Soft Computing manuscript No.
(will be inserted by the editor)

Partially supervised Independent Factor Analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis

Zohra L. Cherfi · Latifa Oukhellou · Etienne Côme · Thierry Denœux · Patrice Aknin

Received: date / Accepted: date

Abstract Using a statistical model in a diagnosis task generally requires a large amount of labeled data. When ground truth information is not available, too expensive or difficult to collect, one has to rely on expert knowledge. In this paper, it is proposed to use partial information from domain experts expressed as belief functions. Expert opinions are combined in this framework and used with measurement data to estimate the parameters of a statistical model using a variant of the EM algorithm. The particular application investigated here concerns the diagnosis of railway track circuits. A noiseless Independent Factor Analysis (IFA) model is postulated, assuming the observed variables extracted from railway track inspection signals to be generated by a linear mixture of independent latent variables linked to the system component states. Usually, learning with this statistical model is performed in an unsupervised way using unlabeled examples only. In this paper, it is proposed to handle this learning process in a soft-supervised way using imperfect information on the system component states. Fusing partially reliable information about cluster membership is shown to significantly improve classification results.

Keywords Belief function theory · Dempster-Shafer theory · Evidence theory · Partially supervised learning · Independent Factor Analysis · Fault diagnosis · Soft labels · EM algorithm

Z. L. Cherfi
GRETTIA, French Institute of Science and Technology for Transport, Development and Networks, Université Paris-Est, Marne-la-Vallée, France
Tel.: +33-1-45-92-56-46
Fax: +33-1-45-92-55-01
E-mail: zohra.cherfi@ifsttar.fr

L. Oukhellou
LISSI, Université Paris-Est Créteil, Créteil, France

E. Côme
GRETTIA, French Institute of Science and Technology for Transport, Development and Networks, Université Paris-Est, Marne-la-Vallée, France

T. Denœux
HEUDIASYC, Université de Technologie de Compiègne, UMR CNRS 6599, Compiègne, France

P. Aknin
GRETTIA, French Institute of Science and Technology for Transport, Development and Networks, Université Paris-Est, Marne-la-Vallée, France

1 Introduction

In the last few years, the diagnosis of complex systems has received growing attention within the *predictive maintenance* framework, also referred to as condition-based maintenance. The idea of predictive maintenance is that continuous monitoring of the system should allow the user to detect malfunctions and to schedule the appropriate maintenance operation accordingly. Therefore, in a predictive context, frequent inspections of systems are deployed in order to collect inspection data. An automatic diagnosis process is then needed for detecting and identifying defect occurrences from the inspection measurements. When a pattern recognition approach is adopted to solve such problems, it involves using machine learning techniques to assign the measured signals to one of several predefined classes of defects [10, 24]. Maintainers can thus be provided with an accurate and systematic analysis of recordings allowing them to plan preventive maintenance appropriately.

Machine learning methods are generally considered within two main paradigms: supervised learning and unsupervised learning [27, 30]. To be effective, these methods require an exhaustive database with data representative of all system states. In most real world applications, a large amount of data is available but their labeling is generally a time-consuming and expensive task [11, 34]. However, in many industrial fields, it can be taken advantage of *expert knowledge* to label the data. In this case, the class labels can be subject to imprecision and epistemic uncertainty. The statistical learning community has considered this problem and proposed several learning schemes, some of them mixing the supervised and unsupervised learning paradigms.

The first situation is that of *semi-supervised learning*, where the learning set is built from a combination of labeled and unlabeled samples [8,12,37]. Adding unlabeled data to a supervised learning set can then be a way to improve the performance of the algorithms with low additional cost. Another situation is that of *partially supervised learning* [2, 3, 19, 28, 31], in which examples are labeled by *sets of classes*. A learning example for which all classes are possible is unlabeled, while it is perfectly labeled if only one class is specified: this framework thus encompasses semi-supervised learning as a special case. Other learning frameworks have been proposed in order to take into account the imperfection of class labels. For instance, models incorporating label noise have been proposed in [4, 35, 36]. In this case, class labels are considered to be pervaded by random errors: they are thus precise but uncertain.

Using imprecise and uncertain class labels can be interesting when they are supplied by one or several experts and when crisp assignments are hard to obtain [20, 26, 32, 52]. Jointly, the labeling by several experts raises the issue of their quality and conformity. A solution to deal with this kind of labels has been proposed in [15, 22]. In this framework, class labels are expressed by Dempster-Shafer belief functions [18, 47]. Two kinds of uncertainty are then considered: *aleatory* uncertainty due to the variability of the variable of interest in the population and *epistemic* uncertainty due to a lack of knowledge on the state of the variable. The proposed model considers these two kinds of uncertainty separately: aleatory uncertainty is represented by a parametric statistical model while epistemic uncertainty is expressed by belief functions representing expert opinions.

This paper presents a fault diagnosis application using partially labeled data to learn a statistical model based on Independent Factor Analysis (IFA) [5, 13]. This generative model assumes the observed variables extracted from the inspection signal to be generated from a linear mixture of independent latent variables linked to the states of the system components. Learning of this statistical model is usually performed in an unsupervised way: the model parameters and latent variables are then learned exclusively from the observed data [1, 5, 40].

The idea investigated in this paper is to incorporate additional information on the class membership of some samples to estimate the parameters of the IFA model,

using an extension of the EM algorithm [15, 22]. Real data related to fault diagnosis in a railway system have been considered. A labeling campaign was organized with the aim of having these signals labeled by different experts, who were allowed to express doubt on their assessments, resulting in "soft labels". The learning was then performed based on the signals labeled by combining these different opinions via the theory of belief functions [18,47]. It is shown that the integration of soft-labeled signals can significantly enhance the information in the data and improve the quality of the diagnosis.

This article is organized as follows. Background material on belief functions is first recalled in Section 2. The IFA model and its fitting using data with soft labels are then addressed in Section 3. Section 4 describes the application under study and introduces the diagnosis problem in greater detail. Data processing and experimental results are finally reported in Section 5, and Section 6 concludes the paper.

2 Background on belief functions

This section provides a brief account of the fundamental notions of the Dempster-Shafer theory of belief functions, also referred to as Evidence Theory. This uncertain reasoning framework was initiated by Dempster [18] and developed by Shafer [47]. It can be seen as an extension of Bayesian Probability Theory. A particular interpretation of Dempster-Shafer theory has been proposed by Smets [51], under the name of the Transferable Belief Model (TBM). The theory of belief functions has proved to be particularly useful to represent and reason with partial information in a wide range of applications, including system diagnosis [43, 54].

2.1 Belief representation

Let $\Omega = \{\omega_1, ..., \omega_n\}$ be a finite *frame of discernment*, defined as a set of exclusive and exhaustive hypotheses about some question Q of interest. Partial information about the answer to question Q can be represented by a *mass function* $m : 2^\Omega \to [0, 1]$ such that:

$$\sum_{A \subseteq \Omega} m(A) = 1, \tag{1}$$

The quantity $m(A)$ represents a measure of the belief that is assigned to subset $A \subseteq \Omega$ given the available evidence and that cannot be committed to any strict subset of A. Every $A \subseteq \Omega$ such that $m(A) > 0$ is called a *focal set* of m. A mass function m is said to be:

- *normalized* if $\emptyset$ is not a focal set (this condition is not imposed in the TBM under the open-world assumption);
- *dogmatic* if Ω is not a focal set;
- *vacuous* if Ω is the only focal set (it then represents total ignorance);
- *simple* if it has at most two focal sets and, if it has two, Ω is one of those;
- *categorical* if it is both simple and dogmatic.

A simple mass function such that $m(A) = 1 - w$ for some $A \neq \Omega$ and $m(\Omega) = w$ can be noted A^w. Thus, the vacuous mass function can be noted A^1 for any $A \subset \Omega$, and a categorical mass function can be noted A^0 for some $A \neq \Omega$.

