Parameter identification in Choquet Integral by the Kullback-Leibler divergence on continuous densities with application to classification fusion

Emmanuel Ramasso[1] **Sylvie Jullien**[2]

[1]FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, AS2M Dep., 25000 Besançon, France
[2]Dynamic3D, Route de Demigny, 71102 Chalon-Sur-Saône, France

Abstract

Classifier fusion is a means to increase accuracy and decision-making of classification systems by designing a set of basis classifiers and then combining their outputs. The combination is made up by non linear functional dependent on fuzzy measures called Choquet integral. It constitues a vast family of aggregation operators including minimum, maximum or weighted sum. The main issue before applying the Choquet integral is to identify the $2^M - 2$ parameters for M classifiers. We follow a previous work by Kojadinovic and one of the authors where the identification is performed using an information-theoritic approach. The underlying probability densities are made smooth by fitting continuous parametric and then the Kullback-Leibler divergence is used to identify fuzzy measures. The proposed framework is applied on widely used datasets.

Keywords: Information fusion, Fuzzy measures, Relative Entropy, Health assessment, Classification

1. Introduction

In most of pattern recognition tasks, a first step consists in extracting relevant features bringing information on classes of interest. Features are then transformed into membership degrees according to the different classes by entities called classifiers.

A classifier is a system that takes as inputs a Q-dimensional vector $X^{\mathrm{T}} = [x_1 \dots x_Q]$ (also called features or attributes) and generates a degree of confidence in the statement "X belongs to class ω_j" for all classes in $\Omega = \{\omega_1 \dots \omega_K\}$.

Multiple Classifier Systems (MCS) [1] are designed when complementary (and sometimes redundant) information sources (here classifiers) are used in order to improve classification accuracy and decision-making. MCS can be also viewed as information fusion systems where inputs are classifiers. MCS can take several forms and among them the parallel one which takes as input $M \times K$ partial degrees of confidence and generates an output, called global confidence degree, made up of K degrees of confidence (one for each class). We denote by $\phi_{m,j}(X)$ the degree of confidence delivered by a classifier $m \in \{1 \dots M\}$ for class $\omega_j \in \Omega$ given the observation X.

Usual combinations of classifier outputs include product, naive Bayes and decision templates among others [1] but most of them can be used provided each output represents an independent source of information. However, the independance assumption is not always satisfied. To face this problem, an approach that considers interactions among classifier outputs such as fuzzy integrals and in particular the Choquet Integral [2, 3] can be used. The explicit interaction coefficients (in the 2-additive form) provide very interesting information on complementarity and redundancy of the fused data which can also be used for subset selection [4].

A fuzzy integral is a type of non-linear functional dependent on fuzzy measures which constitutes a vast family of aggregation operators including many widely used operators (minimum, maximum, weighted sum, ordered weighted sum and so on) [5]. In order to be combined by the Choquet Integral, the commensurability [6] of classifier outputs must be satisfied. That means the classifier outputs must be defined on the same measurement scale.

The combination of all partial confidence degrees provided by classifiers is thus made up by a Choquet Integral which is described in the next section.

2. Choquet capacities and Choquet Integral

Let the M classifiers (sources) be denoted by $\Theta = \{\theta_1, \theta_2 \dots \theta_M\}$. A fuzzy measure μ_k for a given class ω_k weighs the importance of a subset of sources $S \subseteq \Theta$ and is defined by [2, 3]:

$$\begin{array}{rccc} \mu_k & : & 2^\Theta & \rightarrow & [0,1] \\ & & S & \mapsto & \mu_k(S) \end{array} \quad (1)$$

satisfying the following constraints:

- $\mu_k(\emptyset) = 0$ and $\mu_k(\Theta) = 1$
- $S \subseteq T \Rightarrow \mu_k(S) \leq \mu_k(T)$ (monotonicity)

The fuzzy measure is said:

- *additive* when $\mu_k(S \cup T) = \mu_k(S) + \mu_k(T)$, $\forall S, T \subseteq \Theta / S \cap T = \emptyset$ (probability measure),
- *super-additive* when $\mu_k(S \cup T) \geq \mu_k(S) + \mu_k(T)$, $\forall S, T \subseteq \Theta / S \cap T = \emptyset$,

- *sub-additive* when $\mu_k(S \cup T) \leq \mu_k(S) + \mu_k(T)$, $\forall S, T \subseteq \Theta / S \cap T = \emptyset$.

In classification problems, the fuzzy measure is used in order to take into account interactions between sources. One fuzzy measure is tuned for each class and each discrete Choquet Integral aggregates the information provided by the sources as follows [2, 3]:

$$C(\phi_1, \ldots, \phi_M) = \sum_{i=1}^{M} (\phi_{(i)} - \phi_{(i-1)}) \cdot \mu_k(S_{(i)}) \quad (2)$$

where $\mu_k(S_{(i)})$ is the importance of subset of sources $S_{(i)} = \{\theta_{(i)}, ..., \theta_{(M)}\}$ and the value $\phi_{(i)}$ is provided by source $\theta_{(i)}$. The notation (.) indicates a permutation of indices according to the values provided by the sources such as $\phi(1) \leq \phi(2) \leq ... \leq \phi(M) \leq 1$ (and by convention $\phi_{(0)} = 0$). The Choquet integral thus coincides to the weighted arithmetic mean when the fuzzy measure is additive.

