1P−1

ニューラルネットワークによる高次元評価関数の構築と評価
- 超電導データを対象として -

*下川 信祐　大田原 一成
(株) エイ・ティ・アール環境適応通信研究所

1 はじめに

ニューラルネットを用いて，多数の属性変数に対する評価データから評価値を推定する問題を物質デザインの立場 [1, 2] から考察する．

データを基にする予測方法は，（物理学などの）対象理論構築抜きに適用できる点で興味深い．ここでは，（数学上ではない）現実の未知の関数であること，関数関係の成立が期待できること，相当量のデータが公開されていること，(QSAR などに向かない) 強い非線形性がみこまれることなどから，高温超伝導体の相転移温度を対象にする．

高温超伝導体が屡々微量成分の混入（ドープ）によって得られるなど，転移温度は組成比に対して敏感に応ずる特異性・強い非線形性がある．一方，ニューラルネットなどの予測手法は，原理的に正則性（なめらかさ）を基にしている．このため，予測精度の低下や，データフィッテングのための計算量の増大などの困難が生じる．

ここでは，これらの困難に対処するための，いくつかの手法を検討した．まず，データの特異性を弱めるために，非線形変換を施して符号化を変更した．次に，微量な値変化での高微分値を削除するサンプリングを施した．また，データ fitting における計算量を削減するために，誤差評価関数の多重 basin 構造を開放する変形を試みた．予測誤差を評価するとともに誤差の推定についても検討した．誤差の推定から物質の探索戦略が得られる．

2 予測の原理と評価尺度

定義域 X と値域 Y を持つデータ D に対して，これを基にする予測写像 P とは P が D の拡張となることを言う．　真の写像 T が存在して（適当な位相に対して）連続であれば，ある点 $x \in X$ の近傍で D の密度が稠密に向かえば $P(x)$ は $T(x)$ に収束してゆく．

P は D の補間・補外に他ならない．P を構成する手法は，一般に $(C, \mathcal{G}, \mathfrak{a})$ で特徴づけられる．ここで C は，現実の対象や意味を数値に対応させて X と Y の要素を指定する方法を指し符号化と呼ぶ．$\mathcal{G}$ は，予測写像を構成するための計算可能な写像の空間で，ここではニューラルネットを指す．予測写像 P はデータ fitting を実行する最適化問題

$$\min_{g\in\mathcal{G}} V(g) = V(P), V(g)^2 = \sum_{(x,y)\in D} |g(x)-y|^2/|D|$$

を解くことによって得られる．$\mathfrak{a}$ はこの最適化を解くアルゴリズムである．

予測写像を構成する手法は，次の 2 点から評価される：

(i) 予測精度　　(ii) 計算量

(i) では，更に予測誤差の推定が可能であることが望まれる．(ii) は，計算に要する時間的・空間的資源であって，ニューラルネットの規模を含む．

このような手法を直接用いる時，(i),(ii) に最も影響を与えるのは，T の局所的な特異性である．T の微分値の変動（高階微分）が大きければ，fitting 計算量，予測精度何れもが大きく劣化する．たとえば，スカラー値の場合，T が線形近似可能な小領域に D を分割する必要数を N とすれば，3 層ネットの中間層は，$N \dim X$ 個を要する．各点の近傍で高い微分値変動を伴えば，$N \sim \mathrm{Cnst.}^{\dim X}$ であり，そのままでは，計算不可能である．そこで，まず，データ D の局所的性質を吟味する．

3 データの局所的な性質

文献 [3] で公開されているデータの中で，有効なものの全体は D は $|D| = 2500$, 各物質はそれぞれ最大 8 個の元素からなるが，データ全体では $d = 68$ 種の元素が出現する．組成比により表現すると $X = [0,1]^d$, $Y = [0, 150]$ ととれる．

図 1 にデータ 2 点 $(x_i, y_i), (x_j, y_j)$ 間の距離 $|x_i - x_j|$ と差分の相対比 $|y_i - y_j|/\max\{y_i, y_j\}$, 微分商 $|y_i - y_j|/|x_i - x_j|$ （何れも絶対値）の関係を示す．強い大変動 (局所的な大差分値) があり強い特異性が認められる．

4 特異性を和らげる

高温超伝導体は，金属酸化物に微小の別元素の混入（ドープ）により得られることが多い．ドープは大変動を形成する大きな要因と考えられる．

微量成分による大変動を正則化するには，'微量' が非微小量となる変換を施して変化を定義域上局在させなければ良い．そこで，X の各成分同時に次のような非線形写像 φ を施して符号化の変更を行う．

$$\varphi : [0,1] \ni x \mapsto c_1 x^{\alpha_1} + c_2 x^{\alpha_2} + c_3 \in [-2, 2]$$

'Construction and evaluation of a design function on a 3-layered neural network, - in case of a high T_c superconductor data set - '

SHIMOGAWA, Shinsuke (simogawa@acr.atr.co.jp), OHTAWARA, Kazushige (ohtawara@acr.atr.co.jp).
ATR Adaptive Communications Research Labratories.
2-2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto Pref. 619-0288.

