Joint Feature Transformation and Selection Based on Dempster-Shafer Theory

Chunfeng Lian[1,2(✉)], Su Ruan[2], and Thierry Denœux[1]

[1] Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, 60205 Compiègne, France
{chunfeng.lian,thierry.denoeux}@utc.fr
[2] Université de Rouen, QuantIF - EA 4108 LITIS, 76000 Rouen, France
su.ruan@univ-rouen.fr

Abstract. In statistical pattern recognition, feature transformation attempts to change original feature space to a low-dimensional subspace, in which new created features are discriminative and non-redundant, thus improving the predictive power and generalization ability of subsequent classification models. Traditional transformation methods are not designed specifically for tackling data containing unreliable and noisy input features. To deal with these inputs, a new approach based on Dempster-Shafer Theory is proposed in this paper. A specific loss function is constructed to learn the transformation matrix, in which a sparsity term is included to realize joint feature selection during transformation, so as to limit the influence of unreliable input features on the output low-dimensional subspace. The proposed method has been evaluated by several synthetic and real datasets, showing good performance.

Keywords: Belief functions · Dempster-Shafer theory · Feature transformation · Feature selection · Pattern classification

1 Introduction

The performance of pattern classification methods depends crucially on the quality of input features: (1) with a small-sized training pool, a relatively high dimensional feature space increases the complexity of the learning algorithms, thus raising the risk of over-fitting on the training set; (2) it often happens that the input space contains features that are irrelevant, or even at odds with the class labels. These unreliable input features could decrease substantially the classification accuracy of the distance-based learning algorithms (e.g., the K-nearest neighbor rules).

Low-dimensional feature transformation is a feasible solution to the issues discussed above. It attempts to transform the original feature space to a discriminative subspace, in which new features are created for use in model construction. However, since traditional feature transformation methods, e.g., principal component analysis (PCA), neighborhood component analysis (NCA) [5] and large margin nearest neighbor method (LMNN) [18], were not designed specifically

J.P. Carvalho et al. (Eds.): IPMU 2016, Part I, CCIS 610, pp. 253–261, 2016.
DOI: 10.1007/978-3-319-40596-4_22

for tackling data that contains unreliable input features, their performance may severely decline with this kind of imperfect information.

The Dempster-Shafer Theory (DST) [15] is also known as the theory of belief functions or Evidence theory. As a powerful tool for modeling and reasoning with uncertain and/or imprecise information, it has shown remarkable applications in divers fields, such as unsupervised learning [3,13,20], supervised learning [4,6,8,10,11], information fusion [7,9,12,14,17], etc. These facts motivated us to design a new DST-based feature transformation method for data that contains unreliable and noisy features. To this end, a specific cost function consisting of two terms is constructed for learning a low-dimensional transformation matrix. The first term minimizes the imprecision regarding the class membership of each instance. The $\ell_{2,1}$-norm regularization of the transformation matrix acts as the second term. By means of feature selection, it aims to manage the influence of unreliable original features on the output transformation. The proposed cost function is minimized efficiently by a first order method (namely the Beck-Teboulle proximal gradient algorithm [1]). Finally, a low-dimensional transformation of the original feature space is realized to widely separate instances from different classes.

The rest of this paper is organized as follows. The background on DST is recalled in Sect. 2. The proposed method based on DST is then introduced in Sect. 3. In Sect. 4, the proposed method is tested on both synthetic and real-world datasets. Finally, we conclude paper in Sect. 5.

2 Background on Dempster-Shafer Theory

The necessary background on DST is briefly reviewed in this section. 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.

2.1 Evidence Quantification

DST is a formal framework for reasoning under uncertainty based on the modeling of evidence [15]. 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:

$$Bel(A) = \sum_{B \subseteq A} m(B); \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2}$$

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.

2.2 Evidence Combination

In DST, beliefs are elaborated by aggregating different items of evidence. *Dempster's rule of combination* [15], as well as its unnormalized version, i.e., the *conjunctive combination rule* defined in the Transferable Belief Model (TBM) [16], are basic mechanisms for evidence fusion. 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

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - Q} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{3}$$

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.

3 Method

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 instance with V input features, and Y_i is the corresponding class label taking values in a frame of discernment $\Omega = \{\omega_1, \cdots, \omega_c\}$ with an integer $c \geq 2$.

