Feature extraction and classification for rectal bleeding in prostate cancer radiotherapy: A PCA based method

Auréline Fargeas[a,b], Amar Kachenoura[a,b], Oscar Acosta[a,b], Laurent Albera[a,b], Gaël Dréan[a,b], Renaud De Crevoisier[a,b,c]

[a] *INSERM, U1099, Rennes, F-35000, France.*
[b] *Université de Rennes 1, LTSI, Rennes, F-35000, France.*
[c] *Département de Radiotherapie, Centre Eugène Marquis, Rennes, F-35000, France.*

1. Introduction

Rectal bleeding is one of the most important sequelae after prostate cancer radiotherapy and impacts the patients quality of life [1]. Understanding local dose/toxicity relationships is crucial to correlate the treatment outcome with the planning parameters. Several studies have found significant correlations between the parameters derived from rectal Dose-Volume Histograms (DVHs) and the incidence of bleeding such as NTCP models [2]. NCTP models use data in the native space and do not need registration step. More precisely, these methods consist of a reduction of 3D dose distributions to DVHs at the expense of the spatial information about location. Nevertheless, it is important to note some crucial limitations:

- a same DVH may correspond to different dose distributions;
- these methods require the estimation of population specific parameters optimized over a larger database.

In a previous work, we introduced a novel method based on Principal Component Analysis (PCA) [3] to predict rectal bleeding following high-dose prostate cancer radiotherapy. This method exploits the Three-dimensional planned Dose Distribution (3D-pDD) without any parameterization. An inherent problem of outcome modeling is that the analysis with a large number of variables is computationally expensive. Many features may be extracted from data to provide new representations of the populations anatomy. The main goal of features selection is to find an optimal subset from a full set of features which provide relevant information to match or improve the performances of classifiers. In this study, based on the method proposed in [3], we compare the existing approach (namely sequential) with a novel approach (namely combinatory) to select relevant features. In [3], the sequential approach was applied. This selection of features provides higher statistical performances for classifying the rectal bleeding patients. PCA was applied to 3D-pDD of patients treated for prostate cancer by intensity-modulated radiotherapy (IMRT). This allowed for a comparison of the two approaches to select features with respect to their power to make predictions.

2. Materials and methods

2.1. Data and registration

A total of 63 prostate cancer patients who received a total dose of 80 Gy in the prostate by IMRT were included in the study. For each patient, only the CT acquisition, volume delineations and the 3D-pDD were acquired. The delivery was guided by means of an Image-Guided Radiation Therapy (IGRT) protocol, with cone beam CT images or two orthogonal images (kV or MV

imaging devices), using gold fiducial markers in 57% of patients. Eleven of them present rectal bleeding ($\geq$ grade 1) at 2 years. Only the 3D-pDD within the rectum was analyzed. Patients planning CT and dose distributions were elastically registered on a single coordinate system by combining the CTs and organs delineations with the demons algorithm [4], as explained in [5].

2.2. Principal component analysis

Feature extraction methods determine an appropriate subspace of dimensionality R in the original feature space of dimensionality P ($R \leq P$). Linear transforms have been widely used for feature extraction and dimensionality reduction. In this communication, PCA and ICA were used. The purpose of PCA is to derive a relatively small number of decorrelated linear combinations (principal components) of a set of random variables while retaining as much of the information from the original variables as possible. Typically, the PCA for the random vector $\mathbf{x}^m = [x_1^m, \cdots, x_N^m]^T$ consists in looking for an overdetermined ($N \times P$) (i.e. $P \leq N$) orthonormal linear transform $\mathbf{W}$ such that the P components of the vector $\mathbf{z}^m = [z_1^m, \cdots, z_P^m]^T = \mathbf{W}^T\mathbf{x}^m$ are mutually uncorrelated. Thus, if we denote $\mathbf{W} = [\mathbf{e}^1, \cdots, \mathbf{e}^f]^T$ as the eigenvectors of $\mathbf{R}_\mathbf{x}$ (the covariance matrix of $\mathbf{x}_m$), corresponding to the eigenvalues ($\lambda_1, \cdots, \lambda_P$) where $\lambda_1 \geq \cdots \geq \lambda_P$, the first principal component of $\mathbf{x}_m$ is $\mathbf{z}_1^m = \mathbf{e}_1^T\mathbf{x}^m$. Likewise, the P-th principal component is obtained as $\mathbf{z}_P^m = \mathbf{e}_P^T\mathbf{x}^m$. In our case, each individuals 3D-pDD can be represented as a 1D vector by vertically concatenating the transpose rows of all the slices. Thus, we obtained a ($N \times M$) matrix $\mathbf{X}$ for all the individuals, where $M = 63$ represents the number of individuals (the variables) and N the number of voxels (the number of observations carried out). Only the voxels lying within

