The application of artificial neural networks to metabolomics

Kevin M Mendez, Stacey N Reinke, David I Broadhurst

The Centre for Integrative Metabolomics and Computational Biology

Australia Edith Cowan University

Background:

Artificial Neural Networks (ANNs) were first introduced to metabolomics in the late 1990's, but rapidly fell out of favour, primarily due to "black box" lack of interpretability and the relatively high computational requirements compared to the popular alternative Partial Least Squares Regression (PLS), which has rapidly grown in popularity (Fig 1). Move forward 20 years, and ANNs, particularly in the form of 'deep learning', have had a resurgence of popularity in the general scientific community. The driving force behind this popularity is the increased availability of data, increased computer power, and a societal shift in acceptance of Artificial Intelligence.

The **aim** of this study is to compare the predictability of ANNs for binary classification against other popular machine learning methods across 11 open-access metabolomic datasets.

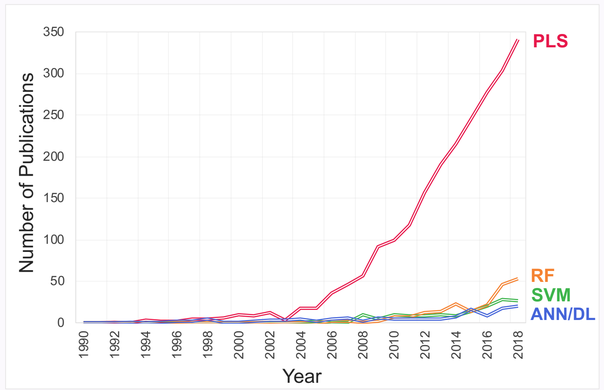

Figure 1: Number of publications per year (from Web of Science) with the key term metabolite*, metabolom* or metabonom* and the key term partial least squares or projection to latent structure (red), random forest (orange), support vector machine (green), or artificial neural network or deep learning (blue).

Methods:

Datasets:

We selected 11 previously published metabolomics datasets. Each had a binary outcome to enable simple performance comparison. Data was selected from a cross-section of popular analytical platforms: NMR, GC-MS & LC-MS with sample sizes ranging from n=58 to n=2139. All included data sets are available for downloaded from open-access data repositories.

Models:

1. PLS-Discriminatory Analysis (PLS-DA)
2. Principal Component Regression (PCR)
3. Principal Component Logistic Regression (PCLR)
4. Radial Basis Function Support Vector Machine (RBF-SVM)
5. Random Forest (RF)
6. ANN Linear Sigmoidal (ANN-LS)
7. ANN Sigmoidal Sigmoidal (ANN-SS)

All analysis was performed using the Python programming language, and published using the Jupyter web app.

Workflow:

- For each model, the hyperparameters optimisation used k-fold cross-validation
- Evaluation of each model was done using bootstrap resampling (n=100)
- The validation metric used is the out-of-bag Area under the Curve (AUC)
- Example of this workflow is shown in Fig 2 (using ANN-SS)
- Scan the associated QR code to re-run the workflow (in the cloud) using Binder

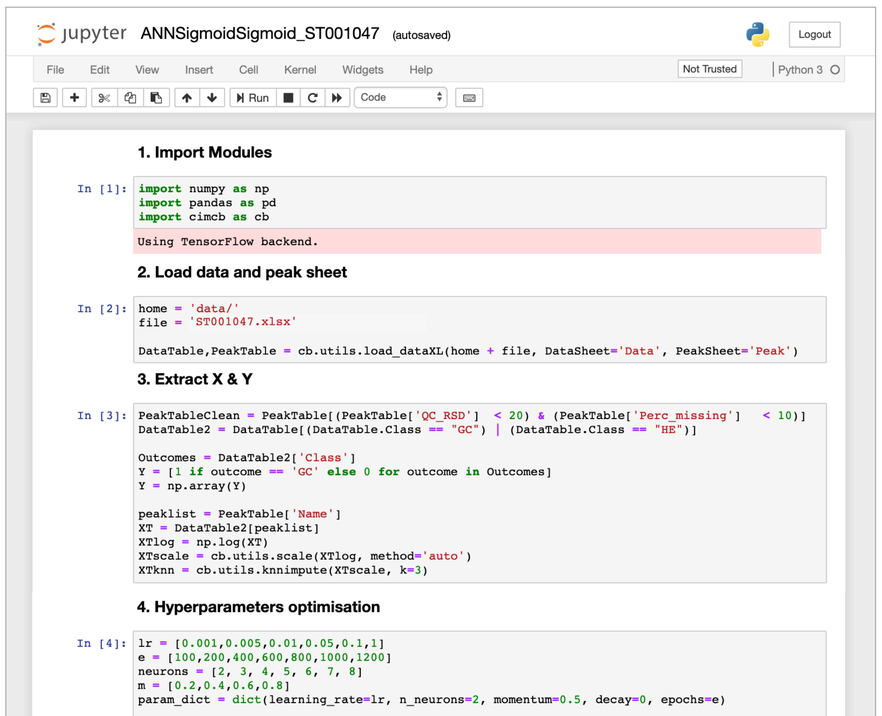

Figure 2: Example of the data analysis workflow with Jupyter Notebooks for dataset ST001047 with ANN-SS method (scan QR code to re-run the workflow using Binder).

