r-ERBFN : an Extension of the Evidential RBFN Accounting for the Dependence Between Positive and Negative Evidence

Frédéric Pichon[1], Serigne Diène[1], Thierry Denœux[2,3], Sébastien Ramel[1], and David Mercier[1]

[1] Univ. Artois, EA 3926 LGI2A, Béthune, F-62400, France
firstname.lastname@univ-artois.fr
[2] Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
[3] Institut universitaire de France, Paris, France
thierry.denoeux@utc.fr

Abstract. Recently, it was shown that a radial basis function network (RBFN) with a softmax output layer amounts to pooling by Dempster's rule positive and negative evidence for each class, and approximating the resulting belief function by a probability distribution using the plausibility transform. This so-called latent belief function offers a richer uncertainty quantification than the probabilistic output of the RBFN. In this paper, we show that there exists actually a set of latent belief functions for a RBFN. This set is obtained by considering all possible dependence structures, which are described by correlations, between the positive and negative evidence for each class. Furthermore, we show that performance can be enhanced by optimizing the correlations brought to light.

Keywords: Belief function · Dempster's rule · Dependence · Evidential classification.

1 Introduction

Evidential classifiers, the most well-known being arguably the evidential k-nearest neighbor classifier [4] and its prototype-based improvement [5], are classifiers whose predictive uncertainty about the unknown class $\boldsymbol{\theta} \in \Theta = \{\theta_1, \ldots, \theta_K\}$ of an instance with feature vector $\mathbf{x}$ is represented by a belief function $Bel_{\mathbf{x}}$ [3, 21]. They allow the distinction between aleatory uncertainty and epistemic uncertainty [11], which is akin to the distinction between conflicting evidence and lack of evidence [17]. Such a distinction is important in situations where the final decision can be postponed (e.g. classification with a reject option) or where additional data can be gathered (e.g. active learning) [17]. Moreover, their fine uncertainty quantification can also be exploited to enhance the predictions of a deep neural network architecture, such as a CNN, as first shown in [25].

Of particular interest in this paper is the evidential classifier introduced recently in [16, 9], as an alternative approach to the prototype-based evidential

classifier [5] having similar properties. This classifier was obtained by applying ideas developed in [7], to a radial basis function network (RBFN) with a softmax output layer (or with an output layer containing a single unit with logistic activation function in the case of binary classification). It was used in [16] to enhance the predictions of a UNet model [19] for a task of lymphoma segmentation from 3D PET-CT images.

In essence, this classifier, called hereafter the evidential RBFN (ERBFN), reveals a predictive, so-called latent, belief function $Bel_{\mathbf{x}}$ underlying the probabilistic prediction $P_{\mathbf{x}}$ of a given (trained) RBFN with a softmax output layer. This belief function underlies the probabilistic prediction in the sense that its transformation into a probability distribution using the plausibility transformation [2] is exactly $P_{\mathbf{x}}$. $Bel_{\mathbf{x}}$ is obtained by, first, defining positive and negative pieces of evidence for each class based on the parameters of the RBFN and on $\mathbf{x}$, and, then, pooling them by Dempster's rule.

In the ERBFN, positive and negative evidence for a given class are considered independent. However, they are obtained from the same set of values and therefore the independence assumption may be questioned. As shown in this paper, this assumption is actually inconsequential insofar as any possible dependence structure yields a predictive latent belief function, that is, a predictive belief function whose plausibility transformation is $P_{\mathbf{x}}$. However, this dependence structure, which as will be seen can be characterized following [14] by a correlation, does have an impact on the predictive belief function and therefore does matter.

To select the dependence structure, i.e., correlation, for each class, different approaches can be followed depending on the available information. When the only information available is the given RBFN with its (trained) parameters, then the best attitude is to be cautious, that is, one should select the correlations leading to the most uncertain (least informative) predictive belief function. This is known as following the least commitment (or maximum uncertainty) principle [18, 24], which plays a role in Dempster-Shafer Theory (DST) similar to the principle of maximum entropy in probability theory. We note that such an approach leads to a simple and sound solution (if one uses the informational ordering considered in [7]), which is not reported here due to lack of space.

When, in addition to the trained RBFN, some learning data are available, which is a situation that is likely in practice and is the one considered in this paper, then it becomes possible to search for the correlations that will yield the best performance, according to some uncertainty quantification quality criterion. Classical prediction quality criteria, such as error rate, are not very well adapted as given some labelled data, they can only evaluate the quality of crisp (precise and certain) predictions. We propose to optimize the correlations with respect to the classification equivalent of the evidential uncertainty quantification quality criterion introduced recently by Denœux in regression [10, 12]. Its rationale is that the uncertainty quantification is all the better if high degrees of belief tend to be assigned to the true classes and low degrees of belief are assigned to the complements of the true classes, i.e., high degrees of plausibility are assigned to

the true classes. It generalizes the cross entropy loss in probability theory. As will be seen, such an optimization leads to predictive latent belief functions that tend to have better uncertainty quantification than the one of [16, 9].

This paper is organized as follows. First, necessary background on the Dempster-Shafer theory of belief functions is provided in Section 2. Then, in Section 3, a means to represent the dependence structure between positive and negative evidence for a proposition by a correlation is presented and used to unveil a new result concerning so-called separable belief functions. This result is then exploited in Section 4 to introduce a new evidential classifier, called **r**-ERBFN, which is an extension of the ERBFN. Its additional parameters are a correlation r in the binary classification case and a vector $\mathbf{r}$ of K correlations in the multi-class classification case, allowing to account for the dependence between positive and negative evidence for each class. This classifier allows us to reveal alternative latent belief functions to that of the ERBFN for a given RBFN. A criterion for selecting a particular latent belief function among the available ones is described in Section 5. Experiments on real data are reported in Section 6. Finally, Section 7 concludes the paper. Proofs are omitted due to lack of space.

