Dissimilarity Metric Learning in the Belief Function Framework

Chunfeng Lian, Su Ruan, Thierry Denoeux

Dissimilarity Metric Learning in the Belief Function Framework

Chunfeng Lian, Su Ruan*, and Thierry Denœux

***Abstract*—The Evidential K-Nearest-Neighbor (EK-NN) method provided a global treatment of imperfect knowledge regarding the class membership of training patterns. It has outperformed traditional K-NN rules in many applications, but still shares some of their basic limitations, e.g., 1) classification accuracy depends heavily on how to quantify the dissimilarity between different patterns and 2) no guarantee for satisfactory performance when training patterns contain unreliable (imprecise and/or uncertain) input features. In this paper, we propose to address these issues by learning a suitable metric, using a low-dimensional transformation of the input space, so as to maximize both the accuracy and efficiency of the EK-NN classification. To this end, a novel loss function to learn the dissimilarity metric is constructed. It consists of two terms: the first one quantifies the imprecision regarding the class membership of each training pattern; while, by means of feature selection, the second one controls the influence of unreliable input features on the output linear transformation. The proposed method has been compared with some other metric learning methods on several synthetic and real data sets. It consistently led to comparable performance with regard to testing accuracy and class structure visualization.**

***Index Terms*—Dempster-Shafer Theory, Dissimilarity Metric Learning, Evidential K-NN, Feature Transformation, Feature Selection, Pattern Classification, Dimensionality Reduction.**

I. Introduction

THE K-nearest neighbor (K-NN) rule [1] is one of the most well-known pattern classification algorithms. As a case-based learning method without need of any prior assumptions [2], the K-NN classifier has been widely used in practice thanks to its simplicity. The original voting K-NN [1] assigns an object into the class represented by its majority nearest neighbors in the training set, while the information concerning the dissimilarity (distance) between the object and its neighbors is neglected. Then, the weighted K-NN [3] has been proposed, in which this dissimilarity is imported into the classification procedure. However, in the case of uncertain and imprecise data, many samples may be corrupted with noise or located in highly overlapping areas; consequently, it becomes difficult for these classical K-NN classifiers to obtain satisfactory classification results.

Asterisk indicates corresponding author.

Chunfeng Lian is with the Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France, and also with the Université de Rouen, QuantIF, LITIS, Rouen, France (e-mail: chunfeng.lian@utc.fr).

*Su Ruan is with the Université de Rouen, QuantIF, LITIS, Rouen, France (e-mail: su.ruan@univ-rouen.fr).

Thierry Denœux is with the Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France (e-mail: thierry.denoeux@hds.utc.fr).

The Dempster-Shafer theory (DST), also known as the theory of belief functions or Evidence theory, is an extension of both probability theory and the set-membership approach [4]. As a powerful framework for modeling and reasoning with uncertain and/or imprecise information [5]–[7], DST has shown remarkable applications in divers fields, such as model parameter estimation [8]–[10], unsupervised learning [11]–[13], supervised learning [14]–[22], ensemble learning [23], [24], information fusion [25]–[29], etc. To endow the K-NN method with the capability to handle uncertain information, Denœux has extended it in the belief function framework. An Evidential K-NN (EK-NN) rule has been proposed in [14], and further optimized in [15]. The EK-NN rule provides a global treatment of partial knowledge regarding the class membership of training patterns. Ambiguity and distance reject options are also taken into account based on the concepts of lower and upper expected losses [30].

The EK-NN method has outperformed other traditional K-NN methods in many situations when using the same information [15], whereas they still have some identical features: 1) the performances of the K-NN rules are strongly influenced by the chosen dissimilarity between different patterns. Better than directly using the simple Euclidean distance measure (such as in the original EK-NN), an adaptive dissimilarity metric tailored for the application should ensure better classification performance; 2) the efficiency of the K-NN rules substantially decrease when the dimensionality of the input data increases.

We propose a solution based on dissimilarity metric learning to deal with these inherent drawbacks of the K-NN classifications. Given an input space $\mathbf{X}$, the metric learning problem can be formulated as finding a transformation matrix A, such that the dissimilarity between any two patterns can be defined in the transformed space $Z = A\mathbf{X}$ [31]. Various studies have demonstrated that a properly learnt dissimilarity measure can dramatically boost the performance of the distance-based learning methods [32]–[37]. Even with a linear transformation of the input space [38]–[40], the K-NN classification can reach significant improvement. In [38], Goldberger et al. proposed a metric learning method called Neighborhood Component Analysis (NCA), which maximizes the expected leave-one-out classification accuracy from a stochastic version of the K-NN classification. Based on a softmax probability distribution defined in the transformed space, NCA labels each query instance by the majority vote of all training samples. As a main advantage of NCA, a continuous and differentiable cost function in respect of the linear transformation matrix A is deduced. This cost function can be minimized by gradient descent. The learnt matrix A can also be forced to be low-rank,

thus accelerating K-NN test and facilitating class structure visualization. Though the cost used in NCA is differentiable, it seems to be sensitive to the initialization. Inspired by NCA, Weinberger et al. proposed a Large Margin Nearest Neighbor (LMNN) method to learn a Mahalanobis distance metric for K-NN classification [40]. LMNN attempts to classify the K nearest neighbors as the same class label, under the constraint that different classes should be separated by a large margin. The learning problem is formulated as a semi-definite programming problem. The corresponding cost function consists of two terms; the first term penalizes large dissimilarities between instances with the same class label in a predefined neighborhood; while as a hinge loss, the second term penalizes small dissimilarities between instances with different class labels in the whole training pool. As a convex function in respect of the matrix A, the cost function of LMNN can be optimized efficiently.

Different from the global learning methods such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), both NCA and LMNN can adapt to the local structure of the application at hand. By learning a local dissimilarity metric, they effectively improved the K-NN classification accuracy in many situations. However, since they were not designed specifically for tackling data that contains unreliable input features, their performance may severely decline with this kind of imperfect information.

In this paper, our goal is to maximize the accuracy and efficiency of the EK-NN classifier on data that contains unreliable input features. To this end, we propose to learn an adaptive dissimilarity metric from this kind of imperfect data in the belief function framework. By using samples in the training pool as independent items of evidence, the belief regarding the class membership of each instance is modeled and refined using DST. A specific cost function consisting of two terms is constructed for learning a low-dimensional transformation matrix A. The first term attempts to minimize the imprecision regarding the class membership of each instance. The $\ell_{2,1}$-norm regularization of A acts as the second term, considering its good property for feature selection as already shown in muti-task learning [41], multiclass classification [42], semi-supervised learning [43], etc. By means of feature selection, it aims to manage the influence of unreliable input features on the output transformation. The proposed cost function is solved efficiently by a first order method (namely the proximal forward-backward splitting algorithm [44]). The influence of the sparsity regularization is tuned according to the application at hand. Finally, a low-dimensional transformation of the input space is realized to widely separate instances from different classes, therefore increasing the classification accuracy and reducing the searching time of the EK-NN classifier simultaneously.