One approximation of Eq. 2 called *2-additive Choquet Integral* is often used and consists in considering a 2-order additive capacity which takes into account both the weights of each source and the interaction between pairs. The weight ν_i of a source θ_i (for the detection of class ω_k) and the coefficient I_{ij} of interaction between both sources θ_i and θ_j can be obtained from the fuzzy measure μ_k by [2, 3]:

$$\nu_i = \sum_{T \subseteq \Theta \setminus i} \frac{(M - |T| - 1)!|T|!}{M!} \times \Big(\mu_k(T \cup \{i\}) - \mu_k(T)\Big) \quad (3a)$$

$$I_{ij} = \sum_{T \subseteq \Theta \setminus \{i,j\}} \frac{(M - |T| - 2)!|T|!}{(M-1)!} \times \Big(\mu_k(T \cup \{i,j\}) - \mu_k(T \cup \{i\}) - \mu_k(T \cup \{j\}) + \mu_k(T)\Big) \quad (3b)$$

These parameters are interesting for interpreting the fuzzy measure and also to highlight which sources are important and how they interact. When interactions between two sources are positive, the sources are said complementary while they are said redundant when interactions are negative.

The problem of Choquet Integral parameters identification was treated by several authors [4, 7]. In the context of classification as considered here (where the classes are known), the method proposed by Grabisch [3] and called Heuristic Least Mean Square (HLMS) is often used. However it requires the global scores (the real output of the Choquet Integral) to perform the optimization of the fuzzy measure. Recently, two information theoritic methods based on entropy [8, 6] and on relative entropy [9] were proposed. The former is purely unsupervised and requires only degrees of confidence of classifiers while the latter requires the ground truth, i.e. the real class of each pattern. The relative entropy-based approach is supervised but requires less prior information than HLMS.

3. Identification of Choquet Integral parameters based on discrete relative entropy

3.1. A probabilistic view

Each fuzzy value $\mu_k(S)$ expresses the relative importance of a subset S for distinguishing class ω_k *from the others* [8, 6]. In order to identify them, the authors in [9] proposed to use the relative entropy, also called Kullback-Leibler divergence (KL) [10], which is a measure of divergence between two densities. It could be interpreted as the expected discrimination information between two hypotheses and thus appears very natural for the identification of fuzzy measures.

To compute KL, one needs first to compute:

- the distribution (say P_k^S) of confidence degrees in class ω_k conditional to class ω_k,
- and the distribution (say $P_{\overline{k}}^S$) of confidence degrees in class ω_k conditional to the other classes ($\overline{\omega_k} = \Omega \setminus \omega_k$),

both given a subset of sources S. These distributions characterize the input data (confidence degrees) and the greater is the difference (calculated by KL) between them, the higher is the discrimination power.

Identifying fuzzy measure using a probabilistic approach was introduced in [8, 6] where the author proposed an unsupervised entropy-based method. When the class is known for each input pattern, the KL-based approach proposed in [9] should be used. It fully exploits the available information provided by the training dataset and, as expected, increases the discrimination power.

3.2. Relative entropy

We assume all confidence degrees to be commensurable values in $[0, 1]$ which is generally true in classification. Let $P_k^\Theta(Y)$ with $Y = (\phi_{1,k}(X), \phi_{2,k}(X) \ldots \phi_{M,k}(X)) \in [0,1]^{|\Theta|}$ (resp. $P_{\overline{k}}^\Theta(Y)$) be the probability that classifiers $1, 2 \ldots M$ *jointly* provide the values $\phi_{1,k}(X)$,$\phi_{2,k}(X)$,... and $\phi_{M,k}(X)$ given the ground truth is class ω_k (resp. given $\overline{\omega_k}$) and observation X. In [9], the distributions were assumed *discrete* and the relative entropy (KL) of both distributions was thus given by:

$$\mathcal{D}(P_k^\Theta || P_{\overline{k}}^\Theta) = \sum_Y P_k^\Theta(Y) \log\left(\frac{P_k^\Theta(Y)}{P_{\overline{k}}^\Theta(Y)}\right) \quad (4)$$

For the sake of simplicity, we will denote by $R_k(S)$ the KL value given by $\mathcal{D}(P_k^S || P_{\overline{k}}^S)$ for a given subset of sources $S \subseteq \Theta$:

$$R_k(S) \equiv \mathcal{D}(P_k^S || P_{\overline{k}}^S) \quad (5)$$

Note that when the distributions P_k^{Θ} and $P_{\overline{k}}^{\Theta}$ are computed, the distributions P_k^S and $P_{\overline{k}}^S$ for $S \subset \Theta$ are obtained by marginalizing out the components $\theta \in \Theta, \theta \notin S$.

To compute Eq. 4, the support of the distribution P_k^S must be included in the support of the distribution $P_{\overline{k}}^S$ otherwise the relative entropy diverges towards infinity. In order to respect this constraint, the skew divergence was used in [9].