2 Background on Dempster-Shafer theory

2.1 Evidence representation

In Dempster-Shafer theory [3, 21], a piece of evidence about the true (unknown) answer $\boldsymbol{\theta}$ to some question is represented by a *mass function*, which is a mapping $m : 2^\Theta \to [0, 1]$ such that $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$, with $\Theta = \{\theta_1, \ldots, \theta_K\}$ the set of possible answers to the question. The mass $m(A)$, for some $A \subseteq \Theta$, represents the probability that the evidence supports exactly the proposition $\boldsymbol{\theta} \in A$ (and nothing else more or less specific). Any subset $A \subseteq \Theta$ such that $m(A) > 0$ is called a *focal set* of m. If Θ is a focal set, then m is *non dogmatic*.

The *vacuous* mass function has Θ as only focal set; it corresponds to a totally uninformative piece of evidence. A mass function m whose focal sets are singletons only, is said to be *Bayesian*; it corresponds to the probability distribution $p : \Theta \to [0, 1]$ such that $p(\theta) = m(\{\theta\})$ for all $\theta \in \Theta$.

A mass function that has the form $m(A) = 1 - d$, $m(\Theta) = d$, for some $A \subset \Theta$ such that $A \neq \emptyset$ and some $d \in [0, 1]$, is said to be *simple*. The quantity d is called the degree of *diffidence* in A [13]. The quantity $w := -\ln(d)$ is called the *weight of evidence* [21]. Such a mass function may be conveniently denoted by A^d or, equivalently, by A_w. It represents a piece of evidence that can be interpreted in two ways, with respective probabilities $1 - d$ and d: according to the first interpretation, the evidence tells that $\boldsymbol{\theta} \in A$, and in the second interpretation, the evidence is useless, i.e., it tells $\boldsymbol{\theta} \in \Theta$.

More generally, a mass function "involves a probability model for the evidence bearing on [the] question" [22]. This model is the following (see, e.g., [8, 23]). The piece of evidence can be interpreted in different ways with given probabilities, with Ω the (finite) set of interpretations and P the probability measure on Ω.

If interpretation $\omega \in \Omega$ holds, the evidence tells that $\boldsymbol{\theta} \in \Gamma(\omega)$, with $\Gamma(\omega)$ a nonempty subset of Θ. The tuple $(\Omega, 2^\Omega, P, \Gamma)$ is called a source [6] and is formally a random set. It induces the mass function m such that $m(A) = P(\{\omega \in \Omega : \Gamma(\omega) = A\})$, for all $A \in 2^\Theta \backslash \{\emptyset\}$.

Given a mass function m and any $A \subseteq \Theta$, the probability that the evidence implies $\boldsymbol{\theta} \in A$ is $Bel(A) := \sum_{B \subseteq A} m(B)$ and that it does not contradict $\boldsymbol{\theta} \in A$ is $Pl(A) := \sum_{B \cap A \neq \emptyset} m(B)$. Functions $Bel : 2^\Theta \to [0,1]$ and $Pl : 2^\Theta \to [0,1]$ are called the belief and plausibility functions, respectively, and are in one-to-one correspondence with m. The contour function $\pi : \Theta \to [0,1]$ is the restriction of the plausibility function to singletons, i.e., $\pi(\theta) = Pl(\{\theta\})$, for all $\theta \in \Theta$.

2.2 Evidence combination

Let $(\Omega_1, 2^{\Omega_1}, P_1, \Gamma_1)$ and $(\Omega_2, 2^{\Omega_2}, P_2, \Gamma_2)$, with $\Gamma_i : \Omega_i \to 2^\Theta \backslash \{\emptyset\}$, $i = 1, 2$, be two sources representing two pieces of evidence about $\boldsymbol{\theta}$ and inducing mass functions m_1 and m_2, respectively. Assume these sources to be independent, i.e., the joint probability $P_{12}(\omega_1, \omega_2)$ that the pair of interpretations $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$ holds is equal to $P_1(\omega_1) \cdot P_2(\omega_2)$.

Let us make the subsequent assumption that the sources are reliable and let $\Gamma_\cap(\omega_1, \omega_2) := \Gamma_1(\omega_1) \cap \Gamma_2(\omega_2)$ for all $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$. According to this assumption, if interpretations ω_1 and ω_2 both hold, then we know for sure that $\boldsymbol{\theta} \in \Gamma_\cap(\omega_1, \omega_2)$, and if $\Gamma_\cap(\omega_1, \omega_2) = \emptyset$, we know that ω_1 and ω_2 cannot hold simultaneously, and therefore the probability that a particular event in $\Omega_1 \times \Omega_2$ holds is obtained by conditioning P_{12} on the event $\Theta_\cap = \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : \Gamma_\cap(\omega_1, \omega_2) \neq \emptyset\}$.

Let $P_\cap$ be the probability measure on $\Omega_1 \times \Omega_2$ resulting from the conditioning of P_{12} on the event $\Theta_\cap$. Under the assumptions that the pieces of evidence represented by mass functions m_1 and m_2 are independent and reliable, our knowledge about $\boldsymbol{\theta}$ can then be represented by the mass function denoted $m_1 \oplus m_2$, called the orthogonal sum of m_1 and m_2, and induced by the random set $(\Omega_1 \times \Omega_2, 2^{\Omega_1 \times \Omega_2}, P_\cap, \Gamma_\cap)$. It is easy to show that

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Theta, A \neq \emptyset,$$

and $(m_1 \oplus m_2)(\emptyset) = 0$, with $\kappa := \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ the *degree of conflict* between m_1 and m_2. The orthogonal sum is well defined if $\kappa < 1$.

The binary operation $\oplus$ is called Dempster's rule. It satisfies several properties. It is commutative, associative and has the vacuous mass function as only neutral element. Furthermore, given two simple mass functions A_{w_1} and A_{w_2}, their orthogonal sum is the simple mass function $A_{w_1 + w_2}$.

