# Dempster-Shafer Theory based Outcome Prediction in Cancer Therapy

Chunfeng Lian[1,2], Su Ruan[2], Pierre Vera[3,2], Thierry Denœux[1]

[1]Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc UMR 7253, 60319, 60203 Compiègne, France

[2]Université de Rouen, QuantIF-LITIS EA 4108, 76000 Rouen, France

[3]Centre Henri-Becquerel, Department of Nuclear Medicine, 76038 Rouen, France

chunfeng.lian@utc.fr, su.ruan@univ-rouen.fr,
pierre.vera@chb.unicancer.fr, thierry.denoeux@hds.utc.fr

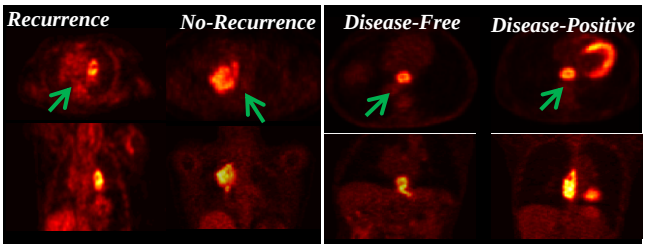
**Résumé** − La prédiction de l'évolution d'un cancer permet d'adapter la planification d'un traitement thérapeutique, ce qui rend le traitement plus efficace. Dans cet article, nous proposons d'utiliser les caractéristiques d'image extraites á partir des images PET et des caractéristiques cliniques pour la prédiction. Compte tenu des imprécisions et des incertitudes dans les deux sources d'information, une nouvelle méthode de prédiction basée sur la théorie des fonctions de croyance est proposée. Tout d'abord, une fonction de perte avec une régularisation parcimonieuse est construite comme critère pour apprendre une métrique de dissimilarité entre les vecteurs de caractéristiques des patients. Une réduction de dimension est ensuite effectuée. Grâce à la contrainte parcimonieuse, l'influence des caractéristiques imprécises est réduite pour la prise de décision. Enfin, la métrique apprise est intégrée dans le classificateur EK-NN (Evidential K-Nearest-Neighbor) pour effectuer la prédiction. La méthode proposée a montré de bonne performance sur deux jeux de données relatives aux tumeurs du poumon et de l'œsophage.

**Abstract** − As a vital task in cancer therapy, outcome prediction is the foundation for tailoring and adapting a treatment planning. In this paper, we propose to use image features extracted from PET and clinical characteristics. Considering that both information sources are imprecise or noisy, a novel prediction model based on Dempster-Shafer theory has been developed. Firstly, a specific loss function with sparse regularization was designed for learning an adaptive dissimilarity metric between feature vectors of labeled patients. Through minimizing this loss function, a linear low-dimensional transformation of the input features is then achieved; meanwhile, thanks to the sparse penalty, the influence of imprecise input features can also be reduced via feature selection. Finally, the learnt dissimilarity metric is used with the Evidential $K$-Nearest-Neighbor (EK-NN) classifier to predict the outcome. We evaluated the proposed method on two clinical data sets concerning to lung and esophageal tumors, showing good performance.

## 1 Introduction

Accurate outcome prediction prior to or even during the cancer treatment is of great clinical value, upon which more effective treatment planning can be updated. Medical imaging plays a fundamental role in this task, since it allows noninvasive monitoring of tumor lesions [5]. Some research has proven that functional information provided by fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) is predictable for response of therapy [10, 8]. Abounding image features can be extracted from FDG-PET, such as standardized uptake values, e.g., $\text{SUV}_{max}$, $\text{SUV}_{peak}$ and $\text{SUV}_{mean}$, total lesion glycolysis (TLG) and metabolic tumor volume (MTV) [10]. Moreover, texture analysis through PET images may also provide complementary predictive values [11]. The quantification of these features before and during cancer therapy has been claimed to be predictable for treatment response [5]. Nevertheless, their application is still hampered by some practical difficulties. First, compared to a relatively large amount of interesting features, we often have just a small sample of observations in clinical study. As a consequence, the predictive power of traditional statistical machine learning algorithms breaks down as the dimensionality of feature space increases. Secondly, due to system noise and limited resolution of PET, as well as partly subjective quantification of clinical characteristics, some of these interesting features are imprecise.