3.3. From relative entropy to Choquet capacities

The relative entropy has to satisfy the conditions presented Section 2 in order to be interpreted as a Choquet capacity. For that, the relative entropy $R_k(S)$ for a subset S is normalized as in Kojadinovic's method [8, 6] by the entropy of the whole set of sources $R_k(\Theta)$:

$$\mu_k(S) = \frac{R_k(S)}{R_k(\Theta)} \tag{6}$$

Moreover, the relative entropy is zero when the set S is empty but also when both distributions are identical. Therefore, a source that provides the same degrees of support for a sought-after class ω_k and for the other classes $\overline{\omega_k}$ is assigned a low importance value since it can not distinguish class ω_k from the others. This is exactly the mean of discrimination power.

The relative entropy has also to satisfy the monotonicity constraint (Section 2), i.e. given two sources θ_i and θ_j, the relative entropy has to satisfy the following equations:

$$\mu(\{\theta_i, \theta_j\}) \geq \mu(\{\theta_i\}) \tag{7a}$$
$$\mu(\{\theta_i, \theta_j\}) \geq \mu(\{\theta_j\}) \tag{7b}$$

In order to check these constraints, one can rewrite the relative entropy as [11, 8, 6]:

$$R_k(\{\theta_i, \theta_j\}) = R_k(\{\theta_i\}) + R_k(\{\theta_j \mid \theta_i\}) \tag{8}$$

that is always positive and therefore has a monotonic behavior [11, 8, 6, 9]:

$$R_k(\{\theta_i, \theta_j\}) \geq R_k(\{\theta_i\}) \tag{9a}$$
$$R_k(\{\theta_i, \theta_j\}) \geq R_k(\{\theta_j\}) \tag{9b}$$

This reasoning can be extended easily to larger subsets of sources. Therefore, the normalized relative entropy satisfies all the constraints in order to be interpreted as a Choquet capacity.

3.4. Modeling positive and negative interactions

When sources θ_i and θ_j, that provide distributions P_k^S and $P_{\overline{k}}^S$, are independent, the relative entropy has an additive behavior [11, 8, 6]:

$$R_k(\{\theta_i, \theta_j\}) = R_k(\{\theta_i\}) + R_k(\{\theta_j\}) \tag{10}$$

When sources θ_i and θ_j are interacting one each other, the relative entropy can be expressed by:

$$\begin{aligned} R_k(\{\theta_i, \theta_j\}) &= R_k(\{\theta_i\}) + R_k(\{\theta_j\}) \\ &+ \Big(R_k(\{\theta_j|\theta_i\}) - R_k(\{\theta_j\}) \Big) \end{aligned} \tag{11}$$

where the last term can be negative or positive according to sources θ_i and θ_j implying that the identified Choquet capacities can be super-additive or sub-additive. Therefore the proposed method is able to model and identify both *positive* and *negative* interactions whereas Kojadinovic's approach can only identify negative ones.

4. Identification of Choquet Integral parameters based on continuous relative entropy

4.1. On using a continuous approach

The core of the KL-based method is the evaluation of the multidimensional probability distributions (P_k^S and $P_{\overline{k}}^S$). In [8, 6, 9], the distributions were computed using discretization of confidence degrees (and histograms). We rather propose to remain in the continuous space (the space of the degrees of confidence) and to use parametric continuous densities for confidence degrees modelling. These densities allow to:

- Ensuring an infinite support for the distributions and therefore avoiding using artificial methods to solve the problem of minimum support.
- Avoiding the necessity of finding the optimal number of bins for the histograms. This could be a serious problem for high dimensional data such as in image processing or in complex systems diagnosis.
- Obtaining a more precise paving of the input space and therefore generating smooth distributions and improving the computation of the relative entropy by summing over more data points sampled from the continuous densities.
- Simplifying the computation of marginalizations (according to the family of densities).

Fig. 1 depicts the problem of finding the number of bins for discrete histograms. We drew 100 points from a mixture of five Gaussians with equal probability and with means 0, 10, 25, 35, 50 and unit variance. We then computed histograms (depicted in the first three figures) with 10, 50 and 100 bins and we also run an EM in order to identify automatically the parameters of a continuous density made up of five components. The results are very different with a preference given to the last figure.

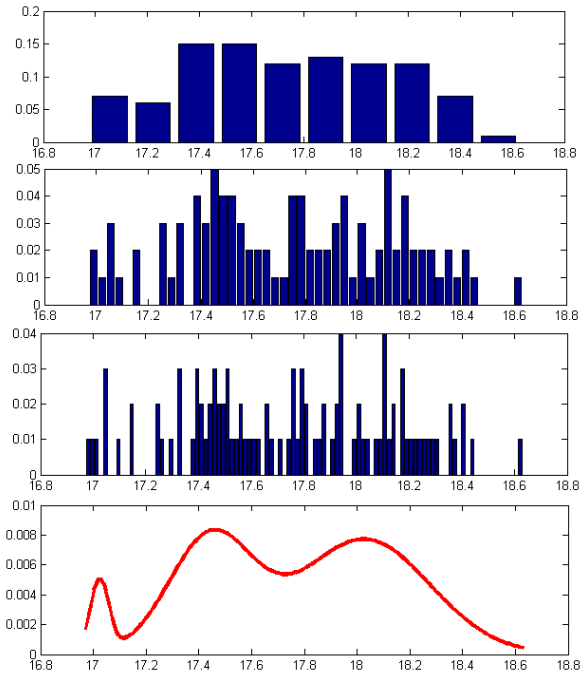

Figure 1: Role of the number of bins (see comments in the text).