Another property of Dempster's rule is related to the plausibility transformation method [2], which allows us to approximate a mass function m by a Bayesian mass function p_m obtained by normalizing the contour function π of m:

$$p_m(\{\theta_k\}) := \frac{\pi(\theta_k)}{\sum_{\ell=1}^K \pi(\theta_\ell)}, \quad k = 1, \ldots, K.$$

Given two mass functions m_1 and m_2, we have $p_{m_1 \oplus m_2} = p_{m_1} \oplus p_{m_2}$ [27], i.e., combination and approximation commute. In the remainder of this paper, the approximation of a mass function according to the plausibility transformation method is simply referred to as, for short, its approximation.

Dempster's rule allows us to define the notion of a separable mass function: a mass function m is *separable* if it can be obtained as the combination by Dempster's rule of simple mass functions. Furthermore, if m is non dogmatic, then m can be canonically decomposed as [21]:

$$m = \bigoplus_{\emptyset \neq A \subset \Theta} A^{d(A)}, \tag{1}$$

with $d(\cdot)$ a mapping from $2^\Theta \backslash \{\emptyset, \Theta\}$ to $(0, 1]$ called diffidence function [13].

Finally, let us remark that the orthogonal sum $m_1 \oplus m_2$ of two mass functions m_1 and m_2 relies on the assumption that they are induced by independent sources, which amounts to specifying the joint probability measure P_{12} on $\Omega_1 \times \Omega_2$ to be the product measure $P_1 \otimes P_2$. However, in principle, any dependence structure, and thus any P_{12} having P_1 and P_2 as marginals, can be selected. This is illustrated by Shafer in [22, Example 3], which is a case of non independence between sources inducing simple mass functions. Another example of such a case is provided by Example 1[4].

Example 1. Assume m_1 and m_2 are simple mass functions, induced by sources $(\Omega_i, 2_i^\Omega, P_i, \Gamma_i)$, with $\Omega_i = \{0, 1\}$, $P_i(0) = 0.2$, $\Gamma_i(0) = A_i$ and $\Gamma_i(1) = \Theta$ for some $A_i \subset \Theta$, $i = 1, 2$. Let S_i be the random variable, with state space Ω_i, representing the interpretation for the i-th source. Then, specifying $P_{12}(0, 0) = 0.2$ and $P_{12}(1, 1) = 0.8$, models the dependency $S_2 = S_1$ (we have $P_{12}(S_2 = 0|S_1 = 0) = 1$ and $P_{12}(S_2 = 1|S_1 = 1) = 1$).

In Section 3, we will see that any possible dependence structure between two simple mass functions can be characterized by a correlation.

3 Dependence between positive and negative evidence

Positive and negative items of evidence with respect to a class, as defined in [7], and more generally with respect to a proposition $\boldsymbol{\theta} \in A$, are nothing but simple mass functions with focal set A and focal set $\overline{A}$, respectively. Combining them by Dempster's rule corresponds to assuming that they are independent. In Section 3.2, we extend their combination to any possible dependence structure, which we characterize by a correlation. This is obtained as a particular case of the more general problem of combining two simple mass functions $A_1^{d_1}$ and $A_2^{d_2}$ having some dependence structure, which leads to a generalization of Dempster's rule for combining simple mass functions (Section 3.1). Then, in a second step

[4] Example 1 is based on the probabilistic dependence structure considered in [23, Example 1].

(Section 3.3), we use this rule to unveil a new result concerning the approximation of separable (non dogmatic) mass functions, which is instrumental for our extension of the ERBFN.

3.1 Correlation-based specification of the dependence

Let us assume that we have two (non dogmatic) simple mass functions $m_1 = A_1^{d_1}$ and $m_2 = A_2^{d_2}$, for some $A_i \subset \Theta$ and $d_i \in (0,1]$, $i = 1,2$, induced by two sources $(\Omega_i, 2_i^{\Omega}, P_i, \Gamma_i)$, with $\Omega_i = \{0,1\}$, $P_i(1) = d_i$, $\Gamma_i(0) = A_i$ and $\Gamma_i(1) = \Theta$, $i = 1,2$. Let S_i be the random variable, with state space Ω_i, representing the interpretation for the i-th source. As explained in Section 2.2, specifying the dependence structure between these items of evidence amounts to specifying a joint probability measure P_{12} on $\Omega_1 \times \Omega_2$, with marginals P_1 and P_2.

It is easy to see that, given d_1 and d_2, P_{12} is completely characterized simply by providing $d_{12} := P_{12}(S_1 = 1, S_2 = 1)$. Indeed, we have

$$\begin{aligned}
P_{12}(S_1 = 1, S_2 = 1) &= d_{12},\\
P_{12}(S_1 = 1, S_2 = 0) &= P_1(S_1 = 1) - P_{12}(S_1 = 1, S_2 = 1)\\
&= d_1 - d_{12},\\
P_{12}(S_1 = 0, S_2 = 1) &= P_2(S_2 = 1) - P_{12}(S_1 = 1, S_2 = 1)\\
&= d_2 - d_{12},\\
P_{12}(S_1 = 0, S_2 = 0) &= 1 - (d_{12} + d_1 - d_{12} + d_2 - d_{12})\\
&= 1 - d_1 - d_2 + d_{12}. \qquad (2)
\end{aligned}$$

