ビッグデータを活用したばらつきの定量化およびモデリング

金子 弘 昌*

1. はじめに

医薬品設計・高機能高分子設計などの分子設計・材料設計では、目的の物性・活性をもつ化学構造や目的の特性を実現するための材料が得られる実験パラメータを効率的に設計することが求められる。そのために重要となるのが物性（もしくは活性）を推定する数値モデルである。分子設計においては、すでに物性が測定された化合物データベースを用いて、化学構造を数値化した変数（記述子）を X 、物性を y として、 X と y との間で物性推定モデル $y=f(X)$ が構築される。物性推定モデル f を用いることで、まだ物性が測定されていない化合物やコンピュータ上のみに存在する仮想的な化学構造に対して物性値を推定できる。材料設計においては、 X は実験パラメータとなる。既存の実験結果を用いることで、 X から y を推定する数値モデル $y=f(X)$ を構築することができる。モデル構築手法としては、線形の partial least squares (PLS)¹⁾ や非線形の support vector regression (SVR)²⁾ が有名である。その他の回帰分析手法については、こちらのウェブサイト^[1]を参照されたい。

物性推定モデルのメリットの一つに、多くのデータを用いることで、物性・活性を X で説明できるばらつきと X で説明できないばらつきとに分離できることが挙げられる。 X は化合物の構造的特徴や実験パラメータであるため、物性推定モデルにより説明できる物性・活性の情報は、構造的な特徴や実験パラメータに由来すると考えられる。一方で説明できない情報は、現状の構造的な特徴や実験パラメータと関連のない情報である。実験誤差や測定ノイズ等もここに含まれる。以下の式のように物性・活性のばらつきを2つに分離して考えることで、化学構造や実験条件により発現する化合物や材料の物性・活性を構造的な特徴や実験パラメータの観点から考察することが可能となる。

(物性・活性のばらつき) = ((1) 化学構造・実験パラメータに由来するばらつき) + ((2) 実験誤差・測定ノイズ等によるばらつき) [1]

物性推定モデルを構築するためのサンプルが少ないときには、物性・活性のばらつきを(1)のばらつきと(2)のばらつきに分離することは統計的に難しい。一般的にモデルのパラメータが(1)だけでなく(2)も表現するように最適化されてしまうためである。サンプル数が小さいときの対処方法については後述するが、大量のサンプル（ビッグデータ）を用いることで、(1)のばらつきと(2)のばらつきとの分離がしやすくなる。多くのサンプルの物性値・活性値を考慮してパラメータをデータに適合するため、安定したパラメータの値を計算できるためである。(1)のばらつきと(2)のばらつきとを分ける回帰分析について、こちらの URL^[1]に様々な種類の手法に関する詳細が記載されている。

2. ポリマー物性を対象にした物性推定モデル

ポリマー物性として屈折率 (refractive index, RI) およびガラス転移温度 (glass transition temperature, Tg) を対象にした、物性推定モデルの構築およびモデルを用いた新規モノマーの設計の例を示す。ディスプレイやレンズなど様々な機器で用いられている材料において、RI は重要な特性の一つである。高分子材料は一般には RI が低いため、高い RI を持つ高分子が望まれている。さらに高分子として安定している必要があるため、Tg もある程度高くなければならない。

PolyInfo^[2] から RI が測定されたポリマー 186 サンプル、Tg が測定されたポリマー 173 サンプルをダウンロードして使用した。 X として RDKit^[3] により計算されたモノマーの 196 個の構造記述子のうち、全モノマーの 90% 以上で同じ値をもつ記述子を削除して残る 95 個を用いた。

物性推定モデルの構築には SVR を用いた。RI, Tg それぞれすべてサンプルをモデル構築用データとして、double cross validation (DCV)³⁾ によりモデルの性能を検証した。DCV によりモデル構築に用いられていないサンプルに対する推定性能を検証できる。

実測値と DCV による推定値とのプロットを図 1 に示す。なお図 1 に対応する決定係数 r^2 は RI, Tg ともに 0.75 であった。 y の情報量のうち 75% を X によって表現できており、また図 1 より実測値と推定値が対角線付近に固まっていることから、良好な精度で RI, Tg を推定できるモデル構築を達成したことがわかる。

モデル構築に使用したモノマー構造から、Breaking of Retrosynthetically Interesting Chemical Substructure