4.2. Modelling inputs by continuous densities

4.2.1. Modelling

We assume that the joint probability density function related to P_k^Θ (and similarly for $P_{\overline{k}}^\Theta$) has a continuous and parametric form. For example, we consider mixtures of Gaussians which are very general and have interesting properties:

$$P_k^\Theta(Y) = \sum_{a=1}^{L_k} c_{k,a} \cdot f_{k,a}(Y) \tag{12}$$

where

$$f_{k,a}(Y) = \mathcal{N}\left(Y|\alpha_{k,a}, \Sigma_{k,a}\right) \tag{13}$$

with $Y = (\phi_{1,k}(X), \phi_{2,k}(X) \dots \phi_{M,k}(X)) \in [0,1]^{|\Theta|}$ the joint observation of degrees of confidence, $c_{k,a}$ the mixing coefficient of component a (among L_k, with $\sum_a c_{k,a} = 1$) and $\mathcal{N}(Y|\alpha_{k,a}, \Sigma_{k,a})$ a multidimensional normal density with mean $\alpha_{k,a}$ and covariance $\Sigma_{k,a}$ (positive definite):

$$\mathcal{N}(Y|\alpha_{k,a}, \Sigma_{k,a}) = \frac{\exp\left(-\frac{1}{2}\left(Y-\alpha_{k,a}\right)^{\mathrm{T}} \Sigma_{k,a}^{-1}\left(Y-\alpha_{k,a}\right)\right)}{(2\pi)^{M/2}|\Sigma_{k,a}|^{1/2}} \tag{14}$$

The dimension of the parameters is the same as Θ which is M. The same expression holds for $P_{\overline{k}}^\Theta$:

$$P_{\overline{k}}^\Theta(Y) = \sum_{b=1}^{L_b} c_{\overline{k},b} \cdot g_{\overline{k},b}(Y) \tag{15}$$

with different parameters indexed by subscripts $(\overline{k}, b)$.

4.2.2. Learning parameters of densities

The parameters of the densities can be estimated automatically by standard methods such as the Expectation-Maximization algorithm (EM [12]) where L, the number of components, can also be estimated.

When the parameters of the distributions P_k^Θ and $P_{\overline{k}}^\Theta$ have been specified, it is easy to compute P_k^S and $P_{\overline{k}}^S$ for subsets $S \subset \Theta$ by marginalization. In case the joint density related to P_k^Θ is represented by a mixture of multivariate Gaussians, the marginal is also a mixture of multivariate Gaussians where some components (marginalized out) have been eliminated. In particular, the $|S|$ components of the mean vector of the marginal are the means of the variables in S and its covariance matrix is composed of the pairwise covariances of the same variables.

4.3. Continuous relative entropy

For two unimodal multivariate normal densities f_k and $g_{\overline{k}}$ (with $L_a = L_b = 1$), the KL has an exact closed form [13]:

$$\begin{aligned} R_k^{\mathrm{ex}}(S; f_k; g_{\overline{k}}) = \frac{1}{2}\Bigg(&\log\left(\frac{|\Sigma_{\overline{k}}|}{|\Sigma_k|}\right) + \dots \\ &\mathrm{Tr}\left(\Sigma_{\overline{k}}^{-1}\Sigma_k\right) - M + \dots \\ &\left(\mu_k - \mu_{\overline{k}}\right)^{\mathrm{T}} \Sigma_{\overline{k}}^{-1}\left(\mu_k - \mu_{\overline{k}}\right)\Bigg) \end{aligned} \tag{16}$$

When densities are multimodal, the continuous relative entropy is obtained by integrating on the support of P_k^S, $\mathrm{Supp}(P_k^S) = \{Y : P_k^S(Y) > 0\}$:

$$R_k(S) = \int_{Y \in \mathrm{Supp}(P_k^S)} P_k^S(Y) \log\left(\frac{P_k^S(Y)}{P_{\overline{k}}^S(Y)}\right) dY \tag{17}$$

To evaluate this expression, several methods can be used [13]. In this paper, we have used Monte Carlo sampling (MC) and variational approximation (VA).

The MC method consists in drawing samples from the mixture associated to P_k^S. For that, a component is chosen randomly using the distribution $c_{k,\cdot}$. A continuous sample is then drawn from the associated Gaussian component and the density is evaluated. Given $\{Y_i, i = 1 \dots N\}$ the set of i.i.d. sampled points, we can approximate the integral (17) by its MC estimate:

$$R_k^{\mathrm{MC}}(S) = \frac{1}{N}\sum_i \log\left(\frac{P_k^S(Y_i)}{P_{\overline{k}}^S(Y_i)}\right) \to \mathcal{D}(P_k^S || P_{\overline{k}}^S) \tag{18}$$

The precision of the evaluation of KL depends obviously on the number of simulations.

In the VA method, the integral is approximated

by the following expression [13]:

$$R_k^{\mathrm{VA}}(S) = \sum_{a=1}^{L_a} c_{k,a} \cdot \log \left(\frac{\sum_{a'=1}^{L_a} c_{k,a'} \cdot e^{-R_k^{\mathrm{ex}}(S; f_{k,a}; f_{k,a'})}}{\sum_{b=1}^{L_b} c_{\overline{k},b} \cdot e^{-R_k^{\mathrm{ex}}(S; f_{k,a}; g_{\overline{k},b}))}} \right) \tag{19}$$

where $R_k^{\mathrm{ex}}(S; f_{k,a}; g_{\overline{k},b})$ is the exact value of KL between component a of f_k and component b of $g_{\overline{k}}$ given by Eq. 16.