Thanks to Fréchet [15], we know that $d_{12} \in [\max(0, d_1 + d_2 - 1), \min(d_1, d_2)]$, and thus any dependence structure between the two pieces of evidence can be specified by choosing a number in this latter interval. Moreover, specifying the probability $d_{12} = P_{12}(S_1 = 1, S_2 = 1)$, given $d_1 = P_1(S_1 = 1)$ and $d_2 = P_2(S_2 = 1)$, actually amounts to specifying the dependence between events $S_1 = 1$ and $S_2 = 1$. Following [14], this dependence can be completely characterized and without loss of information by a scalar $r \in [-1,1]$, representing the correlation between the events. A model of correlation between two events of respective probabilities p_1 and p_2 with correlation $r \in [-1,1]$ is provided in [14]: it is based on the Frank family of copulas and it is such that the probability p_{12} of their conjunction is equal for correlation r to $p_{12} = F(p_1, p_2, r)$ with

$$F(p_1, p_2, r) = \begin{cases} \min(p_1, p_2) & \text{if } r = 1,\\ p_1 \cdot p_2 & \text{if } r = 0,\\ \max(0, p_1 + p_2 - 1) & \text{if } r = -1,\\ \log_s[1 + (s^{p_1} - 1)(s^{p_2} - 1)/(s - 1)] & \text{otherwise,} \end{cases}$$

where $s = \tan(\pi(1 - r)/4)$. This family is continuous and strictly increasing in r, i.e. for $r < r'$, we have $F(p_1, p_2, r) \leq F(p_1, p_2, r')$ for all $(p_1, p_2) \in [0,1]^2$ and there exist $(p_1, p_2) \in [0,1]^2$ such that $F(p_1, p_2, r) < F(p_1, p_2, r')$. The cases

$r = -1$, $r = 0$ and $r = 1$ correspond to particular dependence structures: opposite dependence, independence, and perfect dependence, respectively [14]; we can notice that P_{12} in Example 1 is obtained for correlation $r = 1$.

In short, the dependence structure between two sources underlying two simple mass functions is characterized by a correlation $r \in [-1, 1]$. Now, if we assume further that these sources are reliable, and that their dependence is specified by r, our knowledge about $\boldsymbol{\theta}$ can be represented by the mass function denoted $A_1^{d_1} \oplus_r A_2^{d_2}$ and induced by the random set $(\Omega_1 \times \Omega_2, 2^{\Omega_1 \times \Omega_2}, P_\cap^r, \Gamma_\cap)$, where $P_\cap^r$ is the probability measure P_{12} defined by (2) with $d_{12} = F(d_1, d_2, r)$ and conditioned on the event $\Theta_\cap$. The binary operation $\oplus_r$ is a generalization of Dempster's rule for the combination of two simple mass functions ($\oplus$ is recovered for $r = 0$).

3.2 Dependent positive and negative evidence

Consider the special case of Section 3.1, where $A_1 = A$ for some $A \subset \Theta$, $A \neq \emptyset$, and $A_2 = \overline{A_1}$, i.e, mass functions m_1 and m_2 represent positive and negative items of evidence, respectively, with respect to proposition $\boldsymbol{\theta} \in A$. Assuming these items of evidence to be reliable and their dependence to be specified by some correlation $r \in [-1, 1]$, our knowledge about $\boldsymbol{\theta}$ is then represented by mass function $A^{d_1} \oplus_r \overline{A}^{d_2}$.

Proposition 1. *We have*

$$(A^{d_1} \oplus_r \overline{A}^{d_2})(A) = \frac{d_2 - F(d_1, d_2, r)}{d_1 + d_2 - F(d_1, d_2, r)},$$
$$(A^{d_1} \oplus_r \overline{A}^{d_2})(\overline{A}) = \frac{d_1 - F(d_1, d_2, r)}{d_1 + d_2 - F(d_1, d_2, r)},$$
$$(A^{d_1} \oplus_r \overline{A}^{d_2})(\Theta) = \frac{F(d_1, d_2, r)}{d_1 + d_2 - F(d_1, d_2, r)},$$

and $(A^{d_1} \oplus_r \overline{A}^{d_2})(B) = 0$ *for all* $B \in 2^\Theta \backslash \{A, \overline{A}, \Theta\}$.

3.3 Introducing dependence between positive and negative evidence in a separable mass function

Consider the canonical decomposition (1) of a non dogmatic separable mass function m. Let A be some strict non empty subset of Θ. We have $d(A) \leq 1$ and $d(\overline{A}) \leq 1$. In other words, the mass function m involves a (possibly vacuous) positive evidence and a (possibly vacuous) negative evidence for the proposition $\boldsymbol{\theta} \in A$. More generally, it may be remarked that m involves (possibly vacuous) positive and negative evidence for $2^{|\Theta|-1} - 1$ propositions.

It is then clear that there exist $n \leq 2^{|\Theta|-1} - 1$ distinct, strict and non empty subsets $A_1, \ldots, A_n$ of Θ, with $A_i \neq \overline{A_j}$ for all $i \neq j$, such that m can be rewritten as

$$m = \bigoplus_{i=1}^{n} (A_i^{d_i^+} \oplus \overline{A_i}^{d_i^-}) \tag{3}$$

with $d_i^+ = d(A_i)$ and $d_i^- = d(\overline{A_i})$. Although this is inconsequential for our subsequent developments, we remark that expression (3) is obviously not unique. In particular, the list of subsets $A_1, \ldots, A_n$ can, or not, include a subset A_i such that $d(A_i) = 1$ and $d(\overline{A_i}) = 1$, without changing the fact that (3) holds, given that the vacuous mass function is a neutral element for Dempster's rule. Furthermore, if both $d(A) < 1$ and $d(\overline{A}) < 1$ for some subset A, then either A or $\overline{A}$ can arbitrarily be chosen to be one of the subsets A_i.

In any case, Equation (3) brings to light that a non dogmatic separable mass function relies on the combination of independent positive and negative pieces of evidence for n propositions $\boldsymbol{\theta} \in A_i$, $i = 1, \ldots, n$.

Theorem 1. *Let m be the mass function given by (3). Let $\mathbf{r} := (r_1, \ldots, r_n) \in [-1, 1]^n$. Let $m_{\mathbf{r}} := \bigoplus_{i=1}^{n} (A_i^{d_i^+} \oplus_{r_i} \overline{A_i}^{d_i^-})$. We have $p_m = p_{m_{\mathbf{r}}}$ with p_m and $p_{m_{\mathbf{r}}}$ the approximations of m and $m_{\mathbf{r}}$, respectively.*

Theorem 1 shows that whatever the dependence structure, i.e., correlation r_i, chosen between the positive and negative evidence for proposition $\boldsymbol{\theta} \in A_i$, $i = 1, \ldots, n$, the approximation of the resulting mass function does not depend on this choice.