The information contained in a mass function m can be equivalently represented in several different ways, such as the *belief* and *plausibility* functions defined, respectively, as follows:

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

and

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

The quantity $Bel(A)$ is interpreted as a total degree of justified support assigned to A, while $Pl(A)$ is an upper bound on the degree of support that could be assigned to A if more specific information became available. The function $pl : \Omega \to [0, 1]$ defined by $pl(\omega) = Pl(\{\omega\})$ for all $\omega \in \Omega$ is referred to as the *contour function*.

2.2 Information combination

Conjunctive combination

Smets introduced the conjunctive rule of combination to combine several mass functions defined on the same frame of discernment [48]. For this rule to be used, the different mass functions must be based upon independent pieces of evidence. Let m_1 and m_2 be two mass functions obtained from two different reliable sources. The mass function that results from their conjunctive combination, denoted by $m_1 \textcircled{\cap} m_2$, is defined as:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \tag{4}$$

for all $A \subseteq \Omega$. This rule is commutative and associative, and it admits the vacuous mass function as neutral element. It has the effect of focusing masses of belief to those hypotheses that are jointly supported by both sources. The mass $(m_1 \textcircled{\cap} m_2)(\emptyset)$ assigned to the empty set may be interpreted as a *degree of conflict* between the two sources.

Several extensions of the conjunctive rule have been proposed for handling the conflict among partially inconsistent sources of evidence [23, 55]. For instance, Yager's rule of combination [55] assumes that, in case of conflict, the result is not reliable but the solution must be in the frame of discernment Ω: the mass $(m_1 \textcircled{\cap} m_2)(\emptyset)$ is thus redistributed to Ω, resulting in the normalized mass function $(m_1 \textcircled{\cap} m_2)^*$ defined as

$$(m_1 \textcircled{\cap} m_2)^*(A) = (m_1 \textcircled{\cap} m_2)(A), \quad \forall A \subset \Omega, A \neq \emptyset, \tag{5}$$

$$(m_1 \textcircled{\cap} m_2)^*(\Omega) = (m_1 \textcircled{\cap} m_2)(\Omega) + (m_1 \textcircled{\cap} m_2)(\emptyset), \tag{6}$$

$$(m_1 \textcircled{\cap} m_2)^*(\emptyset) = 0. \tag{7}$$

Disjunctive combination

When it can only be assumed that *at least one* source is reliable, without knowing which one, but sources are still considered as independent, the disjunctive rule is appropriate [49]. It is defined as:

$$(m_1 \textcircled{\cup} m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C) \tag{8}$$

for all $A \subseteq \Omega$. This rule is commutative and associative.

Cautious conjunctive combination

In the conjunctive and disjunctive rules of combination, the data sources are assumed to be independent. The cautious rule of combination $\textcircled{\wedge}$ was introduced in [21] to combine mass functions provided by non independent sources. This rule avoids double-counting information shared by two sources (see also [46]).

Although the cautious rule can be applied to any non dogmatic mass function, it will be recalled here only in the case of *separable* mass function, i.e., mass functions that can be decomposed as the conjunctive combination of simple mass functions [47, 50]. Let m_1 and m_2 be two such mass functions. They can be written as:

$$m_1 = \textcircled{\cap}_{A\subset\Omega} A^{w_1(A)}$$

and

$$m_2 = \textcircled{\cap}_{A\subset\Omega} A^{w_2(A)},$$

where $A^{w_1(A)}$ and $A^{w_2(A)}$ are simple mass functions, $w_1(A) \in (0,1]$ and $w_2(A) \in (0,1]$ for all $A \subset \Omega$. Their combination using the cautious rule is defined as:

$$(m_1 \textcircled{\wedge} m_2)(A) = \textcircled{\cap}_{A\subset\Omega} A^{w_1(A)\wedge w_2(A)}, \tag{9}$$

where $\wedge$ denotes the minimum operator. This rule is commutative, associative and idempotent, i.e., it verifies $m \textcircled{\wedge} m = m$ for all m.

2.3 Cognitive independence

Let Ω and Θ be two finite frames of discernment and let $m^{\Omega\times\Theta}$ be a mass function on the product frame $\Omega \times \Theta$. The *marginal* mass function on Ω is defined as

$$m^{\Omega\times\Theta\downarrow\Omega}(A) = \sum_{C\downarrow\Omega=A} m^{\Omega\times\Theta}(C),$$

for all $A \subseteq \Omega$, where $C \downarrow \Omega$ denotes the projection of $C \subseteq \Omega \times \Theta$ on Ω.

Let m^{Ω} and m^{Θ} denote, respectively, the marginal mass functions on Ω and Θ and let Pl^{Ω} and Pl^{Θ} denote the corresponding plausibility functions. The frames Ω and Θ are said to be *cognitively independent* [47, page 149] with respect to $m^{\Omega\times\Theta}$ if the following equalities hold:

$$Pl^{\Omega\times\Theta}(A \times B) = Pl^{\Omega}(A)Pl^{\Theta}(B), \tag{10}$$

for all $A \subseteq \Omega$ and $B \subseteq \Theta$. As shown by Shafer [47], this property means that new evidence on one variable does not affect our beliefs in the other variable.

3 Statistical model and learning method

3.1 Independent Factor Analysis

IFA was introduced in [5]. It originates from both standard factor analysis (FA) in applied statistics [6] and independent component analysis (ICA) in signal processing [7, 17]. IFA is based on a generative model that makes it possible to recover independent latent components from their observed linear mixtures. In its noiseless formulation (used throughout this paper), the IFA model can be expressed as:

$$\mathbf{y} = H\,\mathbf{z}, \tag{11}$$

where H is a nonsingular square matrix of size S, $\mathbf{y}$ is the observed random vector whose elements are the S mixtures and $\mathbf{z}$ the random vector whose elements are the S latent components. Thanks to the noiseless setting, a deterministic relationship between the distributions of observed and latent variables can be expressed as:

$$f^{\mathcal{Y}}(\mathbf{y}) = \frac{1}{|\det(H)|} f^{\mathcal{Z}}(H^{-1}\mathbf{y}), \tag{12}$$

where $f^{\mathcal{Y}}$ and $f^{\mathcal{Z}}$ denote, respectively, the probability density functions (pdf's) of $\mathbf{y}$ and $\mathbf{z}$. In the IFA model [5, 40], each source density is a mixture of Gaussians (MOG), so that a wide class of densities can be approximated, and latent component are assumed to be independent. The pdf of $\mathbf{z}$ is thus given by:

$$f^{\mathcal{Z}}(\mathbf{z}) = \prod_{j=1}^{S} \sum_{k=1}^{K^j} \pi_k^j \varphi(z^j; \mu_k^j, \nu_k^j), \tag{13}$$

where z^j denotes the j-th component of vector $\mathbf{z}$, $\varphi(.; \mu, \nu)$ denotes the Gaussian pdf with mean μ and variance ν; π_k^j, μ_k^j and ν_k^j are the proportion, mean and variance of component k for source j, and K^j is the number of components for source j. In the classical unsupervised setting used in IFA, the problem is to estimate both the mixing matrix H and the MOG parameters from the observed variables $\mathbf{y}$ alone. Considering an iid random sample $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of size N, the log-likelihood has the following form:

$$\log L(\boldsymbol{\Psi}; \mathbf{Y}) = -N \log(|\det(H)|) + \sum_{i=1}^{N} \sum_{j=1}^{S} \log \left(\sum_{k=1}^{K^j} \pi_k^j \varphi((H^{-1}\mathbf{y}_i)^j; \mu_k^j, \nu_k^j) \right), \tag{14}$$

where $\boldsymbol{\Psi}$ is the IFA parameter vector $\boldsymbol{\Psi} = (H, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S)$, $\boldsymbol{\pi}^j$ the vector of cluster proportions of source j summing to one, $\boldsymbol{\mu}^j$ and $\boldsymbol{\nu}^j$ the vectors of size K^j containing the means and the variances of each cluster.