The rest of this paper is organized as follows. The background on DST and the EK-NN classification rule is recalled in Section II. The proposed metric learning method based on DST is then introduced in Section III. In Section IV, the proposed method is tested on both synthetic and real data sets, and some comparison with other methods is presented. Finally, we conclude paper in Section V.

II. Background

The necessary background on DST and the EK-NN classification rule are briefly reviewed in Sections II-A and II-B, respectively.

A. Dempster-Shafer Theory

DST is also known as the theory of belief functions or evidence theory. As a generalization of both probability theory and the set-membership approaches, DST has two main components, i.e., quantification of a piece of evidence and combination of different items of evidence.

1) Evidence Quantification: DST is a formal framework for reasoning under uncertainty based on the modeling of evidence [4]. Let ω be a variable taking values in a finite domain $\Omega = \{\omega_1, \cdots, \omega_c\}$, called the *frame of discernment*. An item of evidence regarding the actual value of ω can be represented by a *mass function* m on Ω, defined from the powerset 2^Ω to the interval $[0, 1]$, such that

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

Each number $m(A)$ denotes a *degree of belief* attached to the hypothesis that "$\omega \in A$". Function m is said to be normalized if $m(\emptyset) = 0$, which is assumed in this paper. Any subset A with $m(A) > 0$ is called a *focal element* of mass function m. If all focal elements are singletons, m is said to be *Bayesian*; it is then equivalent to a probability distribution. A mass function m with only one focal element is said to be *categorical* and is equivalent to a set.

Corresponding to a normalized mass function m, we can associate *belief* and *plausibility* functions from 2^Ω to $[0, 1]$ defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B); \tag{2}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{3}$$

Quantity $Bel(A)$ (also known as *credibility*) can be interpreted as the degree to which the evidence supports A, while $Pl(A)$ can be interpreted as the degree to which the evidence is not contradictory to A. Functions Bel and Pl are linked by the relation $Pl(A) = 1 - Bel(\overline{A})$. They are in one-to-one correspondence with mass function m.

2) Evidence Combination: In DST, beliefs are elaborated by aggregating different items of evidence. The basic mechanism for evidence combination is *Dempster's rule* of combination [4]. Since Dempster's rule cannot well manage high conflicts between different pieces of evidence, various alternatives to it have been developed to tackle this problem under different situations, e.g., the TBM conjunctive and disjunctive combination rules [45], Yager's rule [5], Dubois-Prade's rule [6], the weighted average [46], [47], and the cautious and bold disjunctive rules [48] etc. In addition, the discounting strategy has been used in some other methods to deal with the conflicts, and a new dissimilarity measure consisting of both the conflict and distance has been introduced in [47] to determine the discounting factor of each source of

evidence to be combined. The conflicts have also been used for detecting the change occurrences in the fusion of multi-temple information [49], like in the change detection of remote sensing. Nevertheless, these alternative methods usually increase the complexity for applications, Dempster's rule still remains the most popular one for combining independent evidence.

Let m_1 and m_2 be two mass functions derived from independent items of evidence. They can be fused via Dempster's rule to induce a new mass function $m_1 \oplus m_2$ defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-Q} \sum_{B \cap C = A} m_1(B) m_2(C), \quad (4)$$

where $Q = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ measures the *degree of conflict* between evidence m_1 and m_2.

B. Evidential K-NN Classification Rule

An EK-NN classifier was proposed in [14] based on DST. Depending on the informativeness of the training samples with respect to the class membership of the query pattern, the EK-NN classifier computes a mass function over the whole frame of classes, and provides a global treatment of imperfect training knowledge with uncertainty.

Let $\{(X_i, Y_i) | i = 1, \cdots, N\}$ be a collection of N training pairs, in which $X_i = [x_1, \cdots, x_V]^T$ is the ith training sample with V features and $Y_i \in \{\omega_1, \cdots, \omega_c\}$ is the corresponding class label. Given a query instance X^t, its class membership can be determined through the following steps:

- Each neighbor of X^t is considered as an item of evidence that supports certain hypotheses regarding the class membership of X^t. Let X_j be one of its K nearest neighbors with class label $Y_j = \omega_q$. The mass function induced by X_j, which supports the assertion that X^t also belongs to ω_q is

$$\begin{cases} m_{t,j}(\{\omega_q\}) & = \alpha \exp(-\gamma_q d_{t,j}^2), \\ m_{t,j}(\Omega) & = 1 - \alpha \exp(-\gamma_q d_{t,j}^2), \end{cases} \quad (5)$$

  where $d_{t,j}$ is the distance between X_j and X^t, while α and γ are two tuning parameters that can be optimized via the method proposed in [15].
- Dempster's rule (4) is then executed to combine all neighbors' knowledge and obtain a global mass function for X^t. The lower and upper bounds for the belief of any specific hypothesis are then quantified via the credibility (2) and plausibility (3) values, respectively. In the case of {0,1} losses, the final decision on the class label of X^t can be made alternatively through maximizing the credibility, the plausibility, or the pignistic probability, as defined by Smets [7].

As an adaptive version of the EK-NN classifier, a neural network classifier based on DST has been proposed in [16]. Some other alternatives to the EK-NN method have also been developed. For instance, the credal classification methods [18]–[20] have been proposed by Liu et al. to deal with the overlapping classes in different cases. These methods permit the objects to be associated with not only the single classes but also meta-classes (i.e., disjunction of several classes) with different masses of belief, thus endowing the ability to specify the imprecision of classification.

III. Evidential Dissimilarity Metric Learning

A new approach, called evidential dissimilarity metric learning (EDML), is proposed in this section. By learning an adaptive dissimilarity measure on training samples that contain unreliable input features, EDML aims to maximize the performance of the EK-NN classifier.

A. Criterion of EDML

Let $\{(X_i, Y_i) | i = 1, \cdots, N\}$ be a collection of N training pairs, in which $X_i = [x_1, \cdots, x_V]^T$ is the ith observation with V input features, and Y_i is the corresponding class label taking values in a frame of discernment $\Omega = \{\omega_1, \cdots, \omega_c\}$. Assume the dissimilarity between instances X_i and X_j is quantified by a squared distance measure:

$$d^2(X_i, X_j) = (X_i - X_j)^T A^T A (X_i - X_j). \quad (6)$$

Then, EDML attempts to find an optimal matrix $A \in \mathbf{R}^{v \times V}$ under the constraint $v \ll V$. Such a linear transformation of the input space can boost the performance of the EK-NN classifier, since important features will have a strong impact when calculating the distance; classification is also faster in the low-dimensional transformed space.

To learn such a matrix A, we regard each X_i as a query instance. Then, the squared distance between X_i and X_j (i.e., $d^2(X_i, X_j)$) is used in (5), so as to represent the partial knowledge concerning the class membership of X_i that offered by training sample $(X_j, Y_j = \omega_q)$. Parameters α and γ used in (5) are restricted to be one for simplification.