4 The r-ERBFN classifier

In this section, we start by introducing an evidential classifier, which is an extension of the ERBFN classifier [16, 9] accounting for the dependence between positive and negative evidence for each class (Section 4.1). Then, we show that, similarly as the ERBFN reveals a latent mass function for a given RBFN, this classifier also produces a latent mass function for this RBFN (Section 4.2).

4.1 Model

Let $\mathbf{x} \in \mathcal{X}$ be the feature vector of some instance with unknown class $\boldsymbol{\theta} \in \Theta = \{\theta_1, \ldots, \theta_K\}$. Let $\mathbf{p}_j \in \mathcal{X}$, $j = 1, \ldots, J$, be J prototypes. Let $s_j = \exp(-\gamma_j d_j)$ be the degree of similarity between $\mathbf{x}$ and $\mathbf{p}_j$, where $d_j = ||\mathbf{x} - \mathbf{p}_j||$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{p}_j$ and $\gamma_j > 0$ is a parameter.

Case $K = 2$ Let $v_j \in \mathbb{R}$ be a parameter associated to prototype $\mathbf{p}_j$. Let $w_j = s_j v_j$. Let $w_j^+ = \max(0, w_j)$ and $w_j^- = \max(0, -w_j)$ be the positive and negative parts, respectively, of w_j. Let m_j^+ be the simple mass function with focal set $\{\theta_1\}$ and weight of evidence w_j^+, i.e., $m_j^+ = \{\theta_1\}_{w_j^+}$. Let m_j^- be the simple mass function with focal set $\{\theta_2\}$ and weight of evidence w_j^-, i.e., $m_j^- = \{\theta_2\}_{w_j^-}$. In other words, prototype $\mathbf{p}_j$ induces a positive evidence m_j^+ and a negative evidence m_j^- for class θ_1.

Let $m^+ := \bigoplus_{j=1}^{J} m_j^+$ be the overall, i.e., given all prototypes, positive evidence for class θ_1. Similarly, let $m^- := \bigoplus_{j=1}^{J} m_j^-$ be the overall negative evidence

for class θ_1. We have $m^+ = \{\theta_1\}^{d^+}$, with $d^+ = \exp(-w^+)$ where $w^+ = \sum_{j=1}^{J} w_j^+$, and $m^- = \{\theta_2\}^{d^-}$, with $d^- = \exp(-w^-)$ where $w^- = \sum_{j=1}^{J} w_j^-$.

It can be remarked that mass functions m_j^+, $j = 1, \ldots, J$, are completely determined by distinct values w_j, $j = 1, \ldots, J$, i.e., changing the value w_j for some j does not affect mass functions m_k^+, $k \neq j$, therefore it seems reasonable to assume that they are independent between themselves, hence the definition of m^+. The same can be said about mass functions m_j^-, $j = 1, \ldots, J$. On the contrary, we can remark that changing the value w_j for some j will affect in general *both* m^+ and m^-. Hence, when pooling the overall positive and negative evidence for θ_1 in order to obtain our overall evidence – represented by some mass function $m_{\mathbf{x}}$ – with respect to the class of the instance, it seems safer to assume that there is some dependence between them. As we have seen, such a dependence can be characterized by a correlation $r \in [-1, 1]$, leading to the following definition:

Definition 1 (r-ERBFN). *The output of the r-ERBFN classifier is the mass function $m_{\mathbf{x},r}$ defined as*

$$m_{\mathbf{x},r} := m^+ \oplus_r m^-. \tag{4}$$

Remark 1. The 0-ERBFN is nothing but the ERBFN classifier introduced in [16, Section 3.2]. It corresponds to assuming that the overall positive and negative evidence for θ_1 are independent.

Case $K > 2$ Let $v_{jk} \in \mathbb{R}$ be a parameter associated to prototype $\mathbf{p}_j$ and to class θ_k. Let $w_{jk} = s_j v_{jk}$. Let $w_{jk}^+ = \max(0, w_{jk})$ and $w_{jk}^- = \max(0, -w_{jk})$. Let $m_{jk}^+ = \{\theta_k\}_{w_{jk}^+}$ and $m_{jk}^- = \overline{\{\theta_k\}}_{w_{jk}^-}$. In other words, prototype $\mathbf{p}_j$ induces a positive evidence m_{jk}^+ and a negative evidence m_{jk}^- for class θ_k.

Let $m_k^+ := \bigoplus_{j=1}^{J} m_{jk}^+$, respectively $m_k^- := \bigoplus_{j=1}^{J} m_{jk}^-$, be the overall positive, respectively negative, evidence for class θ_k. We have $m_k^+ = \{\theta_k\}^{d_k^+}$, with $d_k^+ = \exp(-w_k^+)$ where $w_k^+ = \sum_{j=1}^{J} w_{jk}^+$, and $m_k^- = \overline{\{\theta_k\}}^{d_k^-}$, with $d_k^- = \exp(-w_k^-)$ where $w_k^- = \sum_{j=1}^{J} w_{jk}^-$.

Using a similar reasoning as that in the case $K = 2$, we can safely assume that: mass functions m_{jk}^+ (resp. m_{jk}^-), $j = 1, \ldots, J$, are independent; mass functions m_k^+ and m_k^- are not independent. The dependence between these latter mass functions can be characterized by a correlation r_k. Our overall evidence for class θ_k is then represented by mass function $m_k := m_k^+ \oplus_{r_k} m_k^-$.

