2O08

統計化学による任意の有機溶媒・溶質についての溶解度予測

○右田啓哉、倉田浩二郎、花村遼、荒川正幹、船津公人
東京大学大学院工学系研究科（〒113-8656 東京都文京区本郷 7-3-1）

1. 緒言

有機化合物の様々な溶媒に対する溶解度は、分子材料の設計やその評価のために最も重要な物性の一つである。例えば有機化合物の分離において、溶解度の違いを利用した分離操作の適切な設計を可能にし、また人体や環境中での化合物の動態を知る上でも溶解度が重要である。溶解度の値を得るためによく用いられる手法の一つに QSPR（Quantitative Structure-Property Relationships）がある。これは既知の物性データの統計的な解析などを通じて定量的なモデルを構築する手法の総称で、QSPR モデルで未知の物性値を迅速に評価できるため、候補のスクリーニングなどに適している。

しかし溶解度について汎用的かつ精度のよい QSPR モデルの構築は難しく、常套手段としては溶媒ごとの解析を行うか、化学構造の相互作用についての熱力学的な関係式を明示的に考える。この場合、サンプル数が少ない溶媒について推算ができないことや、複雑な化学構造に対する予測精度が悪くなるという欠点がある。本研究ではこれらを克服するために、SVR[1] (Support Vector Regression)に基づく方法論を提案し、２成分系の常温における溶解度データを用いて検討を行う。

2. モデリング手法

・SVR とサンプルの類似度による溶解度予測モデル

SVR は回帰分析の一種であり、カーネルトリックによる柔軟な非線形モデリングが特徴である。これはデータにおける複雑な数理的関係を、サンプル間の類似度（や内積）として暗黙的に表現することにより、取扱いの容易な問題へと変換することである。類似度を計算するための関数をカーネル関数といい、予測性の良いモデルを構築するために重要である。SVR のアルゴリズムの結果として得られる予測モデル f(x)は、訓練サンプル数 n、重み a、定数 b、２つのデータ x と x'のカーネル関数 K(x, x')によって次式で表わされる。

$$f(x)=\sum_{i=1}^{n} a_i K(x,x_i)+b$$

本研究では SVR を溶解度の予測に適用するために、2 成分からなる溶液の類似度を構成する。まず溶媒の類似度 $K_{solvent}$ と溶質の類似度 K_{solute} をガウスカーネルなどによって個別に計算し、その積 $K_{solvent} \times K_{solute}$ を溶液の類似度 $K_{solution}$ として定義する。$K_{solution}$ は SVR モデリングのアルゴリズムでそのまま扱うことができ、予測の際も同様である。こうすることで溶解度モデルの説明変数としては溶質と溶媒の説明変数のみを用いればよいことになり、化学構造の相互作用についての物理化学的な関係式を明示的には含まずとも、柔軟にモデルを構築することができる。

3. 溶解度データ

CrossFire Beilstein[2]から溶解度データを抽出・整理し、代表的な 18 種の有機溶媒と分子量 1000 以下の 2343 個の溶質を対象とした。溶解度データは 20~25 度の室温付近で測定されたものであり、単位は gL^{-1} で常用対数をとった。溶質の説明変数には DRAGON[3]による分子構造記述子 929 個を用い、溶媒に関しては極性、モル分子容量、誘電率、水素結合のアクセプター数とドナー数を用いた。

4. 結果・考察

・SVR による予測モデルの構築

前処理として 18 種の溶媒ごとに PLS モデリングと遺伝的アルゴリズムを適用し、線形モデルを構築しながら有用な記述子の選択を行った[4]。さらに 3 つ以上の溶媒データセットで選ばれた計 77 個の記述子を溶質の類似度の算出に用いた。 ε -SVR のパラメータを定めるためにクロスバリデーションによる最適化を行った結果、C=16、ε =0.10、ガウスカーネル $K(x,x') = \exp(-\gamma\|x-x'\|^2)$ のγは溶媒で $4.8*10^{-2}$、溶質で $1.5*10^{-2}$ となった。モデルの要因解析を行った結果では、溶質の説明変数としては親水性、構造の複雑さや各種の相互作用に関わる説明変数が、溶媒では極性とモル分子容量などが特に影響しており、定性的な理解とよく一致した。