Maximum likelihood estimation of the model parameters can be achieved by an alternating optimization strategy. The gradient algorithm [1] is indeed well suited to optimize the log-likelihood function with respect to the mixing matrix H when the parameters of the source marginal densities are frozen. Conversely, with H kept fixed, an EM algorithm can be used to optimize the likelihood function with respect to the parameters of each source. These remarks have led to the development of a Generalized EM algorithm (GEM) able to simultaneously maximize the likelihood function with respect to all the model parameters [38].

3.2 Soft-supervised learning in IFA

The IFA model is often considered within an unsupervised learning framework. This section considers the learning of this model in a soft-supervised learning context where partial knowledge of the cluster membership of some samples is available in the form of belief functions. In the general case, we will consider a learning set of the form:

$$\mathbf{M} = \{(\mathbf{y}_1, m_1^1, \ldots, m_1^S), \ldots, (\mathbf{y}_N, m_N^1, \ldots, m_N^S)\}, \tag{15}$$

where $m_i^1, \ldots, m_i^S$ is a set of mass functions encoding uncertain knowledge on the cluster membership of sample i for each one of the S sources. Each mass function m_i^j is defined on the frame of discernment $\mathcal{U}^j = \{c_1, \ldots, c_{K^j}\}$ composed of all possible clusters for source j. The unsupervised case is recovered as a special case where all mass functions are vacuous, while the supervised case would correspond to the situation where each mass function is focused on a singleton.

We can remark that, in the model considered here, two kinds of uncertainty are present: *aleatory* uncertainty induced by the random data generation process of each realization $\mathbf{y}_i$, and *epistemic* uncertainty induced by the imperfect perception of cluster membership. This kind of estimation problem was initially addressed in the specific case of mixture models in [53] and [33]. It was formalized in [15] and it received a general formulation in [22], where a generalization of the likelihood function was introduced together with an extension of the EM algorithm for its maximization.

Let us denote by $\mathbf{x}_i = (\mathbf{y}_i, u_i^1, \ldots, u_i^S)$ the completed data where $\mathbf{y}_i \in \mathbb{R}^S$ are the observed variables and $u_i^j \in \mathcal{U}^j,\ \forall j \in \{1, \ldots, S\}$ are the cluster membership variables which are ill-known. As in classical IFA, stochastic independence between the $\mathbf{x}_i$ will be assumed:

$$f(\mathbf{X}; \boldsymbol{\Psi}) = \prod_{i=1}^{N} f(\mathbf{x}_i; \boldsymbol{\Psi}), \tag{16}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is the complete sample vector and $f(\mathbf{x}_i; \boldsymbol{\Psi})$ is the pdf of a complete observation according to the IFA model:

$$f(\mathbf{x}_i; \boldsymbol{\Psi}) = \frac{1}{|\det(H)|} \prod_{j=1}^{S} \prod_{k=1}^{K^j} \left(\pi_k^j \varphi((H^{-1}\mathbf{y})^j; \mu_k^j, \nu_k^j) \right)^{\mathbb{1}\{u_i^j = c_k\}}. \tag{17}$$

Additionally, the following cognitive independence assumption (10) will be made:

$$pl(\mathbf{X}) = \prod_{i=1}^{N} pl_i(\mathbf{x}_i) = \prod_{i=1}^{N} \prod_{j=1}^{S} pl_i^j(u_i^j), \tag{18}$$

where $pl(\mathbf{X})$ is the plausibility that the complete sample vector is equal to $\mathbf{X}$, $pl_i(\mathbf{x}_i)$ is the plausibility that the complete data for instance i is $\mathbf{x}_i$ and $pl_i^j(u_i^j)$ is the plausibility that source j for example i was generated from component u_i^j.

It should be stressed that assumptions (16) and (18) are unrelated: the former is a property of the random data generation process, while the latter pertains to the uncertain observation process. Under these two assumptions and following [22], the observed data log likelihood can be written as:

$$\log L(\boldsymbol{\Psi}; \mathbf{M}) = \sum_{i=1}^{N} \log \mathbb{E}_{\boldsymbol{\Psi}}[pl_i(\mathbf{x}_i)] = \sum_{i=1}^{N} \log \int_{\mathcal{X}} f(\mathbf{x}_i; \boldsymbol{\Psi}) pl_i(\mathbf{x}_i) d\mathbf{x} \tag{19}$$

$$= -N \log(|\det(H)|) + \sum_{i=1}^{N} \sum_{j=1}^{S} \log \left(\sum_{k=1}^{K^j} pl_{ik}^j \pi_k^j \varphi((H^{-1}\mathbf{y}_i)^j; \mu_k^j, \nu_k^j) \right) \tag{20}$$

where $pl_{ik}^j = pl_i^j(c_k)$ is the plausibility (computed from soft label m_i^j) that sample i belongs to cluster k of latent variable j.

This criterion must be maximized with respect to $\boldsymbol{\Psi}$ to compute parameter estimates. The EM algorithm can be extended to perform this task. In this extended setting it is referred to as E^{2}M for Evidential EM [22]. The next section presents this extension applied to the IFA model.

3.3 Evidential EM algorithm for soft-supervised IFA

As the classical EM algorithm, the E^{2}M algorithm uses the complete data log-likelihood, which has the following expression in the case of the IFA model:

$$\log L(\boldsymbol{\Psi}; \mathbf{X}) = -N \log(|\det(H)|) + \sum_{i=1}^{N} \sum_{j=1}^{S} \sum_{k=1}^{K^j} \mathbb{1}\{u_i^j = c_k\} \log \left(\pi_k^j \varphi((H^{-1}\mathbf{y}_i)^j; \mu_k^j, \nu_k^j) \right). \tag{21}$$

Let $f(\mathbf{X}|\mathbf{M}; \boldsymbol{\Psi}^{(q)})$ denote the conditional pdf obtained by combining $\mathbf{M}$ with the complete data density function $f(\mathbf{X}; \boldsymbol{\Psi})$ using Dempster's rule [22]. The conditional

expectation of $\log L(\boldsymbol{\Psi}; \mathbf{X})$ with respect to $f(\mathbf{X}|\mathbf{M}; \boldsymbol{\Psi}^{(q)})$ defines the auxiliary function $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$ that will be maximized during the M-step of the algorithm:

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \mathbb{E}_{\boldsymbol{\Psi}^{(q)}} \left[\log L(\boldsymbol{\Psi}; \mathbf{X})|\mathbf{M}\right] \tag{22}$$

$$= -N \log(|\det(H)|) + \sum_{i=1}^{N} \sum_{j=1}^{S} \sum_{k=1}^{K^j} t_{ik}^{j(q)} \log \left(\pi_k^j \varphi((H^{-1}\mathbf{y}_i)^j; \mu_k^j, \nu_k^j) \right) \tag{23}$$

where $t_{ik}^{j(q)}$ is the posterior probability that sample i belongs to cluster k for latent variable j given the crisp observations $\mathbf{y}_i$, the imprecise label m_i^j and the current estimate of the parameter vector $\boldsymbol{\Psi}^{(q)}$, and q is the iteration counter. At the E-step, the posterior probabilities are computed as follows:

$$t_{ik}^{j(q)} = \frac{pl_{ik}^j \pi_k^{j\,(q)} \varphi(z_i^{j(q)}; \mu_k^{j\,(q)}, \nu_k^{j\,(q)})}{\sum_{k'=1}^{K^j} pl_{ik'}^j \pi_{k'}^{j\,(q)} \varphi(z_i^{j(q)}; \mu_{k'}^{j\,(q)}, \nu_{k'}^{j\,(q)})}, \tag{24}$$

with $z_i^{j(q)} = ((H^{(q)})^{-1}\mathbf{y}_i)^j$. These quantities are the only terms that need to be computed during the E-step of the algorithm and they differ from the usual posteriors solely by the presence of the pl_{ik}^j terms. During the M-step, the maximization of $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$ leads to analytical solutions similar to the classical updated formulas for the proportions, means and variances of the clusters:

$$\pi_k^{j(q+1)} = \frac{1}{N} \sum_{i=1}^{N} t_{ik}^{j(q)}, \tag{25}$$

$$\mu_k^{j\,(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{j(q)} z_i^{j(q)}}{\sum_{i=1}^{N} t_{ik}^{j(q)}}, \tag{26}$$

$$\nu_k^{j\,(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{j}(q) \left(z_i^{j(q)} - \mu_k^{j\,(q+1)} \right)^2}{\sum_{i=1}^{N} t_{ik}^{j(q)}}. \tag{27}$$