Let Γ_q $(q = 1, \ldots, c)$ be the set of training samples (except X_i) belonging to the same class ω_q. Since the corresponding mass functions point to the same hypothesis (i.e., $Y_i = \omega_q$), they can be combined via Dempster's rule (namely (4)) to deduce a global mass function for all training samples in Γ_q:

$$\begin{cases} m_i^{\Gamma_q}(\{\omega_q\}) & = 1 - \prod_{j \in \Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \\ m_i^{\Gamma_q}(\Omega) & = \prod_{j \in \Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \end{cases}. \quad (7)$$

For $q = 1, \ldots, c$, the global mass function $m_i^{\Gamma_q}$ quantifies the evidence refined from the training pool that support the assertion $Y_i = \omega_q$. The mass of belief $m_i^{\Gamma_q}(\Omega)$ measures the imprecision of this evidence. In other words, it can be regarded as the calculation of the unreliability of the hypothesis $Y_i = \omega_q$. *If the actual value of Y_i is ω_q, the corresponding imprecision should then close to zero, i.e., $m_i^{\Gamma_q}(\Omega) \approx 0$; in contrast, imprecision pertaining to other hypotheses should close to one, i.e., $m_i^{\Gamma_r}(\Omega) \approx 1$, for $\forall r \neq q$.* According to this assumption, we propose to represent the prediction loss for training sample (X_i, Y_i) as

$$loss_i(A) = \sum_{q=1}^{c} t_{i,q} \cdot \left\{ 1 - m_i^{\Gamma_q}(\{\omega_q\}) \cdot \prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega) \right\}^2, \quad (8)$$

where $t_{i,q}$ is the qth element of a binary vector $t_i = \{t_{i,1}, \ldots, t_{i,c}\}$, with $t_{i,q} = 1$ if and only if $Y_i = \omega_q$. When $Y_i = \omega_q$ is true, minimizing $loss_i(A)$ can force both $m_i^{\Gamma_q}(\{\omega_q\}) = 1 - m_i^{\Gamma_q}(\Omega)$ and $\prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega)$ to approach one as far as possible, thus achieving the goal to maximize the

reliability of the right hypothesis ($Y_i = \omega_q$) but minimize the reliability of other assertions. As the result, the learnt matrix A can lead X_i only close to samples from the same class in the transformed space, thus protecting the classification performance of the EK-NN method.

Therefore, for all training samples, the loss function in respect of the transformation matrix A can be finally defined as

$$l(A) = \frac{1}{N}\sum_{i=1}^{N} loss_i(A) + \lambda||A||_{2,1}, \tag{9}$$

where $loss_i(A)$ represents the learning cost for training sample (X_i, Y_i) that quantified by (8). The $\ell_{2,1}$-norm sparsity regularization

$$||A||_{2,1} = \sum_{j=1}^{V}\left(\sum_{i=1}^{v} A_{i,j}^2\right)^{1/2} \tag{10}$$

is imported to select input features. By forcing columns of the transformation matrix A to be zero during the learning procedure, this sparsity term only selects the most reliable input features to calculate the linear transformation, thus controlling the influence of unreliable input features on the output low-dimensional transformed space. Scalar λ is a hyper-parameter that controls the influence of this regularization. Generally speaking, a too small λ may fail to limit the influence of unreliable input features; while, a too large λ may also delete informative features.

B. Optimization

Since $loss_i$ (8) is differentiable in respect of matrix A while $||A||_{2,1}$ (10) is partly smooth (it is non-smooth when and only when $A = 0$), the proximal Forward-Backward splitting (FBS) algorithms [44], [50], which belong to the class of first order methods, are efficient alternatives to solve the proposed loss function (9). More specifically, as an improved version of the classical FBS methods, the Beck-Teboulle proximal gradient algorithm [51] is used in this paper considering its computational simplicity and fast convergence rate.

In general, each iteration of the FBS algorithms can be broken up into a gradient descent step using $\frac{1}{N}\sum_{i=1}^{N} loss_i(A)$, followed by a proximal operation using $||A||_{2,1}$. According to (6)-(8), the derivative of $loss_i$ concerning A (i.e., $\partial loss_i/\partial A$) can be deduced as

$$\begin{aligned}\frac{\partial loss_i}{\partial A} =& \sum_{q=1}^{c} 2t_{i,q}\left\{1 - m_i^{\Gamma_q}(\{\omega_q\})\prod_{r\neq q}^{c} m_i^{\Gamma_r}(\Omega)\right\} \\ &\left\{-\frac{\partial m_i^{\Gamma_q}(\{\omega_q\})}{\partial A}\prod_{r\neq q}^{c} m_i^{\Gamma_r}(\Omega)\right. \\ &\left. - m_i^{\Gamma_q}(\{\omega_q\})\sum_{r\neq q}^{c}\frac{\partial m_i^{\Gamma_r}(\Omega)}{\partial A}\prod_{s\neq r,q}^{c} m_i^{\Gamma_s}(\Omega)\right\}.\end{aligned} \tag{11}$$

In which, value $m_i^{\Gamma_q}$ is calculated via (7), and for $\forall q = 1, \ldots, c$,

$$\frac{\partial m_i^{\Gamma_q}(\{\omega_q\})}{\partial A} = -\sum_{j\in\Gamma_q}\frac{\partial m_{ij}(\Omega)}{\partial A}\prod_{l\in\Gamma_q\setminus j} m_{il}(\Omega); \tag{12}$$

Algorithm 1: Beck-Teboulle proximal gradient algorithm [51]

Initialize $A^{(0)} \in \mathbf{R}^{v\times V}$ and $\beta > 0$, set $H^{(0)} = A^{(0)}$ and $t^{(0)} = 1$;
for $n = 0, 1, 2, \ldots$ **do**
  $G^{(n)} = H^{(n)} - \frac{\beta^{-1}}{N}\sum_{i=1}^{N}\frac{\partial loss_i}{\partial A}|_{A=H^{(n)}}$;
  $A^{(n+1)} = prox_{\beta^{-1},||A||_{2,1}}G^{(n)} =$
  $\arg\min_{A^*}\left\{\lambda||A^*||_{2,1} + \frac{\beta}{2}||A^* - G^{(n)}||^2\right\}$;
  $t^{(n+1)} = [\sqrt{4{t^{(n)}}^2 + 1} + 1]/2$;
  $\delta^{(n)} = 1 + [t^{(n)} - 1]/t^{(n+1)}$;
  $H^{(n+1)} = A^{(n)} + \delta^{(n)}[A^{(n+1)} - A^{(n)}]$;
end

While, mass m_{ij} is determined using (5) and (6), and

$$\frac{\partial m_{ij}(\Omega)}{\partial A} = 2m_{ij}(\{\omega_q\})A(X_i - X_j)(X_i - X_j)^T. \tag{13}$$

Based on (11)-(13), the Beck-Teboulle proximal gradient algorithm executes as the form shown in Algorithm 1, so as to deduce an optimal or at least sub-optimal low-dimensional transformation matrix A. To facilitate the optimization procedure, classical metric learning methods (e.g., PCA) can also be used to generate the initialization (i.e. $A^{(0)}$) for the proposed method. The learnt matrix A is then applied in (6) to measure the dissimilarity between different instances, and finally used in the EK-NN classification.