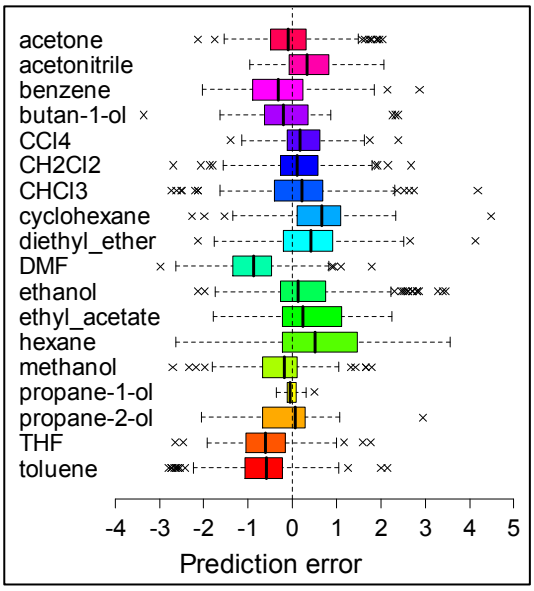

図 1. 溶媒を抜かした場合の予測値の誤差

・モデルの予測能力の評価

データに含まれない溶媒に対する予測能力を評価するために、全 18 種の溶媒のうち 17 種の溶媒に対する溶解度データからモデルを構築し、残りの一つの溶媒に対する溶解度を予測した。図 1 は溶媒ごとの全予測値の誤差についてのボックスプロットを描いたものである。溶媒ごとの平均絶対誤差は殆どが 0.6～0.8 で全体では 0.73 であり、精度よく予測できることがわかった。傾向としては溶質や溶媒の近傍における類似データの個数が予測精度と大きく関係していた。さらにモデルの全体的な予測精度を 5-fold クロスバリデーションによって評価したところ（図 2)、平均絶対誤差が 0.55 となり、十分な予測精度のモデルが構築できたといえる。

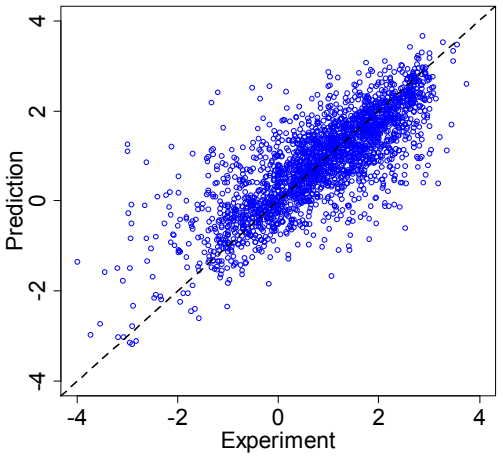

図 2. 5-fold CV の実測値-予測値プロット

5. 結言

有機化学構造の溶解度を対象として、任意の溶媒と溶質に対して適用可能な予測モデルを構築するための QSPR 手法を開発した。溶媒と溶質それぞれについて類似度を計算したのち、それらを合成し溶液の類似度を構成することで SVR によるモデリングを行った。構築されたモデルについて、データに含まれない溶媒に対する予測性やクロスバリデーションによる予測能力の評価を行い、本手法の有用性を確認した。細かい点で改良していく余地はあるものの、このようなモデリング手法は複数の要素が複雑に作用し合う事象に広く適用できると考えられ、他の対象への応用が今後大いに期待できる。

[参考文献]

[1] V. Vapnik, The Nature of Statistical Learning Theory, Springer, NY, (1995).

[2] http://www.crossfirebeilstein.com/

[3] http://www.talete.mi.it/help/dragon_help/DRAGONUserManual.html

[4] Kimura et al., Journal of Chemical Information and Computer Sciences, 38, 276-282 (1998).