ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-253

Comparison of Feature Selection Methods for Predicting RT-Induced Toxicity

Francisco J. NÚÑEZ-BENJUMEA[a], Jesús MORENO-CONDE[a], Sara GONZÁLEZ-GARCÍA[a], Alberto MORENO-CONDE[a], José L. LÓPEZ-GUERRA[b], María J. ORTIZ-GORDILLO[b], Carlos L. PARRA-CALDERÓN[a]

[a] *Biomedical Informatics, Biomedical Engineering and Health Economics, Institute of Biomedicine of Seville, IBIS / Virgen del Rocío University Hospital / CSIC / University of Seville. Seville, Spain.*

[b] *Radiotherapy Oncology Unit, Virgen del Rocío University Hospital, Seville, Spain.*

Abstract. This work addresses a scoping review of Feature Selection (FS) methods applied to a Lung Cancer dataset to elucidate parameters' relevance when predicting radiotherapy (RT) induced toxicity. Subsetting-based and Ranking-based FS methods were implemented along with 4 advanced classifiers to predict the onset of RT-induced acute esophagitis, cough, pneumonitis and dyspnea. Their prediction performance was measured in terms of the AUC for each model to find the best FS.

Keywords Feature Selection, Lung Cancer, Toxicity, Precision Medicine.

1. Introduction

Lung cancer (LC) causes more than 1.1 million deaths a year worldwide, making it the leading cause of cancer death in men and the second leading cause of cancer death in women [1].

A key challenge in RT is to maximize radiation doses to cancer cells while minimizing damage to surrounding healthy tissue. As severe toxicity in a minority of patients limits the doses that can be safely given to the majority, there is interest in developing a test to measure an individual's radiosensitivity before treatment. Variation in sensitivity to radiation depends on multiple factors and recent progress in data mining raises the possibility of customized analysis to characterize individual profiles that predict patient response to RT [2].

Taking into account the huge amount of medical variables that may affect to either acute or chronic toxicity onset [3], it turns out that, to provide real advances in routine care, is essential to carry out a strategy that allows reducing the dimensionality of this set of variables without diminishing its predictive capacity.

2. Materials and methods

2.1. Lung cancer dataset

The LC dataset was developed in the frame of the S31 project (Carlos III Institute of Health of Spain, grant number PI13/01155) and it includes clinical information gathered during routine care. Since the objective of this work was to predict the RT-induced acute toxicity before treatment, only the items included in the First consultation report subset were considered as potential predictors.

Within the LC dataset, 425 patients eventually received RT during their treatment plan. These subjects reported 12 different subtypes of RT-induced acute toxicity, however our work only focused on the prediction of the Esophagitis, Cough, Dyspnea and Pneumonitis in order to provide a balanced number of samples to the FS methods.

2.2. Feature selection methods and classifiers

The following FS methods were used:

- Ranking methods: Minimun Redundancy Maximun Relevance (mRMR), Relief (Rel), Random Forest (RF), Information Gain (IG)
- Subsetting methods: Correlated-based Feature Selection (CFS), Boruta (Bor), Chi-squared filtering (ChiSq)

Furthermore, a polling-based FS method has also been implemented for both ranking and subsetting methods.

The resulting features yielded by each one of the FS methods implemented have been used to train the following classifiers: Support Vector Machine (SVM), Artificial Neural Network (ANN), Linear Regression (LR) and Naïve Bayes (NB).

3. Results and discussion

On average, the best subsetting-based FS method was found to be the CFS (average AUC=0.644), while the best ranking-based FS method was found to be the RF (average AUC=0.642). On the other hand, the results shown that the LR classifier outperforms the others investigated (average AUC=0.679). These results endorse the possibility to implement in a real environment a system for predicting the risk of radiation-induced toxicity aligned with the Precision Medicine and Learning Healthcare System paradigms.

References

[1] International Agency for Research on Cancer. *World Cancer Report 2014*. WHO Press, World Health Organization, Geneve, Switzerland, 2014.

[2] Y. Luo, I. El Naqa, D.L. McShan, D. Ray, I. Lohse, M.M. Matuszak, et al, Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis, *Radiotherapy and Oncology* **123(1)** (2017), 85-92.

[3] I.S. Grills, D. Yan, A.A. Martinez, F.A. Vicini, J.W. Wong, L.L. Kestin, Potential for reduced toxicity and dose escalation in the treatment of inoperable non–small-cell lung cancer: A comparison of intensity-modulated radiation therapy (IMRT), 3D conformal radiation, and elective nodal irradiation, *International Journal of Radiation Oncology* **57(3)** (2003), 875-890.