Feature Transformation. To realize a linear transformation of the input feature space, we need to learn a matrix $A \in \mathbf{R}^{v \times V}$, by which the squared distance between any two instances (e.g., X_i and X_j) is quantified as

$$d^2(X_i, X_j) = ||AX_i - AX_j||_2^{\,2}. \tag{4}$$

The size of the transformation matrix A should satisfy the constraint $v \ll V$, so as to output a low-dimensional transformation.

To learn such a matrix A, we successively set each X_i, $i \in \{1, \ldots, N\}$, as a query instance. Then, other samples in the training pool can be regarded as

independent items of evidence that support different hypotheses concerning the class membership of X_i. The evidence offered by the training sample $(X_j, Y_j = \omega_q)$, $j \neq i$ and $q \in \{1, \ldots, c\}$, asserts that X_i is also originated from the class ω_q. However, this piece of evidence is partially reliable. It is inversely proportional to the dissimilarity between X_i and X_j, and can be quantified as a mass function

$$\begin{cases} m_{ij}(\{\omega_q\}) & = \exp\left(-d^2(X_i, X_j)\right) \\ m_{ij}(\Omega) & = 1 - \exp\left(-d^2(X_i, X_j)\right) \end{cases}, \tag{5}$$

where the distance, i.e., $d^2(X_i, X_j)$, is measured by (4). 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 (i.e., (3)) to deduce a global mass function for all training samples in Γ_q:

$$\begin{cases} m_i^{\Gamma_q}(\{\omega_q\}) & = 1 - \prod_{j \in \Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \\ m_i^{\Gamma_q}(\Omega) & = \prod_{j \in \Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \end{cases}. \tag{6}$$

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 hypothesis. If the actual value of Y_i is ω_q, this 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., $m_i^{\Gamma_r}(\Omega) \approx 1$, $\forall r \neq q$. According to this assumption, we propose to represent the prediction loss for training sample (X_i, Y_i) as a function of the matrix A, namely

$$loss_i(A) = \sum_{q=1}^{c} t_{i,q} \cdot \left\{ 1 - m_i^{\Gamma_q}(\{\omega_q\}) \cdot \prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega) \right\}^2, \tag{7}$$

where $t_{i,q}$ is the qth element of a binary vector $t_i = [t_{i,1}, \ldots, t_{i,c}]$, with $t_{i,q} = 1$ iff $Y_i = \omega_q$. 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.

Feature Selection. To control the influence of unreliable input features in the transformed feature subspace, the $l_{2,1}$-norm sparsity regularization of A, namely

$$||A||_{2,1} = \sum_{j=1}^{V} \left(\sum_{i=1}^{v} A_{i,j}^2 \right)^{1/2}, \tag{8}$$

is adopted to realize the joint selection and transformation of input features. By forcing columns of A to be zero during the learning procedure, this sparsity term

can only select the most reliable input features to calculate the low-dimensional transformation.

Finally, based on all training samples, the loss function to learn the matrix A is defined as

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

where λ is a hyper-parameter that controls the influence of the sparsity penalty.

Optimization. Considering that $loss_i$ (7) is differentiable concerning A, while $||A||_{2,1}$ (8) is partly smooth (it is non-smooth iff $A = 0$), the Beck-Teboulle proximal gradient algorithm [1], which belongs to the class of first-order optimization methods, is used in this paper to find the solution of (9).

4 Experimental Results

In this section, the proposed method was evaluated by a synthetic dataset and two real-world datasets. The Evidential K-nearest-neighbor (EK-NN) classification rule [2] was selected to classify the testing samples after feature transformation.

4.1 Evaluation by Synthetic Datasets

The synthetic dataset was generated using a process similar to the one described in [19]. It contains n_r relevant features uniformly and independently distributed between $[-1, 1]$. The output label of each instance is determined by

$$y = \begin{cases} \omega_1 & \text{if } \max_i(x_i) > 2^{1-\frac{1}{n_r}} - 1 \\ \omega_2 & \text{otherwise} \end{cases}, \tag{10}$$

where x_i is the ith relevant feature. Besides the relevant features, there are n_u irrelevant (noisy) features also uniformly distributed between $[-1, 1]$, without

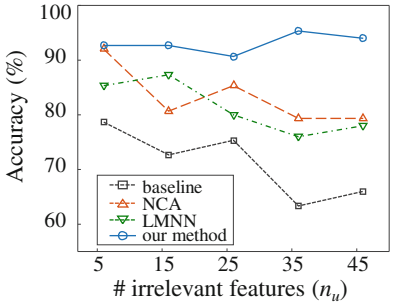

Fig. 1. Testing accuracy of the EK-NN classifier based on different feature transformation methods. Performance in the input feature space is presented as the baseline.