Complexity of the gradient calculation: Given m_{ij} and $m_i^{\Gamma_q}$, for $i = 1, \ldots, N$, the calculation of $\partial loss_i/\partial A$ can be performed in the following order:

1) calculation of $\partial m_{ij}(\Omega)/\partial A$ for $j = 1, \ldots, N-1$ using (13). Since $(X_i - X_j)(X_i - X_j)^T$ can be determined beforehand, this step requires $(N-1)vV^2$ arithmetic operations;
2) calculation of $\partial m_i^{\Gamma_q}/\partial A$ for $q = 1, \ldots, c$ using (12), which requires $\sum_{q=1}^{c}|\Gamma_q|vV = (N-1)vV$ arithmetic operations;
3) calculation of the last term in (11), which requires cvV operations. Since t_i has and only has one nonzero element, the calculation of (11) needs cvV operations.

Based on above steps, the complexity for calculating $\partial loss_i/\partial A$ is $O\left(vV^2N + vV(N+c)\right)$.

IV. Experimental Results

The presented experiments consist of four parts. In the first part, the proposed method, namely EDML, was evaluated on a synthetic data set. The proportion of unreliable (noisy and imprecise) features in this synthetic data was varied to assess the robustness of EDML under different situations. In the second part, EDML was evaluated on several real data sets. The corresponding classification accuracy was compared with some other metric learning methods. The parameters used in the proposed method were also studied. Finally, we further compared the two-dimensional visualization performance of different metric learning methods, so as to evaluate whether

TABLE I
CLASSIFICATION ACCURACY (BOTH TRAINING AND TESTING, IN %) OF THE EK-NN BASED ON DIFFERENT METRIC LEARNING METHODS. IN THE STUDIED SYNTHETIC DATA SETS, $n_r = 2$ AND $n_i = 2$. EDML-FS AND EDML DENOTE, RESPECTIVELY, THE PROPOSED METHOD WITH/WITHOUT THE $\ell_{2,1}$-NORM SPARSITY REGULARIZATION. PERFORMANCE OF THE SVM AND ENN CLASSIFIERS JOINT WITH PCA WERE ALSO PRESENTED AS TWO BASELINES FOR COMPARISON.

	n_u	SVM	ENN	PCA	NCA	LMNN	EDML	EDML-FS
training	6	92.00	91.33	90.67	99.33	98.00	98.67	99.33
	16	83.33	85.33	84.67	100.00	96.00	99.33	100.00
	26	81.33	83.33	74.00	100.00	96.67	100.00	100.00
	36	77.33	76.00	76.00	100.00	100.00	99.33	100.00
	46	76.67	74.67	68.67	100.00	99.33	99.33	100.00
testing	6	85.33	84.00	84.00	91.33	90.00	86.00	**94.67**
	16	74.00	72.00	73.33	84.00	86.67	86.00	**92.00**
	26	66.67	69.33	64.67	78.67	78.67	84.67	**90.00**
	36	69.33	69.67	62.00	70.00	78.00	76.67	**95.33**
	46	64.67	66.00	57.33	82.67	76.67	76.67	**94.00**

the proposed method can effectively separate instances from different classes in low-dimensional subspaces.

A. Performance on Synthetic Data

The studied synthetic data sets were generated using a process similar to the one described in [52]. The feature space contains n_r relevant features uniformly and independently distributed between -1 and +1. To obtain a high non-linear discriminant surface, the output label for a given instance is defined as

$$y = \begin{cases} \omega_1 & \text{if } \max_i(x_i) > 2^{1-\frac{1}{n_r}} - 1, \\ \omega_2 & \text{otherwise,} \end{cases} \tag{14}$$

where x_i is the ith relevant feature. Besides the relevant features, there are n_u irrelevant (noisy) features also uniformly distributed between -1 and +1, without any relation with the class label; and also n_i imprecise features copied as the cubic of the relevant features.

The numbers of relevant, irrelevant and imprecise features were set, respectively, as $n_r = 2$, $n_u \in \{6, 16, 26, 36, 46\}$ and $n_i = 2$ to simulate five different situations. Under each situation, we generated 150 training instances and the same number of testing instances. PCA, NCA, LMNN and the proposed EDML methods were executed to learn a two-dimensional dissimilarity metric A (i.e., $\in \mathbf{R}^{n_r \times (n_r + n_u + n_i)}$) on the training set. The obtained metric A was then used in the EK-NN to classify both the training and testing samples. As two baselines, results obtained by the SVM and Evidential Neural Network (ENN) [16] classifiers joint with PCA were also included for comparison.

Parameters of each method used in this experiment (four metric learning methods, i.e., PCA, NCA, LMNN and EDML, and three classifiers, i.e., EK-NN, ENN and SVM) can be summarized as follows:

- For LMNN, as suggested by [40], parameters K and μ were set as $K = 3$ and $\mu = 0.5$.
- For the proposed EDML, a rough grid search strategy was used to select an appropriate λ from $\{0.005, 0.007, 0.009\}$ according to the training performance.
- For the EK-NN classifier, parameters α and γ were optimized via the operation proposed in [15]. The number of nearest neighbors was set as $K = 3$.
- For the SVM, the gaussian kernel was used with the radial basis $\sigma = 1$.
- For the ENN classifier, the number of prototypes per class was set as 5.

It is worth illustrating that the parameters of the compared methods were always kept the same in the sequel experiments.

Finally, the training and testing (more important) accuracy (in %) obtained by different metric learning methods are summarized in Table I, in which EDML (manually set $\lambda = 0$) and EDML-FS (namely EDML joint with Feature Selection) represent, respectively, the proposed method without/with the sparsity regularization. As can be seen, the proposed EDML-FS led to higher testing accuracy than other methods under all the five different situations. It is also worth noting that the difference increased following the augment of unreliable input features, which reveals that the proposed method is stable and immune to severely deteriorated input information.

TABLE II
PROPERTIES OF THE FIVE REAL DATA SETS STUDIED IN SECTION IV-B.

data sets	classes	input features	instances
Wine	3	13	178
Seeds	3	7	210
Soybean	4	35	47
LSVT	2	309	126
Faces	40	100	400

B. Performance on Real Data

The proposed method was further evaluated using five real data sets of varying input features and classes. Four of these data sets (Wine, Seeds, Soybean-small and LSVT voice rehabilitation [53] data) were downloaded from the UCI Machine Learning Repository[1]. The other one is the Olivetti face recognition data set[2]. As a preprocessing operation for the Face data, we down-sampled the images to 38×31 pixels and used PCA to further reduce the dimensionality to 100. Properties of all the five data sets are briefly summarized in Table II.

