# Fusion of one-class classifiers in the belief function framework

**Astride Aregui**
Suez Environnement - CIRSEE and
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex France
Email: astride.aregui@hds.utc.fr

**Thierry Denœux**
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex France
Email: thierry.denoeux@hds.utc.fr

***Abstract*—A method is proposed for converting a novelty measure such as produced by one-class SVMs or Kernel Principal Component Analysis (KPCA) into a belief function on a well-defined frame of discernment. This makes it possible to combine one-class classification or novelty detection methods with other information expressed in the same framework such as expert opinions or multi-class classifiers.**

**Keywords:** Novelty detection, one-class classification, Transferable Belief Model, Demspter-Shafer theory, evidence theory.

## I. Introduction

In pattern classification applications, one generally assumes the existence of a learning set containing examples for the different classes under study. The learning task then consists in building a decision rule whereby new examples can be classified as belonging to one of the classes. This problem has been formalized in different frameworks, including belief function theory, also referred to as Dempster-Shafer theory (see, e.g, [6], [7], [9]). Among the advantages of the belief function framework is the possibility to easily combine classifiers based on different features, different learning sets, and possibly different granularities in the definition of classes [1].

However, there are situations in which, although several classes are known to exist, the learning set is composed of data from only one class. This is the case, for instance, for some diagnosis or system monitoring applications, in which measurements are available only for the nominal state of the system under study. The learning task is then to determine, based on this data, whether a new observation comes from the same distribution as the learning data. This problem is generally referred to as one-class classification, or novelty detection. Many methods have been proposed to tackle this problem (see a survey in [18], [19]), and new methods have recently drawn attention, such as one-class support vector machines (SVMs) [23] and Kernel Principal Component Analysis (KPCA) [12], [16]. However, until recently, this problem had not been studied in the Dempster-Shafer framework. Such a task is undertaken in this paper. Building on previous work reported in [2], we show how to convert the outputs of one-class classifiers such as one-class SVMs or KPCA into belief functions. Expressing one-class and multi-class classifiers in a common framework allows to provide simple solutions to different fusion problems, such as the combination of several one-class classifiers based on different features or different learning algorithms, or the combination of one-class and multi-class classifiers built from different sets of data.

The rest of the paper is organized as follows. The basic concepts of belief function theory and, more specifically, the Transferable Belief Model (TBM) are first recalled in Section II. Recent approaches to one-class classification are briefly reviewed in Section III. Our method is then presented in Section IV, which constitutes the core of the paper. Finally, experimental results are then presented in Section V, and Section VI concludes the paper.

## II. Belief Function Theory

In this section, the main concepts of Dempster-Shafer theory [25] will be briefly reviewed. The intepretation of belief functions used throughout this paper is that of Smets' Transferable Belief Model (TBM) [28].

### A. The Foundations

Let $\Omega$ denote a finite set (termed *frame of discernment*), and $\omega$ a variable taking value in $\Omega$. Given some evidential corpus EC, the knowledge held by a given agent at a given time regarding the actual value of variable $\omega$ can be modeled by a so called *basic belief assignment* (bba) $m^{\Omega}[EC]$ on $\Omega$, defined as a mapping from $2^{\Omega}$ to $[0,1]$ verifying:

$$\sum_{A\subseteq\Omega} m^{\Omega}[EC](A) = 1. \tag{1}$$

The quantity $m^{\Omega}[EC](A)$ represents the part of the agent's belief allocated to the hypothesis that $\omega$ takes some value in $A$ [25], [28]. When there is no ambiguity regarding the domain or evidential corpus, the notation $m^{\Omega}[EC]$ may be simplified to $m^{\Omega}$ or $m$. A subset $A$ of $\Omega$, whose associated mass is non zero is termed a *focal set* of $m$. A mass $m$ is said to be categorical if it has a single focal set.

Equivalent representations of $m$ include:

- the *belief* function, representing the amount of belief in $A$ that is entirely justified by the evidential corpus EC:

$$bel(A) = \sum_{\emptyset\neq B\subseteq A} m(B), \quad \forall A \subseteq \Omega\ ; \tag{2}$$

- the *plausibility* function, corresponding to the amount of belief that is not in contradiction with $A$ given EC:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega. \tag{3}$$

Given two distinct pieces of evidence $m_1$ and $m_2$, given by two different sources, the conjunctive combination $m_{1\textcircled{\cap}2}$ of $m_1$ and $m_2$ can be defined as follows:

$$m_{1\textcircled{\cap}2}(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega. \tag{4}$$

The absence of knowledge is easily represented in the TBM framework by the so called *vacuous belief function* defined by: $m(\Omega) = 1$ or, equivalently, $pl(A) = 1$, $\forall A \subseteq \Omega$, $A \neq \emptyset$.

More generally speaking, there exists several measures of the degree of information of a belief function. Each defines a partial order, and it is possible to sort a series of bba from the least to the most informative, with respect to a given order. The Least Committment Principle (LCP) dictates that, amongst the bba satisfying some constraints (i.e. the set of belief functions compatible with the available information), the least informative -also termed least committed- bba always be chosen. This reflects a cautious attitude. It conveys the idea that no more credit should ever be given to an hypothesis than is strictly accounted for by available evidence. The LCP plays a role similar to the principle of maximum entropy in Bayesian Probability Theory.