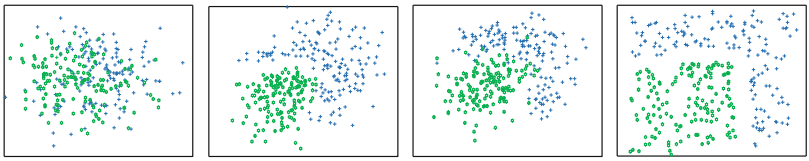

Fig. 2. Two-dimensional transformation results obtained by (a) PCA, (b) NCA, (c) LMNN and (d) our method, respectively.

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, 150 training instances and 150 testing instances were simulated. The proposed method was compared with PCA, NCA [5] and LMNN [18]. Each of the compared methods was used to learn a two-dimensional transformation (i.e., to learn a matrix $A \in \mathbf{R}^{n_r \times (n_r + n_u + n_i)}$) on the training dataset. After that, the EK-NN was used to classify the testing samples in the transformed subspace. The parameters used in the EK-NN classifier was determined by the method proposed in [21], and the number of nearest neighbors was set as $K = 3$.

Finally, the testing accuracy (in %) for different methods with respect to changing number of unreliable features are summarized in Fig. 1, in which the results obtained by the input features are also presented as the baselines for comparison. As can be seen, our method has higher testing accuracy than other methods under all the five different situations. It is also worth to note that the difference increases following the augment of unreliable input features, which reveals that the proposed method is stable and immune to severely deteriorated input information.

Apart from the classification performance, we also visualized a synthetic dataset (fifty input features with $n_r = 2$, $n_i = 2$ and $n_u = 46$) in the 2-D subspace, so as to evaluate whether the proposed method can effectively separate instances from different classes after feature transformation. The proposed method was still compared with PCA, NCA, and LMNN. As shown in Fig. 2, it outputs the largest margin between different classes as compared to the other three methods.

4.2 Evaluation by Real-World Datasets

The proposed method was further evaluated using two real-world datasets offered by oncologists[1]:

[1] They are with the Department of Nuclear Medicine, Centre Henri Becquerel, 76038 Rouen, France.

(1) Lung Tumor Dataset: This dataset contains twenty-five lung tumor patients (instances) treated with chemo-radiotherapy (CRT). For each patient, fifty-two intensity and texture features were extracted from the positron emission tomography (PET) images acquired before and during the treatment. The class label for each patient was *recurrence* or *no-recurrence*, which was clinically assessed at one year after the end of CRT.

(2) Esophageal Tumor Dataset: This dataset contains thirty-six esophageal tumor patients (instances) treated with chemo-radiotherapy. For each patient, twenty-nine features were extracted from the PET images and the clinical documents. The class label for each patient was *disease-free* or *disease-positive*, which was clinically assessed at one month after the end of CRT.

The two real-world datasets are briefly summarized in Table 1. Comparing to a limited number of instances (which is often encountered in the medical domain), a relatively large amount of input features were gathered for each clinical dataset. In addition, due to system noise and limited resolution of PET imaging, some features calculated from the PET images are unreliable, or even at odds with the class labels.

Since the datasets are small-sized, the leave-one-out cross-validation (LOOCV) was adopted to assess the performance. The proposed method was compared with the other three feature transformation methods, i.e., PCA, NCA and LMNN. For all the compared methods, the dimensionality of the output subspace was chosen between two to five. The best output dimension was determined according to the average testing accuracy. Finally, the average training accuracy and testing accuracy (more important) obtained by different methods

Table 1. Description of the real-world datasets.

Dataset	Classes	Instances	Features
Lung tumor	2	25	52
Esoph. tumor	2	36	29

Table 2. Comparing the performance ($ave \pm std$) of our method with the other three feature transformation methods (PCA, NCA and LMNN). The results obtained by the EK-NN in the original feature space is served as the baseline for comparison.

