong classifications have a greater plausibility than the correct ones which means the plausibility is a bad indicator of the results confidence (and correctness). Conversely, when A_{eval} is greater than 1, then the plausibility is a good indicator of the reliability of the classification decision.

To conclude, let us underline the three main differences of the proposed approach with usual methods used to combine binary classifiers:

- The evidential calibration allows identifying "rare" scores for a given classifier and taking into account the fact that the samples that were not represented well in the training set present a lower confidence level.
- Thanks to the operator of deconditioning, we are able to handle the classifiers classically considered as "non-reliable" for some specific classes such as the classifiers OVO (e.g. classifier {1}vs{2} when considering class labeled 3). In classic decoding (such as loss-based one) a constant penalty is assigned[7]. Using compound hypothesis in the framework of BFT, we are now able to interpret the results of this classifier.
- In the case where hypothesis *A* (or *B*) handled by a given simpler classifier represents a compound class, we do not longer have to assume that the included singleton classes are equiprobable but we are able, when the information provided by another simple classifier allows it and according to the transferable belief model[9], to transfer completely our belief to the right subset of *A*.

3. EXPERIMENTS

In this section we describe experiments performed on real waste hyperspectral data provided by Veolia laboratories. We had two objectives: evaluating different strategies to construct the ECOC matrix in the evidential framework and evaluating the interest of combining different data derived from initial hyperspectral data. Indeed, we noticed some complementarities in terms of classification performance between classifiers when considering different input data: either the first derivative of the hyperspectral spectra or the second derivative. Note that, in a similar way, Bai et al. have used[13] different features as inputs of simple classifiers to get better and more independent classifiers.

3.1 Experimental configuration

As with any classification method, the performance of a SVM classifier strongly depends on the features extracted from raw input data. Classical preprocessing on each pixel spectrum involves filtering and derivation at different orders. Specifically, the Savitsky-Golay filter is widely used[18,19] for hyperspectral data analysis. This filter fits a low degree polynomial on data within a sliding window having fixed size. It allows us to smooth the data and to compute the derivatives from the fitted polynomials. Considering different derivative orders (typically 0, 1 and 2) appears all the more justified since, for classification, not the whole spectrum is considered but only some selected features, in order to reduce both the data complexity and the correlation between the bands. Indeed, keeping in mind the constraint of the processing time, dimensionality reduction has to be considered for the classifier input data. Usually, a principal component analysis (PCA) is performed[20,21] on the filtered spectra or on the first derivative of hyperspectral spectra.

Specifically to our application case, preprocessing involves the computation of different derivative orders (1 and 2) of the spectrum by the Savitsky-Golay filter and then, for each of these derivatives, the computation of the PCA provides

the input data for the SVM classifiers. The number of selected components, that is set to represent 99% of the information, varies between 3 to about 20 whereas the whole spectrum dimensionality is about 275. In the following, the input data denoted D1 and D2 where subscript denotes the derivation order are thus the outputs of PCA applied to derivative orders 1 and 2.

3.2 Datasets

The sample sets used for these experiments have been collected in the Veolia laboratories using a starter-kit hyperspectral sensor (about 275 wavelengths) configured for lab experiments with halogen lamps and a 30 cm large linear stage, as well as specimen boards with small material samples: four boards called Paper, Plastic1, Plastic2a, Plastic2b. We are looking for 9 classes, namely 7 polymers classes (not listed here for paper shortness) and 2 fibrous classes (paper and cardboard). The hyperspectral images offered the possibility to identify manually the regions of interest from specimen boards to locate the different materials, and three different datasets were then automatically extracted. Each sample corresponds to a spectrum observed at a given pixel of a board.

- The 'training dataset' has 1000 samples per class and is used for SVM training. It allows for the estimation of each SVM classifier 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 belief functions from the SVM scores via Xu's evidential calibration.
- The last dataset has 1000 samples per class and is used for test and performance estimation. In addition to these samples from previous boards, the test dataset also include a board, called 'Superposition' and exclusively used for validation, that presents real objects stacked on top of each other to provide more realistic conditions.