As in standard IFA, the maximization of Q with respect to the mixing matrix must be performed using gradient ascent:

$$H^{(q+1)} = H^{(q)} + \tau \Delta H^{(q)}, \tag{28}$$

where τ is the learning rate that can be adjusted by a linear search method [41], and $\Delta H^{(q)}$ is the IFA learning rule for the mixing matrix H [5]:

$$\Delta H^{(q)} = ([H^{(q)}]^{-1})^t \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i^{(q)}\right) (\mathbf{z}_i^{(q)})^t - \mathbf{I} \right), \tag{29}$$

where $\mathbf{g}\left(\mathbf{z}_i^{(q)}\right)$ is the S-dimensional vector computed using the posterior probabilities and the model parameters:

$$\mathbf{g}\left(z_i^{j(q)}\right) = \sum_{k=1}^{K^j} t_{ik}^{j(q)} \frac{z_i^{j(q)} - \mu_k^{j(q)}}{\nu_k^{j(q)}}. \tag{30}$$

Finally, to account for scale indeterminacy in the IFA model, rows of H together with the latent variables should be scaled by the mixture parameters to constrain the latent variables to have unit variance. This transformation leaves the likelihood unchanged and must be performed after each M-step [5].

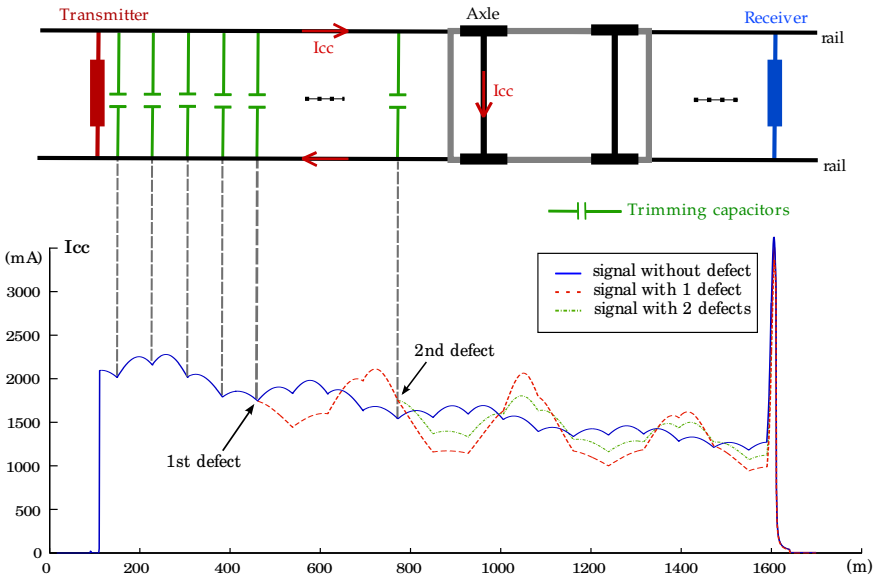

Fig. 1 Track circuit representation and examples of inspection signals (Icc)

4 Application

The application considered in this paper concerns fault diagnosis of railway track circuits. The addressed problem will first be introduced in Section 4.1 and the model construction will be described in Section 4.2.

4.1 Problem description

The track circuit is an essential component of the automatic train control system [43]. Its main function is to detect the presence or absence of vehicle traffic on a given section of the railway track. For this purpose, the railway track is divided into different sections (Fig. 1); each one of them is equipped with a specific track circuit consisting of:

- A *transmitter* connected to one of the two section ends, which supplies a frequency modulated alternating current;
- The two rails that can be considered as a transmission line;
- A *receiver* at the other end of the track section, which essentially consists of a trap circuit used to avoid the transmission of information to the neighboring section;
- *Trimming capacitors* connected between the two rails at constant spacing to compensate the inductive behavior of the track.

A train is detected when the wheels and axles short-circuit the track. It induces the loss of the track circuit signal and the drop of the received signal below a threshold indicates that the section is occupied.

On French high-speed lines, the track circuit is also a fundamental component of the track-to-vehicle transmission system. It uses a specific carrier frequency to transmit coded data to the train regarding, for example, the maximum authorized speed on a given section on the basis of safety constraints.

The different parts of this system can be subject to malfunctions due to aging, atmospheric conditions or track maintenance operations. Faults must be detected as soon as possible to maintain the system at the required safety and availability levels.

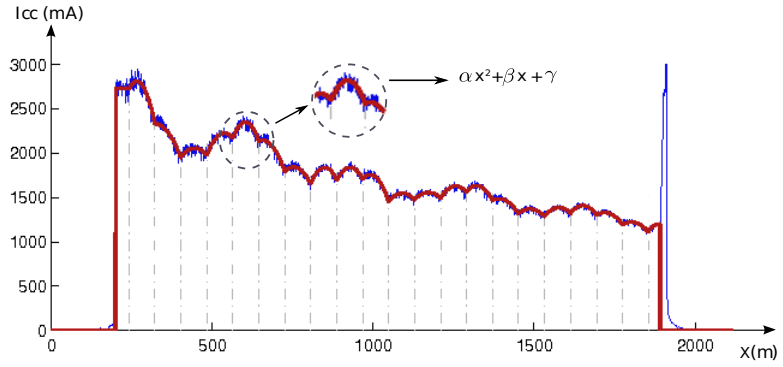

Fig. 2 Inspection signal parametrization

In the most extreme cases, an abnormal attenuation of the transmitted signal may induce important signaling problems (a section can be considered as occupied even if it is not). A diagnosis system is needed to avoid such undesirable situation and inform maintainers about failures on the basis of inspection signals analysis [42, 43].

Information about the condition of track circuits is obtained using an inspection vehicle that delivers a measurement signal (denoted as Icc) linked to electrical characteristics of the system. Fig. 1 shows examples of inspection signals simulated along a 1500 m track circuit: one of them corresponds to a fault-free system, while the others correspond to a signal with one and two defective capacitors, respectively. The problem considered here concerns the diagnosis of track circuit from real inspection signals, focusing on trimming capacitor faults.

4.2 Implementation of the IFA model

The proposed method is based on the following two observations:

- The presence of a defect in a capacitor only affects the signal *downstream* (i.e., between the capacitor and the receiver), leaving the other part of the signal unchanged (Fig. 1);
- The inspection signal has a specific structure, which is a succession of local arches that can be approximated by quadratic polynomials $\alpha x^2 + \beta x + \gamma$ (Fig. 2).

As the inspection signal at one location along the track circuit is affected by the unobservable states of all capacitors located between that location and the transmitter, it can be seen as a result of a mixing process and described using the IFA model presented in Section 3.1. The corresponding generative model is shown in Fig. 3.