The training and testing instances were randomly generated with 70/30 splitting, and repeated 50 times. Under each

[1]Please see at https://archive.ics.uci.edu/ml/index.html
[2]Please see at http://www.uk.research.att.com/facedatabase.html.

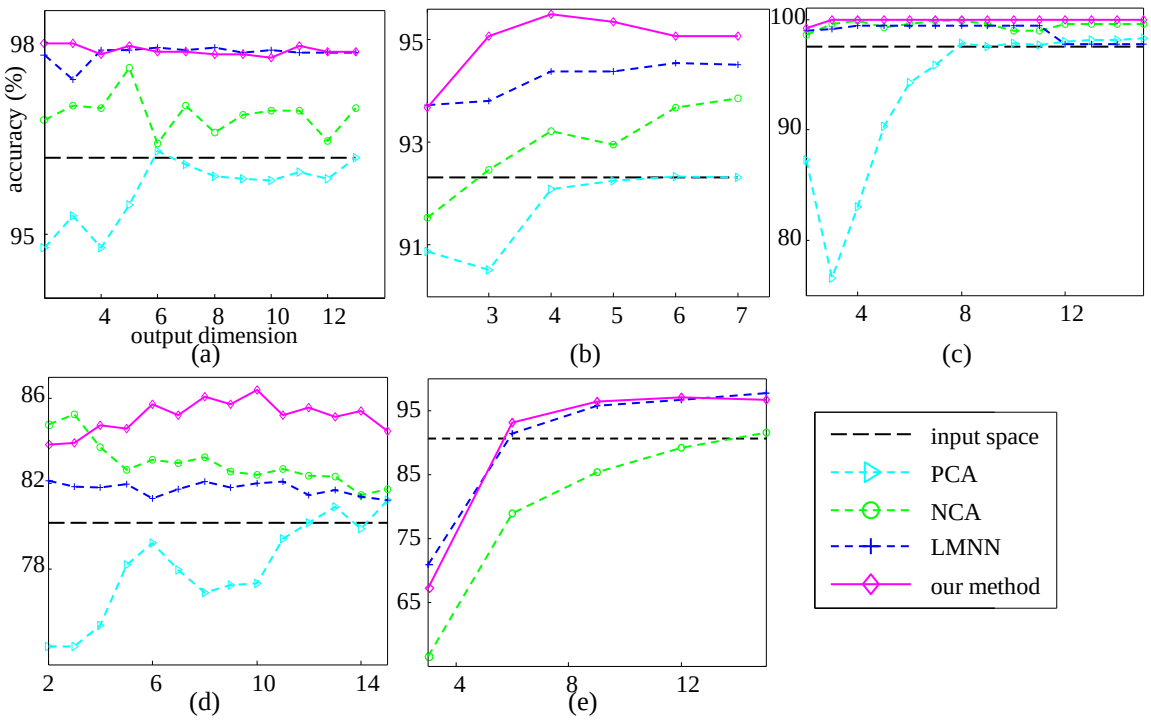

Fig. 1. Average testing accuracy obtained by different metric learning methods: (a) Wine data, (b) Seeds data, (c) Soybean-small data, (d) LSVT data and (e) Faces data. In each subfigure, the horizontal axis represents the output dimension (i.e., v) of the learnt transformation A, while the vertical axis represents the corresponding classification accuracy (in %).

TABLE III
THE BEST TRAINING AND THE CORRESPONDING TESTING ACCURACY (AVE±STD, IN %) OBTAINED BY DIFFERENT METHODS WITH $v \in \{2, 3, \ldots, 15\}$. EDML-FS AND EDML DENOTE, RESPECTIVELY, THE PROPOSED METHOD WITH/WITHOUT THE SPARSITY REGULARIZATION. RESULTS OF THE SVM ANDENN JOINT WITH PCA WERE ALSO PRESENTED AS TWO BASELINES FOR COMPARISON.

		SVM	ENN	PCA	NCA	LMNN	EDML	EDML-FS
training	Wine	99.95±0.19	100.00±0.00	97.65±1.24	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	Seeds	99.58±0.47	98.74±0.644	92.90±1.87	97.04±1.33	95.86±1.50	99.13±0.70	98.52±1.27
	Soybean	100.00±0.00	100.00±0.00	98.85±1.72	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	LSVT	99.52±0.83	93.82±1.28	80.71±3.94	98.63±1.69	100.00±0.00	96.45±1.46	97.14±1.33
	Faces	83.43±2.69	99.99±0.05	90.28±1.34	99.87±0.24	100.00±0.00	99.37±0.37	99.71±0.26
testing	Wine	96.15±2.50	97.51±2.30	95.90±2.63	96.97±2.23	97.88±1.68	96.46±2.95	**97.98±1.86**
	Seeds	91.78±2.93	92.92±2.96	92.25±3.14	93.85±2.88	94.54±2.78	94.48±2.45	**95.49±1.84**
	Soybean	**100.00±0.00**	98.00±3.83	98.28±3.97	99.86±1.00	99.44±2.80	**100.00±0.00**	**100.00±0.00**
	LSVT	80.42±4.82	85.21±3.94	80.92±5.76	82.57±6.35	82.13±5.40	85.03±4.98	**86.09±4.63**
	Faces	65.88±4.05	85.23±3.08	89.48±2.20	89.18±3.50	**97.63±1.66**	93.40±2.13	97.08±1.55

random split, we used PCA, NCA, LMNN and the proposed EDML methods, respectively, to learn a low-dimensional dissimilarity metric A (i.e., $\in \mathbf{R}^{v \times V}$ with $v \leq V$) on the training data; then used it in the EK-NN to classify both the training and testing instances. Parameters of compared methods were the same as that used in the last experiment (namely Section IV-A). For the proposed method, the hyper-parameter λ was still determined by a rough grid search strategy according to the training performance. On average, good results were obtained with λ between $[0.0005, 0.01]$ for the five data studied in this experiment.

The value of the output dimension v was orderly set as $\{2, 3, \ldots, 15\}$. Then, the average testing accuracy with different v was calculated and is shown in Fig. 1. As can be seen, the proposed method consistently performed well on these data sets as compared with other methods. More specifically, LMNN (blue line) and EDML (magenta line) had comparable testing accuracy on Wine and Faces data sets; NCA (green line), LMNN and EDML resulted in almost the same performance (EDML was slightly better) on the soybean-small data set; and EDML yielded the best performance on the other two data (Seeds and LSVT).