3.3 Experimental results

The different ECOC building methods we have considered are the following: the classic OVO and OVA strategies, dense random codes, expert-knowledge and data-driven ECOC (CMSECOC[12]). The CMSECOC were built using the 1vsALL confusion matrix with a percentage of the 10% closest classes to build the super-classes. Those methods have been applied to input data D1 and D2.

Figure 1 shows the difference between correct classification rates (CCR) versus the boards for the 15 tested strategies. The reference CCR are those obtained by the OVO often seen as an efficient strategy because of the number of involved simple classifiers, 36 in our case. According to Figure 1, we note that the ordering of the different classifiers according to performance may differ: for instance, for Plastic2a board, the OVO is indeed among the first (best) classifiers, whereas for Superposition board, the OVO is among the last (worst) classifiers. However, we can exhibit some ECOCs that perform correctly on all boards such as a random KVR one based on 10 simple classifiers and an expert knowledge-based one. We also note that best performance is not necessary achieved by the strategies involving the highest numbers of simple classifiers.

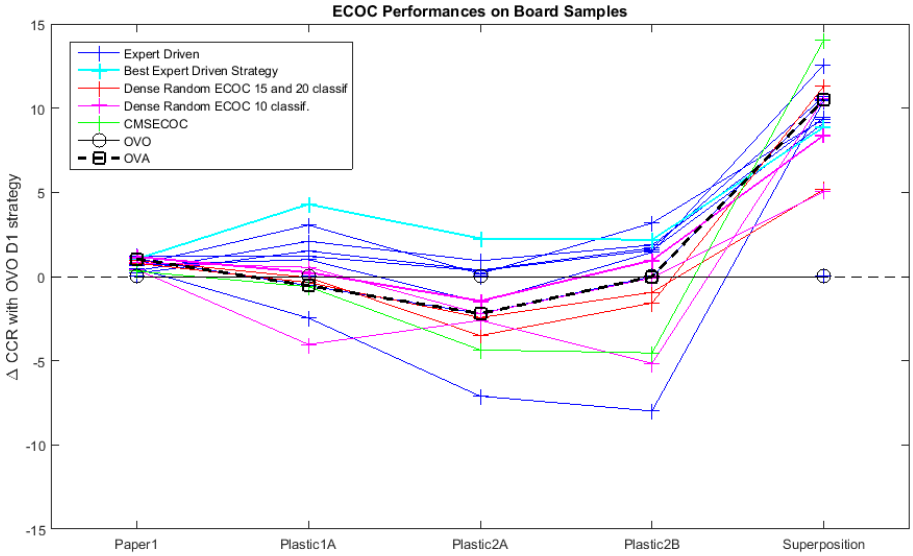

Figure 1: Difference between CCR obtained by a given evidential ECOC and OVO versus the considered boards; case of D1 data.

Then, we perform fusion between bbas derived from evidential ECOC using either D1 or D2 data. Fusion has been tested either between the same ECOC used for D1 and D2, or between two ECOC respectively providing the best result for D1 and the best result for D2 (with respect to the test boards), and or between the ECOC respectively provided by the data-driven strategy for D1 and for D2. We underline that, since we use an associative rule to combine the bbas (both the conjunctive and the cautious rules are associative), we are able to combine partial results (such as those obtained considering a given input data, D1 or D2 in our case) regardless of combination ordering and number.

Figure 2 shows the CCR versus the proposed A_{eval} criterion for the different evidential ECOC strategies only considering D1 or D2 or combining D1 and D2. Comparing D1 and D2 performance, we note that using D1 data allows us to achieve much better results. However, the combination of both data (D1 and D2) outperforms the only-D1 approach. Such an improvement provided by fusion generally implies two things: first, a complementarity of the errors of the two combined datasets (here outputs using either D1 or D2 data) and second, a pertinent modeling of the information provided by the two datasets involving in particular the modeling of the imprecision. Thus, we conclude that the proposed BFT model is pertinent. Not shown on the figure, we have also observed that the approach consisting in combining the best classifier strategies for each derivative order does not provide systematically the best results. This highlights the fact that performance is not the only criterion for choosing classifiers to combine: without an effective complementarity and pertinent modeling of the imprecision, there may be no real improvement brought by fusion process. Finally, from Figure 2, we note that the fusion process also improves the A_{eval} criterion that implies that the belief values themselves (not only the argmax of them) could be interpreted for auto-evaluation of the classification result.