4.4. Final algorithm

The overall algorithm for computing the fuzzy measure is as follows:

Require: $\mathcal{L}_k$ the set of confidence degrees in class ω_k of M classifiers given the ground truth is class ω_k

Require: $\mathcal{L}_{\overline{k}}$ the set of confidence degrees in class ω_k of M classifiers given the ground truth are classes different from ω_k

Ensure: the fuzzy measure μ_k for class ω_k

1: $P_k^{\Theta} \leftarrow$ Estimate the parameters of the densities on $\mathcal{L}_k$
2: $P_{\overline{k}}^{\Theta} \leftarrow$ Estimate the parameters of the densities on $\mathcal{L}_{\overline{k}}$
3: $\mu_k(\Theta) \leftarrow 1$, $\mu_k(\emptyset) \leftarrow 0$
4: $R_k(\Theta) \leftarrow$ Apply Eq. 17 with P_k^{Θ} and $P_{\overline{k}}^{\Theta}$
5: **for all** $S \subset \Theta$ **do**
6: $\quad P_k^S \leftarrow$ marginalize P_k^{Θ} on S
7: $\quad P_{\overline{k}}^S \leftarrow$ marginalize $P_{\overline{k}}^{\Theta}$ on S
8: $\quad R_k(S) \leftarrow$ Apply Eq. 17 with P_k^S and $P_{\overline{k}}^S$
9: $\quad \mu_k(S) \leftarrow \dfrac{R_k(S)}{R_k(\Theta)}$ (Eq. 6)
10: **end for**

From μ_k, one can compute the weight of each source (i.e. classifier) and their interactions using Eq. 3b. These values can help an end-user or any people interested in knowing which classifiers contribute to the final results as well as how they interact.

5. Experiments

A toy example is first presented. Then, the proposed method is evaluated on two datasets from UCI [14]: vehicle and image segmentation. Classifiers used were the following: Evidential Neural Network (EvNN) [15] (with 4 prototypes for each class), Evidential Nearest Neighborhood (EvKNN) [16] (with $K = 5$) and Support Vector Machines (SVM) [12] (with a Gaussian Kernel of size 2.2). Classifiers were learnt using 1-vs-1 strategy for each class, and the final scores were obtained by using a weighted vote. SVM scores were transformed into probabilities using a sigmoid transfer function. Note that classifier parameters were not "optimized" for each dataset, since the goal is here to assess the fusion process. The KL was assessed using the MC method using $1e6$ samples.

5.1. A toy example

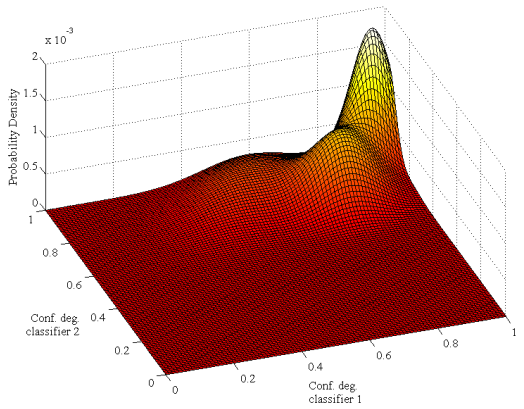

Figure 2: A density P_k^{Θ}.

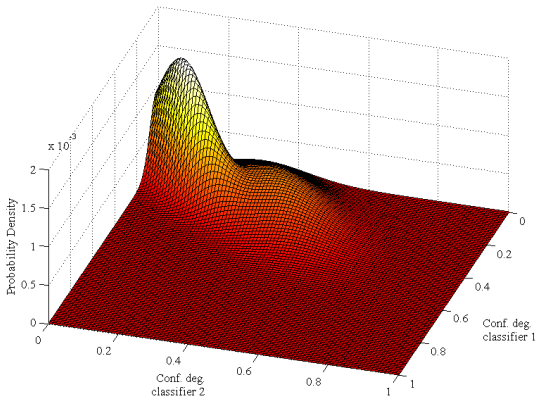

Figure 3: A density $P_{\overline{k}}^{\Theta}$.

As an example, let consider two classifiers θ_1 and θ_2 with confidence degrees in ω_k, given the ground truth is ω_k, being distributed according to Fig. 2, and according to Fig. 3 for confidence degrees in ω_k given the ground truth is another class $\overline{\omega_k}$. From these densities, we are looking for characterizing the importance of the coalition $\{\theta_1, \theta_2\}$ in distinguishing ω_k from $\overline{\omega_k}$.

In Fig. 2, the confidence degrees of classifier 1 are globally close to unity given class ω_k. That means classifier 1 often provides high scores for ω_k given the ground truth is ω_k. Classifier 2 however seems to provide some results close to 0.5 meaning classifier 2 is frequently not certain about the predicted class. Given the ground truth is $\overline{\omega_k}$ (Fig. 3), classifier outputs are globally close to 0 for ω_k. That means classifiers generally provide low values for ω_k when the ground truth is $\overline{\omega_k}$ as expected.