In the TBM, decision making is based on the *pignistic transformation* [28], which converts a mass function $m$ into a probability function *BetP* defined as:

$$BetP(\omega_k) = \sum_{A \subseteq \Omega, \omega_k \in A} \frac{m(A)}{(1 - m(\emptyset))\,|A|}, \quad \forall \omega_k \in \Omega. \tag{5}$$

### B. Operations on Product Frames

Let us now assume that we have two frames of discernment $\mathcal{T}$ and $\Omega$. In Section IV, $\mathcal{T}$ will represent the domain of a novelty measure $T$ defined as a function of measurements made on a system, and $\Omega = \{\omega_0, \ldots, \omega_K\}$ will represent a set of system states.

Let $m^{\mathcal{T}\times\Omega}$ denote a bba defined on the Cartesian product $\mathcal{T} \times \Omega$ of the two variables $T$ and $\omega$. The *marginal bba* $m^{\mathcal{T}\times\Omega\downarrow\mathcal{T}}$ on $\mathcal{T}$ is defined, for all $S \subseteq \mathcal{T}$ as:

$$m^{\mathcal{T}\times\Omega\downarrow\mathcal{T}}(S) = \sum_{\{A \subseteq \mathcal{T}\times\Omega \;\mid\; \mathrm{Proj}(A\downarrow\mathcal{T}) = S\}} m^{\mathcal{T}\times\Omega}(A), \tag{6}$$

where $\mathrm{Proj}(A\downarrow\mathcal{T})$ denotes the projection of $A$ onto $\mathcal{T}$:

$$\mathrm{Proj}(A \downarrow \mathcal{T}) = \{t \in \mathcal{T} \,|\, \exists \omega \in \Omega,\ (t, \omega) \in A)\}. \tag{7}$$

The inverse operation is the *vacuous extension*. Let $m^{\Omega}$ be a bba on $\Omega$. Its vacuous extension on $\mathcal{T} \times \Omega$ is defined as:

$$m^{\Omega\uparrow\mathcal{T}\times\Omega}(A) = \begin{cases} m^{\Omega}(B) & \text{if } A = B \times \mathcal{T} \text{ for some } B \subseteq \Omega \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

When a given hypothesis $h \subseteq \Omega$ is ascertained, the beliefs are altered to reflect the new state of knowledge. The conditioning operation consists in combining masses conjunctively with a categorical bba supporting hypothesis $h$. Hence, the mass of belief allocated to $S \subseteq \mathcal{T}$ knowing that hypothesis $h \subseteq \Omega$ holds, i.e. $m^{\Omega}_h(h) = 1$, is:

$$m^{\mathcal{T}}[h] = \left(m^{\mathcal{T}\times\Omega} \textcircled{\cap} m_h^{\Omega\uparrow\mathcal{T}\times\Omega}\right)^{\downarrow\mathcal{T}}. \tag{9}$$

Now let $m^{\mathcal{T}}[h]$ be the bba on $\mathcal{T}$ conditioned with respect to $h \subseteq \Omega$. Assume we now learn that $h$ may finally not hold and all previous states of knowledge have been lost. Masses associated with any non-empty set $S$ of $\mathcal{T}$ are then transferred to $(S \times h) \cup (\mathcal{T} \times (\Omega \setminus h))$. The *ballooning extension* process [26], opposite of the conditioning operation, thus yields:

$$m^{\mathcal{T}}[h]^{\Uparrow(\mathcal{T}\times\Omega)}(A) = \begin{cases} m^{\mathcal{T}}[h](S) & \text{if } A = (S \times h) \cup (\mathcal{T} \times (\Omega \setminus h)), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

### C. The General Bayesian Theorem

The General Bayesian Theorem (GBT) was introduced by Smets [26]. It generalizes Bayes' theorem in that, whenever the belief functions are Bayesian, and we also have a Bayesian prior on the classes, the two theorems are exactly equivalent. However, the power of the GBT lies in the fact that it does not require any prior knowledge on $\Omega$ (for instance, no prior class probabilities).

Let us suppose we know all the conditional bbas $m^{\mathcal{T}}[\omega_k]$, $k = 0, \ldots, K$, we have no prior knowledge on $\Omega$, and we observe $t_* \subseteq \mathcal{T}$. From that, we would like to derive our belief regarding the system state, knowing the value of statistic $T$. In other words, we seek $m^{\Omega}[t_*]$. The GBT allows us to find the answer in three steps.

We shall first calculate the ballooning extension of each of the functions $m^{\mathcal{T}}[\omega_k]$, that is to say, "decondition" them in order to obtain a belief on $\mathcal{T} \times \Omega$. The obtained bbas $m^{\mathcal{T}}[\omega_k]^{\Uparrow\mathcal{T}\times\Omega}$ are distinct, as the original $m^{\mathcal{T}}[\omega_k]$ were distinct. Hence, the $m^{\mathcal{T}}[\omega_k]^{\Uparrow\mathcal{T}\times\Omega}$ can be combined by applying the conjunctive combination rule: this will be the second step. We now have a global and unconditional belief function on $\mathcal{T}\times\Omega$. Conditioning with respect to $t_*$ returns the belief function we need, namely $m^{\Omega}[t_*]$. Thus:

$$m^{\Omega}[t_*] = \left(\textcircled{\cap}_{k=0}^{K} m^{\mathcal{T}}[\omega_k]^{\Uparrow\mathcal{T}\times\Omega}\right)[t_*]. \tag{11}$$

