Editorial

Probabilistic Universal Model Approximator (PUMA): A Novel Algorithm for Visualizing Classification Models

Editorial

Statement of the Problem: Analysis of data is the most the most challenging step in metabolomics experiments. In part, this is related to the enormous amount of data generated by metabolomics analytical methods. Chemometrics, especially principal components analysis (PCA) and orthogonal partial least squares discriminant analysis (OPLS-DA) have been the most commonly used methods for analyzing metabolomics data. Recently, the increase in complexity of metabolomics data sets further increased the reliance on more sophisticated supervised classification algorithms (e.g. support vector machine (SVM) and random forest (RF)) for analyzing metabolomics data. However, the lack of intuitive visualizations and the abstract nature of the output of these algorithms substantially impact their interpretabilit. Methodology & Theoretical Orientation: Here we propose a novel algorithm, the "probabilistic universal model approximator" (PUMA). Based on PCA probabilistic mapping, PUMA projects a 3D surface that joins the points with a 50% probability of class assignment and overlays this surface on a 3D PCA scores scatterplot to delineate how the model of interest defines the interface that separates between classes. Findings: Unlike the over-optimistic OPLS-DA plots, PUMA plots are based on PCA scores scatter plots hence they impartially capture most of the data set variance irrespective of its correlation to the Y matrix (classes). PUMA's modular design allows the examination of a myriad of classification models (such as SVM, RF, KNN, XGB, … etc.) and its interactive 3D output allows visually inspecting the performance of the model of interest (Fig. 1). Conclusion & Significance: Careful interpretation of the output of complex machine-learning classification algorithms is becoming a necessity, given the increasing reliance on such algorithms and the abstract output they produce. PUMA is an innovative visualization approach that intuitively aid in judgement of classification algorithms' performance and is envisaged to increase their transparency.