To further analyze the experimental results obtained on these real data sets, we computed the average training performance as a criterion to select the best output dimension v (from $\{2, 3, \ldots, 15\}$) for the learnt dissimilarity metric A. The best training accuracy and the corresponding testing accuracy (more important) for each method are summarized in Table III, in which results obtained by the SVM and ENN joint with PCA are also presented as two baselines for comparison. As in the former subsection, EDML (manually set $\lambda = 0$) and EDML-FS represented the proposed method without/with

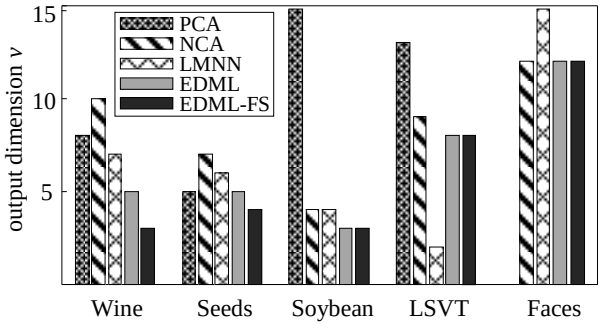

Fig. 2. The best output dimension v (between 2 and 15) according to the training performance obtained by different methods on the five real data sets.

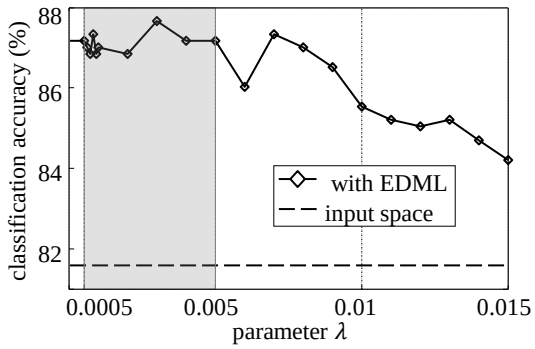

Fig. 3. Average testing accuracy on the LSVT data set with regard to the hyper-parameter λ. The output dimension was set as $v = 5$. The horizontal line represents the accuracy obtained in the input space.

the sparsity regularization. From Table III, it can be found that EDML-FS consistently yielded better performance than other methods on the first four data sets, especially on the LSVT data. This is mainly because the proposed method only selected the most informative features (from all the three hundred input features) to calculate the linear transformation. LMNN outperformed our method on the Face data set, but only with a slight difference. In addition, we can also see that, thanks to the $\ell_{2,1}$-norm sparsity regularization, EDML-FS performed better than EDML.

C. Parameter Analysis

1) Output Dimension: As discussed above, the best output dimension v (from $\{2, 3, \ldots, 15\}$) for the five real data sets studied in the last subsection was determined according to the training performance. Therefore, besides the classification accuracy presented in Table III, the corresponding output dimension obtained by different methods on these real data sets was also summarized and is shown in Fig 2.

2) Regularization Parameter: the hyper-parameter λ in the loss function (10) controls the effect of the sparsity regularization on the output low-dimensional transformation. It should be tuned specifically for each data set at hand. Generally speaking, a too small λ may fails to limit the influence of unreliable input features, while a too large λ may also removes many significant input features. On average, good results were

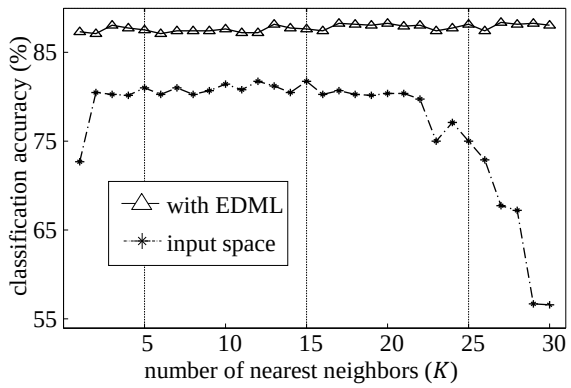

Fig. 4. Average accuracy of the EK-NN classification on the LSVT data set with regard to the number of nearest neighbors K. The output dimension was set as $v = 5$.

obtained with λ between $[0.0005, 0.01]$ for all the five data studied in the last subsection.

In this experiment, the LSVT data set was used as an example to further analysis the influence of the parameter λ. The training and testing data were generated with 70/30 splitting, and repeated 20 times. We orderly selected a λ from $\{0, 0.0005, 0.0006, \ldots, 0.001, 0.002, \ldots, 0.015\}$ to learn a low-dimensional transformation (with $v = 5$) of the input space. Then, the EK-NN classifier (with $K = 3$) was used to classify the testing instances on the transformed space. For all the 20 random splits, the average testing accuracy (in %) with regard to λ is finally shown in Fig. 3, in which the horizontal line represents the average accuracy of the EK-NN classification in the input space. As can be seen, relatively high performance on this data set is obtained with λ between $[0.0005, 0.01]$. The classification is less sensitive in the region $[0.0005, 0.005]$ than in other regions of λ.

3) Number of Nearest Neighbors: we also studied the parameter K of the EK-NN classification with the dissimilarity metric learnt by the proposed method. Still on the LSVT data set, the training and testing data were generated with 70/30 splitting, and repeated 20 times. Under each random split, we used the proposed method to learn a low-dimensional transformation of the input space. The output dimension and the regularization parameter were set as $v = 5$ and $\lambda = 0.002$. Then, the EK-NN classifier with $K = \{1, 2, \ldots, 30\}$ was orderly executed to classify the testing instances in the transformed space. As for comparison, the EK-NN classifier with the same K was also directly executed in the input space to classify the testing instances. The average testing accuracy with regard to K is finally summarized in Fig.4. It can be found that, with a metric learnt by the proposed method, the EK-NN classification always has higher accuracy on this data set than directly using the Euclidian distance in the input space. In addition, we can also see that the proposed method is robust to the parameter K.