In order to quantify the importance $\mu_k(\{\theta_1, \theta_2\})$ of the coalition $\{\theta_1, \theta_2\}$ given ω_k, we compute the

divergence between both distributions. The higher is the divergence, the higher is the importance of $\{\theta_1, \theta_2\}$ for distinguishing ω_k from the other classes. In this example, densities were obtained using two mixtures with the following parameters:

- given ω_k, $\alpha_{k,1} = [0.1,\ 0.1]$, $\Sigma_{k,1} = [.01,\ 8;\ 8,\ .01]$, $\alpha_{k,2} = [0.2,\ 0.4]$ and $\Sigma_{k,2} = [.02,\ .01;\ .01,\ .02]$ (with equal mixing coefficients $c_{k,a}$).
- given $\overline{\omega_k}$, $\alpha_{\overline{k},1} = [0.8,\ 0.8]$, $\Sigma_{\overline{k},1} = [8,\ 1;\ 1,\ 8]$, $\alpha_{\overline{k},2} = [0.55,\ 0.9]$, $\Sigma_{\overline{k},2} = [.02,\ 8;\ 8,\ .02]$, $\alpha_{\overline{k},3} = [0.95,\ 0.97]$, $\Sigma_{\overline{k},3} = [4,\ 1;\ 1,\ 4]$ (with equal mixing coefficients $c_{k,a}$).

With these parameters, Eq. 17 leads to $R_k(\{\theta_1, \theta_2\}) \approx 9.44$ (with $N = 1.10^6$).

5.2. Vehicle dataset

The UCI's vehicle dataset is a four-classes problem composed of 946 examples almost uniformly distributed between classes. The goal is to classify data into one of the following types of vehicle: OPEL (ω_1), SAAB (ω_2), BUS (ω_3) and VAN (ω_4). The half of the dataset was used for classifier training, and the other half for testing.

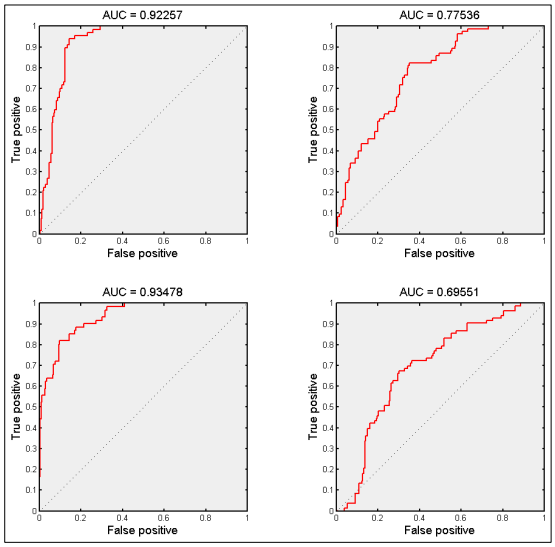

Figure 4: EvNN classification results. Top: classes 1 and 2, bottom: classes 3 and 4.

Figures 4-6 pictorially described ROC curves computed for each class given results of classifiers. Also are displayed the Area Under the Curve (AUC) which reflects the efficiency in detecting the class. These curves can be compared with ROC curves of Figure 7 computed from the results of the fusion process proposed in this paper. Table 1 also gives the obtained fuzzy measures, while interaction and classifier weights computed from them (as detailed previously) are provided in Tables 2 and 3.

ROC curves clearly show the complementarity of individual classifiers. For example, class ω_1 is better recognized using EvKNN (Fig. 5) with almost 98%

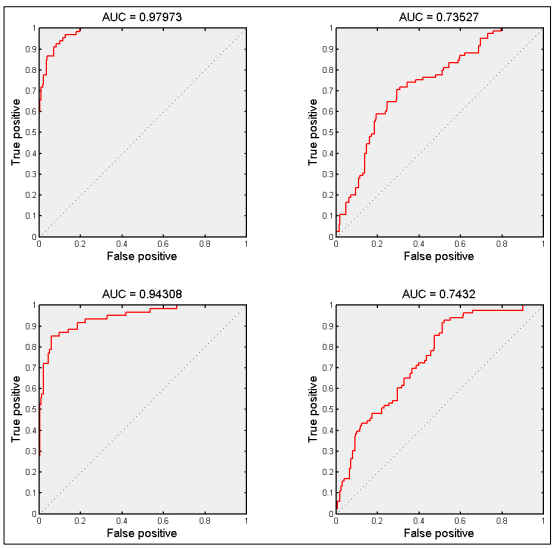

Figure 5: EvKNN classification results. Top: classes 1 and 2, bottom: classes 3 and 4.

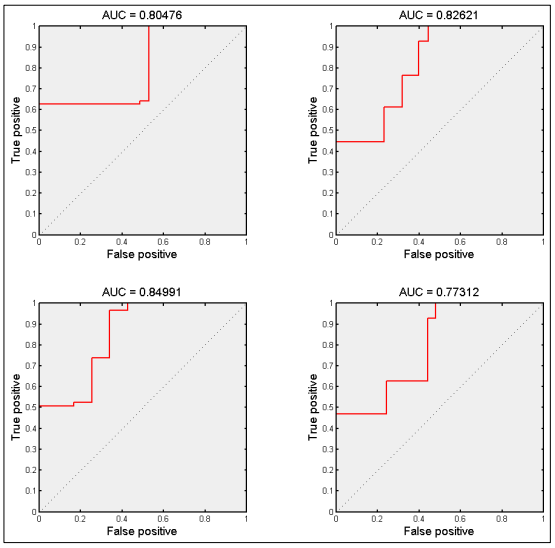

Figure 6: SVM classification results. Top: classes 1 and 2, bottom: classes 3 and 4.