It can be shown that $m^{\Omega}[t_*]$ can be expressed as follows:

$$m^{\Omega}[t_*](A) = \prod_{\omega_k \in A} pl^{\mathcal{T}}[\omega_k](t_*) \prod_{\omega_k \in \overline{A}} (1 - pl^{\mathcal{T}}[\omega_k](t_*)). \tag{12}$$

### D. Belief Functions on $\mathbb{R}$

The above concepts may be extended to the case where the frame of discernment is the set $\mathbb{R}$ of real numbers. In the simplest approach, a bba is defined as above, with the constraint that the set $\mathcal{F}(m) = \{A_1, \ldots, A_n\}$ of focal sets be finite. Typically, focals sets are chosen among intervals or, more generally, Borel sets [11], [22], [31], [32]. Denoting $m_i = m(A_i)$, with $\sum_{i=1}^{n} m_i = 1$, and assuming $A_i \neq \emptyset$ for all $i$, Equations (2)-(3) become:

$$bel(A) = \sum_{\emptyset \neq A_i \subseteq A} m_i, \tag{13}$$

and

$$pl(A) = \sum_{A_i \cap A \neq \emptyset} m_i. \tag{14}$$

A more complex generalization of the finite case is obtained if one replaces the concept of bba by that of basic belief density (bbd) [27], but it will not be needed here.

## III. One-class Classification

We do not intend here to carry out an exhaustive review of the existing one-class classifiers (or novelty detection methods) but rather introduce the principle of the most common ones. The reader is referred to [4], [18], [19] for a survey on novelty detection. We will focus on non-parametric methods, and especially on kernel methods, which show excellent performance. We will in particular describe one-class support vector machines (SVM) and kernel principal component analysis (KPCA).

### A. One-class SVM

One-class SVMs were introduced by Schölkopf [23] and Tax and Duin [29] as a way to estimate the support of a distribution. The underlying idea is that there is no need to estimate the exact density of a population in order to be able to determine whether a new measurement originates from the same distribution or not. The specification of the support of the distribution, i.e. the region of space containing a large fraction of points drawn from that distribution, is sufficient for most applications and much more computationally efficient than full density estimation.

Schölkopf introduced the method as the determination of the hyperplane that separates the training data from the origin with maximal margin. This is done through the definition of a function $f$ that is positive in the support of the distribution and negative elsewhere. Given a learning set $x_1, \ldots, x_n$, it can be shown that an optimal function may be defined as:

$$f(x) = \sum_{i=1}^{n} (\alpha_i k(x_i, x) - b), \tag{15}$$

where $b$ is a scalar parameter called bias, $0 \leq \alpha_i \leq (\nu n)^{-1}$, $\sum_i \alpha_i = 1$ and $k(\cdot, \cdot)$ is a kernel function. Function $f$ can be determined by solving a quadratic programming problem. A pattern $x$ is rejected if $f(x)$ is negative (or smaller than some threshold).

It can be shown that hyperparameter $\nu$ is both an upper bound on the fraction of outliers (i.e. errors) and a lower bound on the fraction of support vectors thus controlling the trade-off between precision and generalisation capacity. Note that, when $\nu = 1$ and the kernel can be normalized as a density in input space, then (15) is exactly equivalent to a Parzen-windows density estimate [24].

*Example 1:* Figure 1 shows a simple two-dimensional data set of $n = 100$ learning vectors, with a contour plot of function $-f(x)$ computed using (15), with a Gaussian kernel $k(x, y) = \exp(\|x-y\|^2/(2\sigma^2))$. Parameter $\nu$ was set to 0.5, and the kernel bandwidth was defined as half the mean Euclidean distance between two training vectors, as suggested in [5]. We can see that the support of the distribution is well approximated by contour lines of $f(x)$. A novelty detection rule may be implemented by rejected patterns for which $-f(x)$ is higher than some threshold.

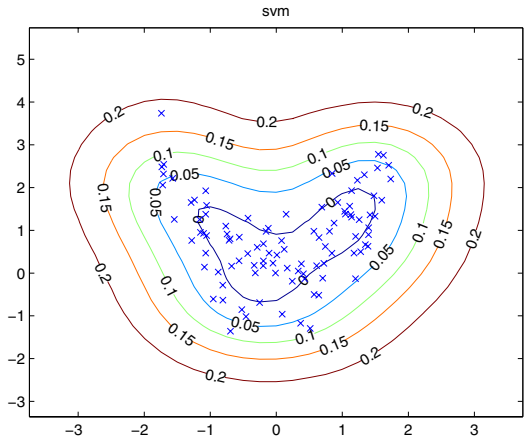


Figure 1. Data set of Example 1, and contour plot of the SVM novelty measure $-f(x)$.

### B. Kernel Principal Component Analysis

Principal Component Analysis (PCA) allows the handling of high dimensional data through linear orthogonal projection on a lower dimensional space, thus providing a more compact representation. The optimal subspace has a significantly lower dimension than the original space but retains most of the variance of the original data. It is obtained by the identification of the dependences between the observations. However, simple PCA relies on the hypothesis of a linear correlation between the data, which is obviously not always the case. The introduction of a kernel function solves the problem: it acts as though the data had undergone a prior transformation from the original space into a feature, Hilbert space, in which they are linearly correlated. The resulting technique is termed Kernel Principal Component Analysis (KPCA).

Moreover, PCA based novelty detection is traditionally performed via the monitoring of different types of error, amongst which the squared prediction error (SPE), Hotelling's statistic (or T2) and the reconstruction error. A pattern is considered novel if any of the monitored errors is above some threshold.