Quantification and Modeling of Variance Using Big Data

* Kaneko, Hiromasa

明治大学 理工学部 応用化学科

川崎市多摩区東三田 1-1-1 (〒214-8571)

hkaneko@meiji.ac.jp

2018. 11. 19 受理

図1 各ポリマー物性の実測値と推定値とのプロット

図2 ポリマー物性の良好な推定値を示した化学構造の例 (RI: 1.66, Tg: 150.58°C)

(BRICS)⁴⁾によって新たな化学構造を1000個生成し、先ほど構築したSVRモデルを用いてRI, Tgの値を推定した。RI, Tgの推定値がともに高い値を示した化学構造の例を図2に示す。図2の構造に限らず、良好な物性を示した構造にはベンゼン環やニトロ基が存在していた。これらの官能基は高いRIかつ安定なポリマーを重合するには重要であると考えられる。

3. モデルの適用範囲 (Applicability Domain, AD)

物性推定モデルが構築されれば、どのようなサンプルでもモデルに入力すれば、物性の値を推定できる。しかし、ビッグデータを使用して物性推定モデルを構築したとしても、新しいサンプル（物性が未知のサンプル）の物性推定を行う際には注意しなければならない。物性推定モデルを適切に運用するためにはモデルの適用範囲 (applicability domain, AD)^{5), [4]}を設定する必要がある。ADとは、物性推定モデルが本来の推定性能を発揮できるデータ領域のことである。物性値を推定したいサンプルがAD内であればその推定値を信頼できるが、AD外であれば推定値は信頼できない。物性推定モデルを用いて物性値を推定する際には、入力するデータがADの中か外かを考えなければならない。

ADを設定する方法には大きく分けて、範囲⁶⁾、トレーニングデータの中心からの距離⁷⁾、データ密度⁷⁾、アンサンブル学習⁸⁾が存在する。範囲はXの各変数に上限・下限を設定する方法である。最大値・最小値や四分位点などで設定でき非常にシンプルという点はメリットであるが、変数間の関係を考慮できないなどのデメリットもある。ただし、主成分分析 (principal component analysis, PCA) などの低次元化手法を行った後に範囲を設定することで、変数間の関係も考慮できる。トレーニングデータの中心からの

距離をADの指標とすることで、距離が小さいときには推定値を信頼でき、距離が大きときには推定値を信頼できない、といった検討が可能になる。ユークリッド距離の代わりにマハラノビス距離を用いることで、変数間の相関関係を考慮したADの設定ができる。データ密度によるADは、推定したいサンプルの周辺にどのくらいのトレーニングサンプルが固まっているかを指標にする。データ密度の推定には、k最近傍法 (k-nearest neighbor, kNN) や one-class support vector machine^{9), [5]}を用いることができる。データ密度と推定誤差との関係を計算しておくことで、新たなサンプルのyの推定誤差を見積もることが可能となる。アンサンブル学習ではサンプルや変数をランダムに選択して複数の回帰モデルを構築する。新たなサンプルのyの値を推定するときには、複数のモデルから得られる推定値を統合することでモデルの推定性能の向上を達成する。さらに、推定値の標準偏差によりADを検討できる。標準偏差が小さいほど推定誤差は小さいと見積もることが可能となる。以上のように物性推定モデルとADとを組み合わせることで、推定したいサンプルごとのyの推定値のばらつきを表現できる。

4. アンサンブル学習によるADを考慮した物性推定

アンサンブル学習により、yの値を推定したいサンプルの誤差を事前に推定した例⁸⁾を紹介する。用いたデータは水溶解度が測定された1,290化合物¹⁰⁾であり、yは水溶解度の対数値である。Xは構造記述子であり、Dragon 6.0^[6]で計算された2232記述子である。実際にはすべての化合物で同じ値を持つ記述子を削除し、残った1691記述子を使用した。

ここではアンサンブル学習によりADの中と外とでyの値の誤差を推定できるかどうかを検討したため、表1のようにモデル構築用データとモデル検証用データを設定した。

アンサンブル学習として、41記述子をランダムに300回選択してそれぞれSVRによりサブモデルを構築した。各サンプルのyの推定値を300の推定値の平均としたときの、モデル検証用データ（データ1）のyの実測値と推定値とのプロットを図3に示す。表1より同様の原子種からなる化合物はAD内と考えられ、良好にyの値を推定できたことがわかる。

続いて、すべてのサンプルに対する、300の推定値の標準偏差 (standard deviation, STD) とyの推定誤差の絶対値との関係を図4に示す。点はモデル構築用データ、白抜きの丸はモデル検証用データ（データ1）、アスタリスクはモデル検証用データ（データ2）であることを表す。モデル検証用データ1と比較してモデル検証用データ2のSTDが大きいことがわかる。またSTDが大きいほど推定誤差の絶対値も大きくなることが確認された。アンサンブル学習における推定値の標準偏差を用いることで、ADを

