Clustering methods for selecting representative samples in chemical databases

Felipe V. Calderan
Marcos G. Quiles
Juarez L. F. da Silva

Institute of Science and Technology, Federal University of São Paulo (UNIFESP), São José dos Campos, SP, Brazil.

Machine learning (ML) methods have demonstrated their role in Materials Sciences (MS) researches [1]. These ML methods, from unsupervised to supervised algorithms, have been applied to solve several tasks in MS, such as property prediction, design of new compounds, surrogate models in molecular dynamics simulations, among others [2]. However, besides the actual advances in the field, the use of ML models in MS is still in its infancy.

Aiming to contribute further with the field, we are investigating clustering algorithms [3] for selecting representative samples from a given dataset of molecules, and then providing visual insights to facilitate the analysis of the MS specialist.

Our preliminary results have shown that it's viable to cluster a dataset while having constraints specified by the specialist. This has been achieved through feature weighting [4] by optimization methods and may be very powerful, since it allows biased clusters with specific characteristics to be built. We are now looking forward to run more tests and eventually realease the application as a toolbox for MS.

Keywords: Data clustering, data visualization

References:

[1]. Butler, K. T. et al.; Machine learning for molecular and materials science. *Nature* **2018**, 559, 547-555.

[2]. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C.; Machine learning in materials informatics: recent applications and prospects. *NPJ Computational Materials* **2017**, 54, 1-13.

[3]. Jain, A. K.; Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **2010**, 31, 651-666.

[4]. Modha, D.S., Spangler, W.S. Feature Weighting in k-Means Clustering. *Machine Learning* **2003**, 52, 217–237.