Dimensionality reduction is a feasible solution to the issue discussed above. However, traditional methods, including feature transformation methods (e.g., [4]) and feature selection (e.g., [12]), are not designed to work for imperfect data tainted with uncertainty. As a powerful framework for representing and reasoning with uncertainty and imprecise information, Dempster-Shafer theory (DST) [9] has been increasingly applied in statistical pat-




(a) Lung Tumors (b) Esophageal Tumors

FIG. 1: Example tumor uptakes on FDG-PET imaging; (a) recurrence and no-recurrence instances of lung tumor; (b) disease-free and disease-positive instances of esophageal tumor.

tern recognition [3, 6]. These facts motivated us to design a DST-based prediction method for imprecise input features and small observation samples.

To this end, and different with our former feature selection approach [7], a new feature transformation method based on DST is proposed in this paper. We construct a specific loss function with sparse penalty to learn an adaptive low-rank distance metric for representing dissimilarity between different patients' feature vectors. A linear low-dimensional transformation of input features is then achieved through minimizing this loss function. Simultaneously, using the $\ell_{2,1}$-norm regularization of learnt dissimilarity metric in the loss function, feature selection is also realized to reduce the influence of imprecise features. At last, we apply the learnt dissimilarity metric in the evidential $K$-nearest-neighbor (EK-NN) classifier [2] to predict the treatment outcome.

The rest of this paper is organized as follows. The fundamental background on DST is reviewed in Section 2. The proposed method is introduced in Section 3, after which some experimental results are presented in Section 4. Finally, Section 5 concludes this paper.

## 2 Dempster-Shafer Theory

DST is a framework for reasoning under uncertainty based on the modeling of evidence [9]. More precisely, let $\omega$ 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 $\omega$ can be represented by a *mass function* $m$ from $2^\Omega$ to [0,1], such that $\sum_{A\subseteq\Omega} m(A) = 1$. 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.

Corresponding to a normalized mass function $m$, the *belief* and *plausibility function* for all $A \subseteq \Omega$ are further defined as:

$$Bel(A) = \sum_{B\subseteq A} m(B); \quad Pl(A) = \sum_{B\cap A\neq\emptyset} m(B). \qquad (1)$$

Quantity $Bel(A)$ represents the degree to which the evidence *supports* $A$, while $Pl(A)$ represents the degree to which the evidence is *not contradictory* to $A$.

Different items of evidence can be aggregated to elaborate beliefs in DST. Let $m_1$ and $m_2$ be two mass functions derived from independent items of evidence. They can be combined via *Dempster's rule* to generate a refined mass function:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-Q} \sum_{B\cap C=A} m_1(B)m_2(C) \qquad (2)$$

for all $A \in 2^\Omega \setminus \emptyset$, where $Q = \sum_{B\cap C=\emptyset} m_1(B)m_2(C)$ measures the *degree of conflict* between $m_1$ and $m_2$.

## 3 Method

Let $\{(X_i, Y_i)|i = 1, \cdots, n\}$ be a collection of $n$ labeled patients, in which $X_i = [x_1, \cdots, x_v]^T$ is the $i$th observation with $v$ input features, and $Y_i$ is the corresponding label taking values in a frame of discernment $\Omega = \{\omega_1, \cdots, \omega_c\}$.

Firstly, we need to learn a dissimilarity metric $d(X_i, X_j)$, so as to maximize the prediction performance of the EK-NN classifier on future testing patient. We regard this problem as learning a transformation matrix $A \in \mathbf{R}^{h\times v}$ , from which the distance $d(X_i, X_j)$ is defined as

$$d(X_i, X_j) = (X_i - X_j)^T A^T A(X_i - X_j). \qquad (3)$$

Matrix $A$ is further restricted to be of low-rank $h$ (i.e., $h \ll v$), such that a low-dimensional linear transformation of the input feature space can be learnt, making EK-NN classifier more efficient.

In the DST framework, if $X_i$ is a query instance, then other labeled points in the training data set can be viewed as partial knowledge regarding $X_i$'s prediction label. So, each point $X_j$ $(\neq i)$ with $Y_j = \omega_q$ is a piece of evidence that increases the belief that $X_i$ also belongs to $\omega_q$. This piece of evidence can be quantified as a mass function

$$\begin{cases} m_{ij}(\omega_q) & = \exp\left(-d(X_i, X_j)\right) \\ m_{ij}(\Omega) & = 1 - \exp\left(-d(X_i, X_j)\right) \end{cases}, \qquad (4)$$

where dissimilarity $d(X_i, X_j)$ is measured via Equation (3).

