Software Modules for Data Analytics

Data analytics is very dependent on the availability of algorithms and software. Three recently released software packages are described in more detail here as well as the location where they can be downloaded.

One of the common considerations when building models concerns selecting which data analytics and machine learning (DA/ML) method for the problem at hand.

Software has been released for the automated selection of data analytics and machine learning (DA/ML) methods based on the data characteristics and domain knowledge (https://github.com/vickysun5/SmartProcessAnalytics). Its algorithmic design avoids overfitting and other problems seen in numerous AutoML software packages, while also having the ability to learn interpretable nonlinear models from noisy and limited data. A tutorial on the software is available at

Weike Sun and Richard D. Braatz. Smart process analytics for predictive modeling. Comp. Chem. Eng., 143:107134, 2020, https://doi.org/10.1016/j.compchemeng.2020.107134.

One of the most effective machine learning approach for analyzing process data is feature engineering, where features are transformations of the raw data to improve the model quality before applying a subsequent regression. The latter step usually applies sparse regression, meaning that only the most valuable features are used in model building. Open-source software has been released (https://github.com/vickysun5/ALVENcode) that combines automated nonlinear feature generation and sparse regression to learn interpretable nonlinear models from noisy and limited data. The Algebraic Learning Via Elastic Net (ALVEN) algorithm generates features from families of nonlinear transformations that arise in physical, chemical, and biological systems. ALVEN balances model complexity and prediction accuracy through a two-step feature selection procedure, to produce interpretable nonlinear models while avoiding overfitting. The features incorporated into ALVEN enable the learning from data of true first-principles relationships, including the energy in the mass-spring oscillator ($E=mv^2/2+kx^2/2$), Einstein's formula for mass-energy equivalence ($E=mc^2$), and convective heat and mass transfer relations (e.g., $Nu = aRe^{1/2}Sc^{1/3}$). The software includes the generalization of ALVEN to nonlinear dynamical systems. The model accuracy of the algorithms is compared to well-established machine learning methods in case studies including for data from a 3D printer. The tutorial for the software is available at

https://doi.org/10.1016/j.compchemeng.2020.107103

An introduction was recently published on data analytics for new types of higher order tensorial information streams, which include real-time video, chemical imaging, and hyphenated (e.g., LC-MS) methods. Such data are increasingly available in chemical and biological manufacturing processes and contain valuable information about the process condition and product quality

(https://doi.org/10.1016/j.compchemeng.2020.107099). In this article, different types of higher order data in manufacturing processes are described, and their potential usage is addressed. Then some perspectives are provided on the application of tensorial data analytics to manufacturing processes. The most representative multilinear subspace learning methods are reviewed. Looking into the future, the potential and research needs for tensorial data analytics are briefly discussed.