Hybrid chemometric approaches to increase efficiency of classification and data fusion techniques

Yulia B. Monakhova\textsuperscript{a,b}, Monika Hohmann\textsuperscript{c,d}, Svetlana P. Mushtakova\textsuperscript{b}, Norbert Christoph\textsuperscript{d}, Helmut Wachter\textsuperscript{d}, Bernd Diehl\textsuperscript{b}, Ulrike Holzgrabe\textsuperscript{e}, Douglas N. Rutledge\textsuperscript{e}

\textsuperscript{b} Institute of Chemistry, Saratov State University, Saratov, Russia
\textsuperscript{a} Spectral Service AG, Cologne, Germany
\textsuperscript{c} Institute of Pharmacy and Food Chemistry, University of Würzburg, Würzburg, Germany
\textsuperscript{d} Bavarian Health and Food Safety Authority, Würzburg, Germany
\textsuperscript{e} UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, France

In chemometrics it is common to apply a cascade of multivariate approaches to the dataset under study. For example, PCA and cluster analysis are often applied to detect outliers and provide exploratory analysis of the data before the application of more advanced chemometric approaches such as discriminant analysis. Another promising but yet unexplored research direction is to increase the efficiency of classical chemometric methods by synergetic combination of their advantages within a new approach.

In this regard, independent components analysis (ICA) as a preprocessing tool was combined with linear discriminant analysis methods (LDA and FDA). To illustrate the performance of the ICA/DA methodology, four representative nuclear magnetic resonance (NMR) data sets of wine samples were used. The classification was performed regarding grape variety, year of vintage and geographical origin. The average increase for ICA/DA in comparison with PCA/DA in the percentage of correct classification varied between 6±1 and 8±2%.

Second, a new algorithm was proposed for the supervised analysis of multiblock data structures. It associates the advantages of the ability of common components and specific weights analysis (CCSWA) to explain the maximum of variance in the data and PLS-DA framework to shift results towards the targeted property. The robustness of the classification results comparing with each single methodology (PLS-DA and CCSWA) was shown on the example of the determination of tomato production type (organic/conventional) based on the fused $^1$H NMR, IR and stable isotope data ($^{13}$C/$^{12}$C, $^{15}$N/$^{14}$N, and $^{18}$O/$^{16}$O).

The use of hybrid chemometric approaches such as ICA/DA and PLSDA/CCSWA resulted in improved classification performance comparing with single approaches and increased efficiency of chemometric modelling.