表1 モデル構築用データおよびモデル検証用データ⁸⁾

	モデル構築用データ	モデル検証用データ
H, C, N, Oのみからなる化合物	567	257 (データ1)
その他の化合物	0	466 (データ2)

図3 水溶解度の実測値と推定値のプロット⁸⁾

図4 アンサンブル学習における推定値の標準偏差 (STD) と推定誤差の絶対値との関係. 点はモデル構築用データ, 白抜き丸はモデル検証用データ (データ1), アスタリスクはモデル検証用データ (データ2) を表す⁸⁾

考慮して y の誤差を見積もることが可能である.

5. 半教師あり学習による AD の拡大

これまで述べたように AD の外のサンプルにおける y の推定値は信用できない. しかし, 特にサンプル数が小さい場合に AD は狭くなりやすく, y の値を予測したいサンプルの多くが AD の外になってしまう. そこで半教師あり学習により AD を拡大する手法について紹介する. 半教師あり学習とは, y の測定値のある教師ありデータと y の測定値のない教師なしデータの両方を用いた学習のことである. 例えば, 教師ありデータの X と教師なしデータの X とを結合して主成分分析^[7]等により次元削減を行い, 削減後の次元において教師ありデータを用いて X と y の間で回帰モデルを構築する. これにより以下のようなメリットがあることが確認されている^[11].

- (1) モデルの安定性の向上
- (2) AD の拡大
- (3) X の事前分布の明確化

特に教師ありデータの数が少なく, 教師なしデータの数が多い場合に, 半教師あり学習によるメリットが大きい.

実際のデータを用いて半教師あり学習のメリットを確認した例^[11]を紹介する. 用いたデータは Environmental Toxicity Prediction Challenge 2009 website^[8]からダウンロードした, 環境毒性について測定された 1093 化合物である.

図5 PLS, PCA-PLS, SVR, PCA-SVR を用いたときのモデル検証用データの $pIGC_{50}$ の実測値・推定値プロット^[11]

y は *T. Pyriformis* を対象とした 50% 成長阻害濃度の対数値 ($pIGC_{50}$) である. 100 化合物をランダムに選択してモデル構築用データとし, 残りの化合物をモデル検証用データとした.

回帰分析手法として線形の PLS と非線形の SVR を用い, それぞれ全化合物を用いた PCA による前処理を行った場合 (半教師あり学習: PCA-PLS, PCA-SVR) と行わなかった場合 (教師あり学習: PLS, SVR) とで比較した. PCA の主成分数は累積寄与率が 70% を超えた最初の成分数とした.

PLS, PCA-PLS, SVR, PCA-SVR を用いた際の, モデル検証用データの $pIGC_{50}$ の実測値と推定値とのプロットを図5に示す. PLS, PCA-PLS の結果を比較すると, PLS では推定誤差が非常に大きい化合物が散見されたが, PCA-PLS ではすべての化合物において良好な推定結果が得られた. 半教師あり学習によりモデルの推定結果が安定化することを確認した. 図5(c) (d) より, SVR と比較して PCA-SVR を用いることで, 特に $pIGC_{50}$ の値が大きい化合物や小さい化合物の推定誤差が小さくなったことがわかる. 半教師あり学習により AD が広がることで, y の値が大きいデータや小さいデータも良好に推定できたと考えられる. 半教師あり学習を用いることでモデルの安定性が向上し AD が拡大することが確認された.

6. 実験計画法

これまで分子設計・材料設計において既存のデータがある状況を仮定して話を進めた. 一方で既存のデータがない状況において, 新たに合成する化合物を探索したり新たに実験するための実験パラメータを選択する必要がある場合, 実験計画法を活用することで同じサンプル数で AD をよ

り拡大したり、式[1]の(1)のばらつきと(2)のばらつきとをより分離しやすくしたりすることが可能となる。例えば5つの実験パラメータ（反応温度や反応時間など）があり、それぞれ10の候補の値がある場合、すべての組み合わせは10万通りにもなってしまいますが、すべて実験することは現実的でない。このような状況において実験計画法を用いれば、その中の数回の実験候補から実験パラメータと材料の物性・活性との間の関係性をモデル化することで、所望の物性・活性を実現するための実験候補を効率的に探索できる。