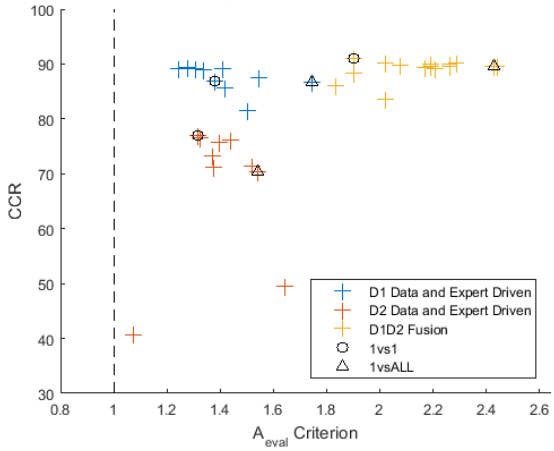

Figure 2: CCR versus A_{eval} criterion; case of D1 data, D2 data and combination of D1 and D2 data.

Finally, Figure 3 shows an example of image results obtained considering D1, D2 and D1-D2 combination. Some pixels ill-classified both in D1 and D2 results are well-classified after combination of these pieces of information.

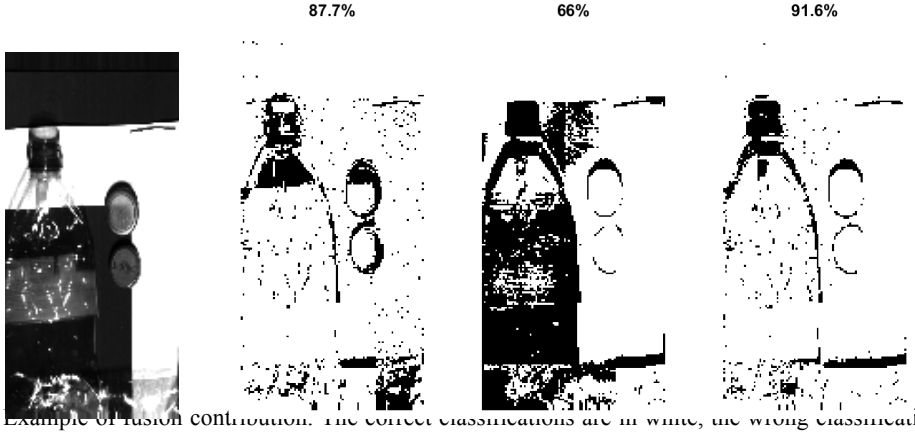

Figure 3: Example of fusion contribution. The correct classifications are in white, the wrong classifications in black. From left to right: picture of the test board Superposition, correct classification for D1 input, D2 input and the fusion of the results.

4. CONCLUSION AND PERSPECTIVES

In this study, we proposed to evaluate classic ECOC building strategies using the belief functions (BF). Indeed we have stressed several points where imprecision appears intrinsically within the ECOC formulation and we know that BF framework allows modeling such imprecision or partial ignorance in addition to modeling uncertainty. Then, we proposed a multi-class classification process involving three main steps: evidential calibration of the binary classifier, projection of the obtained bbas from the binary classifier discernment frame to the multi-class one, and combinations of the obtained bbas. Our classification process also allowed us to combine in a transparent way (in particular associatively) the results obtained considering different datasets. Considering hyperspectral data, SVM classifiers have been advocated by numerous authors so that we focus on these classifiers, even if our approach apply with any binary classifiers that provide a score (Adaboost etc.). First results on waste sorting application show the efficiency of our approach and the interest of combining different features (derivatives from the hyperspectral filtered data in our case). They also show a dependency of the performance on the criteria considered (CCR, number of classifiers for processing time, reliability).