Lee et al. [16] derived formulae for SPE and T2 statistics in the KPCA framework, thus adapting the monitoring technique to highly non-linear data. Recently, Hoffmann [12] also provided a calculation of the reconstruction error adapted to KPCA. The latter proved more efficient than the former two in detecting outliers. We will therefore introduce a method relying on Hoffmann's value of KPCA related reconstruction error. In the sequel, this statistic will be termed kernel reconstruction error or KRE.

Hoffmann demonstrated that:

$$KRE(x) = k(x,x) - \frac{2}{n}\sum_{i=1}^{n} k(x,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j) - \sum_{\ell=1}^{q} f_\ell(x)^2, \quad (16)$$

with

$$f_\ell(x) = \sum_{i=1}^{n} \alpha_{\ell,i}\left[k(x,x_i) - \frac{1}{n}\sum_{r=1}^{n} k(x_i,x_r) - \frac{1}{n}\sum_{r=1}^{n} k(x,x_r) + \frac{1}{n^2}\sum_{r,s=1}^{n} k(x_r,x_s)\right],$$

where $k(\cdot,\cdot)$ is a kernel function such as the Gaussian kernel, $\ell$ is the index of the $\ell^{\text{th}}$ eigenvector of the observation covariance matrix with components $\alpha_{\ell,i}$, $i = 1,\ldots,n$ (with $\ell = 1$ for the eigenvector with the largest eigenvalue), and $q$ is the number of eigenvectors.

The KRE has the property of being small for data drawn from the same distribution as the data that were used to built the KPCA model and greater for data drawn from other distribution. It can thus be used as a novelty measure.

## IV. Formalization in the Belief Function Framework

Given a set of observations $x_1,\ldots,x_n$ drawn from a given distribution, one-class classifiers are used to determine whether an unknown, new point, comes from the same distribution or not. Training the classifier to this task consists in building a novelty measure $T \in \mathcal{T} \subseteq \mathbb{R}$ as a function of $x_1,\ldots,x_n$ using, e.g., (15) or (16), whose value will be small in the region of space containing the data $x_1,\ldots,x_n$, and larger as the distance to this region increases. A new observation is then rejected if the value of $T$ exceeds some threshold.

This method provides a hard decision, but no description of the uncertainty attached to it. Consequently, it is not clear how to combine this information with that provided by other classifiers before a decision is made. Here, we suggest to address this problem in the belief function framework.

Let $\omega_0$ the normal or reference state of the system under study, for which a set $x_1,\ldots,x_n$ of examples is available, and $\omega_1$ the set of all other states, for which no data is available. The frame of discernment is $\Omega = \{\omega_0,\omega_1\}$. Having observed a value $t$ of $T$, we want to define a BBA $m^\Omega[t]$ on $\Omega$, that quantifies our belief about the system state given $t$. Our approach is based on the following three-step procedure:

1) Given the observed sample $t_1,\ldots,t_n$ of $T$ for the training data, build a *predictive belief function* $m^{\mathcal{T}}_*$ that quantifies our belief in future values of $T$ drawn from the same distribution;
2) Build the belief function $m^{\mathcal{T}}[\omega_0]$ that quantifies our belief in $T$ given that the system is in the normal state;
3) Using the GBT, build the belief function $m^\Omega[t]$ that quantifies out belief in the system state, given $T = t$.

These three steps are detailed below.

### A. Step 1: Computing the Predictive Belief Function

Let $t_1,\ldots,t_n$ be the observed values of statistic $T$, and assume that this is a realization of an independent, identically distributed (iid) random sample from a probability distribution with cumulated distribition function (cdf) $F_T$. Based on this information, what can be said regarding a future value of $T$ to be drawn from the same distribution and, in particular, how can this be expressed as a belief function on $\mathcal{T}$?

This problem was solved in [8], for the special case where $\mathcal{T}$ is discrete, using the concept of *predictive belief function*. A belief function $bel^{\mathcal{T}}$ is said to be a predictive belief function for $T$ at level $\alpha$ if it converges towards the true probability distribution of $T$ as the number of observations tends to infinity, and it is less committed (i.e., less informative) than the true distribution with probability $1-\alpha$.

In [8], it is shown how to use multinomial confidence intervals to built a predictive belief function when $\mathcal{T}$ is finite. When $\mathcal{T} = \mathbb{R}$, a similar analysis can be performed using a *confidence band* on $F_T$. Let $\mathbf{T} = (T_1,\ldots,T_n)$ be an iid random sample of $T$, and let $\underline{F}(\cdot;\mathbf{T})$ and $\overline{F}(\cdot;\mathbf{T})$ be two functions computed from $\mathbf{T}$ and such that $\underline{F}(\cdot;\mathbf{T}) \leq \overline{F}(\cdot;\mathbf{T})$. Then the pair $(\underline{F},\overline{F})$ is called a *confidence band at level* $\alpha \in (0,1)$ [17, page 334] iff

$$P\left\{\underline{F}(t;\mathbf{T}) \leq F_T(t) \leq \overline{F}(t;\mathbf{T}),\ \forall t \in \mathbb{R}\right\} = 1-\alpha.$$

A non-parametric confidence band can be built through Kolmogorov's statistic $D_n$. The value of $D_n$ represents the supremum of the difference between the sample cdf $\widehat{F}(\cdot;\mathbf{T})$ and the theoretical cdf $F_T$ at a confidence level $\alpha$:

$$D_n = \sup_t |\widehat{F}(t;\mathbf{T}) - F_T(t)|,$$

where the sample cdf $\widehat{F}(\cdot;\mathbf{T})$ is defined as:

$$\widehat{F}(t;\mathbf{T}) = \begin{cases} 0, & t < T_{(1)} \\ k/n, & T_{(k)} \leq t < T_{(k+1)} \\ 1, & T_{(n)} \leq t, \end{cases} \quad (17)$$

for all $t \in \mathbb{R}$, where $T_{(1)} \leq T_{(i)} \leq \ldots \leq T_{(n)}$ denote the observations arranged in increasing order.

