機械学習による周期表の再発見

草場　穫　　統計科学専攻　博士課程3年

1. 要約

周期表は、周期律に従って化学元素を並べた表である。現在の周期表の原型は、19世紀のロシアの化学者メンデレーエフによって作られた。この素晴らしい成果は、情報の視覚化の最も成功した例の1つと見なすことができる。本研究では、機械学習技術を用いてデータ駆動型周期表の作成を試みた。この目標を達成するために、我々は柔軟かつ離散座標系でデータを視覚化できるGTM（Generative Topographic Mapping）ベースのモデルを開発した。このモデルを使用して、我々は周期則をうまくとらえたいくつかの表を得ることに成功した(図1)。得られた表の解釈は、各特徴量（融点、電気陰性度など）ごとの滑らかなヒートマップを構築することによって得られる(図2)。また、学習されたモデルから各化学元素の新しい記述子（負担率ベクトル）を生成することができる。この記述子に基づく化合物の形成エネルギーの予測は、標準的な周期表に基づく記述子と比較して約33%の誤差減少(MSE)を達成した(図3)。いくつかの困難は残っているが、本研究はメンデレーエフの成果のような洗練された視覚的表現を、既存の機械学習法に幾つかの工夫を加えることによって得る可能性を示した。

2. 結果

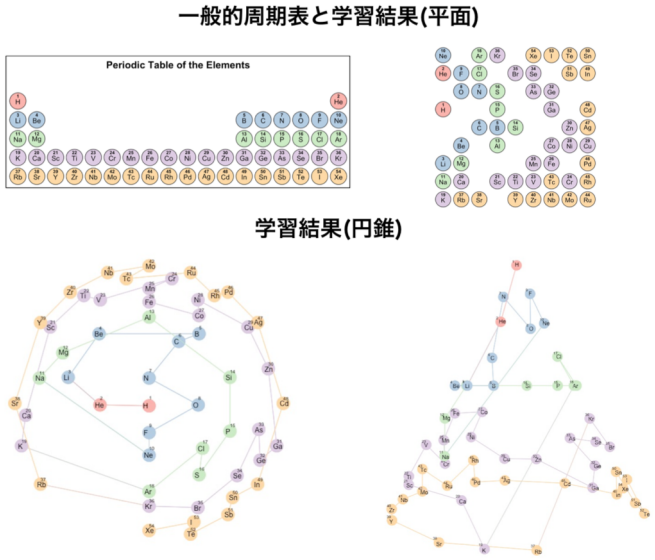


図1　学習によって得られた表 (周期ごとに色分け)

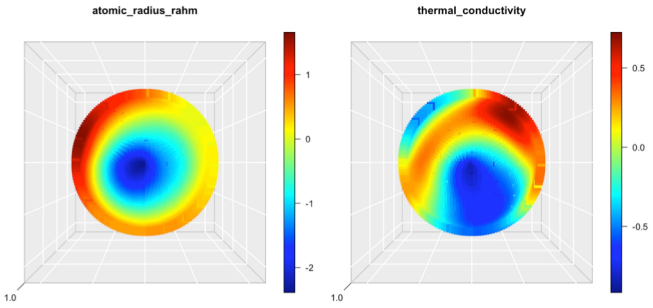


図2　各特徴量ごとのヒートマップ (クリギングにより内挿)