すべての組み合わせとして例えば10万通りからの候補の選び方について説明する。目指すことは、所望の物性・活性を持つ分子もしくは材料を設計することであるが、実験計画法ではその前に目指すこととして $y=f(X)$ のモデル f を構築することと設定している。実験候補も、基本的には f を構築しやすいように選択される。 X の変数間に相関関係があるということは、情報の重複があることを意味する。なるべく少ない実験候補で f を構築しなければならないため、情報の重複が小さい方が効率的である。例えば10万通りの中から10の実験候補を選択しなければならない場合、一つの方法としてランダムに10個選ぶことが挙げられる。もちろんこの方法でも、情報の重複がないような実験候補の組み合わせが選ばれる確率は高い。さらに情報の重複をなくすため、情報の重複もしくはモデルの構築しやすさを表す指標がある。この指標の値を最小化もしくは最大化することで、より情報の重複のない実験候補を選択できる。指標の例は以下の通りである。

- A最適基準： $X^T X$ の逆行列の対角成分の和（最小化）
- D最適基準： $\det(X^T X)$ （最大化）
- E最適基準： $X^T X$ の固有値の最小値（最大化）
- I最適基準： $X^T(X^T X/m)^{-1}X$ の対角成分の平均値（最小化）
- minimax 基準： $X^T X$ の対角成分の最大値（最小化）

ここで、 X は縦の大きさがサンプル数、横の大きさが変数の数である X の行列、 m はサンプル数である。

選択された実験候補について、実際に実験すると y の値が得られる。このデータセットを用いることで f を構築できる。ここで、サンプル間の情報の重複が小さくなるようにサンプルが選択されているため、線形手法であれば、主成分回帰やPLS^[9]のような X の変数間の相関関係を考慮した回帰分析手法でなくても、最小二乗法による重回帰分析^[10]で十分である。 X と y との間の非線形性を考慮したい場合、 X の各変数を二乗した変数や、2つの変数の積を新たな変数として追加して f を構築する。もしくはガウス過程による回帰^[11]やSVR^[12]などの非線形回帰分析手法を用いる。構築された f を用いることで、まだ実験していない候補の X の値を f に入力することで、 y の値を推定できる。もちろん y の推定値の信頼性は上述したADで確認する。 y の推定値を考慮して次の実験候補を選択し、実際に実験して y の値を獲得する。このような、 X と y との間の回帰モデル構築、新たな実験候補の選択、実際の実験を繰り返すことで、効率的に目標の y の値を満たす分子・材料を開発することが可能となる。

7. おわりに

これまで述べたようにビッグデータを活用することで回帰分析により分子や材料の物性・活性 y のばらつきを化学構造や実験パラメータ X で説明することが可能になっている。ただし、 y と X との間の高精度なモデルが構築されたとしても、モデルを運用するためには、モデルが高い性能を発揮できるデータ領域であるADを考慮しなければならない。ビッグデータであればADは広くなることが期待されるが、実際にはビッグデータではなく、サンプル数が小さい材料も多い。このような場合は、半教師あり学習やアンサンブル学習によりADが拡大することにより多くの新規サンプルにおける y の値を推定可能となる。また X と y との非線形関係を適切に表現することでもADは拡大する^[13]。

まだ実験データがない材料においては、実験計画法を活用して材料開発することで、効率的に目標の物性値をもつ材料を開発することが可能となる。 X と y との関係のモデル化と新たな実験候補の探索、そして実際の実験との繰り返しにより、実験回数の低減ができる。ADを考慮して新たな実験候補を探索する手法として、ベイズ最適化^[14]が提案されている。ベイズ最適化では、 y の値が目標の範囲に入るかどうかではなく、その実験候補により y の目標範囲を満たす確率を計算することで、その確率が高い候補を優先的に選択できる。

