ガウス過程の混合エキスパートモデルによる化学構造からの物性予測とその逆問題

総合研究大学院大学　池端 久貴　統計数理研究所　吉田 亮
(株)地球快適化インスティテュート　磯村 哲
北陸先端科学技術大学院大学　本郷 研太　北陸先端科学技術大学院大学　前園 涼

1. 背景

本研究では、統計的機械学習に基づく有機化合物の分子設計手法を提案する．化学構造からその特性の予測問題（フォワード予測）は古くから研究対象とされており，これまで多くの手法が提案されている．一方，特性から構造のバックワード予測，すなわち分子設計の問題に関しては、あまり多くの研究成果が報告されていない．化合物は，同じような特性を持つにもかかわらず，その化学構造が互いに大きく異なっていることがあり，バックワード予測は容易ではない．まずは予測精度の高いフォワード予測モデルを構築し，それを基にしたバックワード予測モデルの検討を行う．

2. ガウス過程の混合エキスパート

ガウス過程の混合エキスパート[1]は，局所的に特性が大きく変化する領域と小さく変化する領域が混在する非定常なデータに対しても，複数のエキスパートを用いてデータに当てはめることができる．化合物の特性データを観察すると，少なからず非定常性を持つことを確認できる．ガウス過程の混合エキスパートは次のように定式化される．

$$p(\boldsymbol{y}|\boldsymbol{x},\,\theta)=\sum_{c}p(\boldsymbol{y}\,|\boldsymbol{c},\,x,\,\theta)p(\boldsymbol{c}\,|x,\,\phi)=\sum_{c}[\prod_{j}p(\{y_j:c_i=j\}|\{x_i:c_i=j\},\,\theta_j)]p(\boldsymbol{c}|\boldsymbol{x},\,\phi)$$

ここで$\boldsymbol{c}=(c_1,\cdots,c_n)$はサンプルの各エキスパートへの割り当て，$p(\boldsymbol{y}\,|\boldsymbol{c},\,\boldsymbol{x},\,\theta)$は各ガウス過程のエキスパートによる予測分布$\boldsymbol{y}$を$\boldsymbol{x}$の順序に並び替えたもの，$p(\boldsymbol{c}|\boldsymbol{x},\,\phi)$は$\boldsymbol{x}$のエキスパートへの割り当て確率を表すGating networkである．また，θとϕはパラメータである．

3. 記述子と予測モデルの比較

化合物の構造から得られる説明変数として，もっとも広く使われているものは，フィンガープリントと呼ばれる記述子である．例えば，PubChemフィンガープリントを用いると，化合物は芳香環やある官能基の有無を表す881次元の二値ベクトルで表現される．

PubChemフィンガープリントを用いて，Lassoとk近傍法と予測性能を比較したところ，ガウス過程の混合エキスパートによる予測モデルが最もRMSEが低かった(図1)．なおデータはPubChemに格納されている化学構造(=*X*)とMMFF94 energy(=*y*)を用いた．ガウス過程に用いるカーネルはガウスカーネルとした．

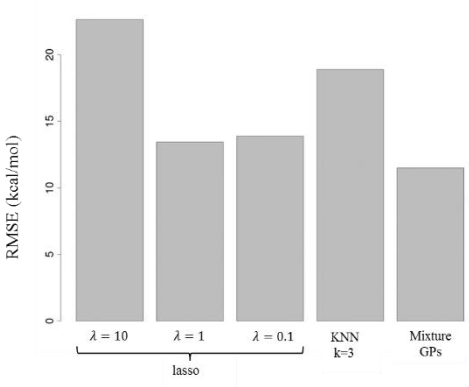

図１：予測性能比較の結果

4. 逆推定

構造から特性へのバックワード予測には，ベイズの定理から以下のように逆確率を得たものを用いる．

$$p(X|Y=y,\boldsymbol{\psi})\propto p(Y=y|X)p(X\,|\,\boldsymbol{\psi})$$

$p(Y=y|X)$はガウス過程の混合エキスパートによるフォワード予測モデルによる条件付確率である．また，$p(X\,|\,\boldsymbol{\psi})$は未知の物質を含めた化学構造の周辺出現確率を$\boldsymbol{\psi}$を用いてパラトリックモデルで表現したものである．化学構造の表現方法として，SMILESのような文字列を用いたものがいくつか提案されており，化学構造の学習に言語モデルを適用することができる．簡単な例として，次のようなn-gramモデルを用いたモデリングが考えられる．

$p(M\,|\,\boldsymbol{\psi})=\prod_{i=n+1}^{N}p(s_i\,|\,s_{i-n},\,\cdots,\,s_{i-1},\,\boldsymbol{\psi})$，where $M=\{s_1,\,\cdots,\,s_N\}$ is a SMILES string.

データベースに格納されている既知物質から$\boldsymbol{\psi}$の推定値$\widehat{\boldsymbol{\psi}}$を得て，$p(X\,|\,\widehat{\boldsymbol{\psi}})$を$X$の周辺分布とみなすことにより，未知の化合物を対象にバックワード予測が可能となる．

参考文献

[1] Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In Advances in Neural Information Processing Systems 14, pages 881–888. MIT Press, 2002.