Chemometrics and Intelligent Laboratory Systems 166 (2017) 1–6

Fuzzy clustering as rational partition method for QSAR

Alfonso Pérez-Garrido[ab], Francisco Girón-Rodríguez[c], Andrés Bueno-Crespo[b], Jesús Soto[b], Horacio Pérez-Sánchez[b], Aliuska Morales Helguera[d].

[a] Bioinformatics and High Performance Computing Research Group (BIO-HPC), Universidad Católica San Antonio, Guadalupe, Murcia, Spain
[b] Pharmacy Dpt., Universidad Católica de Murcia, Guadalupe, Murcia, Spain
[c] Food Technology and Nutrition Dpt., Universidad Católica de Murcia, Guadalupe, Murcia, Spain
[d] Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, Cuba

Abstract
Various methods are used to make the partition of data sets for QSAR development and model validation. In this work we used a fuzzy minimals partitioning and we compare this methodology with another rational partition methods like k-means clustering (KMS) and Minimal Test Set Dissimilarity (MTSD). For the development of QSAR models Ordinary Least Squares (OLS) and Extreme Learning Machine (ELM) methods were used. The generated QSAR equations were validated by the coefficient of determination of the internal leave one out (LOO) cross validation method Q^2_{LOO} and then the coefficient of the external test set Q^2_{ext} was compared between partition methods. The results of this comparison showed that using fuzzy minimal for big and structurally diverse data sets gave an applicability domain similar to KMS and a better predictability models than both methods, KMS and MTSD.

Keywords:
Data partition
Fuzzy clustering
Regression
Extreme Learning Machine
K-means clustering
Minimal Test Set Dissimilarity
QSAR
Validation