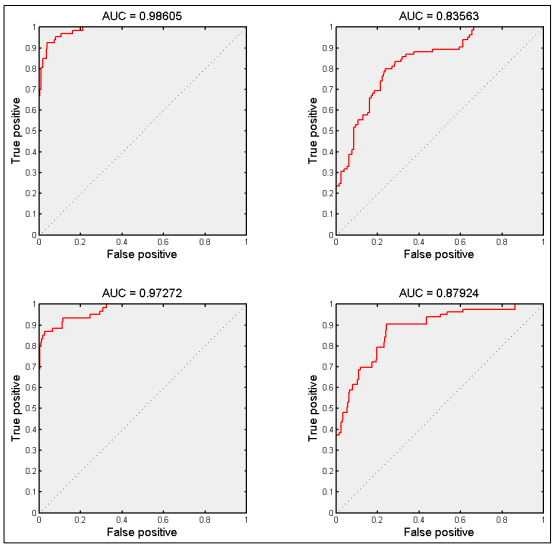

Figure 7: Fusion by Choquet Integral. Top: classes 1 and 2, bottom: classes 3 and 4.

of good classification. Class ω_2 is better recognized using SVM (Fig. 6) with an accuracy close to 83%. Class ω_3 is well recognized (about 93%) by EvNN (Fig. 4) and EvKNN (Fig. 5). Lastly, class ω_4 is better detected by SVM (Fig. 6).

set	ω_1	ω_2	ω_3	ω_4
1	0.22	0.39	0.23	0.24
2	0.66	0.57	0.34	0.37
12	0.75	0.70	0.44	0.65
3	0.36	0.29	0.20	0.44
13	0.50	0.60	0.43	0.62
23	0.89	0.84	0.83	0.65

Table 1: The fuzzy measure for each class learnt by the proposed algorithm for the “vehicle" dataset.

set	ω_1	ω_2	ω_3	ω_4
I_{12}	-0.08	-0.20	-0.10	$+0.10$
I_{13}	-0.03	-0.02	$+0.03$	$+0.01$
I_{23}	-0.08	$+0.03$	$+0.33$	-0.10

Table 2: Interaction values associated to the fuzzy measures of Table 1.

set	ω_1	ω_2	ω_3	ω_4
ν_1	0.15	0.26	0.19	0.28
ν_2	0.56	0.47	0.45	0.35
ν_3	0.29	0.28	0.37	0.37

Table 3: Weight values associated to the fuzzy measure of Table 1.

As shown in Figure 7, the proposed fusion process draw benefits of all these classifiers, providing AUCs close to 99%, 84%, 97% and 88% for class ω_1, ω_2, ω_3 and ω_4 respectively (improvement close to 10%). Interaction indexes can explain this result. Indeed, class ω_1, that is well detected by all classifiers, is represented by a fuzzy measure with negative interactions because of redundancy. The highest redundancy is detected for class ω_2 between EvNN and EvKNN ($I_{12} = -0.20$) while the highest complementarity is detected for class ω_3 between EvNN and SVM classifiers ($I_{23} = +0.33$). Weights are also the highest for classifiers with the best accuracies, except for class ω_2.

5.3. Image segmentation dataset

The UCI's image segmentation dataset is a seven-classes problem composed of 210 training examples and 2100 for testing. Initially, features are 19-dimensional but we reduced them to 6 dimensions and we kept features [1 3 4 8 9 17]. The goal is thus to classify data into one of the following types of vehicle: Brickface (ω_1), Sky (ω_2), Foliage (ω_3), Cement (ω_4), Window (ω_5), Path (ω_6) and Grass (ω_7).

For this dataset, we present the results in the form of confusion matrices (Table 4-7) where the ground truth is on columns while results of classifiers are on lines. Tables 4- 6 are confusion matrices of individual classifiers. We here degraded on purpose the results of the first classifier (EvNN) on classes ω_1, ω_3, ω_5 and ω_7 (by adding noise on the parameters trained by the algorithm). As a result, the confusion matrix presents low detection rate for these classes (close to 0 for some of them). We then applied the fusion process.

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_1	0	0	0	6	18	0	0
ω_2	67	264	0	67	15	0	0
ω_3	0	0	41	8	7	0	0
ω_4	233	36	259	219	256	0	0
ω_5	0	0	0	0	3	0	0
ω_6	0	0	0	0	1	300	1
ω_7	0	0	0	0	0	0	299

Table 4: Confusion matrix of EvNN classifier (global acc.: 53.6%, degraded on purpose).

