Consensus Feature Selection by Multi-Response Sparse SVR

Dechuan Zhuang[1], Curt M. Breneman[1], Kristin P. Bennett[2], Steven M. Cramer[3]
[1]Department of Chemistry and Chemical Biology, [2]Department of Mathematics, [3]Department of Chemical Engineering
Rensselaer Polytechnic Institute, Troy, NY

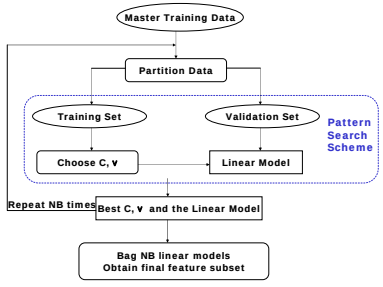
Introduction

Feature selection in Quantitative Structure Retention Relationship (QSRR) modeling can provide valuable insight into physical mechanism of ion-exchange displacement chromatography system.

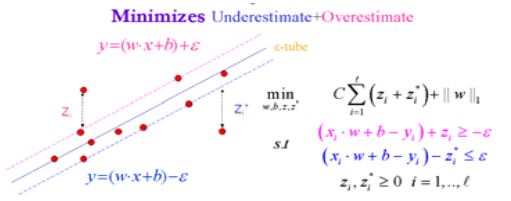
In this study, a new method, which is based on sparse support vector regression, has been proposed and implemented for consensus descriptor selection under multiple responses. By minimizing both the empirical risks of multiple models and the l_1-norm of a vector that holds upper bounds of the weights of features, all responses are considered simultaneously. With traditional MOE descriptors and electron-density based TAE descriptors, sparse support vector regression under multiple responses is employed to obtain a common set of descriptors that are important to all responses.

A visualization tool called "starplot" was used to show the details of the relative importance of selected descriptors. By observing the shapes of those star plots and looking into the physical meaning of the descriptors represented, we are able to explore more details of the underlying mechanism. Nonlinear SVR models were then carried out to evaluate the predictability of the models. A statistical technique known as "bagging" (Bootstrap Aggregation) was used to ensure robustness of the models.

Feature Selection



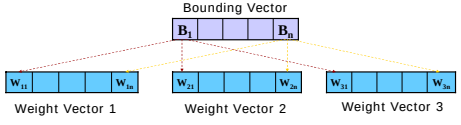
Sparse Support Vector Regression



- Generalization error can be bounded by minimizing empirical risk and model complexity
  - Empirical risk is calculated by ε-insensitive function
  - Regularization factor $|w|_1$ controls model complexity
  - Hyper-parameter C controls tradeoff between risk and complexity
- Original input space is mapped into a higher dimensional space by means of kernel functions
- Feature Selection through sparse SVR in the original input space
  - Drives the weights of irrelevant descriptors to zero by l_1-norm
- Bootstrap aggregation is employed to achieve more robust models

Multi-Response Sparse SVR

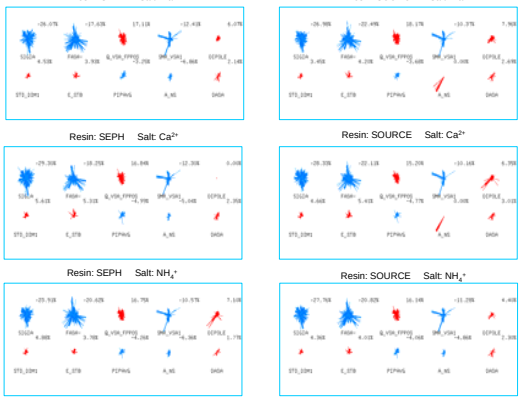
$$
\begin{aligned}
min.\quad & \sum_{k=1}^{m}\left(C_k v_k \varepsilon_k + b_k + \sum_{i=1}^{n} u_{ki} + \sum_{i=1}^{n} v_{ki} + \frac{C_k}{l}\sum_{i=1}^{l}\xi_i\right) + D\sum_{i=1}^{n} B_i \\
& y_{kj} - \sum_{i=1}^{n}(u_{ki} - v_{ki})x_{ji} - b_k \le \varepsilon_k + \xi_{kj} \\
& \sum_{i=1}^{n}(u_{ki} - v_{ki})x_{ji} + b_k - y_{kj} \le \varepsilon_k + \xi^*_{kj} \\
s.t.\quad & u_{ki} \le B_i \\
& v_{ki} \le B_i \\
& u_{ki}, v_{ki}, \xi_{kj}, \xi^*_{kj}, \varepsilon_k \ge 0, \quad j = 1,\ldots,l \quad i = 1,\ldots,n \quad k = 1,\ldots,m
\end{aligned}
$$



- Training multiple models together, linking them by a vector that controls the global model complexity
- Setting an upper bound for each component of all the weight vectors and then minimize the sum of the bounds
- Driving the weights of each component of the weight vectors to be in the same range, this approach is more likely to obtain a common set of selected features with different weights

Case Study

Case 1: Cation-Exchange Chromatography



Interpretation

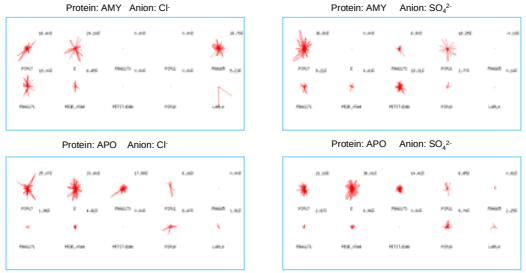
Descriptor	Sign	Definition
SIGIA	-	Average of the surface integral of kinetic electronic energy density
FASA-	-	Water accessible area of atoms with negative partial charges over that of all
Q_VSA_FFPOS	+	Fractional positive polar van der waals surface area
SMR_VSA1	-	Molecular refractivity weighted by subdivided surface area
DIPOLE	+	Dipole moment calculated from the partial charges of the molecule
STD_DIM1	+	Square root of the largest eigenvalue of the covariance matrix of the atomic coordinates
E_STB	+	Bond stretch-bend cross-term potential energy
PIPAVG	-	Average value of the Politzer Ionization Potential
A_NS	-	Number of sulfur atoms
DASA	+	Absolute difference between the positively and negatively charged water accessible areas

Classes of selected features

Polarizability	Electrostatic Effect	Shape & Charge Distribution
SIGIA	FASA-	DIPOLE
SMR_VSA1	Q_VSA_FFPOS	STD_DIM1
PIPAVG	DASA	E_STB
	A_NS	A_NS

- The more negatively charged the protein, the lower the retention in cation-exchange chromatography.
- Flatter proteins have greater retention. Uneven charge distribution on proteins increase the retention.

Case 2: Displacer Anion-Exchange Chromatography



Predictability