The distribution of $D_n$ does not depend on $F_T$. It was computed for fixed $n$ by Kolmogorov [14], who also computed the asymptotic distribution of $D_n$. Let $d_{n,\alpha}$ denote the critical value of $D_n$ defined as $P(D_n > d_{n,\alpha}) = \alpha$. Thus,

$$P\left\{\widehat{F}(t;\mathbf{T}) - d_{n,\alpha} \leq F_T(t) \leq \widehat{F}(t,\mathbf{T}) + d_{n,\alpha}, \forall t \in \mathbb{R}\right\} = 1-\alpha, \quad (18)$$

which implies that $\widehat{F} \pm d_{n,\alpha}$ defines a confidence bound at level $1-\alpha$ [13, page 481]. This band may be narrowed by using the inequalities $0 \leq F_T(t) \leq 1$. The following bounds thus hold:

$$\underline{F}(t;\mathbf{T}) = \max(0, \widehat{F}(t;\mathbf{T}) - d_{n,\alpha}), \quad (19)$$

$$\overline{F}(t;\mathbf{T}) = \min(1, \widehat{F}(t;\mathbf{T}) + d_{n,\alpha}). \quad (20)$$

For $n \geq 20$ and $\alpha = 0.05$, $d_{n,\alpha}$ may be approximated by $1.36/\sqrt{n}$. IN the sequel, the notations $\underline{F}(t;\mathbf{T})$, $\overline{F}(t;\mathbf{T})$ and $\widehat{F}(t;\mathbf{T})$ will be simplified to $\underline{F}(t)$, $\overline{F}(t)$ and $\widehat{F}(t)$, respectively.

Furthermore, a confidence band can be seen as defining a set $\mathcal{P}$ of probability distributions. As shown by Kriegler and Held [15], the lower envelope of $\mathcal{P}$ is a belief function $bel_*^{\mathcal{T}}$ on $\mathbb{R}$. The focal sets of $bel_*^{\mathcal{T}}$ are intervals whose bounds are defined by sample points. In particular,

$$bel_*^{\mathcal{T}}((-\infty, t]) = \underline{F}(t), \tag{21}$$

for all $t \in \mathbb{R}$. In [3], it is shown that $bel_*^{\mathcal{T}}$ is a predictive belief function for $T$ at level $1 - \alpha$.

*Example 2:* Let us consider again the data of Example 1. Let $T = -f(x)$, where $f(x)$ is defined as the output of the one-class SVM defined by (15). Let $t_1, \ldots, t_{100}$ the values of the $T$ statistics for the $n = 100$ learning examples. Figure 2 shows the sample cdf as well as the Kolmogorov confidence band for this sample, with $\alpha = 0.05$.

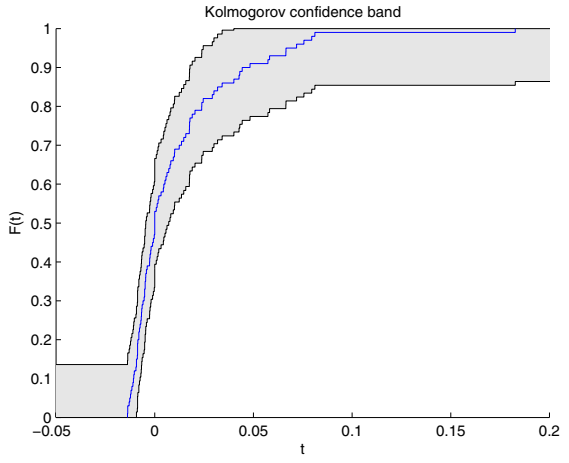


Figure 2. Sample cdf and Kolmogorov confidence band (with $\alpha = 0.05$) for the sample of $n = 100$ output values of the one-class SVM novelty measure obtained from the data of Example 1.

### B. Step 2: Constructing $pl^{\mathcal{T}}[\omega_0]$

The predictive belief function $bel_*^{\mathcal{T}}$ computed in the previous step represents our belief in future values of $T$ drawn from exactly the same distribution as the learning sample, which corresponds to observations of $T$ gathered while the system was in a normal state, under some well-defined experimental conditions $EC$.

However, by construction of statistic $T$, values of $T$ smaller than those encountered in the training data do not indicate departure from the normal state. Consequently, assuming the system to be in the normal state $\omega_0$, $T$ can be expected to take values smaller than (or equal to) those encountered un the training set. THis may be formalized in the TBM as follows:

$$pl_0^{\mathcal{T}}((t, +\infty)) \leq pl_*^{\mathcal{T}}((t, +\infty)), \quad \forall t \in \mathbb{R}. \tag{22}$$

where:

- $bel_*^{\mathcal{T}}$ is the predictive belief function computed in the previous step, in experimental conditions $EC$,
- $bel_0^{\mathcal{T}} = bel^{\mathcal{T}}[\omega_0]$ denotes the belief function on $T$ knowning that the system is in the normal state $\omega_0$, and $pl_0^{\mathcal{T}} = pl^{\mathcal{T}}[\omega_0]$ the corresponding plausibility function.