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_1	152	9	26	69	61	0	0
ω_2	38	223	0	55	17	0	0
ω_3	22	1	176	12	45	0	0
ω_4	45	55	13	140	6	30	0
ω_5	43	12	85	24	171	0	14
ω_6	0	0	0	0	0	270	0
ω_7	0	0	0	0	0	0	286

Table 5: Confusion matrix of EvKNN classifier (global acc.: 67.5%)

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_1	98	0	1	3	4	0	0
ω_2	4	78	0	3	0	0	0
ω_3	5	0	59	0	0	0	0
ω_4	18	6	3	49	1	0	0
ω_5	0	0	14	0	94	0	0
ω_6	0	0	0	0	0	111	0
ω_7	175	216	223	245	201	189	300

Table 6: Confusion matrix of SVM classifier (global acc.: 37.6%)

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_1	97	3	3	1	11	0	0
ω_2	55	234	0	58	15	0	0
ω_3	0	0	156	6	20	0	0
ω_4	148	63	62	231	113	3	0
ω_5	0	0	79	4	141	0	7
ω_6	0	0	0	0	0	297	0
ω_7	0	0	0	0	0	0	293

Table 7: Confusion matrix after fusion using MC sampling (global acc.: 69%)

Table 7 shows the confusion matrix of the fusion process result. This matrix clearly shows that the proposed method is able to draw benefits from individual classifiers. Table 8 is the confusion matrix obtained by the fusion process based on the variational approximation of the KL divergence which shows that, for this application, the approximation

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_1	95	3	8	3	25	0	0
ω_2	75	226	0	68	21	0	0
ω_3	5	0	145	6	25	0	0
ω_4	125	71	81	222	100	3	0
ω_5	0	0	66	1	129	0	5
ω_6	0	0	0	0	0	297	2
ω_7	0	0	0	0	0	0	293

Table 8: Confusion matrix after fusion using variational approximation (global acc.: 66%)

is satisfying.

Interaction values obtained in this application are shown in Table 9. For classes ω_1, ω_3 and ω_5 interactions between classifiers emphasize complementarity. In particular, classifiers EvNN and EvKNN present the strongest complementarities. This is represented in confusion matrix where these classifiers mix sometimes several classes while classifier 3 confuses between class ω_7 and the others.

In general, an efficient classifier has also a relatively high weight (Tab. 10), and if it is not the case, interaction values provide compensation.

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
I_{12}	0.32	−0.29	0.15	−0.22	0.03	−0.17	−0.28
I_{13}	0.06	0.00	0.06	−0.04	0.03	−0.02	−0.03
I_{23}	−0.09	0.07	0.02	−0.09	0.05	0.05	−0.02

Table 9: Interaction values for application 2.

class	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ν_1	0.32	0.54	0.35	0.69	0.33	0.52	0.48
ν_2	0.54	0.28	0.50	0.24	0.50	0.33	0.48
ν_3	0.14	0.18	0.15	0.07	0.17	0.15	0.04

Table 10: Weight values for application 2.

6. Conclusion

We proposed an information-theoritic approach relying on Kullback-Leibler divergence for fuzzy measures identification in the context of supervised classification. The use of well known parametric and continuous functions for the representation of confidence degrees allows to simplifying the estimation of joint densities and marginalization. We shown its application on widely used datasets where the fuzzy measure brought a lot of useful information concerning classifier importance and interactions. Results also emphasized that the proposed fusion process allows, on the one hand, to improve classification results and, on the other hand, to be robust to classifiers mistakes.

Further investigations concern the study of algorithms used for learning distribution parameters which are of key of importance.

References

[1] L.I. Kuncheva, J.C. Bezdek, and R.P.W. Duin. Decision templates for multiple classifier fusion. *Pattern Recognition*, 34:299–314, 2001.

[2] M. Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89:445–456, 1996.

[3] M. Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. *IEEE Int. Conf. on Fuzzy Systems*, 1:145–150, 1995.

[4] M. Grabisch. Fuzzy integral for classification and feature extraction. *Fuzzy Measures and Integrals:Theory and Applications*, pages 415–434, 1998.

[5] M. Grabisch, S.A. Orlovski, and R.R. Yager. Fuzzy aggregation of numerical preferences. In *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, 1998.

[6] I. Kojadinovic. Unsupervised aggregation of commensurate correlated attributes by means of the choquet integral and entropy functionals. *Int. Jour. of Intelligent Systems*, pages 128–154, 2008.

[7] M. Grabisch, I. Kojadinovic, and P. Meyer. A review of methods for capacity identification in choquet integral based multi-attribute utility theory, applications of the R package. *European Jour. of Operational Research*, 189:766–785, 2008.

[8] I. Kojadinovic. Estimation of the weights of interacting criteria from the set of profiles by means of information-theoretic functionals. *European Journal of Operational Research*, 155:741–751, 2004.

[9] S. Jullien, G. Mauris, L. Valet, Ph. Bolon, and S. Teyssier. Identification of choquet integral's parameters based on relative entropy and applied to classification of tomographic images. In *IPMU*, pages 1360–1367, 2008.

[10] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[11] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

[12] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, August 2006.

[13] Hershey and Olsen. Approximating the kullback-leibler divergence between gaussian mixture models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[14] A. Frank and A. Asuncion. UCI machine learning repository, 2010. University of California, Irvine, School of Information and Computer Sciences.

[15] T. Denoeux. An evidence-theoric neural network classifier. *IEEE Trans. on Systems, Man and Cybernetics*, 3:712–717, 2000.

[16] T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 5:804–813, 1995.