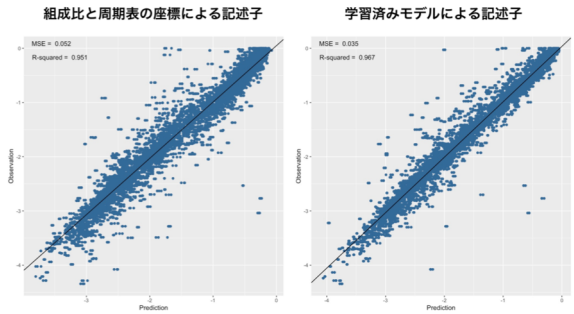


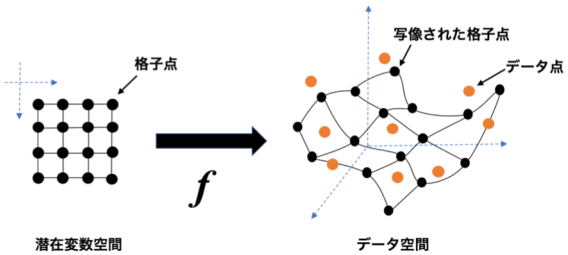
図3　ランダムフォレストによる化合物の形成エネルギーの予測

3.手法

周期表は図1に示されるように、離散的に並んだ表の格子中に一つずつ対応する元素が並べられている。周期表のような表現を学習によって得るためには、離散座標系でデータを視覚化できるような手法が必要である。また各元素の持つ規則性(周期律)は非線形であるため、この関係性を捉えるためには柔軟なモデルが必要である。先行研究として自己組織化マップを用いたものがあるが、周期律を上手く捉えた表を作るといった意味ではあまり上手くいっていない (Lemes and Dal Pino, 2011)。さらに、既存の様々な次元圧縮法 (t-SNE, PCA, Kernel PCA, ISOMAP, LLE) による解析も行なったが、これらはそもそも表型の視覚化を与えない上、周期律を捉えるという意味でもあまり良い結果が出なかった。

そこで本研究では近年提案された、柔軟なデータ視覚化法であるGTM-LDLV (図4)をベースにした学習アルゴリズムを開発した。離散座標上での視覚化を得たいので、GTM-LDLVの潜在変数$\boldsymbol{Z}$として、各行の総和が1以下かつ各列の総和が1に等しいという制約を満たす物のみ考える。しかし、この制約が成り立つためには、格子点数$\boldsymbol{K}$とデータ数Nに対して$\boldsymbol{K} \geqq N$が成り立つ必要がある。初めから$\boldsymbol{K} \geqq N$を満たす数の格子点を用意して推定を行なった場合、推定が安定しない上に多くの計算時間を必要とする。よって**Algorithm1**のような3段階の推定によってモデルを学習させる。3段階目の推定では、制約付きの$\boldsymbol{Z}$は写像された格子点 $\boldsymbol{f}(\boldsymbol{u}_k)$ と観測データ点 $\boldsymbol{x}_n$ の1to1の二部グラフマッチングにより計算される。

Algorithm1の2段階目の推定と同様の手続きにより、各格子点に対応する特徴量をガウス過程により内挿することができ、これにより図2のような滑らかなヒートマップを構成することができる。この図により得られた表がどのような特徴を捉えているか解釈することができる。また、各データ点ごとに写像された格子点に対する負担率ベクトルが計算でき、これらを元素の新しい記述子として使用することができる。この記述子に基づく形成エネルギーの予測は組成比と周期表の座標による記述子を用いたものと比較して約33%の誤差減少(MSE)を達成した。



Generative model : $p(X,Z\,|\,g,H,\beta)=K^{-N}\prod_{n=1}^{N}\prod_{k=1}^{K}N(x_n\,|\,y_k,\beta^{-1}I)^{z_{kn}}$　Length parameter : $p(r)=N(r\,|\,0,C_{st}(\zeta_r))$

Projected nodes : $y_k=f(u_k)=g(u_k)h(u_k)$　Variance parameter: $p(g)=N^+(g\,|\,0,C_{st}(\zeta_g))$

$p(H\,|\,r)=\prod_{d=1}^{D}N(h_{(d)}\,|\,0,C_h)$　variance: $p(\beta)=Gam(\beta\,|\,d_{\beta 0},s_{\beta 0})$

