Feature Selection of Non-Small Cell Lung Cancer Nodules

Axel Masquelin, Andrea Elhajj, Samantha Connolly, Thayer Alshaabi

Through the use of data-characterization algorithms, the field of radiomics provides a wealth of mineable, quantitative data that is extracted from medical images. The data from these radiomic features unveil disease characteristics that cannot be seen by simple visual inspection of these medical scans. Feature selection algorithms may certainly be used to identify the features with the greatest relevance and predictive power to ultimately create a model for disease diagnosis and outcome prediction. This research investigates a data-driven feature selection method using symbolic genetic programming (GP) in tandem with support vector machines (SVM) and random forests (RF) to build a predictive model for non-small cell lung cancer diagnosis. We compare the effectiveness of using GP, SVM, and RF as stand-alone predictive models versus using a pipeline approach where GP is used as a feature selector followed by a SVM or RF to create a classifier for non-small cell lung cancer diagnosis from radiomics data. Our pipeline approach performed similarly well as SVM or RF alone, but fewer features were utilized in the pipeline approach. This suggests that selecting features using GP before creating a model using SVMs or RFs leads to more generalized results. Additionally, reducing the number features used by the classifier tends to decrease the likelihood of overfitting the given set of samples whereas building a predictive model with all available features may show slightly better performance by overfitting the given data.

Keywords: Feature Selection, Genetic Programming, Cell Lung Cancer Diagnosis, Radiomics, NLST Dataset.