The above equation means that, for any $t$, the plausibility that $T$ will exceed $t$ when the system is in a normal state may not exceed the plausibility that $T$ will exceed $t$ when the system is in a normal state *and* in experimental conditions $EC$ (i.e. the experimental conditions are the same as those that prevailed when the training set was collected). This property may be referred to as *cognitive inequality*, as it boils down to stochastic inequality when $pl_0^{\mathcal{T}}$ and $pl_*^{\mathcal{T}}$ are probability measures [2].

Equation (22) defines a set of constraints that should be statisfied by $bel_0^{\mathcal{T}}$. In the TBM, the *least commitment principle* dictates to select the *least committed* belief function, among those compatible with a set of constraints [26]. A bba $m_1^{\mathcal{T}}$ may be said to be less committed than a bba $m_2^{\mathcal{T}}$ if $pl_1^{\mathcal{T}}(A) \geq pl_2^{\mathcal{T}}(A)$, for all $A \subseteq \mathcal{T}$. To build the least committed belief function compatible with (22), we first observe that

$$pl_*^{\mathcal{T}}((t, +\infty)) = 1 - bel_*^{\mathcal{T}}((-\infty, t]) = 1 - \underline{F}(t),$$

where $\underline{F}$ is the step function defined by (19). Let $t_{(i_0)}, \ldots t_{(n)}$ be the points of discontinuity of $\underline{F}$. We have the following proposition.

*Proposition 1:* The least committed belief function $bel_0^{\mathcal{T}}$ compatible with constraints (22) is defined by the following bba:

$$\begin{aligned} m_0^{\mathcal{T}}((-\infty, t_{(i_0)})) &= \underline{F}(t_{(i_0)}) \\ m_0^{\mathcal{T}}((-\infty, t_{(i)})) &= \underline{F}(t_{(i)}) - \underline{F}(t_{(i-1)}),\ i = i_0 + 1, \ldots, n \\ m^{\mathcal{T}}((-\infty, +\infty)) &= 1 - \underline{F}(t_{(n)}). \end{aligned}$$

*Sketch of proof:* The solution can be built incrementally by assigning fractions of the unit mass to intervals so as to maximize $pl_*^{\mathcal{T}}((t, +\infty))$ for all $t$ under constraints (22):

- The largest mass that can be assigned to $\mathbb{R}$ is $1 - \underline{F}(t_{(n)})$, which ensures the condition $pl_0^{\mathcal{T}}((t, +\infty)) = 1 - \underline{F}(t)$ be satisfied for all $t \geq t_{(n)}$.
- Assigning a mass $\underline{F}(t_{(n)}) - \underline{F}(t_{(n-1)})$ to $(-\infty, t_{(n)})$, we get, for all $t_{(n-1)} \leq t < t_{(n)}$,

$$\begin{aligned} pl_0^{\mathcal{T}}((t, +\infty)) &= 1 - \underline{F}(t_{(n)}) + \underline{F}(t_{(n)}) - \underline{F}(t_{(n-1)}) \\ &= 1 - \underline{F}(t_{(n-1)}) = 1 - \underline{F}(t). \end{aligned}$$

- The process can be iterated until the whole mass has been assigned. □

We observe that the focal sets are nested: consequently, the plausibility function is a possibility measure. The corresponding plausibility distribution has a very simple expression

$$pl_0^{\mathcal{T}}(t) = 1 - \underline{F}(t), \quad \forall t \in \mathbb{R},$$

and we have $pl^{\mathcal{T}}(A) = \sup_{t \in A} pl_0^{\mathcal{T}}(t)$ for all $A \subseteq \mathbb{R}$.

*Example 3:* Figure 3 shows a plot of $pl_0^{\mathcal{T}}(t) = 1 - \underline{F}(t)$ as a function of $t$ for the data of Examples 1 and 2. Negative values of $T$ are completely plausible under $\omega_0$, whereas the plausibility decreases while $t$ increases.

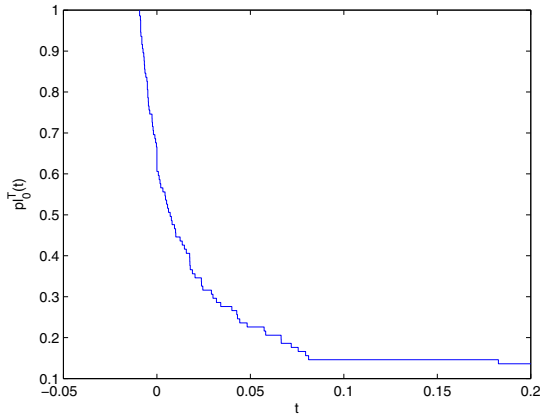


Figure 3. Plot of $pl_0^{\mathcal{T}}(t) = 1 - \underline{F}(t)$ as a function of $t$ for the data of Examples 1 and 2.

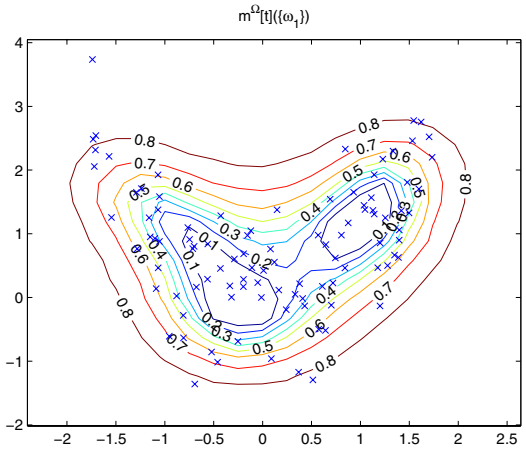


Figure 4. Contour plot of $m^{\Omega}[t](\{\omega_1\})$ for the data of Examples 1-3.