The variables y_i^j are features extracted from each inspection signal i by approximating each arch j by a quadratic polynomial:

$$y_i^j = (\hat{\beta}_i^j, \hat{\gamma}_i^j). \tag{31}$$

We note that only two coefficients are needed because the third coefficient is linearly related to the three coefficients of the previous polynomial. The size of each observation vector $\mathbf{y}_i$ is thus $2S$, where S is the number of capacitors in the track circuit.

As the IFA model requires using as many continuous latent variables as observed ones, we need to define $2S$ latent variables. They are defined in our model as:

$$z_i^j = (c_i^j, \varepsilon_i^j), \tag{32}$$

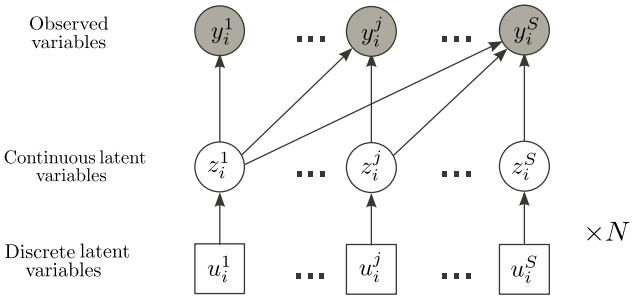

Fig. 3 Generative model for the diagnosis of track circuits represented by a graphical model

where c_i^j is the capacitance of capacitor j for circuit i and ε_i^j is a noise variable. Each variable c_i^j is assumed to have a MOG distribution with three components corresponding to three states of the capacitor: fault-free, medium defect and major defect. The state of capacitor j for circuit i is encoded by a discrete latent variable u_i^j. The noise variables ε_i^j are assumed to be normally distributed.

In the IFA model, the vectors $\mathbf{y}_i$ and $\mathbf{z}_i$ of observed and latent variables are assumed to be linked by relation (11). As there is no influence of a trimming capacitor state on parts of the inspection signal located upstream along the track circuit (Fig. 1), some terms of matrix H are constrained to be equal to zero in our implementation of the model [14, 16].

5 Results and discussion

This section presents experiments carried out to validate the two main ideas explored throughout this paper, i.e., the integration of soft labels for estimating the parameters of the IFA model and the fusion of expert opinions in the belief function framework.

The diagnosis system was assessed using real signals provided by the French National Railway Company (SNCF) and obtained during inspections carried out on a 333 km high-speed line during a two year period, at a frequency of one inspection every two weeks. Although a large amount of data was collected, no ground truth information about the state of capacitors could be obtained because of the high cost of collecting such information. Moreover, given the current scheduled maintenance policy, most provided signals were fault-free. Therefore, only a small proportion of the available signals was used in the experiments. Signals presenting defects and considered as more relevant were selected primarily. However, this selection did not prevent the representation of fault-free cases in the dataset because there is usually no more than one or two defective capacitors in each track circuit.

Overall, 422 real signals were presented to four experts for labeling. Imprecise and uncertain labels were elicited and represented as simple belief functions. The labels from individual experts were then combined using each of the rules described in Section 2.2 and the IFA model was fit using the combined labels thanks to the algorithm described in Section 3.3. Each of these steps is described in greater detail below.

5.1 Soft label elicitation

A database composed of 422 real signals was used to elicit class labels from experts. Three classes were considered, corresponding to the three operating modes of the capacitors, namely: fault free, medium defect and major defect. All the inspection

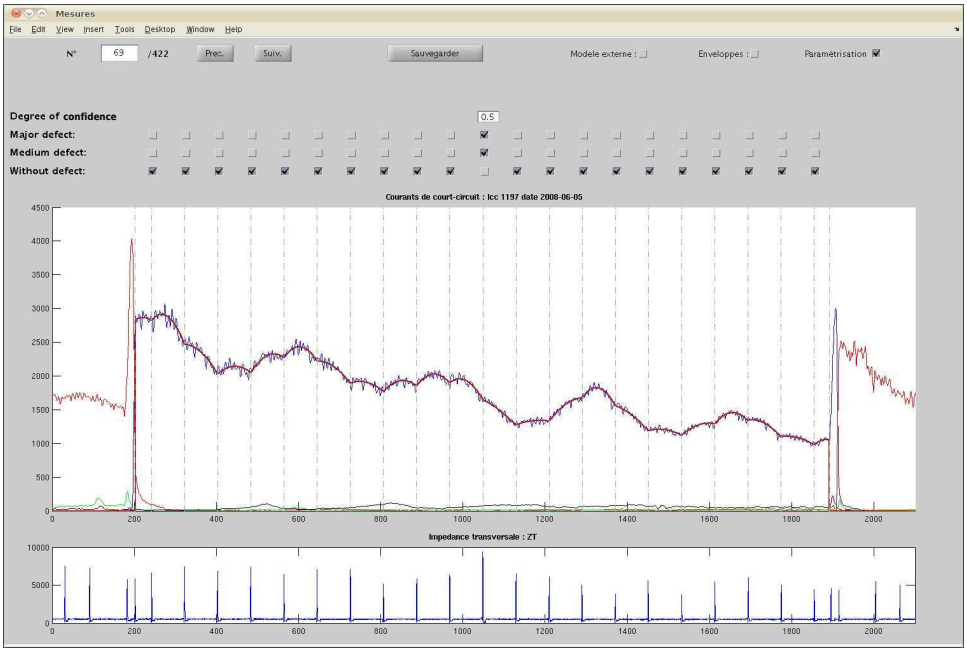

Fig. 4 Interface for labeling inspection signals

signals included into the database were shown separately to four experts thanks to a dedicated software application (Fig. 4). Two kinds of inspection signals were presented to the experts: the Icc signal considered as the observation in the statistical model and another signal that is not used in the diagnosis task but can be helpful during the labeling operation (see the lower part of Fig. 4). Using this graphical interface experts were asked to:

1. Visualize the inspection signals one by one;
2. Indicate, for each capacitor, a set of possible classes;
3. Specify a *degree of confidence* in their decision.

5.2 Combination in the belief function framework

Expert opinions elicited as described above were expressed in the frame

$$\Omega = \{\omega_0, \omega_1, \omega_2\}, \tag{33}$$

where ω_0, ω_1 and ω_2 stand for "fault free", "medium defect" and "major defect", respectively. The opinion of each expert about each capacitor was represented by a simple mass function on Ω, based on the set of possible classes and the confidence degree given by the expert (the mass assigned to Ω was one minus the confidence degree).

In the example shown in Fig. 4, one of the experts expressed an imprecise opinion about the 11-th capacitor's operating state by selecting the two defective classes of major and medium defects. In addition, he gave a confidence of 0.5 on this labeling. This information is represented by the assignment of a mass equal to 0.5 to $\{\omega_1, \omega_2\}$ and the rest 0.5 to Ω. The first column of Table 1 shows the corresponding mass function.