If we make the assumption that the the prototypes $\mathbf{p}_j$ (together with their associated parameters γ_j) have been identified, i.e., are fixed, then we can remark that mass functions m_k, $k = 1, \ldots, K$, are determined by distinct sets of values : m_k is determined by the set $\{v_{jk} : 1 \leq j \leq J\}$ whereas $m_{k'}$, $k' \neq k$, is determined by the set $\{v_{jk'} : 1 \leq j \leq J\}$. Hence, under this assumption, mass functions m_k, $k = 1, \ldots, K$, can be considered independent, leading to the following definition:

Definition 2 (r-ERBFN). *The output of the* $\mathbf{r}$*-ERBFN classifier, with* $\mathbf{r} = (r_1 \ldots, r_K)$*, is the mass function* $m_{\mathbf{x},\mathbf{r}}$ *defined as*

$$m_{\mathbf{x},\mathbf{r}} := \bigoplus_{k=1}^{K} \left(m_k^+ \oplus_{r_k} m_k^- \right). \tag{5}$$

Remark 2. The **0**-ERBFN is nothing but the classifier described in [9]. It corresponds to assuming that the overall positive and negative evidence for each class are independent.

4.2 Latent mass function

Case $K = 2$ We recall that a RBFN with a logistic output unit is a probabilistic classifier for a binary classification problem ($\Theta = \{\theta_1, \theta_2\}$). It is a neural network composed of a hidden layer containing J hidden units, each hidden unit j, $j = 1, \ldots, J$, being characterized by a prototype $\mathbf{p}_j$ and a scale parameter $\gamma_j > 0$. The activation of hidden unit j is $s_j = \exp(-\gamma_j d_j)$, where $d_j = ||\mathbf{x} - \mathbf{p}_j||$ with $\mathbf{x}$ the feature vector of an instance. Furthermore, let v_j be the weight of the connection between hidden unit j and the logistic output unit. Then, the probabilistic prediction $P_{\mathbf{x}}$ of this classifier is

$$P_{\mathbf{x}}(\theta_1) = \frac{1}{1 + \exp(-\sum_{j=1}^{J} s_j v_j)}. \tag{6}$$

Now, consider a r-ERBFN, for some $r \in [-1, 1]$, whose parameters $\mathbf{p}_j$, γ_j and v_j, have been identified to that of a given RBFN with a logistic output unit. Let $m_{\mathbf{x},r}$ denote the output mass function defined by (4) of this r-ERBFN.

Theorem 2. *For all* $r \in [-1, 1]$*, the approximation* $p_{m_{\mathbf{x},r}}$ *of* $m_{\mathbf{x},r}$ *satisfies*

$$p_{m_{\mathbf{x},r}}(\{\theta_1\}) = P_{\mathbf{x}}(\theta_1). \tag{7}$$

Theorem 2 shows that the output $m_{\mathbf{x},r}$ of a r-ERBFN, whose parameters have been identified to that of a given RBFN, is a latent mass function for the probabilistic prediction $P_{\mathbf{x}}$ of this RBFN, for all $r \in [-1, 1] \backslash \{0\}$, in the same way as is the output of the 0-ERBFN.

Case $K > 2$ A RBFN with a softmax output layer is a probabilistic classifier for a multi-class classification problem, whose parameters are prototypes $\mathbf{p}_j$ and scale parameters $\gamma_j > 0$, $j = 1, \ldots, J$, as well as weights v_{jk} connecting hidden unit j and output unit k, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Its probabilistic prediction $P_{\mathbf{x}}$ is

$$P_{\mathbf{x}}(\theta_k) = \frac{\exp(\sum_{j=1}^{J} s_j v_{jk})}{\sum_{\ell=1}^{K} \exp(\sum_{j=1}^{J} s_j v_{j\ell})}. \tag{8}$$

Consider a $\mathbf{r}$-ERBFN, for some $\mathbf{r} := (r_1, \ldots, r_K) \in [-1, 1]^K$, whose parameters $\mathbf{p}_j$, γ_j and v_{jk}, have been identified to that of a given RBFN with a softmax output layer. Let $m_{\mathbf{x},\mathbf{r}}$ denote the output mass function defined by (5) of this $\mathbf{r}$-ERBFN.

Theorem 3. *For all* $\mathbf{r} \in [-1,1]^K$*, the approximation* $p_{m_{\mathbf{x},\mathbf{r}}}$ *of* $m_{\mathbf{x},\mathbf{r}}$ *satisfies*

$$p_{m_{\mathbf{x},\mathbf{r}}}(\{\theta_k\}) = P_{\mathbf{x}}(\theta_k), \quad \forall \theta_k \in \Theta. \tag{9}$$

Proof. It has been established in [9] that Eq. (9) holds for $\mathbf{r} = \mathbf{0}$. The theorem follows from Theorem 1. □

We have thus proved that in the multi-category case also, there exists a set of latent mass functions for the probabilistic prediction of a RBFN, of which the one identified in [9] is a particular member, obtained for $\mathbf{r} = \mathbf{0}$.

5 Identification of the correlations

Assume a RBFN having a softmax output layer[5] and whose parameter values are given. Let us consider the **r**-ERBFN and identify its prototypes, parameters γ_j and v_{jk} to those of this RBFN. To compute the output mass function of the **r**-ERBFN, it remains to identify the correlations **r**. Note that this amounts to selecting a particular latent mass function among the set of latent mass functions brought to light in Section 4.2.

In order to select a given **r**, one can consider its prediction error (or loss). When a prediction is probabilistic, i.e., in the form of a probability distribution $P_{\mathbf{x}}$, its loss is typically evaluated by the negative log-likelihood (or cross-entropy)

$$\mathcal{L}(y, P_{\mathbf{x}}) = -\ln P_{\mathbf{x}}(y), \tag{10}$$

with y the true class of the instance with feature vector $\mathbf{x}$. Minimizing (10) is equivalent to maximizing the probability of the true class.