ANN-SS PLS

Results:

The out-of-bag AUC with 95% confidence intervals (CI) for each combination of dataset and model is shown in Table 1. The results show that the two non-linear machine learning methods, ANN-SS & RBF-SVM, outperform all the linear methods. However, the bootstrap 95% confidence intervals indicate that there is no significant improvement over PLS-DA for any dataset. All methods perform well for large data sets. SVM and Random Forests were highly prone to severe overfitting in the training data (data not shown).

Conclusion:

The non-linear machine learning algorithms proved to be generally superior to the linear methods. However, the confidence intervals suggest that this improvement is not significant for these data sets. As such, the simplest model should be preferred. In this case, PLS-DA performed consistently well. The larger discussion point is that all the methods were prone to overfitting, but with the regression methods performing better. Only the very large data sets were observed to have similar in-bag and out-of-bag ROC curves. These results provide further evidence that the quality and size of the data is more important than the modelling method.

Table 1: Out-of-bag AUC using bootstrap resampling (n=100) for the following datasets and methods (scan QR codes for more detail).

Datasets	Platform	No. of Samples	No. of Peaks	PLS-DA	PCR	PCLR	RBF-SVM	RF	ANN-LS	ANN-SS
ST001047	NMR	140	149	0.93 (0.82, 0.99)	0.89 (0.73, 0.97)	0.79 (0.62, 0.93)	0.92 (0.84, 0.98)	0.89 (0.67, 0.99)	0.78 (0.45, 0.93)	0.93 (0.82, 0.98)
MTBLS90	LC	968	189	0.82 (0.80, 0.85)	0.81 (0.78, 0.85)	0.79 (0.76, 0.83)	0.84 (0.81, 0.88)	0.81 (0.76, 0.85)	0.82 (0.79, 0.85)	0.83 (0.80, 0.87)
MTBLS93	LC	2139	202	0.97 (0.96, 0.98)	0.96 (0.95, 0.97)	0.90 (0.88, 0.91)	0.97 (0.96, 0.98)	0.92 (0.89, 0.94)	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)
MTBLS92	LC	447	240	0.70 (0.64, 0.76)	0.68 (0.61, 0.74)	0.65 (0.60, 0.70)	0.72 (0.65, 0.78)	0.73 (0.67, 0.80)	0.70 (0.63, 0.78)	0.70 (0.62, 0.77)
MTBLS24	NMR	106	701	0.76 (0.61, 0.89)	0.76 (0.58, 0.87)	0.70 (0.57, 0.84)	0.83 (0.72, 0.94)	0.85 (0.73, 0.93)	0.80 (0.67, 0.89)	0.78 (0.63, 0.91)
ST000496	GC	100	69	0.96 (0.88, 1.00)	0.89 (0.70, 0.98)	0.89 (0.77, 0.97)	0.96 (0.86, 1.00)	0.79 (0.62, 0.94)	0.93 (0.78, 0.99)	0.96 (0.87, 1.00)
ST000369	GC	163	180	0.77 (0.63, 0.86)	0.70 (0.54, 0.80)	0.65 (0.55, 0.74)	0.82 (0.71, 0.90)	0.73 (0.59, 0.84)	0.72 (0.59, 0.81)	0.83 (0.73, 0.90)
MTBLS161U	NMR	58	30	0.72 (0.51, 0.91)	0.76 (0.55, 0.94)	0.73 (0.48, 0.89)	0.85 (0.66, 0.98)	0.70 (0.46, 0.84)	0.76 (0.56, 0.91)	0.80 (0.57, 0.93)
MTBLS161S	NMR	58	30	0.88 (0.70, 0.99)	0.75 (0.54, 0.90)	0.69 (0.54, 0.87)	0.85 (0.68, 0.97)	0.78 (0.60, 0.94)	0.77 (0.53, 0.94)	0.85 (0.64, 0.96)
MTBLS547	LC	118	42	0.93 (0.82, 0.99)	0.92 (0.80, 0.99)	0.85 (0.74, 0.94)	0.97 (0.89, 1.00)	0.92 (0.77, 0.98)	0.93 (0.80, 0.99)	0.98 (0.87, 1.00)
MTBLS404	LC	184	120	0.94 (0.88, 0.98)	0.91 (0.84, 0.96)	0.81 (0.69, 0.89)	0.95 (0.88, 0.99)	0.81 (0.68, 0.91)	0.88 (0.72, 0.95)	0.95 (0.88, 0.98)

Kevin M Mendez | MSc Research Student | Centre for Integrative Metabolomics & Computational Biology (www.cimcb.info) | School of Science | Edith Cowan University | Perth, Australia | email: k.mendez@ecu.edu.au | twitter: @KevinMMendez