Dempster-Shafer Theory Based Feature Selection with Sparse Constraint for Outcome Prediction in Cancer Therapy

Chunfeng Lian[1,2], Su Ruan[2], Thierry Denœux[1], Hua Li[4], and Pierre Vera[2,3]

[1] Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, 60205 Compiègne, France
[2] Université de Rouen, QuantIF - EA 4108 LITIS, 76000 Rouen, France
[3] Centre Henri-Becquerel, Department of Nuclear Medicine, 76038 Rouen, France
[4] Washington University School of Medicine, Department of Radiation Oncology, Saint Louis, 63110 MO, USA

Abstract. As a pivotal 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 is developed. Firstly, a specific loss function with sparse regularization is 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.

Keywords: Outcome Prediction, PET, Feature Selection, Sparse Constraint, Dempster-Shafer Theory.

1 Introduction

Accurately predicting the treatment outcome prior to or even during cancer therapy is of great clinical value. It facilitates the adaptation of a treatment planning for individual patient. Medical imaging plays a fundamental role in assessing the response of a treatment, as it can monitor and follow-up the evolution of tumor lesions non-invasively [8]. Up to now, the metabolic uptake information provided by fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) has been proven to be predictable for pathologic response of a treatment in several cancers, e.g., lung tumor [6,11] and esophageal tumor [15]. Abounding image features can be extracted from FDG-PET, such as standardized uptake values (SUVs), like SUV_{max}, SUV_{peak} and SUV_{mean}, that describe metabolic uptake

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 695–702, 2015.
DOI: 10.1007/978-3-319-24574-4_83

in a region of interest, total lesion glycolysis (TLG) and metabolic tumor volume (MTV) [15]. In addition, some complementary analysis of PET images, e.g., image texture analysis [16], can also provide supplementary evidence for outcome evaluation. The quantification of these features before and during cancer therapy has been claimed to be predictable for treatment response [8]. Nevertheless, their further application is still hampered by some practical difficulties. First, compared to a relatively large amount of interesting features from the point of view of clinicians, we often have just a small sample of observations in clinical study. As a consequence, the predictive power of traditional statistical machine learning algorithms, e.g., K-nearest neighbor (K-NN) classifier, break down as the dimensionality of feature space increases. Secondly, due to system noise and limited resolution of PET imaging, the effect of small tumor volumes [1], as well as partly subjective quantification of clinical characteristics, some of these (texture, intensity and clinical) features are imprecise.

Dimensionality reduction is a feasible solution to the issues discussed above. However, traditional methods, including feature transformation methods, e.g., kernel principal component analysis (K-PCA) [13] and neighbourhood components analysis (NCA) [7], and feature selection, e.g., univariate and multivariate selection [12,17], 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) [14] has been increasingly applied in statistical pattern recognition [5,10] and information fusion for cancer therapy [2,9]. These facts motivated us to design a new DST-based prediction method for imprecise input features and small observation samples.

In this paper, we firstly develop 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 [4] 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 then introduced in Section 3, after which some experimental results are presented in Section 4. Finally, Section 5 concludes this paper.

2 Backgrounds on Dempster-Shafer Theory

DST is also known as evidence theory or theory of belief functions. As a generalization of both probability theory and set-membership approach, it acts as a framework for reasoning under uncertainty based on the modeling of evidence [14].

Specifically speaking, we assume 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 from 2^{Ω} 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:

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

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:

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

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 these two pieces of evidence.

3 Method

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

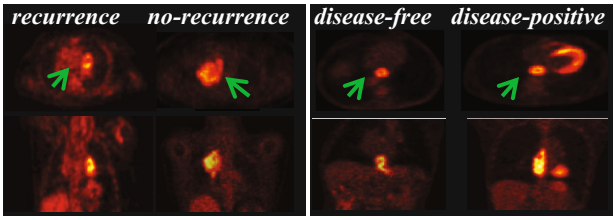
Firstly, we need to learn a dissimilarity metric $d(X_i, X_j)$ on this training data set, so as to maximize the prediction performance of the EK-NN classifier on future testing patient. Alternatively, we regard this problem as learning a transformation matrix $A \in \mathbf{R}^{h\times v}$, from which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

$$d(X_i, X_j) = (AX_i - AX_j)^T(AX_i - AX_j) = (X_i - X_j)^T A^T A(X_i - X_j). \tag{3}$$

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 the 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. More precisely, 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 ω_q. However, this piece of evidence is not 100% certainty. It is inversely proportional to the dissimilarity between X_i and X_j, and can be quantified as a mass function

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



(a) Lung Tumors (b) Esophageal Tumors

Fig. 1. Examples of tumor uptakes on FDG-PET imaging from different views; (a) recurrence and no-recurrence instances before treatment of lung tumor; (b) disease-free and disease-positive instances before treatment of esophageal tumor.

where dissimilarity $d(X_i, X_j)$ is measured via Eq. (3). This mass function can be expressed by saying that, based on (X_j, Y_j), the belief can only partly be committed to ω_q, while the complementary of it is uncertainty, and can only be assigned to the whole frame Ω.