### C. Step 3: Constructing $m^{\Omega}[t]$

The belief function $bel_0^{\mathcal{T}}[\omega_0]$ built in the previous step quantifies our beliefs on $T$, given that the system is in state $\omega_0$. Since no data is available regarding state $\omega_1$, our belief on $T$ given $\omega_1$ is vacuous, i.e., $pl^{\mathcal{T}}[\omega_1](A) = 1$, for all $A \subseteq \mathbb{R}$. The GBT presented in Section II-C allows us to compute our belief on $\Omega$ given that $T \in t_*$ for any $t_* \subseteq \mathcal{T}$. Using (12), we get:

$$m^{\Omega}[t_*](\{\omega_0\}) = 0 \quad (23)$$
$$m^{\Omega}[t_*](\{\omega_1\}) = 1 - pl_0^{\mathcal{T}}(t_*) \quad (24)$$
$$m^{\Omega}[t_*](\Omega) = pl_0^{\mathcal{T}}(t_*). \quad (25)$$

In the special case where $t_* = \{t\}$, we get

$$m^{\Omega}[t_*](\{\omega_0\}) = 0 \quad (26)$$
$$m^{\Omega}[t_*](\{\omega_1\}) = \underline{F}(t) \quad (27)$$
$$m^{\Omega}[t_*](\Omega) = 1 - \underline{F}(t). \quad (28)$$

Note that this result has a simple interpretation: a large value of $T$ supports the hypothesis that the system is not in the normal state. The degree of support increases as a function of $t$. On the contrary, a small value of $T$, similar to those obtained when the system is in normal conditions, may occur either when the system is in a normal state, or when the system is in an abnormal state that does not affect the values of $T$. Therefore, small values of $T$ are highly plausible under both $\omega_0$ and $\omega_1$ and they do not support any specific hypothesis.

*Example 4:* Figure 4 shows a contour plot of $m^{\Omega}[t](\{\omega_1\})$ for the data of Examples 1-3. The mass assigned to $\omega_1$ is small in the region covered by data from class $\omega_0$, and increases with the distance to that class.

## V. Classifier Fusion Example

### A. Problem Description

The data set considered here is the breast-cancer data obtained from the UCI machine-leaning repository [20]. The patterns in this data set belong to two classes: benign and malignant. Each pattern consists of nine cytological characteristics graded with an integer from 1 to 10. As in [12], a uniform noise in $[-0.05, 0.05]$ was added to each value to avoid numerical errors because of the discrete values. After removing patterns with missing characteristic values, the final data set consisted of 683 patterns. This data set was split into a training set of 300 patterns (200 benign, 100 malignant), and a test set of 383 patterns (244 benign, 139 malignant).

In order to illustrate the ability of our method to combine one-class classifier outputs with other information, we considered the following problem. We assumed that the first six characteristics were available for benign data only, whereas the other three characteristics were available for patterns from both classes. We note that this is a common situation: in many applications, more measurements are available for the class that occurs more frequently. The problem is thus to merge:

- the outputs from a one-class classifier trained on observations of the first six characteristics for benign cases only,
- with the the outputs from a two-class classifier trained on observations of the other three characteristics for benign and malignant cases.

### B. Results

A one-class classifier was built using the KPCA method presented in Section III-B. The kernel bandwidth was determined by the direct pluggin method [21] [30, page 71]. The $T$ statistic was KRE computed using (16). A belief function on $\Omega = \{\omega_0, \omega_1\}$, with $\omega_0$ and $\omega_1$ corresponding, respectively, to to the benign and malignant class, was computed using the method introduced in Section IV.

In parallel, a two-class classifier based on characteristics 7, 8 and 9 was trained using the evidential neural network method introduced in [7]. This method is grounded on belief function theory, and produces for each input pattern a belief function on $\Omega$.

For each pattern, the belief functions computed by the one-class and two-class classifiers were combined using the TBM

conjunctive rule (4), and the resulting bba was transformed into a pignistic probability function on $\Omega$ using (5).

Figures 5 and 6 show test estimates of the Receiver Operating Characteristic (ROC) curves for the one-class, two-class and combined classifiers.

ROC curves are a well-known, widely-used mean of representing the performance of a classifier. In a two-class problem (faulty or normal system), they represent the portion $\alpha$ of errors detected while the system is in a normal state (called false positive) against the portion $1-\beta$ of faults detected when the system actually is faulty (termed true positive). $\beta$ represents the percentage of faults that are not detected and should be. The ideal classifiers would minimize both $\alpha$ and $\beta$. However, it can be shown that, whatever the classifier, $\alpha$ always increases when $\beta$ decreases and vice versa. Hence, the best classifier is the one that makes the best compromise, i.e. that minimizes $\beta$ for a given value of $\alpha$ [10].

On our example, it can be observed that, although the one-class classifier has poor performance when considered alone, combining it with the two-class classifier does result in significantly improved performances. Such a combination has been made possible by expressing the outputs from both classifiers in a common framework.