In the case of the **r**-ERBFN, the prediction is evidential, i.e., in the form of a mass function $m_{\mathbf{x}}$. Following [10, 12], since in this case we no longer have a single probability for the true class but two numbers - a degree of belief $Bel_{\mathbf{x}}(\{y\})$ and a degree of plausibility $Pl_{\mathbf{x}}(\{y\})$ - we can consider the following generalized negative log-likelihood (GNLL)

$$\mathcal{L}(y, m_{\mathbf{x}}) = -\frac{1}{2}\ln Bel_{\mathbf{x}}(\{y\}) - \frac{1}{2}\ln Pl_{\mathbf{x}}(\{y\}). \tag{11}$$

Minimizing (11) amounts to seeking high degrees of belief and of plausibility for the true class. Moreover, we may notice that if $m_{\mathbf{x}}$ is Bayesian, i.e., corresponds to a probability distribution, then we have $Bel_{\mathbf{x}}(\{y\}) = Pl_{\mathbf{x}}(\{y\}) = m_{\mathbf{x}}(\{y\})$, and loss (11) reduces, as may be required, to (10).

Given a loss of the form (11) and some learning data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i$ is the feature vector of instance i and y_i is its true class, we may then fit over this learning set, the correlation vector **r**, i.e., we can search for the vector of correlations $\hat{\mathbf{r}}$ that optimizes the total GNLL over this learning data:

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r} \in [-1,1]^K} \sum_{i=1}^{n} \mathcal{L}(y_i, m_{\mathbf{x}_i,\mathbf{r}}).$$

[5] We focus for short on the case $K > 2$ in this section, but our developments also hold for $K = 2$.

The resulting optimized correlation vector $\hat{\mathbf{r}}$ may then be used to compute the predictive latent mass function for any test feature vector $\mathbf{x}$.

6 Experiments

The previous section has put forward the $\hat{\mathbf{r}}$-ERBFN, as a classifier producing sensible predictive latent mass functions for a RBFN, according to the principle of minimizing the loss. The purpose of this section is to illustrate using some numerical experiments, the interest of this classifier with respect to the original proposal from [16, 9], which corresponds to the $\mathbf{0}$-ERBFN. First, we describe how we trained the RBFN in our experiments (Section 6.1). Then, we provide the remainder of our experimental protocol and the results obtained (Section 6.2).

6.1 Training of the RBFN

In all of our experiments, the parameters $\mathbf{p}_j$, $\gamma_j > 0$ and v_{jk} of the considered $\mathbf{r}$-ERBFN classifiers were identified to that of a RBFN with a softmax output layer (or a logistic output unit in the case of a binary classification problem) learnt over the training dataset following two phase learning as described in [20]. Precisely, in the first phase, for each class, three prototypes were obtained as the centers of the clusters resulting from applying the (constrained[6] [1, Algorithm 2.2]) K-means clustering procedure to the examples of the class. Furthermore, the scale parameter γ_j associated to prototype $\mathbf{p}_j$ was set to $\gamma_j = 1/(2\sigma_j^2)$ where (kernel width) σ_j was the mean of the distances between the prototype $\mathbf{p}_j$ and the training examples in its associated cluster. In the second phase, the connection weights v_{jk} between hidden units and output units were learnt by minimisation of the $L2$ regularized cross entropy loss, using gradient descent (with learning rate and regularization coefficient both set to 10^{-3} and with 10^3 epochs).

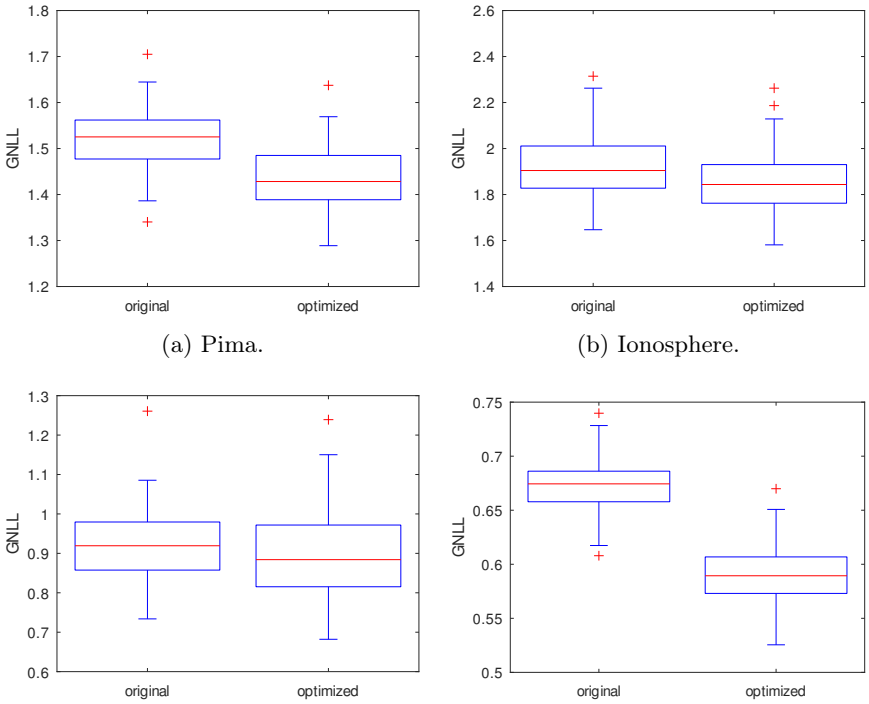
6.2 Experimental settings and results

We used four real datasets[7] considered in related work [11]: Pima (7 features, 2 classes, 532 instances), Ionosphere (33 features, 2 classes, 351 instances), Glass identification (9 features, 6 classes, 214 instances), Vowel identification (10 features, 6 classes, 540 instances). For each dataset, we proceeded similarly as in [12, Section 5]. Specifically, the data were split randomly (using stratified random sampling) into training, validation and test sets containing, respectively, 60%, 20% and 20% of the instances. The training set was used to learn the RBFN as presented in Sect. 6.1, the validation set was used to optimize $\mathbf{r}$ as described in Sect. 5, and the test set was used to evaluate the performance, according to the average GNLL, of $\mathbf{r} = \hat{\mathbf{r}}$ as well as of $\mathbf{r} = \mathbf{0}$. This process was repeated 50 times.