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

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

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

$$loss_i = \sum_{q=1}^{c} t_{i,q} \cdot \left(1 - m_i^{\Gamma_q}(\omega_q) \cdot \prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega)\right)^2, \tag{6}$$

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 $Y_i = \omega_q$.

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

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

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

As a differentiable function regarding matrix A, Eq. (7) is then minimized using a quasi-Newton method [3]. After that, we apply the learnt matrix A in Eq. (3), and use the EK-NN classifier to predict the treatment outcome of future testing patients.

Table 1. Comparing prediction accuracy (ave±std, in %) of different methods. 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
EK-NN	69.50±4.46	60.00±50.00	63.73±2.14	61.11±49.44
SVM	100.00±0.00	76.00±43.60	100.00±0.00	63.89±48.71
T-test	99.67±1.15	72.00±45.83	75.56±2.10	66.67±47.81
IG	86.50±3.86	68.00±47.61	88.57±2.37	75.00±43.92
SFS	95.67±2.24	84.00±37.42	85.63±1.60	52.78±50.56
SFFS	64.33±3.62	72.00±45.83	59.68±6.79	80.56±40.14
PCA	88.33±1.70	80.00±40.82	59.60±5.81	55.56±50.40
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
K-PCA	81.33±4.36	80.00±40.82	71.19±5.89	72.22±45.43
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**

4 Experimental Results

We compared the proposed method (called evidential low-dimensional transformation with feature selection, i.e., ELT-FS) with several feature transformation methods, namely PCA, linear discriminant analysis (LDA), NCA and K-PCA; and several feature selection methods, namely T-test, Information Gain (IG), Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) [12]. We used two real patient data sets:

1) Lung Tumor Data: Twenty-five patients with stage II-III non-small cell lung cancer treated with curative intent chemo-radiotherapy were considered. Totally 52 SUV-based (SUV_{max}, SUV_{mean}, SUV_{peak}, MTV and TLG) and texture-based (gray level size zone matrices (GLSZM) [16]) features were extracted from longitudinal PET images before and during the treatment. The extraction of GLSZM-based features consists of two main steps: firstly, homogenous areas were identified within the tumor, and then a matrix linking the size of each of these homogeneous areas to its intensity was constructed; after that, features characterizing regional heterogeneity were then calculated from this matrix, such as parameters that quantify the presence of a high-intensity large-area

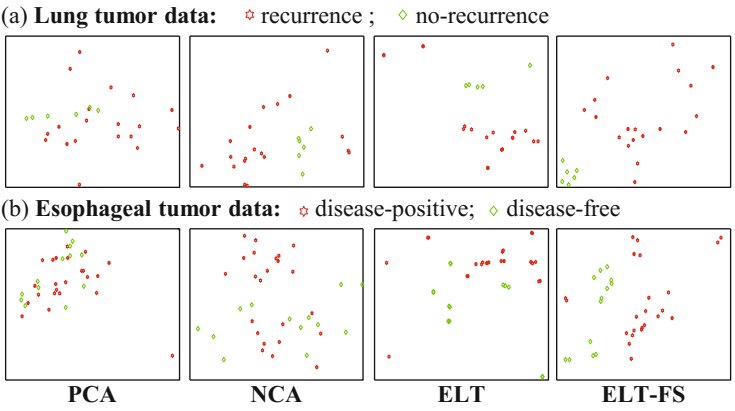


Fig. 2. Two-dimensional transformation results of PCA, NCA, our ELT (without feature selection, i.e., $\lambda = 0$) and ELT-FS.

emphasis or a low-intensity small-area emphasis. The definition of recurrence for patients at one year after the treatment was primarily clinical with biopsy and FDG-PET/CT. Local or distant *recurrence* was diagnosed on nineteen patients, while *no recurrence* was reported on the remaining six patients (example images can be seen in Fig. 1(a)).

2) Esophageal Tumor Data: Thirty-six patients with esophageal squamous cell carcinomas treated with chemo-radiotherapy were studied. Totally 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) were extracted based on the baseline PET images. The disease-free evaluations include a clinical examination with PET/CT and biopsies. Finally, thirteen patients were labeled *disease-free* when neither loco regional nor distant tumor recurrence is detected, while the remaining twenty-three patients were diagnosed as *disease-positive* (as shown in Fig. 1(b)).

