How machine learning can improve your interpretation of SIMS data

Satoka Aoyagi[1]

(1) Faculty of Science and Technology, Seikei University, 3-3-1 Kichijoji-kitamachi, Musashino, Tokyo

Machine learning [1-4] can be powerful for extracting information from complex datasets. In order to apply machine learning effectively to our datasets, we first need to understand what they are. Unsupervised learning methods are helpful in providing an outline of a target dataset, after which supervised learning methods can be applied to extract particular information that is needed. The application of machine learning can also reveal how complex but rich information secondary ion mass spectrometry (SIMS) datasets have. Even extremely complex datasets can be interpreted more easily if they are classified into simpler categories. Moreover, if the classified ones are still difficult to understand, mass peaks can be identified by using supervised learning methods.

In terms of identifying organic materials, we now have more options, such as time-of-flight mass spectrometry, tandem MS (MSMS) and Orbitrap. MSMS spectra can be compared with conventional databases, and many precisely identified mass peaks can be matched to those in mass spectrometry databases. Therefore, the interaction between conventional databases and mass peaks in acquired SIMS spectra will be as important as it is to mass spectrometry fields. Understanding the fundamental mechanisms of machine learning is crucial for every SIMS researcher to effectively deal with complex datasets.

This tutorial will introduce the history of machine learning applications to SIMS datasets, the principles of basic machine learning methods, and several applications. Machine learning is generally classified into three types: unsupervised, supervised, and reinforcement learning. Unsupervised learning methods such as principal component analysis (PCA) and non-negative matrix factorisation (NMF) are commonly used to extract important features from a sample. Furthermore, artificial neural network (ANN) based methods such as autoencoders and self-organising maps are effective for analysing datasets containing nonlinear factors, such as matrix effects [1-3]. Correction of matrix effects is essential for quantitative analysis, as well as for obtaining appropriate mass images. ANN-based methods can be used effectively once the basic structures of ANNs have been understood. Therefore, the material that is useful for understanding perceptron and multilayer perceptron will be shared. Applications of supervised learning systems developed through the Versailles Project on Advanced Materials and Standards (VAMAS) will then be introduced, including quantitative and qualitative analyses of organic SIMS data using an ANN-based supervised learning method (VAMAS TWA2 A31) [3] and a material prediction system using a Random Forest-based supervised learning method developed by extending the peptide prediction system (VAMAS TWA2 A26) [4]. Furthermore, Python codes and Excel files for practical use will be shared.

[1] S. Aoyagi, Application of machine learning to spectrum and image data, J. Vac. Sci. Technol. A 41, 063202 (2023).
[2] S. Aoyagi, K. Matsuda, *Rapid Commun. Mass Spectrom.*, **37**(4), e9445 (2023).
[3] S. Aoyagi et al., *Anal. Chem.*, **95**(40), 15078-15085 (2023).
[4] S. Aoyagi et al., *Anal. Chem.*, **93**(9), 4191-4197 (2021).