applied sciences

MDPI

Article

A Machine Learning Tool to Predict The Response to Neoadjuvant Chemotherapy in Patients With Locally Advanced Cervical Cancer

Francesca Arezzo [1,*], Daniele La Forgia [2], Vincenzo Venerito [3], Marco Moschetta [4], Alberto Stefano Tagliafico [5,6], Claudio Lombardi [1], Vera Loizzi [7], Ettore Cicinelli [1], Gennaro Cormio [1]

[1] Department of Biomedical Sciences and Human Oncology, Obstetrics and Gynecology Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

[2] SSD Radiodiagnostica Senologica, IRCCS Istituto Tumori Giovanni Paolo II", via Orazio Flacco 65 – 70124 Bari, Italy

[3] Department of Emergency and Organ Transplantations – Rheumatology Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

[4] Department of Emergency and Organ Transplantation – Breast Care Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, Bari, 70124, Italy

[5] Department of Health Sciences (DISSAL) - Radiology Section, University of Genova, Via L.B. Alberti 2, Genoa, 16132, Italy

[6] IRCCS Ospedale Policlinico San Martino, 16132 Genova, Italy

[7] Interdisciplinar Department of Medicine, Obstetrics and Gynecology Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

* Correspondence: francescaarezzo@libero.it; Tel. +393274961788

Citation: Arezzo, F.; La Forgia, D.; Venerito, V.; Moschetta, M.; Tagliafico, A.S.; Lombardi, C.; Loizzi, V.; Cicinelli, E.; Cormio, G. A Machine Learning Tool to Predict the Response to Neoadjuvant Chemotherapy in Patients With Locally Advanced Cervical Cancer. *Appl. Sci.* **2021**, *11*, 823. https://doi.org/10.3390/app11020823

Received: 17 December 2020
Accepted: 20 January 2021
Published: 17 January 2021

SUPPLEMENTARY MATERIAL

Feature Selection

For classification with small training samples and high dimensionality, feature selection plays an important role in avoiding overfitting problems and improving classification performance. One of the commonly used feature selection methods for small samples problems is recursive feature elimination (RFE) method[1].

If the relationship between a feature and the output is suspected to be non-linear, tree-based methods (e.g., decision trees, random forest, and extreme gradient boosting (XGBoost)) can be applied to perform feature selection with low complexity. They can model non-linear relations well and do not require much tuning. XGBoost is a gradient boosted decision tree that is designed for speed and performance. In this method, new models are created that predict the errors of prior models to make the final prediction. This method runs several times faster than existing state-of-the-art methods on a single machine. It scales the models in distributed or memory limited settings, which is the main factor that leads to the success of XGBoost. The stopping criteria for XGBoost are as follows: First, for each output y_i, the R^2 of regression performance is measured with all features as input. The corresponding feature importance is also ranked in terms of information gain, which is the relative contribution of each feature to the full model. The second step is to iterate through the ranked list of input attributes and recursively eliminate attributes from the least important feature to the most important feature. After each elimination, the XGBoost model is built with the remaining features, and the corresponding R^2 score is returned. Finally, with the plot of feature number versus R^2 score, the elbow method is used to select the optimal number of features. Elbow method identifies the point at which adding more features does not improve the R score. Boosted tree algorithms such as XGBoost are capable of detecting non-linear effects or interactions. Therefore, these methods have the potential to fit into data with significantly fewer features [2].

Bayesian Ridge Conditional Imputation on Scikit.learn Iterative Imputer

A multivariate imputer estimates each feature from all the others. Deploying this method is considered a powerful strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion [3].

Scikit.learn Iterative Imputer uses Bayesian Ridge regression as default. There exist several strategies to perform Bayesian ridge regression. The Scikit.learn implementation is based on the algorithm described by Tipping [4] where updates of the regularization parameters are done as suggested by MacKay [5]. Iterative Imputer has proven to be the more accurate method of imputation for obstetrics and ginechology datasets [6].

Hyperparameters fine-tuning

A model hyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data.

Grid search is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyperparameter values specified.

In brief, for each algorithm, analysts define a grid consisting of several values to test for any hyperparameter of interest. The grid search performs a K-fold cross-validation splitting a dataset in K partitions and then searching for each split the combination leading to the best performance in a grid of hyperparameters[3].

The hyperparameter grids for each algorithm have been shown below:

Logistic Regression [3]:

- Solver: 'lbfgs', 'liblinear' // Selected: 'lbfgs'

Random Forest [3]:

- Number of trees in the forest: (100, 500 ,100) // Selected: 500

- Maximum depth of the tree: (3, 5, 10)// Selected: 5

K-nearest neighbours [3]:

- Number of neighbors: (5, 10, 15) // Selected: 5

References

1. Zeng X, Chen Y, Tao C, Alphen Dv. Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition. *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* 2009. p. 1205-8.
2. Kamel E, Sheikh S, Huang X. Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. *Energy*. 2020;206:118045.
3. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. 2013:arXiv:1309.0238. Accessed: September 01, 2013.
4. Tipping ME. Sparse bayesian learning and the relevance vector machine. *J Mach Learn Res*. 2001;1:211–44.
5. MacKay DJC. Bayesian Interpolation. In: Smith CR, Erickson GJ, Neudorfer PO, eds. *Maximum Entropy and Bayesian Methods: Seattle, 1991*. Dordrecht: Springer Netherlands; 1992. p. 39-66.
6. Altukhova O. Choice of method imputation missing values for obstetrics clinical data. *Procedia Computer Science*. 2020;176:976-84.