以上のようなデータ解析が活用されることで、これまで埋もれていた各種データの再活用や効率的な材料開発が促進することを期待する。

参 考 文 献

- 1) Wold, S., Sjöström, M. and Eriksson, L. : *Chemometr. Intell. Lab. Syst.*, **58**, 109 (2001)
- 2) Bishop, C. M. : *Pattern recognition and machine learning*, Springer (2006)
- 3) Filzmoser, P., Liebmann, B. and Varmuza, K. : *J. Chemometr.*, **23**, 160 (2009)
- 4) Degen, J., Wegscheid-Gerlach, C., Zaliani, A. and Rarey, M. : *ChemMedchem*, **3**, 1503 (2008)
- 5) Horvath, D., Marcou, G. and Varnek, A. : *J. Chem. Inf. Model.*, **49**, 1762 (2009)
- 6) Kaneko, H., Arakawa, M. and Funatsu, K. : *Comput. Chem. Eng.*, **35**, 1135 (2011)
- 7) Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J. and Mekenyan, O. : *J. Chem. Inf. Model.*, **45**, 839 (2005)
- 8) Kaneko, H. and Funatsu, K. : *J. Chem. Inf. Model.*, **54**, 2469 (2014)
- 9) Kaneko, H. and Funatsu, K. : *Chemometr. Intell. Lab. Syst.*, **58**, 109 (2001)
- 10) Hou, T. J., Xia, K., Zhang, W. and Xu, X. J. : *J. Chem. Inf. Comput. Sci.*, **44**, 266 (2004)
- 11) Kaneko, H. : *Chemometr. Intell. Lab. Syst.*, **182**, 47 (2018)

後注：

- [1] 明治大学データ化学工学研究室（金子研究室），データ解析に関するいろいろな手法・考え方・注意点のまとめ．[Online]．Available：
<https://datachemeng.com/summarydataanalysis/> [Accessed：4-Dec-2018]
- [2] 物質・材料研究機構（NIMS），高分子データベース（PoLyInfo）．[Online]．Available：
<https://polymer.nims.go.jp/> [Accessed：4-Dec-2018]
- [3] RDKit：Open-Source Cheminformatics Software．[Online]．Available：
<https://www.rdkit.org/> [Accessed：4-Dec-2018]
- [4] 明治大学データ化学工学研究室（金子研究室），モデルの適用範囲・モデルの適用領域（Applicability Domain, AD）～回帰モデル・クラス分類モデルを使うとき必須となる概念～．[Online]．Available：
<https://datachemeng.com/applicabilitydomain/> [Accessed：4-Dec-2018]
- [5] 明治大学データ化学工学研究室（金子研究室），One-Class Support Vector Machine (OCSVM) で外れ値・外れサンプルを検出したりデータ密度を推定したりしよう！[Online]．Available：
<https://datachemeng.com/ocsvm/> [Accessed：4-Dec-2018]
- [6] Kode Chemoinformatics．[Online]．Available：
https://chm.kode-solutions.net/products_dragon.php [Accessed：4-Dec-2018]
- [7] 明治大学データ化学工学研究室（金子研究室），主成分分析（Principal Component Analysis, PCA）～データセットの見える化・可視化といったらまずはこれ！～．[Online]．Available：
<https://datachemeng.com/principalcomponentanalysis/> [Accessed：4-Dec-2018]
- [8] CADASTER, Environmental Toxicity Prediction Challenge, 2009. [Online]．Available：
<http://www.cadaster.eu/node/65.html> [Accessed：4-Dec-2018]
- [9] 明治大学データ化学工学研究室（金子研究室），部分的最小二乗回帰（Partial Least Squares Regression, PLS）～回帰分析は最初にこれ！～．[Online]．Available：
<https://datachemeng.com/partialleastsquares/> [Accessed：4-Dec-2018]
- [10] 明治大学データ化学工学研究室（金子研究室），最小二乗法による線形重回帰分析～人工知能・機械学習・統計の基礎の基礎～，2017．[Online]．Available：
<https://datachemeng.com/ordinaryleastsquares/> [Accessed：4-Dec-2018]
- [11] 明治大学データ化学工学研究室（金子研究室），ガウス過程による回帰（Gaussian Process Regression, GPR）～予測値だけでなく予測値のばらつきも計算できる！～．[Online]．Available：
<https://datachemeng.com/gaussianprocessregression/> [Accessed：4-Dec-2018]
- [12] 明治大学データ化学工学研究室（金子研究室），サポートベクター回帰（Support Vector Regression, SVR）～サンプル数10000以下ならこれを使うべし！～．[Online]．Available：
<https://datachemeng.com/supportvectorregression/> [Accessed：4-Dec-2018]
- [13] 明治大学データ化学工学研究室（金子研究室），Xの変数を適切に非線形変換して線形の回帰モデルを作るとモデルの適用範囲・適用領域が広がるかも！[Online]．Available：
<https://datachemeng.com/variabletransformationad/> [Accessed：4-Dec-2018]
- [14] 明治大学データ化学工学研究室（金子研究室），ベイズ最適化（Bayesian Optimization, BO）～実験計画法で使ったり，ハイパーパラメータを最適化したり～．[Online]．Available：
<https://datachemeng.com/bayesianoptimization/> [Accessed：4-Dec-2018]