the rectum were considered ($N = 12,726$). After centering the data, the eigen-decomposition of the covariance matrix $\mathbf{Rx} = \mathbf{XX^T}$ was performed. However, because of the huge size of $\mathbf{Rx}$, its diagonalization is computationally expensive. A usual procedure is to perform the eigen-decomposition of a $M \times M$ covariance matrix $\mathbf{Cx} = \mathbf{X}^T\mathbf{X}$, yielding new P eigenvectors $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_P]$. This leads to an indirect way of computing the original eigenmatrix $\mathbf{W}$ as $\mathbf{W} = \mathbf{XY}$. It has to be noticed that because the rank of $\mathbf{Rx}$ can not exceed M, the maximum number P of meaningful eigenvectors is less or equal to M.

2.3. Classification procedure

The whole procedure is divided in two steps, namely the training step and the classification step.

2.3.1. Training step: features selection

In this section, we aimed at finding the more discriminant features (eigenvectors). That is to say, the one that better classify the population into rectal and non-rectal bleeding groups. We used the Receiver Operating Characteristic (ROC) curve. The ROC curve is the graph that displays the full picture of trade-off between the true positive rate (Sensitivity, Se) and false positive rate (1-Specificity, 1-Sp) across a series of cut-off points as shown in Fig. 1. For each feature, we obtained the optimal cut-off point (named optimal threshold) that best discriminates the two groups. To do so, we maximized the vertical distance from line of equality (Random Guess Line, RGL, Fig. 1) to each point of the curve. After computing the optimal threshold associated to each feature, we performed the top ranked axis from the axis

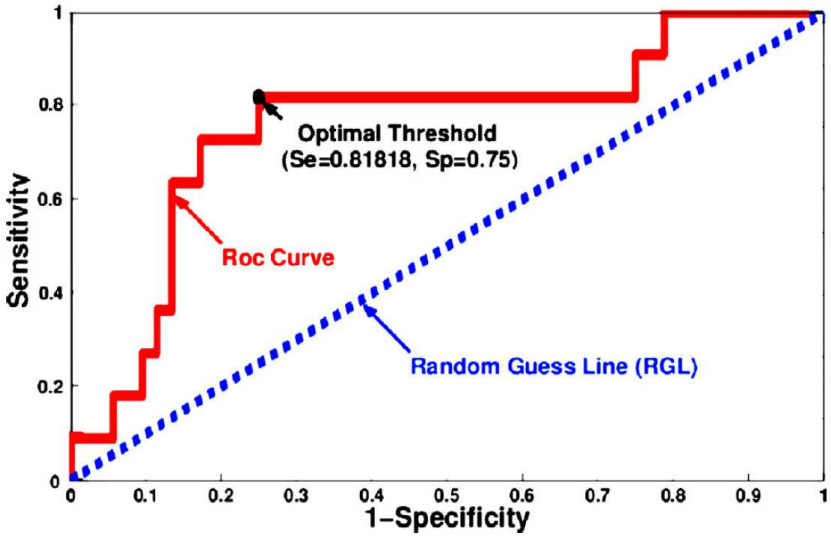

Figure 1: ROC curve and corresponding optimal threshold when using the more relevant principal axis (34th).

associated with the maximal optimal threshold to the minimal one. If the result obtained when using the best feature is not significant, another solution is the exploitation of more than one feature (combinations of features) to discriminate between patients with and without rectal bleeding. We use two different approaches (sequential and combinatory) to choose the n-optimum features set. For the sake of clarity, we explain the two approaches in the case where n = 2 (using the two best features as depicted in Fig. 2).

Sequential approach. The first feature is chosen as the most discriminant axis (the one with the highest vertical distance to the RGL). Then the Sequential Forward Selection (SFS) [6] strategy is used to select the second feature that provides the 2-optimum feature set. In other words, the threshold of the second feature is equal to the optimal threshold obtained from its ROC