Covariance matrix for H: $c_h(u_i,u_j)=\left\{\frac{2l(u_i)l(u_j)}{l^2(u_i)+l^2(u_j)}\right\}^{L/2}\exp\left(-\frac{\|u_i-u_j\|^2}{l^2(u_i)+l^2(u_j)}\right)$　Covariance matrix for g and r: $c_{st}(u_i,u_j;\zeta_0)=V_0\exp\left(-\frac{\|u_i-u_j\|^2}{2l_0}\right)$

$l(u)=\exp(r(u))$

[N. Yamaguchi. GTM with latent variable dependent length-scale and variance. *CACS*, 2013]

図4　GTM-LDLVの概要

Algorithm 1 Proposed model

初期値 $\boldsymbol{\theta}^0=\{\boldsymbol{Z}^0,\beta^0,\boldsymbol{g}^0,\boldsymbol{H}^0,\boldsymbol{r}^0\}$

for $t=1$ to T^1 **do**

$\boldsymbol{Z}^t$ を $p(\boldsymbol{Z}|\boldsymbol{X},\beta^{t-1},\boldsymbol{g}^{t-1},\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$ から発生する.

β^t を $p(\beta|\boldsymbol{X},\boldsymbol{Z}^t,\boldsymbol{g}^{t-1},\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$ から発生する.

$\boldsymbol{g}^t$ を $p(\boldsymbol{g}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$ から発生する.

$\boldsymbol{H}^t$ を $p(\boldsymbol{H}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{g}^t,\boldsymbol{r}^{t-1})$ から発生する.

$\boldsymbol{r}^t$ を $p(\boldsymbol{r}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{g}^t,\boldsymbol{H}^t)$ から発生する.

end for

十分大きな数 T_b に対して $\boldsymbol{\theta}^t=\{\boldsymbol{Z}^t,\beta^t,\boldsymbol{g}^t,\boldsymbol{H}^t,\boldsymbol{r}^t\}$, $t=T_b,T_b+1,\cdots,T^1$ を記録する。

以上の時点におけるアンサンブルにより,GTM-LDLVの各モデルパラメータが推定される。
潜在変数空間上の格子点を, $K \geq N$ を満たすように増加させる。
GTM-LDLVによって得られたパラメータを観測値と考え, ガウス過程により増加させた格子点に対応するパラメータ値を内挿する。
以上のようにして得られたパラメータ $\boldsymbol{\theta}^{itp}=\{\boldsymbol{Z}^{itp},\beta^{itp},\boldsymbol{g}^{itp},\boldsymbol{H}^{itp},\boldsymbol{r}^{itp}\}$ を次のアルゴリズムの初期値として用いる。

for $t=1$ to T^2 **do**

$\boldsymbol{Z}^t \leftarrow \underset{Z\in S}{\operatorname{argmax}}\, p(\boldsymbol{Z}|\boldsymbol{X},\beta^{t-1},\boldsymbol{g}^{t-1},\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$, $S=\{Z|\sum_{n=1}^{N}z_{kn}\leq 1\,(k=1,\cdots,K)\}$ ← 二部グラフマッチングにより推定

$\beta^t \leftarrow \underset{\beta}{\operatorname{argmax}}\, p(\beta|\boldsymbol{X},\boldsymbol{Z}^t,\boldsymbol{g}^{t-1},\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$

$\boldsymbol{g}^t \leftarrow \underset{\boldsymbol{g}}{\operatorname{argmax}}\, p(\boldsymbol{g}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{H}^{t-1},\boldsymbol{r}^{t-1})$

$\boldsymbol{H}^t \leftarrow \underset{\boldsymbol{H}}{\operatorname{argmax}}\, p(\boldsymbol{H}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{g}^t,\boldsymbol{r}^{t-1})$

$\boldsymbol{r}^t \leftarrow \underset{\boldsymbol{r}}{\operatorname{argmax}}\, p(\boldsymbol{r}|\boldsymbol{X},\boldsymbol{Z}^t,\beta^t,\boldsymbol{g}^t,\boldsymbol{H}^t)$

end for