After modeling the evidence for all training samples (except $X_i$) using Equation (4), they are further allocated into different groups $\Gamma_q$ $(q = 1, \ldots, c)$ according to corresponding class labels. Then, after combination using Dempster's rule (Equation (2)), the mass function for each group $\Gamma_q$ is represented as

$$\begin{cases} m_i^{\Gamma_q}(\{\omega_q\}) & = 1 - \prod_{j\in\Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \\ m_i^{\Gamma_q}(\Omega) & = \prod_{j\in\Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right] \end{cases}. \qquad (5)$$

The mass of belief $m_i^{\Gamma_q}(\Omega)$ for group $\Gamma_q$ reflects the imprecision about the hypothesis that $Y_i = \omega_q$. If any hypothesis is true, the corresponding mass function should be more precise. For instance, *if the actual value of* $Y_i$ *is* $\omega_q$,

TAB. 1: Prediction accuracy (both training and testing, in %) of EK-NN ($K = 3$) based on different dissimilarity metrics. ELT-FS* and ELT* denote, respectively, the proposed method with/without the $\ell_{2,1}$-norm sparse regularization.

| Method | **Lung Tumor Data** | | **Esophageal Tumor Data** | |
|---|---|---|---|---|
| | training | testing | training | testing |
| all features | 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 |
| LDA | 100.00±0.00 | 52.00±50.99 | 100.00±0.00 | 55.56±50.40 |
| NCA | 99.50±1.83 | 80.00±40.82 | 94.21±3.24 | 69.44±46.72 |
| ELT* | 95.83±3.80 | **88.00±33.17** | 88.02±4.03 | 63.89±48.71 |
| ELT-FS* | 100.00±0.00 | **88.00±33.17** | 97.46±1.64 | **83.33±37.80** |

*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.,* $m_i^{\Gamma_r}(\Omega) \approx 1$*, for* $\forall r \neq q$. Based on this idea, we propose to represent the prediction loss for training sample $(X_i, Y_i)$ as

$$loss_i = \sum_{q=1}^{c} t_{i,q} \cdot \{1 - [1 - m_i^{\Gamma_q}(\Omega)] \cdot \prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega)\}^2, \quad (6)$$

where $t_{i,q}$ is the $q$th element of a binary vector $t_i = \{t_{i,1}, \ldots, t_{i,c}\}$, with $t_{i,q} = 1$ if and only if $Y_i = \omega_q$.

As a result, for all training samples, the loss function with respect of the transformation matrix $A$ can be expressed as

$$l(A) = \frac{1}{n} \sum_{i=1}^{n} loss_i + \lambda ||A||_{2,1}, \quad (7)$$

where $loss_i$ is calculated using Equation (6). Sparse regularization $||A||_{2,1} = \sum_{i=1}^{v} (\sum_{j=1}^{h} A_{i,j}^2)^{1/2}$ is added to select features in order to limit the influence of imprecise input features during the linear transformation. Scalar $\lambda$ is a hyper-parameter that controls the influence of the regularization term.

A quasi-Newton method [1] is used to minimize Equation (7). After that, we apply the learnt matrix $A$ in Equation (3), and use the EK-NN classifier to predict the treatment outcome of future testing patients.

# 4 Experimental Results

We compared the proposed method (called evidential low-dimensional transformation with feature selection, i.e., ELT-FS) with several linear transformation methods, namely PCA, LDA and neighborhood component analysis (NCA) [4]. We used two real data sets:

*1) Lung Tumor Data:* Twenty-five patients with stage II-III non small cell lung cancer were studied. 52 SUV-based ($SUV_{max}$, $SUV_{mean}$, $SUV_{peak}$, MTV and TLG) and texture-based (gray level size zone matrices (GLSZM) [11]) features were extracted from corresponding PET images. The definition of recurrence of tumor for patients at one year after the treatment is primarily clinical with biopsy and PET/CT. There were 19 patients with label *recurrence*, while the remaining six patients were labeled with *no recurrence* (example images can be seen in Figure 1(a)).

*2) Esophageal Tumor Data:* Thirty-six patients with esophageal squamous cell carcinomas were studied. We have 29 SUV-based ($SUV_{max}$, $SUV_{mean}$, $SUV_{peak}$, MTV and TLG), GLSZM-based and patients' clinical features (gender, tumour stage and location, WHO performance status, dysphagia grade and weight loss from baseline). The disease-free evaluations include a clinical examination with PET/CT and biopsies. So 13 patients were labeled *disease-free* when neither loco regional nor distant tumor recurrence is detected, while the remaining 23 patients were labeled as *disease-positive* (example images can be seen in Figure 1(b)).