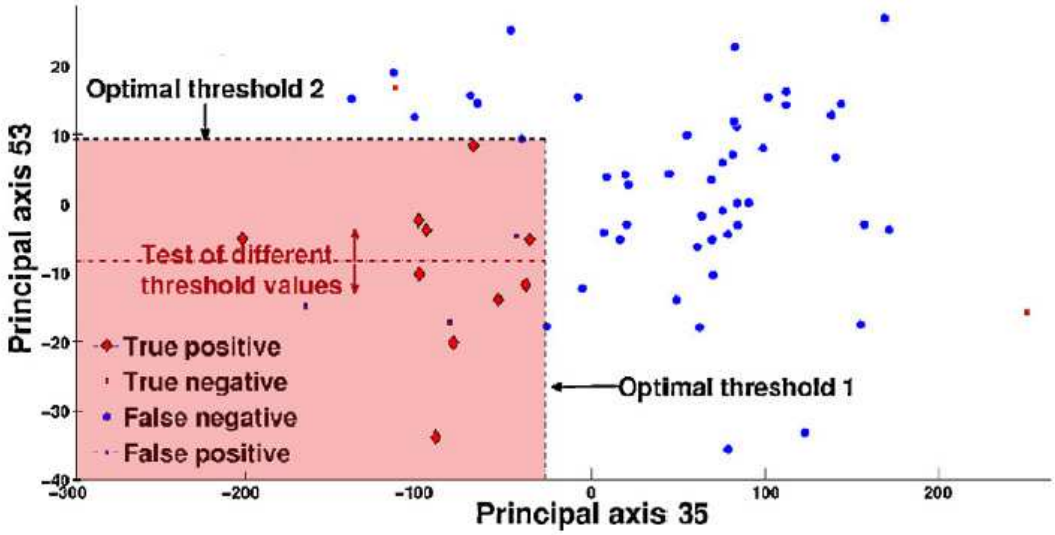

Figure 2: Projection of patients when using the first two more relevant principals axes.

curve (Fig. 2).

Combinatory approach. In this case, the first feature is chosen as for sequential approach. However, to set the threshold of the second feature, we test all the values of its cut-off points (red dotted line in Fig. 2). Then, the second optimal threshold is equal to the cut-off point that gives the best statistical performance. It is interesting to note that, contrary to the sequential approach, when we use a combination of features ($n \geq 1$) the performance never decreases. Indeed, if the second feature does not improve the performance in terms of Se and Sp, we just take into account the results previously obtained using the $n-1$ best features.

2.3.2. Classification step

In order to classify a new patient, we project its 3D-pDD on the subspace spanned by the n vectors base representing the n-optimum features set. Eventually, the patient is classified as having rectal bleeding if its 3D-pDD projection (n dimensional vector) is include in the toxic area (red area, Fig. 2).

3. Results and discussions

Because of the reduced number of patients (especially the number of patients suffering from rectal bleeding, 11 patients), a leave-one-out cross validation is performed to estimate the accuracy of our predictive model. At each round of the cross-validation, one patient is randomly extracted and used as test sample for validating the analysis. Then the proposed procedures are applied to the 62 remaining patients. The performance of the proposed methods is evaluated by computing the sensitivity (Se) and the specificity (Sp). The Se represents the percentage of patients with rectal bleeding who are correctly identified as having rectal bleeding effects. The Sp defines the percentage of patients without rectal bleeding who are correctly identified. Fig. 3 displays the performance of the classifier as a function of the number, n, of exploited features when using (a) the sequential approach and (b) the combinatory approach. When using just the best feature (namely the $34th$), 82% of patients suffering from rectal bleeding and 75% of non-rectal bleeding patients have been well classified ($Se = 0.82$ and $Sp = 0.75$), which means that only two patient with rectal bleeding and 13 non-rectal bleeding patients were badly classified. In the case where we exploited more than one feature,