With the previous labeling process, each signal labeled by the four experts was associated to four mass functions for each capacitor of the corresponding track circuit. The mass functions obtained on each single capacitor were then combined by

Table 1 Mass functions representing soft labels elicited from the four experts and combined mass functions using the conjunctive, disjunctive and cautious rules

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_{\textcircled{\cap}}(A)$	$m_{\textcircled{\cup}}(A)$	$m_{\textcircled{\wedge}}(A)$
$\emptyset$	0	0	0	0	0	0	0
$\{\omega_0\}$	0	0	0	0	0	0	0
$\{\omega_1\}$	0	0.8	0.9	0	0.98	0	0.9
$\{\omega_2\}$	0	0	0	0	0	0	0
$\{\omega_0, \omega_1\}$	0	0	0	0	0	0	0
$\{\omega_0, \omega_2\}$	0	0	0	0	0	0	0
$\{\omega_1, \omega_2\}$	0.5	0	0	0.9	0.019	0.3	0.09
Ω	0.5	0.2	0.1	0.1	0.001	0.7	0.01

Table 2 Contour functions resulting from their combination of mass functions in Table 1 by the conjunctive, disjunctive and cautious rules

ω	$pl_{\textcircled{\cap}}(\omega)$	$pl_{\textcircled{\cup}}(\omega)$	$pl_{\textcircled{\wedge}}(\omega)$
$\{\omega_0\}$	0.001	0.7	0.01
$\{\omega_1\}$	1	1	1
$\{\omega_2\}$	0.02	1	0.1

the conjunctive, disjunctive and cautious conjunctive rules defined, respectively, by (4), (8) and (9). These three rules have been chosen because they are well justified theoretically and they have clear interpretations (see, e.g., [45] for a discussion on various justifications for these rules). In particular, the cautious rule has been shown to yield good results when combining dependent items of evidence [29, 46], which makes it a good candidate for combining the opinions of multiple experts sharing common background knowledge. Without prior information about the dependence relations between expert opinions, there seems to be no way to select one of those rules before performing the experiments. In the case of conjunctive rules, conflicting opinions were handled using Yager's normalization (5)-(7), after the combination of all mass functions.

The contour functions associated to the combined mass functions were then used to estimate the parameters of the IFA model. Examples of combined mass functions and corresponding contour functions are shown in Tables 1 and 2, respectively.

5.3 Performance evaluation

To study the impact of the fusion method, the database of 422 signals was associated to seven different sets of labels obtained from the four individual expert assessments and the three combination rules.

Because of the lack of the ground truth about the real state of capacitors, we chose to take as reference another labeling of the same database obtained thanks to a third-party expertise. This reference labeling was established as a consensus among three experts working in favorable conditions. For each signal in the database, historical data collected during past inspections over two years was provided as a support. Thereby, the labeling of each signal was carried out with a sufficient hindsight that allowed the experts to observe the onset and progression of defects in the capacitors. In this way, the estimated labels were sufficiently precise and reliable to assume that they reflect the true capacitors state. Furthermore, it took two days to get these reference labels, whereas the experts had only a few hours to perform the labeling task. This reference labeling will be referred to as REF in the following.

As shown in Table 3, the number of major or medium defects in the considered database according to the REF labeling is much smaller than the number of fault

Table 3 Distribution of capacitors according to the REF labeling

	fault free	medium defect	major defect
number of capacitors	8102	126	106

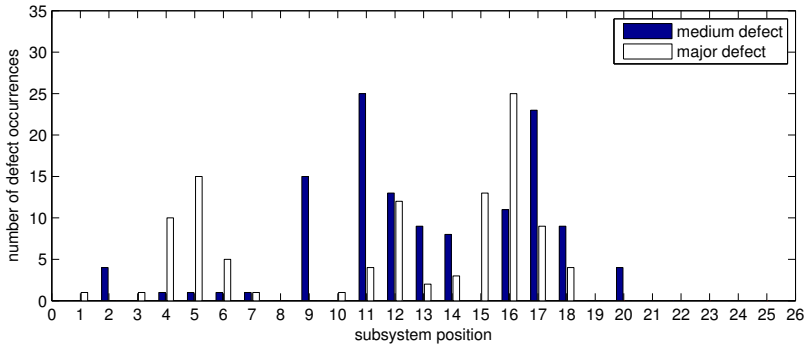

Fig. 5 Number of medium (blue) and major (white) defect occurrences with respect to the capacitor positions for track circuits containing 25 capacitors and according to the REF labeling.

free capacitors, resulting in underrepresentation of the two defect classes. This is even more apparent when considering the number of defects according to the capacitor position. Figure 5 shows the distribution of defects (according to the REF labeling) with respect to the capacitor positions for track circuits containing 25 capacitors. It can be seen that some positions do not have any occurrence of a defective capacitor. This representation highlights the lack of fault occurrences for some capacitor positions and the need for more defect cases to learn the model parameters.

To overcome this problem, the real data were complemented by simulated data generated using an electrical model [42], which allows the simulation of track circuits behavior by modifying some contextual parameters. Five hundred noisy signals with known class labels were generated, corresponding to track circuits of 25 capacitors with different values of the capacitance of each capacitor.

Note that the simulated data were only used to learn the model parameters; they were never used for the evaluation of the diagnosis performances. A cross-validation approach was adopted to assess the performance of each labeling. The database was split randomly into ten subsets; we used nine of them increased by the 500 simulated signals as the training set to learn the IFA model parameters, and the remaining subset was used as the test set to estimate the performances obtained with these parameter estimates. These two steps were repeated 10 times, each time leaving out a different subset for testing. The results were averaged over the 10 test sets.

5.4 Results

The results were analyzed according to the prediction of capacitor states. Confusion matrices between the classes defined by the REF labeling and the estimated classes for all capacitors in the database for each of the four experts and the different combination schemes are reported in Tables 4 and 5, respectively. Decisions d_0, d_1 and d_2 correspond to the estimated class. These decisions were determined for each capacitor according to the maximum posterior probabilities computed on every test sets using the parameters estimated from each labeled database.

Note that a unique performance measure could be computed from each confusion matrix by defining misclassification costs. For instance, the cost of classifying a major

Table 4 Confusion matrices for decisions based on labels elicited from each experts

	ω_0	ω_1	ω_2
d_0	98.8	33.1	2.1
d_1	0.9	51.1	6.9
d_2	0.2	15.8	90.9

(Expert 1)

	ω_0	ω_1	ω_2
d_0	98.9	34.7	3.0
d_1	0.8	58.8	12.2
d_2	0.3	6.5	84.7

(Expert 2)

	ω_0	ω_1	ω_2
d_0	98.7	22.1	2.1
d_1	1.1	63.6	13.8
d_2	0.2	14.3	84.1

(Expert 3)

	ω_0	ω_1	ω_2
d_0	98.8	34.6	3.3
d_1	1.0	49.6	5.8
d_2	0.2	15.8	90.9

(Expert 4)

Table 5 Confusion matrices for decisions based on the four fusion schemes

	ω_0	ω_1	ω_2
d_0	98.9	30.7	2.9
d_1	0.9	58.0	7.7
d_2	0.2	11.3	89.4

(Majority rule)

	ω_0	ω_1	ω_2
d_0	98.9	20.2	2.9
d_1	1.0	64.2	6.5
d_2	0.1	15.6	90.6

(Conjunctive combination)

	ω_0	ω_1	ω_2
d_0	98.9	23.1	2.9
d_1	0.8	62.8	8.0
d_2	0.1	14.1	89.2

(Disjunctive combination)

	ω_0	ω_1	ω_2
d_0	98.9	20.4	2.6
d_1	1.0	65.3	4.9
d_2	0.1	14.2	92.4

(Cautious combination)

defect as medium is obviously less than the cost of classifying it as fault-free. Even though quantitative cost assessments are difficult to obtain in this application, weaker information such as interval-valued or linguistic cost assessments could be used, as suggested in [44]. This is left for further study.

The results reveal good classification performances despite some misclassification between contiguous classes (i.e., between ω_0 and ω_1 and ω_1 and ω_2). The confusion matrices corresponding to individual experts provide some information on expert skills (Table 4). Indeed, experts 1 and 4 better detected major defects, while experts 2 and 3 were more accurate for the detection of medium defects. The combination of expert opinions makes it possible to improve the detection of both types of defects (Table 5). The best results were achieved by the cautious rule, which suggests that the expert opinions cannot be regarded as independent.