このとき，図1は図**2**のように変化する．かなり大変動を除去できている．

しかし，これだけでは，特異性は今だ著しい．そこで，デザインの目的：'大きな相転移温度の物質の探索' に立ち戻り，写像 P の予測目標を組成 x の近傍での T_c 測定の最大値 $(\max_{\varphi(z_i)\sim\varphi(x)} T_c(z_i))$ と設定する．φ の性質から，近傍の中は全て同一の元素種セットであり探索との整合性が確保されている．近傍を距離 $\rho^{\varphi} \leq 0.05$ としてデータから近傍内最大点を取り出すサンプリングを行ったときの変動性が図**3**である．正則化による効果を図に示す．図**4a,b**は，fitting計算のプロセス，図**5a,b**は，fittingと予測誤差の改善を示す．

5 Basin resolution

特異性・非線形性が強いとfittingアルゴリズム $\mathfrak{a}$ の計算量が問題と1なる．$\mathfrak{a}$ の計算の困難は，$V(g)$ が多数のbasinを伴うからである．我々は，この問題に対所するため，問題 $V(g) \to 0$ を比較的簡単なbasin構造を持つ $u(g')$ を用いて $u(g') \to 0$ に変形する方法を見出した．$u(g')$ は内点で0以外の極小値をとらないという性質を持つ．図**6**は，$\mathrm{grad}\, u(g')$ と $\mathrm{grad}\, V(g)$ の比較例でありbasinが開放されている様子が伺える．

6 誤差評価，誤差推定，探索戦略

図**7**は $\min_{z\in D_t} \rho^{\varphi}(x,z)$ の解 $z_0(x)$ を用いて収束誤差を補正した後の予測誤差評価である．ここに，D_t は，サンプリングで得られたデータから，評価用データを除いた教師データである．予測精度は平均相対2乗誤差で4.7%最大誤差(相対値)17.4%であり，近傍探索極大値の予測については，相当の精度が確保された．探索戦略は $\mathrm{grad}(P(x) + \mathrm{err}(x))$ となる．ここで，$\mathrm{err}(x)$ は $P(x)$ の誤差推定関数である．図**8**は，$\rho^{\varphi}_{D_t}(x) := |z_0(x) - x|$(学習データへの距離関数)と精度の関係を示している．予測精度の推定関数 $\mathrm{err}(x) = \mathrm{Cnst}.\rho^{\varphi}_{D_t}(x)$ として利用する時の効率を表している．

参考文献

[1] K. Shinjo et. al., An Engineering Model of Composite Nanomaterials, *ICCE/8*, Tenerife, Spain, August 2001.

[2] K. Ohtawara et. al., A Design Theory of Material Systems, (in press) *Int. J. Mod. Phys.* **C**, 2001.

[3] Y. Asada, SUPERCON, `http://asagiri.nrim.go.jp`, 1999.

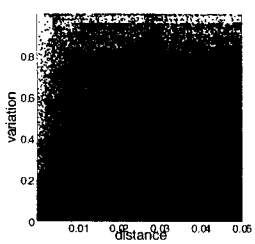

図 **1.** 元データの変動性.

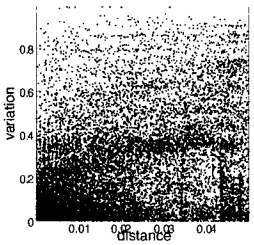

図 **2.** 座標変換後の変動性.

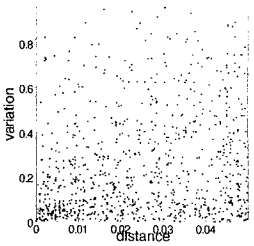

図 **3.** 局所最大データでの変動性.

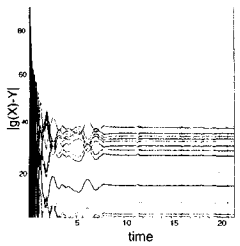

図 **4a.** 収束 - 元.

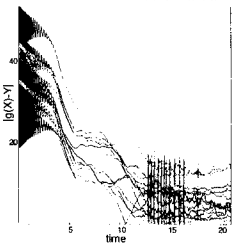

図 **4b.** 収束 - 正則.

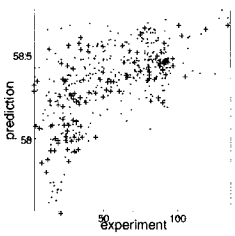

図 **5a.** 収束・予測誤差 - 元.

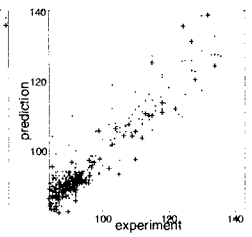

図 **5b.** 収束・予測誤差 - 正則.

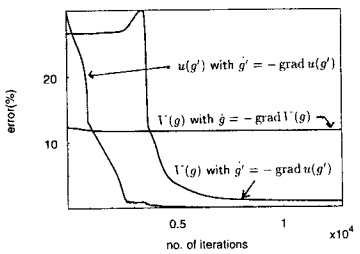

図 **6.** Basin 開放の効果.

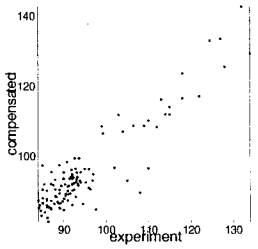

図 **7.** 予測誤差 (収束誤差補性).

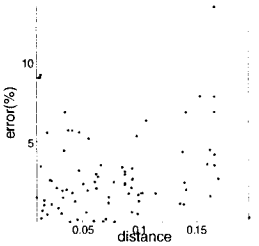

図 **8.** 予測精度と $\rho^{\varphi}_{D_t}$ 距離.