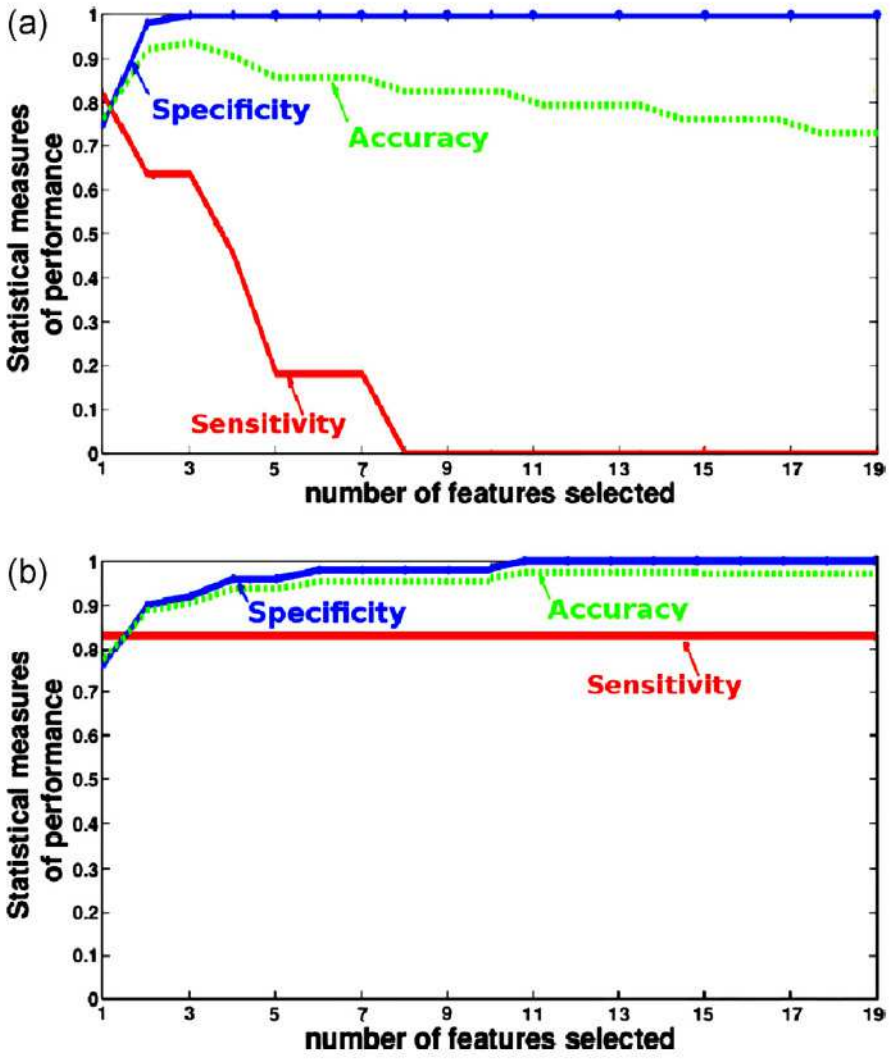

Figure 3: Accuracy, sensitivity and specificity as function of the number of exploited features using: (a) sequential approach, and (b) combinatory approach.

Fig. 3 (a) shows that the sequential approach is more efficient when using only the best feature ($34th$). Indeed, the performance dramatically decreases in terms of sensitivity when using more than one feature. Regarding the

combinatory approach (Fig. 3 (b)), the best results ($Se = 0.82$ and $Sp = 1$) were obtained when more than 15 features were exploited. It means that only two patients were badly classified (and 61 patients were well classified). We note that, with this approach, the statistical measures increase when more than one feature were used. We also remarked that, in this case, the performance of our classifier is stable when more than 15 features were used. This may suggest that there are redundant or irrelevant information in the remained features. The low number of patients (63 patients) did not allow to estimate the population specific parameters optimized over a larger database of well-established pre-diction methods such as NTCP models. Future work includes comparison with these methods on a larger database.

4. Conclusion

In this work, we applied the proposed PCA based methods to classify prostate cancer patients suffering from rectal bleeding. Although the experiments were performed on a reduced data set, the two procedures used to select the relevant features (sequential and combinatory) provide high accuracy and suggest that our methodology is promising to study the local dose/toxicity relationships. We also show that the approach used to select the best combination of principal axes is very important. Application on large database will be performed in future to more evaluate the robustness of the proposed methods.

Acknowledgments

This work was supported by Region Bretagne and has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the Investing for the Future program under reference ANR-10-LABX-07-01.

[1] A. Jackson, Partial irradiation of the rectum, in: Seminars in radiation oncology, Vol. 11, Elsevier, 2001, pp. 215–223.

[2] T. Rancati, C. Fiorino, G. Gagliardi, G. Cattaneo, G. Sanguineti, V. C. Borca, C. Cozzarini, G. Fellin, F. Foppiano, G. Girelli, et al., Fitting late rectal bleeding data using different NTCP models: results from an italian multi-centric study (AIROPROS0101), Radiotherapy and oncology 73 (1) (2004) 21–32.

[3] B. Chen, O. Acosta, A. Kachenoura, J. D. Ospina, G. Drean, A. Simon, J.-J. Bellanger, P. Haigron, R. De Crevoisier, Spatial characterization and classification of rectal bleeding in prostate cancer radiotherapy with a voxel-based principal components analysis model for 3D dose distribution, in: Prostate Cancer Imaging. Image Analysis and Image-Guided Interventions, Springer, 2011, pp. 60–69.

[4] J.-P. Thirion, Image matching as a diffusion process: an analogy with maxwell's demons, Medical image analysis 2 (3) (1998) 243–260.

[5] O. Acosta, G. Drean, J. D. Ospina, A. Simon, P. Haigron, C. Lafond, R. De Crevoisier, Voxel-based population analysis for correlating lo-

cal dose and rectal toxicity in prostate cancer radiotherapy, Physics in medicine and biology 58 (8) (2013) 2581.

[6] D. P. Muni, N. R. Pal, J. Das, Genetic programming for simultaneous feature selection and classifier design, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36 (1) (2006) 106–117.