The confusion between contiguous classes (specially ω_1 and ω_2) can be explained by two factors. First, considering the overall number of capacitors represented in the database, the number of major and medium defects remains too small as compared to fault-free cases for expecting a reliable learning of these two classes. Secondly, the identification of medium defects is a particularly difficult exercise due to the continuous nature of the real states. In critical cases they can be confused with the two contiguous classes (ω_0 and ω_2), which further reduces their detection rate. To confirm this analysis, we computed the degree of conflict resulting from the conjunctive combination of the mass functions obtained from the expert labels and those obtained from the REF labeling. As shown by Fig. 6, the degree of conflict is globally very low (< 0.03), which is consistent with the high number of fault-free capacitors in the

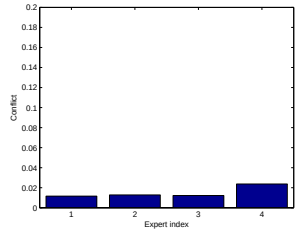

Fig. 6 Global mean degree of conflict between each expert and the REF labeling for the whole database

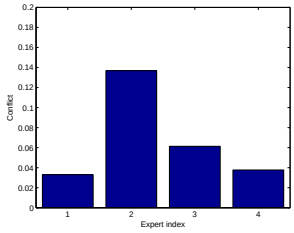

Fig. 7 Mean degree of conflict between each expert and the REF labeling for major defect cases

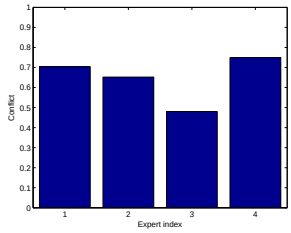

Fig. 8 Mean degree of conflict between each expert and the REF labeling for medium defect cases

Table 6 Confusion matrices for decisions based on simulated data only, with N generated examples

	ω_0	ω_1	ω_2
d_0	90.5	35.1	4.6
d_1	8.4	52.4	17.4
d_2	1.1	12.5	78.0

$N = 500$

	ω_0	ω_1	ω_2
d_0	91.0	28.4	5.0
d_1	8.4	58.1	14.2
d_2	0.6	13.5	80.8

$N = 900$

	ω_0	ω_1	ω_2
d_0	90.8	22.7	3.1
d_1	8.4	59.0	15.6
d_2	0.8	18.4	81.3

$N = 1500$

database (Table 3). Indeed, conflict is higher in the case of major defects (< 0.14), and even more so in the case of medium defects (conflict between 0.45 and 0.8), as shown in Figs. 7 and 8, respectively.

As simulated data is used together with real data in our study, we may wonder whether simulated data alone could be sufficient to achieve good classification performances. Table 6, which displays confusion matrices for classifiers trained using simulated data alone (with $N \in \{500, 900, 1500\}$ samples), shows that this is not the case, even for large N. This demonstrates that the use of real signals with expert labeling is essential to achieve good performances in this diagnosis task.

Finally, a detection matrix can be computed by merging the two defect classes ($\omega_1 \cup \omega_2$). The following performance indicators can then be computed:

- The accuracy (AC), defined the proportion of correct predictions:

$$AC = \frac{\#(d_0, \omega_0) + \#(d_1 \cup d_2, \omega_1 \cup \omega_2)}{N}; \tag{34}$$

- The true positive rate (TP), defined as the proportion of defective capacitors that were correctly identified:

$$TP = \frac{\#(d_1 \cup d_2, \omega_1 \cup \omega_2)}{\#(d_0, \omega_1 \cup \omega_2) + \#(d_1 \cup d_2, \omega_1 \cup \omega_2)}; \tag{35}$$

- The false negative rate (FN) defined as the proportion of defective capacitors that were incorrectly classified as fault-free:

$$FN = \frac{\#(d_0, \omega_1 \cup \omega_2)}{\#(d_0, \omega_1 \cup \omega_2) + \#(d_1 \cup d_2, \omega_1 \cup \omega_2)}. \tag{36}$$

The results reported in Table 7 show that an accuracy of at least 97% is reached by the different sets of labels. However, the different combination schemes outperform each of the individual experts, particularly in terms of true positives and false negatives. The conjunctive and cautious rules yield better results than the disjunctive rule.

Table 7 Accuracy(AC), true positive rate (TP) and false negative rate (FN) corresponding to the learning by each expert labeling and their combination

	Expert 1	Expert 2	Expert 3	Expert 4	Maj.	$\textcircled{\cap}$	$\textcircled{\cup}$	$\textcircled{\wedge}$
AC	97.2%	97, 9%	98.5%	97.2%	98.2%	98.5%	98.1%	98.5%
TP	79.5%	79.5%	84.1%	78.2%	82.6%	87.5%	78.4%	87.9%
FN	20.3%	20.3%	12.8%	21.2%	17.3%	12.5%	13.6%	12.0%

Moreover, the benefit of exploiting label uncertainty for estimating the parameters of the IFA model can be noted in Tables 4, 5 and 7. Indeed, the results obtained if no confidence is considered and expert opinions are combined using the majority rule are much lower than those obtained by the other fusion methods and are not necessarily better than the performances obtained using individual expert labels (see Expert 3 in Tables 4 and 7).

6 Conclusion

The advantages of combining statistical data with knowledge from multiple experts has been demonstrated through a real-world diagnosis application. The proposed approach estimates the parameters of a statistical model using both objective data and uncertain class labels elicited from several experts. Parameter estimates are computed by maximizing a generalized likelihood criterion, using the evidential EM algorithm introduced in [15, 22].

The particular application that was considered concerns the diagnosis of railway track circuits. Experiments were carried out with real signals labeled by four different human experts. Experts' uncertain knowledge about the state of each capacitor was encoded as belief functions, which were pooled using different combination rules. These combined opinions were shown to yield better classification results than those obtained from each individual expert and from the majority rule. The cautious rule of combination introduced in [21] outperformed the conjunctive and disjunctive rules in this problem, which can be explained by the existence of common knowledge shared among the experts. Additionally, the benefits of taking into account degrees of confidence has also been demonstrated, which provides an empirical justification of the use of belief functions to encode expert knowledge in this kind of problem.

This work can be extended in several directions. The approach relies on expert knowledge elicitation in the belief function framework, an important problem that has not received much attention until now [9]. More sophisticated combination schemes could also be considered: for instance, discount rates could be learned from the data to take into account the competence of each individual expert using, e.g., the expert tuning method described in [25] (see also [39]). Finally, the parameter estimation approach based on uncertain data is obviously very general and can be applied to many other problems involving a statistical model and uncertain observations.

Acknowledgements This work was supported by the French National Research Agency (ANR) under project DIAGHIST. The authors thank the French National Railway Company (SNCF) and its experts for their collaboration.

