Machine Learning and Chemometrics: A contradictive approach or a good complement?

F. Westad[1]

[1]CAMO Software AS, Oslo, Norway - fw@camo.com

One can hardly read the news these days without reading articles about how Big Data, Artificial Intelligence and Machine Learning are changing the world, the term disruptive technology is often used. How does this relate to the PAT community, product and process development and real-time process control? Obviously, computers can perform many tasks more efficiently than humans, e.g. searching through millions of records and looking for patterns. One frequently mentioned example is within medical diagnosis where scientific literature, patient journals and laboratory results are combined for diagnostic purposes where computers do it better than humans, who have difficulty in accessing and assessing all this information. Other examples are the ability to predict personality based on like-clicks on Facebook or to suggest travel destinations by analysing previous travel patterns and bookings. All these applications have thousands or millions of objects, which enables various algorithms to find patterns for classification purposes by use of deep learning. On the other side, the challenge within PAT is mostly that the number of *objects* is the limiting factor. How to we sample from our system to generate a good basis for the models with a minimum of effort? One example is prediction of the moisture content in lyophilisates directly through the glass vials with a required prediction uncertainty of < 0.2%. The models to be used within the PAT framework need to be applied on the *individual* object (sample) and a statistical significant correlation is not sufficient in itself: With enough objects, any pairwise correlation will be significant; however the important aspect is the *relevance* given the actual application.

Another important aspect is interpretation. The methods we use for exploratory data analysis, classification and prediction must be interpreted with the user's domain specific knowledge in mind. Therefore, human interaction in the modelling phase is mandatory. One reason is to prevent the inclusion of indirect correlation to make the model "better". These correlations may hold for the objects that were chosen to build the model but not for the future. Another reason is the need to set up the correct validation scheme for validating the model to account for future uncontrollable sources of variation such as raw material supplier, particle size, season, instrument etc. In most practical applications there will always be subgroups of objects in our data tables which invalidates the common procedure of splitting the objects 100 times randomly into calibration and test sets 60/40 [1]. When this is said, as long as the models can be interpreted and validated in the correct manner, there is no conflict between machine learning and chemometrics.

[1] F. Westad, F. Marini, *Analytica Chimica Acta*, **2015**, vol. 893, pp. 14-24.