Figure 1 shows boxplots of test GNLL values for the four datasets.

[6] Enforcing at least two training examples of the given class per cluster.

[7] Pima is available from the `R` package `MASS` [26]. Ionosphere, Glass and Vowel are available from the UCI ML repository `https://archive.ics.uci.edu`. For Vowel, we considered only the first six classes, as in [11].



(a) Pima. (b) Ionosphere.

(c) Glass. (d) Vowel.

Fig. 1. Generalized negative log-likelihood for the Pima (1a), Ionosphere (1b), Glass (1c) and Vowel (1d) datasets for $\mathbf{r} = \mathbf{0}$ (original) and $\mathbf{r} = \hat{\mathbf{r}}$ (optimized).

We can see that for all four datasets, the $\hat{\mathbf{r}}$-ERBFN outperforms the $\mathbf{0}$-ERBFN; the differences are highly significant (p-values of paired t-tests for the comparison of GNLL values were at most[8] 2.6×10^{-11} over all datasets.)

7 Conclusion

This paper has brought to light a set of latent belief functions for a RBNF, extending the latent belief function identified in [16, 9] to all possible dependence structures between positive and negative evidence for each class. These latent belief functions allow some performance improvement in terms of uncertainty quantification. A singular – least informative – belief function exists in this set; it will be described in a future publication.

Acknowledgments. Serigne Diène's PhD work is funded by the Hauts-de-France region and Artois University.

References

1. Bradley, P.S., Bennett, K.P., Demiriz, A.: Constrained k-means clustering. Tech. Rep. MSR-TR-2000-65, Microsoft Research, Redmond (2000), www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2000-65.pdf
2. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. Int. J. of Approximate Reasoning **41**(3), 314–330 (Apr 2006)
3. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics **38**, 325–339 (1967)
4. Denœux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics **25**(5), 804–213 (1995)
5. Denœux, T.: A neural network classifier based on Dempster-Shafer theory. IEEE Trans. on Syst., Man, and Cybernetics - Part A **30**(2), 131–150 (2000)
6. Denœux, T.: Quantifying predictive uncertainty using belief functions: Different approaches and practical construction. In: Kreinovich, V., Sriboonchitta, S., Chakpitak, N. (eds.) Predictive Econometrics and Big Data, Studies in Computational Intelligence, vol. 753, pp. 157–176. Springer (2018)
7. Denœux, T.: Logistic regression, neural networks and Dempster-Shafer theory: a new perspective. Knowledge-Based Systems **176**, 54–67 (2019)
8. Denœux, T.: Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence. Fuzzy Sets & Syst. **424**, 63–91 (2021)
9. Denœux, T.: Théorie des fonctions de croyance et apprentissage automatique, (2022), journée Apprentissage automatique multimodal et fusion d'informations (2ème édition), GdR ISIS, virtual, January 19th, 2022
10. Denœux, T.: Quantifying prediction uncertainty in regression using random fuzzy sets: The ENNreg model. IEEE Trans. on Fuzzy Systems **31**(10), 3690 –3699 (2023)
11. Denœux, T.: Uncertainty quantification in logistic regression using random fuzzy sets and belief functions. Int. J. of Approximate Reasoning **168**, 109159 (2024)

[8] P-value obtained for the Glass dataset.

12. Denœux, T.: Combination of dependent and partially reliable Gaussian random fuzzy numbers. Information Sciences (accepted for publication)
13. Dubois, D., Faux, F., Prade, H.: Prejudice in uncertain information merging: Pushing the fusion paradigm of evidence theory further. Int. J. of Approximate Reasoning **121**, 1 – 22 (2020)
14. Ferson, S., Nelsen, R., Hajagos, J., Berleant, D., Zhang, J., Tucker, W.T., Ginzburg, L., Oberkampf, W.L.: Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis. Tech. Rep. SAND2004-3072, Sandia Nat. Lab., Albuquerque, New Mexico (2004)
15. Fréchet, M.: Généralisations du théorème des probabilités totales. Fundamenta Mathematicae **25**, 379–387 (1935)
16. Huang, L., Ruan, S., Decazes, P., Denœux, T.: Lymphoma segmentation from 3D PET-CT images using a deep evidential network. Int. J. of Approximate Reasoning **149**, 39–60 (2022)
17. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn **110**, 457–506 (2021)
18. Klir, G.J.: Uncertainty and Information: Foundations of Generalized Information Theory. Wiley-IEEE Press (2005)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
20. Schwenker, F., Kestler, H.A., Palm, G.: Three learning phases for radial-basis-function networks. Neural Networks **14**(4), 439–458 (2001)
21. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton, N.J. (1976)
22. Shafer, G.: Probability judgment in artificial intelligence. In: Kanal, L.N., Lemmer, J.F. (eds.) Uncertainty in Artificial Intelligence, Machine Intelligence and Pattern Recognition, vol. 4, pp. 127–135. North-Holland (1986)
23. Shenoy, P.: On distinct belief functions in the Dempster-Shafer theory. In: Miranda, E., Montes, I., Quaeghebeur, E., Vantaggi, B. (eds.) Proc. of the Thirteenth Int. Symposium on Imprecise Probability: Theories and Applications. vol. 215, pp. 426–437. PMLR (2023)
24. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. Int. J. of Approximate Reasoning **9**(1), 1–35 (1993)
25. Tong, Z., Xu, P., Denœux, T.: An evidential classifier based on Dempster-Shafer theory and deep learning. Neurocomputing **450**, 275–293 (2021)
26. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer, New York, fourth edition (2002)
27. Voorbraak, F.: A computationally efficient approximation of Dempster-Shafer theory. Int. J. of Man-Machine Studies **30**(5), 525–536 (1989)