References

1. Amari S, Cichocki A, Yang HH (1996) A New Learning Algorithm for Blind Signal Separation. In Proceedings of the 8th Conference on Advances in Neural Information Processing Systems (NIPS). MIT Press, , pp 756–763

2. Ambroise C, Denoeux T, Govaert G, Smets P, (2001) Learning from an imprecise teacher: probabilistic and evidential approaches. In: Proceedings of the 10th International symposium on applied stochastic models and data analysis (ASMDA), Compiègne, France, pp 100–105.
3. Ambroise C, Govaert G (2000) EM algorithm for partially known labels. In: Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS). Springer. Namur, Belgium, pp 161–166
4. Amini R, Gallinari P (2005) Semi-supervised learning with an imperfect supervisor. Knowl. Inf. Syst. Springer-Verlag New York, 8(4):385–413
5. Attias H (1999) Independent factor analysis, Neural Computation, MIT Press, Cambridge, MA, 11(4):803–851
6. Bartholomew DJ, Martin K (1999). Latent variable models and factor analysis. 2nd edn. Arnold, London
7. Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Computation, MIT Press, Cambridge, MA 7(6):1129–1159
8. Bengio Y, Grandvalet Y (2005) Semi-supervised learning by entropy minimization. In: Proceedings of the 17th Conference on Advances in Neural Information Processing Systems (NIPS). MIT Press, Cambridge, pp 529–536
9. Ben Yaghlane A, Denoeux T, Mellouli K (2006) Elicitation of expert opinions for constructing belief functions. In: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '06), Paris, France, pp 403–411
10. Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York
11. Coelho F, de Pádua Braga A, Natowicz R, Rouzier R (2010) Semi-supervised model applied to the prediction of the response to preoperative chemotherapy for breast cancer. Soft Computing - A Fusion of Foundations, Methodologies and Applications. Springer-Verlag
12. Chapelle O, Scholkopf B, Zien A (2006) Semi-Supervised Learning, MIT Press, Cambridge, MA
13. Côme E, Cherfi Z, Oukhellou L, Aknin P (2008) Semi-supervised IFA with prior knowledge on the mixing process: An application to a railway device diagnosis. In: Proceedings of the 7th ICMLA'08. San Diego, pp 415–420
14. Côme E, Oukhellou L, Denoeux T, Aknin P (2009), Noiseless Independent Factor Analysis with mixing constraints in a semi-supervised framework. Application to railway device fault diagnosis. In: Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN), Limassol, Cyprus, pp 416–425
15. Côme E, Oukhellou L, Denoeux T, Aknin P (2009), Learning from partially supervised data using mixture models and belief functions, Pattern Recognition, 42(3):334–348
16. Côme E, Oukhellou L, Denoeux T, Aknin P (2011) Fault diagnosis of a railway device using semi-supervised independent factor analysis with mixing constraints, Pattern Analysis & Applications (to appear), doi:10.1007/s10044-011-0212-3
17. Comon P (1994) Independent Component Analysis, a new concept?, Signal Processing, Special issue on Higher-Order Statistics, Elsevier, 36(3):287–314
18. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38(2): 325–339
19. Denoeux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Trans. Syst. Man Cybernet. 25(5): 804–813
20. Denoeux T, Zouhal LM (2001) Handling possibilistic labels in pattern classification using evidential reasoning, Fuzzy Sets Syst. 122(3): 47–62.
21. Denoeux T (2008) Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence. Artificial Intelligence, 172:234–264
22. Denoeux T (2010) Maximum likelihood from evidential data: an extension of the EM algorithm. In C. Borgelt et al. (Eds), Combining soft computing and statistical methods in data analysis, AISC 77. Springer , pp 181–188
23. Dubois D, Prade H (1988) Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence, 4(4): 244–264
24. Duda RO, Hart PE, Stork DG (2001) Pattern Classification, 2nd edn, John Wiley and Sons, New York
25. Elouedi Z, Mellouli K, Smets Ph (2004) Assessing sensor reliability for multisensor data fusion within the Transferable Belief Model, IEEE Transactions on Systems, Man and Cybernetics B, 34(1):782–787
26. Elouedi Z, Mellouli K, Smets Ph (2001) Belief decision trees: Theoretical foundations, International Journal of Approximate Reasoning, 28(2-3):91–124
27. Ghahramani Z (2004) Unsupervised Learning. In Bousquet O, Raetsch G, and von Luxburg U (ed) Advanced Lectures on Machine Learning. Springer-Verlag, pp 72-112
28. Grandvalet Y (2002) Logistic regression for partial labels. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Annecy, France, 3:1935–1941

29. Ha-Duong M (2008) Hierarchical fusion of expert opinions in the Transferable Belief Model, application to climate sensitivity, International Journal of Approximate Reasoning 49(3):555–574
30. Hastie T, Tibshirani R, Friedman J (2006). The Elements of Statistical Learning, Data Mining, Inference and Prediction. Statistics. Springer, New York
31. Hüllermeier E, Beringer J (2005) Learning from ambiguously labeled examples, in: Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA-05), Madrid, Spain, pp 168–179
32. Jenhani I, Ben Amor N, Elouedi Z (2007) Decision trees as possibilistic classifiers, Int. J. Approximate Reasoning, 43(8):784–807
33. Jraidi I, Elouedi Z (2007) Belief classification approach based on generalized credal EM, In Mellouli K (Ed), 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU '07), Hammamet, Tunisia, pp. 524–535, Springer
34. Klose A (2004) Extracting fuzzy classification rules from partially labeled data, Soft Computing - A Fusion of Foundations, Methodologies and Applications. Springer-Verlag, 8(6):417–427
35. Lawrence ND, Schölkopf B(2001) Estimating a kernel fisher discriminant in the presence of label noise. In: Proceedings of the 18th International Conference on Machine Learning (ICML). Morgan Kaufmann, San Francisco, pp 306–313.
36. Li Y, Wessels L, De Ridder D, Reinders M (2007) Classification in the presence of class noise using a probabilistic kernel fisher method. Pattern Recognition, 40(12):3349–3357
37. McLachlan GJ (1977) Estimating the linear discriminant function from initial samples containing a small number of unclassified observations, J. Am. Stat. Assoc. 72(358):403–406
38. Mclachlan GJ, Krishnan T (1997) The EM algorithm and Extension. Wiley, New York
39. Mercier D, Quost B, Denoeux T (2008) Refined modeling of sensor reliability in the belief function framework using contextual discounting, Information Fusion 9(2):246–258
40. Moulines E, Cardoso J, Cassiat E (1997) Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, pp 3617–3620
41. Nocedal J, Wright S (1999) Numerical Optimization. Springer Series in Operations Research and Financial Engineering, Springer
42. Oukhellou L, Debiolles A, Aknin P, Vilette F (2004) Automatic diagnostic of track circuit in a predictive maintenance context. International Conference on Railway Engineering. London
43. Oukhellou L, Debiolles A, Denoeux T, Aknin P (2010) Fault diagnosis in railway track circuits using Dempster-Shafer classifier fusion. Engineering Applications of Artificial Intelligence, 23:117–128
44. Palacios A, Sánchez L, Couso, I (2011) Linguistic cost-sensitive learning of genetic fuzzy classifiers for imprecise data. International Journal of Approximate Reasoning, 52(6):841–862
45. Pichon F and Denoeux T (2010) The unnormalized Dempster's rule of combination: a new justification from the Least Commitment Principle and some extensions. Journal of Automated Reasoning 45(1):61–87
46. Quost B, Masson M-H, Denoeux T (2011) Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. International Journal of Approximate Reasoning, 52(3):353–374.
47. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
48. Smets Ph (1990) The combination of evidence in the Transferable Belief Model. IEEE transactions on Pattern Analysis and Machine Intelligence, 12(5):447–458
49. Smets Ph (1993) Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning, 9:1–35
50. Smets Ph (1995) The canonical decomposition of a weighted belief. In Int. Joint Conf. on Artificial Intelligence. Morgan Kaufman, San Mateo, Ca, pp 1896–1901
51. Smets Ph, Kennes R (1994) The Transferable Belief Model. Artificial Intelligence, 66:191–234
52. Vannoorenbergue P, Denoeux T (2002) Handling uncertain labels in multiclass problems using belief decision trees. Proceedings of IPMU'2002, Vol. III, Annecy, France, pp 1919–1926.
53. Vannoorenberghe P, Smets Ph (2005) Partially Supervised Learning by a Credal EM Approach, In Godo L (Ed.), Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU '05), Barcelona, Spain, pp 956–967, Springer.
54. Worden K, Manson G, Denoeux T. (2009) An evidence-based approach to damage location on an aircraft structure. Mechanical Systems and Signal Processing 23(6):1792–1804
55. Yager RR (1987) On the Dempster-Shafer framework and new combination rules. Information Sciences. Elsevier Science Inc., New York, 41(2):93–137