The leave-one-out cross-validation (LOOCV) procedure was used for evaluation. For been compared methods (except NCA, since it was designed specifically for the K-NN classifers), after learning a low-dimensional subspace, the SVM (Gaussian kernel with $\sigma = 1$ was empirically chosen) classifier was used to predict class labels of both training instances and the left testing instance; while the EK-NN classifier (K was empirically set as 3) was used with NCA and the proposed method. Hyper-parameter λ for ELT-FS was determined by a rough grid search strategy. The dimension of output subspace was chosen between two to five according to the minimum average testing error. Finally, the average training and testing accuracy for all methods are summarized in Table 1, in which results obtained by the SVM and EK-NN in the input space, and by our method without feature selection (namely with $\lambda = 0$) are also presented for comparison. It can be observed that the proposed method, especially ELT-FS, leads to higher testing performance than other methods. Although LDA results

in larger training accuracy than our method, the worst testing performance is obtained too. It maybe because the studied data sets were too small, therefore the covariance matrix obtained by LDA has been badly scaled. It is also worth noting that ELT and ELT-FS have the same testing performance on the lung tumor data, while ELT-FS performs much better on the esophageal tumor data than ELT. This result maybe can be explained from two different aspects: firstly, the lung tumor data is easier to be separated than the esophageal tumor data, hence the difference became small; on the other hand, it perhaps also demonstrates that the sparse term can play a real role to improve the prediction under complex situation, such as on the esophageal tumor data.

Furthermore, we visualized the dimension reduction in 2D achieved using 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 methods 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 using PET image features and clinical characteristics. A specific loss function has been designed to tackle uncertainty and imprecision, so as to learn an adaptive dissimilarity metric for the EK-NN classifier. Through minimizing this loss function to obtain a low-rank transformation matrix A, we have realized a low-dimensional linear transformation of input features. Simultaneously, thanks to the $\ell_{2,1}$-norm regularization of A, a feature selection procedure has been implemented to reduce the influence of imprecise input features during prediction. Experimental results obtained on two clinical data sets show that the proposed method performs well as compared to some other methods. 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 regularization hyper-parameter λ. Moreover, to further improve the prediction accuracy, we will attempt to include more features that extracted from other complementary sources of information, such as CT, FLT-PET, FMISO-PET images, etc.

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

References

1. Brooks, F.J., Grigsby, P.W.: The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. Journal of Nuclear Medicine 55(1), 37–42 (2014)
2. Chandana, S., Leung, H., Trpkov, K.: Staging of prostate cancer using automatic feature selection, sampling and Dempster-Shafer fusion. Cancer Informatics 7, 57 (2009)

3. Dennis, Jr., J.E., Schnabel, R.B.: Numerical methods for unconstrained optimization and nonlinear equations, vol. 16. SIAM (1996)
4. Denœux, T.: A K-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25(5), 804–813 (1995)
5. Denœux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Transactions on Knowledge and Data Engineering 25(1), 119–130 (2013)
6. Erasmus, J.J., McAdams, H., Patz, J. E.F., Goodman, P.C., Coleman, R.E.: Thoracic FDG PET: state of the art. Radiographics 18(1), 5–20 (1998)
7. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems, pp. 513–520 (2005)
8. Lambin, P., van Stiphout, R.G., Starmans, M.H., Rios-Velazquez, E., Nalbantov, G., Aerts, H.J., Roelofs, E., van Elmpt, W., Boutros, P.C., Pierluigi, Granone, O.: Predicting outcomes in radiation oncology—multifactorial decision support systems. Nature Reviews Clinical Oncology 10(1), 27–40 (2013)
9. Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I.: Fusion of multi-tracer PET images for dose painting. Medical Image Analysis 18(7), 1247–1259 (2014)
10. Lian, C., Ruan, S., Denœux, T.: An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognition 48(7), 2318–2327 (2015)
11. Mi, H., Petitjean, C., Dubray, B., Vera, P., Ruan, S.: Prediction of lung tumor evolution during radiotherapy in individual patients with PET. IEEE Transactions on Medical Imaging 33(4), 995–1003 (2014)
12. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15(11), 1119–1125 (1994)
13. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 583–588. Springer, Heidelberg (1997)
14. Shafer, G.: A mathematical theory of evidence, vol. 1. Princeton University Press, Princeton (1976)
15. Tan, S., Kligerman, S., Chen, W., Lu, M., Kim, G., Feigenberg, S., D'Souza, W.D., Suntharalingam, M., Lu, W.: 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)
16. Tixier, F., Hatt, M., Le Rest, C.C., Pogam, A.L., Corcos, L., Visvikis, D.: Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. Journal of Nuclear Medicine 53(5), 693–700 (2012)
17. Zhang, N., Ruan, S., Lebonvallet, S., Liao, Q., Zhu, Y.: Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. Computer Vision and Image Understanding 115(2), 256–269 (2011)