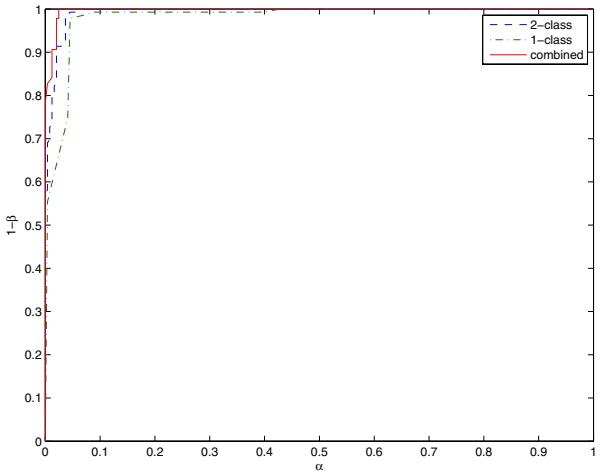


Figure 5. Test estimates of the ROC curves for the three classifiers (two-class: dotted line; one-class: dash-dotted line; combined: continuous line) on the breast cancer problem.

## VI. Conclusion

A new method for converting a novelty measure into a belief function has been presented. Using Kolmogorov's confidence bands and the General Bayesian Theorem, one obtains a belief function with a very simple expression as a function of the sample cdf of the novelty measure values for the training data. Expressing the outputs from one-class classifiers such as one-class SVMs or KPCA in the belief function framework makes it possible to combine them with other information expressed in the same framework, such as other one-class classifiers, evidential multi-class classifiers, or even expert

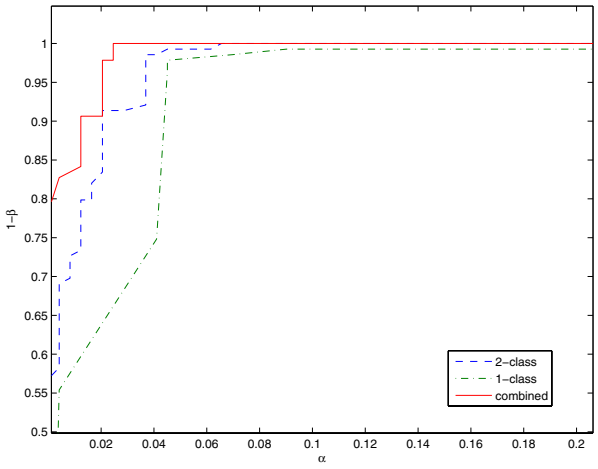


Figure 6. Zoom on the top left hand corner of Figure 5.

opinions. This approach is expected to be particular useful in system diagnosis and process monitoring applications, in which data corresponding to abnormal system states are not always available or are scarce. Results in this application area will be reported in upcoming papers.

## References


[1] A. Appriou. Uncertain data aggregation in classification and tracking processes. In B. Bouchon-Meunier, editor, *Aggregation and Fusion of imperfect information*, pages 231–260. Physica-Verlag, Heidelberg, 1998.

[2] A. Aregui and T. Denœux. Novelty detection in the belief functions framework. In *Proceedings of IPMU '06*, volume 1, pages 412–419, Paris, July 2006.

[3] A. Aregui and T. Denœux. Constructing predictive belief functions from continuous sample data using confidence bands. *Submitted manuscript*, 2007.

[4] C. Campbell. Kernel methods: a survey of current techniques. *Neurocomputing*, 48:63–84, 2002.

[5] M. Davy, F. Desobry, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. *Signal Processing*, 86(8):2009–2025, 2006.

[6] T. Denœux. A *k*-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[7] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[8] T. Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.

[9] T. Denœux and P. Smets. Classification using belief functions: the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics B*, 36(6):1395–1406, 2006.

[10] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, Palo Alto, USA, 2004.

[11] S. Ferson, V. Kreinovitch, L. Ginzburg, D. S. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Technical Report SAND2002-4015, Sandia National Laboratories, Albuquerque, NM, 2003.

[12] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40:863–874, March 2007.

[13] M. Kendall and A. Stuart. *The advanced theory of statistics*, volume 2. Charles Griffin and Co Ltd, London, fourth edition, 1979.

[14] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Istituto Italiano degli Attuari*, 4:83–91, 1933.

[15] E. Kriegler and H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39(2–3):185–209, 2005.
[16] J.-M. Lee, I.-B. Lee, et al. Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59:223–234, 2004.
[17] E. L. Lehman. *Testing statistical hypotheses*. Springer-Verlag, New-York, 2nd edition, 1986.
[18] M. Markou and S. Singh. Novelty detection: a review, part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.
[19] M. Markou and S. Singh. Novelty detection: a review, part 2: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.
[20] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.
[21] B. Park and J. Marron. Comparison of data-driven bandwidth selectors. *J. of Amer. Stat. Assoc.*, 85:66–72, 1990.
[22] S. Petit-Renaud and T. Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004.
[23] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.
[24] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.
[25] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
[26] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.
[27] P. Smets. Belief functions on real numbers. *International Journal of Approximate Reasoning*, 40(3):181–223, 2005.
[28] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
[29] D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.
[30] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.
[31] R. R. Yager. Arithmetic and other operations on Dempster-Shafer structures. *Int. J. Man-Machines Studies*, 25:357–366, 1986.
[32] R. R. Yager. Cumulative distribution functions from Dempster-Shafer belief structures. *IEEE Trans. on Systems, Man and Cybernetics B*, 34(5):2080–2087, 2004.