The leave-one-out cross-validation (LOOCV) procedure was used for evaluation. PCA, LDA, NCA were compared with our ELT-FS. Each method learns a low-dimensional transformation matrix $A$ on training data set. The EK-NN classifier was then used to predict class label of the left testing instance. Parameter $K$ of EK-NN was set as 3. Hyper-parameter $\lambda$ used in our ELT-FS was determined using a rough grid search strategy. For PCA, NCA and ELT-FS, the dimension of transformed feature space was chosen between two to five according to the minimum average testing error. Finally, the average classification accuracy for all methods are summarized in Table 1. Experiments with all features and our method without feature selection (namely ELT with $\lambda = 0$ ) are also presented for comparison. It can be observed that our method, especially ELT-FS, leads to higher testing performance in both cases.

Furthermore, we visualized the dimension reduction in 2D achieved using the PCA, NCA, ELT and ELT-FS methods, as shown in Fig. 2. It can be seen that different classes in both data sets are better separated by our method than using other methods. The best separation is achieved using our method with feature selection (ELT-FS).

# 5 Conclusion

In this study, a novel approach based on DST has been proposed to predict the outcome of a cancer treatment

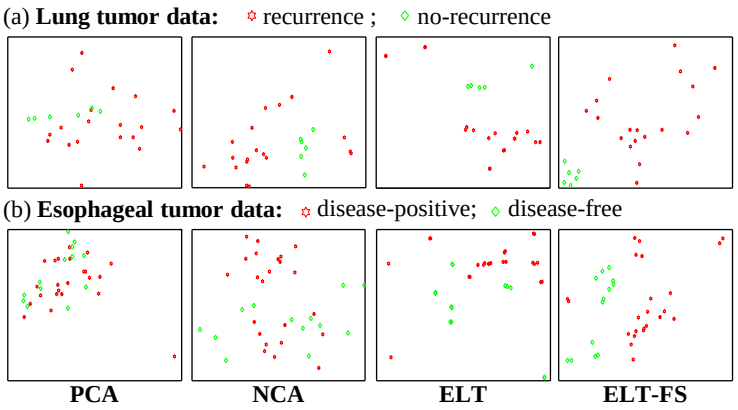



FIG. 2: Two-dimensional transformation results of PCA, NCA, our ELT (without feature selection, i.e., $\lambda = 0$) and ELT-FS. All observations were studied in both data sets.

using PET image features and clinical characteristics. A specific loss function has been designed to take into account problems of uncertainty and imprecision, so as to learn an adaptive dissimilarity metric for EK-NN classifier. We have realized a low-dimensional linear transformation of input feature vectors. Simultaneously, thanks to the $\ell_{2,1}$-norm sparse regularization, a feature selection procedure has been performed to reduce the influence of imprecise input features. Experimental results obtained on two clinical data sets show that the proposed method performs well. In the future, we will further evaluate it on more and larger data sets with different types of tumors, and study the influence of the parameter $\lambda$.

## References

[1] J. E. Dennis Jr and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. SIAM, 1996.

[2] T. Denœux. A K-nearest neighbor classification rule based on Dempster-Shafer theory. *Systems, Man and Cybernetics, IEEE Transactions on*, 25(5):804–813, 1995.

[3] T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):119–130, 2013.

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

[5] P. Lambin, R. G. van Stiphout, M. H. Starmans, E. Rios-Velazquez, G. Nalbantov, H. J. Aerts, E. Roelofs, W. van Elmpt, P. C. Boutros, P. Granone, et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nature Reviews Clinical Oncology*, 10(1):27–40, 2013.

[6] C. Lian, S. Ruan, and T. Denœux. An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recognition*, 48(7):2318–2327, 2015.

[7] C. Lian, S. Ruan, T. Denœux, and P. Vera. Outcome prediction in tumour therapy based on Dempster-Shafer theory. In *International Symposium on Biomedical Imaging*, 2015.

[8] H. Mi, C. Petitjean, B. Dubray, P. Vera, and S. Ruan. Prediction of lung tumor evolution during radiotherapy in individual patients with PET. *Medical Imaging, IEEE Transactions on*, 33(4):995–1003, 2014.

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

[10] S. Tan, S. Kligerman, W. Chen, M. Lu, G. Kim, S. Feigenberg, W. D. D'Souza, M. Suntharalingam, and W. Lu. Spatial-temporal [18 F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 85(5):1375–1382, 2013.

[11] F. Tixier, M. Hatt, C. C. Le Rest, A. Le Pogam, L. Corcos, and D. Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *Journal of Nuclear Medicine*, 53(5):693–700, 2012.

[12] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu. Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. *Computer Vision and Image Understanding*, 115(2):256–269, 2011.