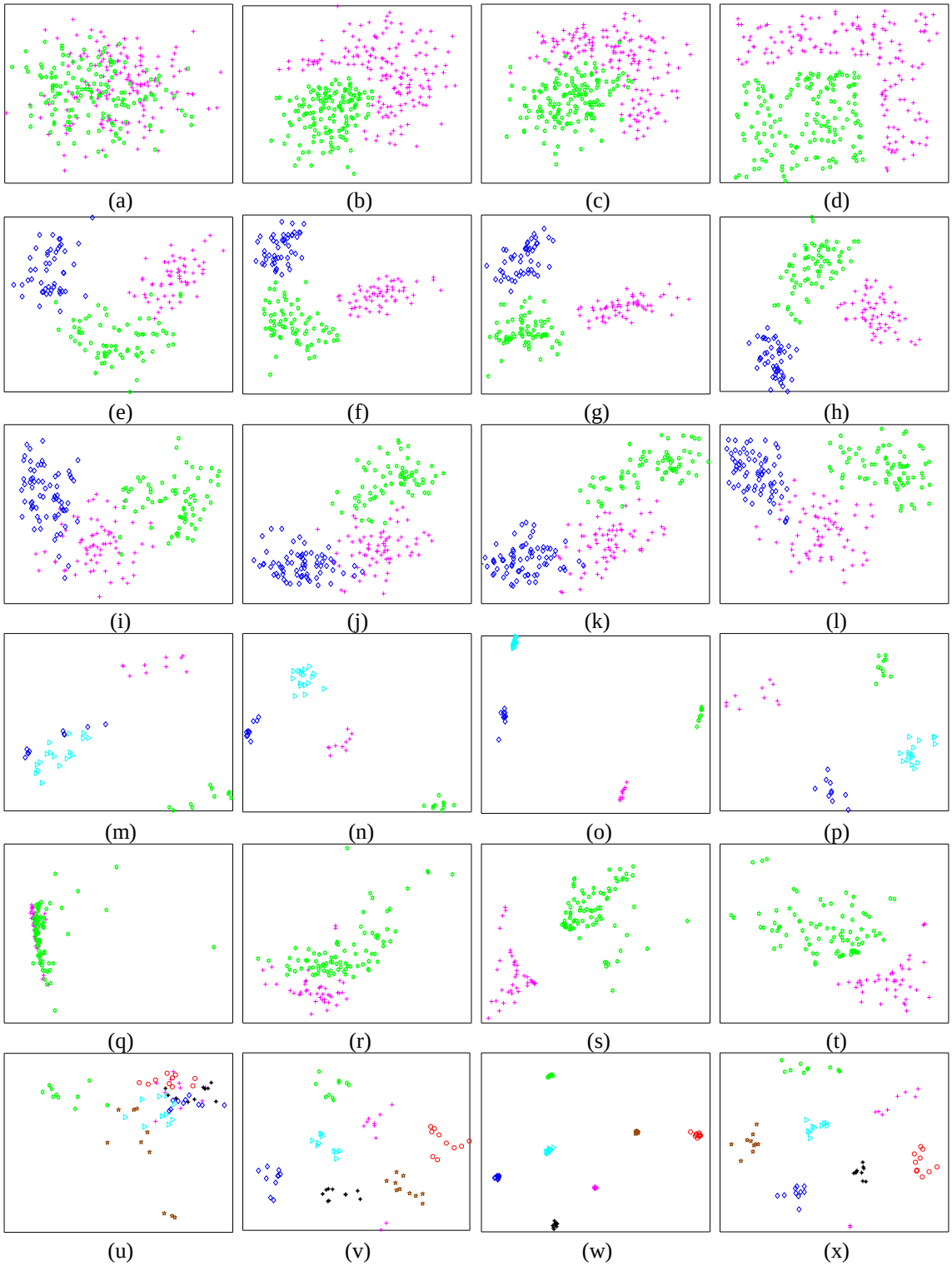

Fig. 5. Two-dimensional transformation results obtained by PCA, NCA, LMNN and the proposed method (orderly from the first to the forth column). (a)-(d) on synthetic data; (e)-(h) on Wine data; (i)-(l) on Seeds data; (m)-(p) on Soybean data; (q)-(t) on LSVT data; (u)-(x) on Faces data;

D. Two-Dimensional Visualization

To further evaluate whether the proposed method can effectively separate instances from different classes in low-dimensional transformation space, we visualized the dimension reduction in 2D, as shown in Fig. 5. PCA, NCA, LMNN and the proposed method were still compared on the same data sets (one synthetic and five real data sets) used in the former subsections. The input feature space for the synthetic data was set as fifty ($n_r = 2$,$n_i = 2$ and $n_u = 46$). For simplicity, only the first seven classes were studied in the Faces data. From the obtained results we can see that instances from different classes were always well separated by our method on all the six data sets. It led to the largest margin on the synthetic data, and the most satisfying separation on the Seeds data. In contrast, NCA did not separate the LSVT data perfectly; while LMNN resulted in large overlaps on the synthetic data.

V. Conclusion

To optimize the performance of the EK-NN classification on imperfect data sets, an approach based on Dempster-Shafer theory has been proposed to learn a dissimilarity metric specifying for the application at hand. By treating other training patterns as different sources of information, the belief concerning the class membership of each query pattern has been quantified and refined in the belief function framework. A specific loss function consisting of two terms has been developed for metric learning under uncertainty, in which the first term is used to minimize the imprecision regarding each instance's class membership, while the second term is the $\ell_{2,1}$-norm sparsity regularization of the low-dimensional transformation matrix. Through a feature selection procedure, it serves to limit the influence of uncertainty and/or imprecise input features. The proposed method has been evaluated on several synthetic and real data sets, consistently showing good performance with regard to classification accuracy and class structure visualization. Moreover, it has also proved that the proposed method is not sensitive to the parameter K.

Future work will focus on two main aspects. First, to improve the efficiency of the proposed method on large data sets, we will further study other more advanced optimization algorithms. Furthermore, it is worth extending the proposed method to learn non-linear transformation of the input space, so as to improve the performance of the EK-NN classifier on high complex data sets.

Acknowledgment

The authors thank all the anonymous referees for their helpful and constructive comments. This work was partly supported by the China Scholarship Council (CSC).

References

[1] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[2] T. Denœux and P. Smets, "Classification using belief functions: relationship between case-based and model-based approaches," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 6, pp. 1395–1406, 2006.

[3] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 325–327, 1976.

[4] G. Shafer, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.

[5] R. R. Yager, "On the Dempster-Shafer framework and new combination rules," *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[6] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," *Computational Intelligence*, vol. 4, no. 3, pp. 244–264, 1988.

[7] P. Smets and R. Kennes, "The transferable belief model," *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[8] T. Denœux, "Maximum likelihood estimation from uncertain data in the belief function framework," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 119–130, 2013.

[9] E. Ramasso and T. Denœux, "Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions." *Fuzzy Systems, IEEE Transactions on*, pp. 1–12, 2013.

[10] Z. Su, Y. Wang, and P. Wang, "Parametric regression analysis of imprecise and uncertain data in the fuzzy belief function framework," *International Journal of Approximate Reasoning*, vol. 54, no. 8, pp. 1217–1242, 2013.

[11] T. Denœux and M.-H. Masson, "EVCLUS: evidential clustering of proximity data," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 1, pp. 95–109, 2004.

[12] M.-H. Masson and T. Denœux, "ECM: An evidential version of the fuzzy C-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

[13] N. Makni, N. Betrouni, and O. Colot, "Introducing spatial neighbourhood in evidential C-means for segmentation of multi-source images: application to prostate multi-parametric mri," *Information Fusion*, vol. 19, pp. 61–72, 2014.

[14] T. Denœux, "A K-nearest neighbor classification rule based on Dempster-Shafer theory," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 25, no. 5, pp. 804–813, 1995.

[15] L. M. Zouhal and T. Denœux, "An evidence-theoretic K-NN rule with parameter optimization," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 28, no. 2, pp. 263–271, 1998.