Method	Lung Tumor Dataset		Esophageal Tumor Dataset	
	Training	Testing	Training	Testing
Original space	69.50±4.46	60.00±50.00	63.73±2.14	61.11±49.44
PCA	81.50±5.25	76.00±43.60	56.90±5.81	58.34±50.00
NCA	99.50±1.83	80.00±40.82	94.21±3.24	69.44±46.72
LMNN	100.00±0.00	68.00±47.61	85.48±4.50	80.56±40.14
Our method	100.00±0.00	**88.00±33.17**	97.46±1.64	**83.33±37.80**

are summarized in Table 2, in which results obtained by the input features are presented as baselines for comparison. As shown in Table 2, the proposed method leads to better testing accuracy than other methods on the studied real-world datasets.

5 Conclusion

An approach based on DST has been proposed to realize joint feature transformation and feature selection from the input space that contains unreliable features. To this end, a loss function consisting of two terms has been constructed, in which the first term attempts to minimize the imprecision regarding each training sample's class membership; while the second term, namely a sparsity regularization of the transformation matrix, serves to limit the influence of unreliable input features on the output feature transformation. The constructed loss function has been minimized by a proximal gradient algorithm to find a satisfactory transformation matrix. After that, the output matrix has been used to accomplish the low-dimensional transformation of the input space. Experimental results obtained on the synthetic dataset and the real-world datasets show that the proposed method can be used to improve the performance of classification methods (e.g., the EK-NN classifier) on low-quality data.

Acknowledgements. This work was partly supported by China Scholarship Council.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
2. Denœux, T.: A K-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. Syst. Man Cybern. **25**(5), 804–813 (1995)
3. Denœux, T., Kanjanatarakul, O., Sriboonchitta, S.: EK-NNclus: a clustering procedure based on the evidential k-nearest neighbor rule. Knowl.-Based Syst. **88**, 57–69 (2015)
4. Denœux, T., Smets, P.: Classification using belief functions: relationship between case-based and model-based approaches. IEEE Trans. Syst. Man Cybern. Part B Cybern. **36**(6), 1395–1406 (2006)
5. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems, pp. 513–520 (2005)
6. Jiao, L., Pan, Q., Denoeux, T., Liang, Y., Feng, X.: Belief rule-based classification system: extension of FRBCS in belief functions framework. Inf. Sci. **309**, 26–49 (2015)
7. Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I.: Fusion of multi-tracer PET images for dose painting. Med. Image Anal. **18**(7), 1247–1259 (2014)
8. Lian, C., Ruan, S., Denœux, T.: An evidential classifier based on feature selection and two-step classification strategy. Pattern Recogn. **48**(7), 2318–2327 (2015)

9. Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P.: Dempster-Shafer theory based feature selection with sparse constraint for outcome prediction in cancer therapy. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 695–702. Springer, Heidelberg (2015)
10. Liu, Z., Pan, Q., Mercier, G., Dezert, J.: A new incomplete pattern classification method based on evidential reasoning. IEEE Trans. Cybern. **45**(4), 635–646 (2015)
11. Ma, L., Destercke, S., Wang, Y.: Online active learning of decision trees with evidential data. Pattern Recogn. **52**, 33–45 (2016)
12. Makni, N., Betrouni, N., Colot, O.: Introducing spatial neighbourhood in evidential C-means for segmentation of multi-source images: application to prostate multi-parametric MRI. Inf. Fusion **19**, 61–72 (2014)
13. Masson, M.H., Denœux, T.: ECM: an evidential version of the fuzzy C-means algorithm. Pattern Recogn. **41**(4), 1384–1397 (2008)
14. Nguyen, T., Boukezzoula, R., Coquin, D., Perrin, S.: Combination of sugeno fuzzy system and evidence theory for NAO robot in colors recognition. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2015)
15. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
16. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. **66**(2), 191–234 (1994)
17. Wang, F., Miron, A., Ainouz, S., Bensrhair, A.: Post-aggregation stereo matching method using Dempster-Shafer theory. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 3783–3787 (2014)
18. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
19. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
20. Zhou, K., Martin, A., Pan, Q., Liu, Z.-G.: Median evidential C-means algorithm and its application to community detection. Knowl.-Based Syst. **74**, 69–88 (2015)
21. Zouhal, L.M., Denœux, T.: An evidence-theoretic K-NN rule with parameter optimization. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **28**(2), 263–271 (1998)