Therefore future work aims at asserting a method to choose the ECOC strategy without testing and training all the possible classifiers and which would take into account the uncertainty. We saw that the CMSECOC approach was interesting, adding redundancy where it is most needed, but the construction of the superclass requires some threshold determination. Besides, such an approach is strongly related to the initial multi-class classifier and does not take into account the individual performance of the simple classifiers used in final multi-class classifier. Then, we will investigate

a possibility to define ECOC approaches that are both data and performance-driven, with the notion of performance not only taking into account the actual CCR but also the auto-evaluation of each simple classifier.

REFERENCES

[1] Cortes, C. and Vapnik, V.,"Support-vector networks", *Machine Learning*, **20**(3), 273–297 (1995).

[2] Vapnik, V. *The Nature of Statistical Learning Theory*, Springer, (1995).

[3] Kuo, B.-C., Ho, H. H., Li, C.H, Hung, C.C., Taur, J.S., "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 (2014).

[4] Melgani, F. and Bruzzone, L., "Classification of hyperspectral remote sensing images with support vector machines", *IEEE Transactions on Geoscience and Remote Sensing*, **42**(8):1778-1790 (2004).

[5] Samiappan, S., Prasad S., Bruce L. M., "Non-uniform random feature selection and kernel density scoring with svm based ensemble classification for hyperspectral image analysis", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **6**(2):792-800 (2013).

[6] Dietterich, T. G. and Bakiri, G., "Solving multiclass learning problems via error-correcting output codes", *Journal of Artificial Intelligence Research*, **2**: 263–286 (1995).

[7] Allwein, E., Schapire, R., Singer, Y., "Reducing multiclass to binary: a unifying approach for margin classifiers", *Machine Learning Research*, **1**:113-141 (2002).

[8] Shafer, G., *A mathematical theory of evidence*, Princeton: Princeton University press (1976).

[9] Smets, P., and Kennes, R., "The transferable belief model", *Artificial intelligence*, **66**(2):191-234 (1994).

[10] Xu Ph., Davoine F., Zha H., Denoeux T., "Evidential calibration of binary svm classifiers", *International Journal of Approximate Reasoning*, **72**:55-70 (2016).

[11] Lachaize, M., Le Hégarat-Mascle S., Aldea E., Maitrot, A., Reynaud R., "SVM Classifier fusion using belief functions: application to hyperspectral data classification", *Proceedings of BELIEF'2016*: 113-122 (2016).

[12] Zhou, J., Yang, Y., ZHANG, M. and XING, H., "Constructing ECOC based on confusion matrix for multiclass learning problems", *Science China Information Sciences*, **59**(1) (2016).

[13] Bai, X., Niwas, S. I., Lin, W., Ju, B. F., Kwoh, C. K., Wang, L., Chew, P. T. , "Learning ECOC code matrix for multiclass classification with application to glaucoma diagnosis", *Journal of medical systems*, **40**(4):1-10 (2016).

[14] Utschick W., Weichselberger W., "Stochastic organization of output codes in multiclass learning problems", *Neural Computation* **13**: 1065–1102 (2001)

[15] Pujol O., Radeva P., Vitria J. "Discriminant ECOC: a heuristic method for application dependent design of error correcting output codes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**: 1001–1007 (2006).

[16] Escalera S., Tax D M J, Pujol O., et al. « Subclass problem-dependent design of error-correcting output codes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30:**1–14

[17] Denœux, T., "Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence", *Artificial Intelligence*, **172**:234-264 (2008).

[18] King, R.L., Ruffin C., LaMastus, F., Shaw D., "The analysis of hyperspectral data using savitzky-golay filtering-practical issues", *Proceedings of IGARSS'99*, **1:**398-400 (1999).

[19] Vaiphasa, C., "Consideration of smoothing techniques for hyperspectral remote sensing", *ISPRS Journal of Photogrammetry and Remote Sensing*, **60**(2):91-99 (2006).

[20] Chen, G, and Qian, S. E., "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage", *IEEE Transactions on Geoscience and remote sensing*, **49**(3):973-980 (2011).

[21] Cavalli, R. M., Licciardi, G. A., Chanussot, J., "Archaeological Structures Using Nonlinear Principal Component Analysis Applied to Airborne Hyperspectral Image", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2013).