[16] T. Denœux, "A neural network classifier based on Dempster-Shafer theory," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 30, no. 2, pp. 131–150, 2000.

[17] Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: theoretical foundations," *International Journal of Approximate Reasoning*, vol. 28, no. 2, pp. 91–124, 2001.

[18] Z. Liu, Q. Pan, and J. Dezert, "A new belief-based K-nearest neighbor classification method," *Pattern Recognition*, vol. 46, no. 3, pp. 834–844, 2013.

[19] Z. Liu, Q. Pan, J. Dezert, and G. Mercier, "Credal classification rule for uncertain data based on belief functions," *Pattern Recognition*, vol. 47, no. 7, pp. 2532–2541, 2014.

[20] Z. Liu, Q. Pan, G. Mercier, and J. Dezert, "A new incomplete pattern classification method based on evidential reasoning," *Cybernetics, IEEE Transactions on*, vol. 45, no. 4, pp. 635–646, 2015.

[21] C. Lian, S. Ruan, and T. Denœux, "An evidential classifier based on feature selection and two-step classification strategy," *Pattern Recognition*, vol. 48, no. 7, pp. 2318–2327, 2015.

[22] L. Jiao, Q. Pan, T. Denoeux, Y. Liang, and X. Feng, "Belief rule-based classification system: Extension of FRBCS in belief functions framework," *Information Sciences*, vol. 309, pp. 26–49, 2015.

[23] H. Altınçay, "Ensembling evidential k-nearest neighbor classifiers through multi-modal perturbation," *Applied Soft Computing*, vol. 7, no. 3, pp. 1072–1083, 2007.

[24] Y. Bi, J. Guan, and D. Bell, "The combination of multiple classifiers using an evidential reasoning approach," *Artificial Intelligence*, vol. 172, no. 15, pp. 1731–1751, 2008.

[25] N. Milisavljevic and I. Bloch, "Sensor fusion in anti-personnel mine detection using a two-level belief function model," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 33, no. 2, pp. 269–283, 2003.

[26] Z. Elouedi, K. Mellouli, and P. Smets, "Assessing sensor reliability for multisensor data fusion within the transferable belief model," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 1, pp. 782–787, 2004.

[27] S. Destercke, D. Dubois, and E. Chojnacki, "Possibilistic information fusion using maximal coherent subsets," *Fuzzy Systems, IEEE Transactions on*, vol. 17, no. 1, pp. 79–92, 2009.

[28] R. R. Yager, "Set measure directed multi-source information fusion," *Fuzzy Systems, IEEE Transactions on*, vol. 19, no. 6, pp. 1031–1039, 2011.

[29] B. Lelandais, S. Ruan, T. Denœux, P. Vera, and I. Gardin, "Fusion of multi-tracer PET images for dose painting," *Medical Image Analysis*, vol. 18, no. 7, pp. 1247–1259, 2014.

[30] T. Denoeux, "Analysis of evidence-theoretic decision rules for pattern classification," *Pattern recognition*, vol. 30, no. 7, pp. 1095–1107, 1997.

[31] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, 2006.

[32] D. G. Lowe, "Similarity metric learning for a variable-kernel classifier," *Neural Computation*, vol. 7, no. 1, pp. 72–85, 1995.

[33] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, 2002, pp. 505–512.

[34] D.-Y. Yeung and H. Chang, "A kernel approach for semisupervised metric learning," *Neural Networks, IEEE Transactions on*, vol. 18, no. 1, pp. 141–149, 2007.

[35] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007, pp. 209–216.

[36] W. Bian and D. Tao, "Constrained empirical risk minimization framework for distance metric learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 8, pp. 1194–1205, 2012.

[37] D. Kedem, S. Tyree, F. Sha, G. R. Lanckriet, and K. Q. Weinberger, "Non-linear metric learning," in *Advances in Neural Information Processing Systems*, 2012, pp. 2573–2581.

[38] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems*, 2005, pp. 513–520.

[39] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *Proceedings of the 21th International Conference on Machine Learning*, 2004, pp. 94–101.

[40] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[41] A. Evgeniou and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems*, 2010, pp. 41–48.

[42] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 11, pp. 1738–1754, 2012.

[43] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 252–264, 2015.

[44] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011, pp. 185–212.

[45] P. Smets, "The combination of evidence in the transferable belief model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 5, pp. 447–458, 1990.

[46] C. K. Murphy, "Combining belief functions when evidence conflicts," *Decision Support Systems*, vol. 29, no. 1, pp. 1–9, 2000.

[47] Z. Liu, J. Dezert, Q. Pan, and G. Mercier, "Combination of sources of evidence with different discounting factors based on a new dissimilarity measure," *Decision Support Systems*, vol. 52, no. 1, pp. 133–141, 2011.

[48] T. Denœux, "Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence," *Artificial Intelligence*, vol. 172, no. 2, pp. 234–264, 2008.

[49] Z. Liu, J. Dezert, G. Mercier, and Q. Pan, "Dynamic evidential reasoning for change detection in remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 5, pp. 1955–1967, 2012.

[50] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

[51] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[52] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *The Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[53] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 22, no. 1, pp. 181–190, 2014.

Chunfeng Lian received the B.S. degree in Electronic and Information Engineering from the Xidian University, Xi'an, China. He is currently pursuing the Ph.D. degree with the Department of Information Processing Engineering, Sorbonne Universités, Université de Technologie de Compiègne, Compiègne, France. His current research interests concern the applications of the belief function theory in pattern recognition and medical image analysis.

Su Ruan received the M.S. and the Ph.D. degrees in Image Processing from the University of Rennes, France, in 1989 and 1993, respectively. She was previously an Associate Professor at the University of Caen, France, from 1993 to 2003, and a Full Professor at the University of Champagne-Ardenne, France, from 2003 to 2010. She has been a Full Professor at the University of Rouen, France, since 2010.

Her main area of research is image processing, particularly in the fields of image segmentation, pattern recognition, and data fusion. Her developments include advanced machine learning techniques, shape models, non rigid registration, graph-based image segmentation and data fusion strategies, applicable to medical imaging.

Thierry Denœux received the Bachelor's and the Ph.D. degrees from the Ecole des Ponts ParisTech, Paris, France, in 1985 and 1989, respectively. He is currently a Full Professor (Classe Exceptionnelle) with the Department of Information Processing Engineering, Sorbonne Universités, Université de Technologie de Compiègne, Compiègne, France. His research interests concern the management of uncertainty in intelligent systems. His main contributions are in the theory of belief functions with applications to pattern recognition, data mining, and information fusion.

Dr. Denœux is the Editor-in-Chief of the International Journal of Approximate Reasoning, and was an Associate Editor of the IEEE Transactions on Fuzzy Systems, from 2012 to 2014. He is the President of the Belief Functions and Applications Society.