多次元パラメータ階層型データのためのパラメータ選択法

伊藤 貴之○（お茶の水女子大学／京都大学）
清 豪　　小山田 耕二　　酒井 晃二　　岩下 武史　　金澤 正憲（京都大学）

Automatic Dominant Parameter Determination Technique for Visualization of Multi Parameter Hierarchical Data

Takayuki ITOH　Takeru KIYOSHI　Koji KOYAMADA
Koji SAKAI　Takeshi IWASHITA　Masanori KANAZAWA

ABSTRACT

This paper presents a visualization technique of hierarchical data which each leaf and non-leaf node has multi parameters. The technique determines the dominant two parameters from the multi parameters, by applying response surface technique. By assigning the two parameters to horizontal and vertical axes of display spaces, the technique represents the dependency among the dominant parameters of hierarchical data. The technique applies HeiankyoView, a visualization technique for large-scale hierarchical data. The paper introduces some visualization results proofing the effectiveness of the presented technique, and a scientific application that the presented technique is to be effectively used.

Keywords: Visualization, HeiankyoView, Multi Parameter Hierarchical Data, Response Surface.

1. はじめに

本報告では「多次元パラメータ階層型データ」を、各ノードが多次元パラメータ値をもつ階層型データ、と定義する（図 1 参照）。多次元パラメータ階層型データは、例えば企業の組織構造で階層化された人事情報や、IP アドレスで階層化されたネットワーク計算機情報などに代表されるように、日常生活でもよく見られるデータ構造である。

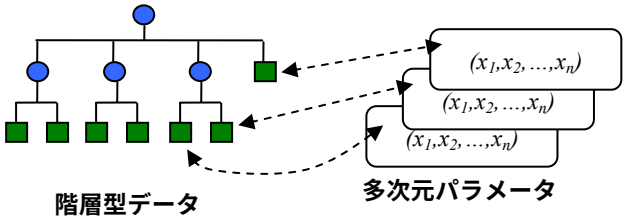

図 1 多次元パラメータ階層型データの定義。

筆者らは科学技術シミュレーションの目的で多次元パラメータ階層型データを扱っている。一般的に科学技術シミュレーションでは､信頼できる計算解を得るために、入力パラメータの調整を必要とする場合が多い。この入力パラメータ調整の過程を支援するために、入力パラメータと計算解の相関性を理解することが必要であり、その手段として情報可視化が有効であると考えられる[1]。ここで本報告では、計算解の誤差 y を関数 $y = S(x_1, x_2, ..., x_n) - C$ と定式化する。ここで S は、入力パラメータ x_1 ～x_n から得られる計算解であり、C は測定結果などから得られる理想解である。この定式において、入力パラメータの調整とは、y の最小値を導くパラメータ組 x_1 ～x_n を発見することに相当する。文献[2]にて筆者らは、以下の手順によって適切な入力パラメータ値を得る手法を提案している。

1. 入力パラメータ値の範囲を設定する
2. 入力パラメータ値の範囲内にて少数かつ有効な入力パラメータ値をいくつかサンプリングする。
3. その入力パラメータ値を用いてシミュレーションを実行する。
4. 得られた計算解を参照して、入力パラメータ値の範囲を絞り込む。
5. 2.～4.を、最適解が得られるまで反復する。

以上の処理によって m 回のシミュレーションを実行した結果として、計算解の誤差 y および入力パラメータ x_1 ～x_n のセットが m 個生成される。筆者らは m 個の数値セットを、シミュレーション実行時の入力パラメータ値の範囲を用いて階層化することで、多次元パラメータ階層型データを生成している。本報告は、この多次元パラメータ階層型データの有効な可視化について議論するものである。

筆者らは、大規模階層型データの可視化を目的とした手法「平安京ビュー」[3]を提案している。本報告の目的は、階層型データを構成する各ノードに割り当てられた入力パラメータ値 x_1 ～x_n に対する返り値 y の相関性を、「平安京ビュー」を用いて視覚的に発見することである。

このとき一般的に、必ずしも全ての入力パラメータが y に大きな影響を与えるとは限らない。よって x_1 ～x_n の中から最も y に大きな影響を与えている入力パラメータを特定し、その入力パラメータを可視化に用いることが有効であると考えられる。

本報告は、関数 $y = S(x_1, x_2, ..., x_n) - C$ の返り値 y に対して、最も大きな影響を与えている 2 個の入力パラメータ x_i と x_j を、応答曲面法を用いて特定する手法を提案する。さらに本報告では、この 2 個の入力パラメータを、「平安京ビュー」による階層型データ可視化に活用する。具体的には、提案手法を用いて特定した 2 個のパラメータを、画面空間の横軸と縦軸に対応つけ、返り値 y を画面空間の奥行き軸に対応つけることで、2 個のパラメータの数値変化に対する返り値 y の数値変化を表現する。

2. 平安京ビュー

「平安京ビュー」[3]は、長方形の入れ子構造を用いて階層型データを表現する可視化手法である。これは筆者ら自身によって過去に提案された「データ宝石箱」[4]という可視化手法に対して、画面配置アルゴリズムを改良した手法である。 図 2 に、「平安京ビュー」を用いた大規模階層型データの可視化例を示す。

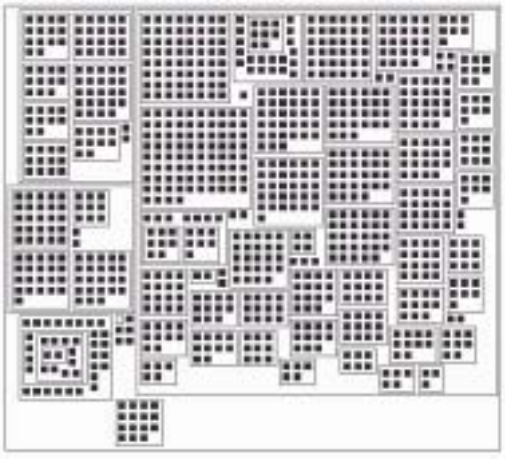

図 2. 「平安京ビュー」による大規模階層型データの可視化の例。

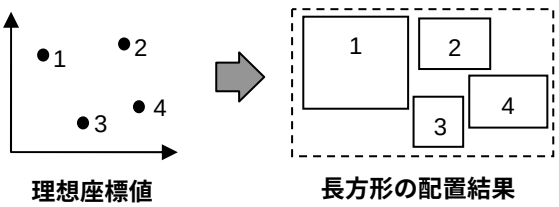

図 3. 長方形の理想座標値と配置結果。

「平安京ビュー」は図 2 に示すような可視化結果を得るために、以下の 2 条件を満たすような長方形画面配置アルゴリズムを採用している。

[条件 1] 長方形同士が重なり合わない。

[条件 2] 長方形群による画面占有面積を最小化する。

以上の 2 条件は、長方形の画面上の位置を自在に制御するものではない。本報告では、以下の**[条件 3]**を付加することで、図 3 に示すように、各々の長方形の画面上の位置を制御する。

[条件 3] あらかじめ個々の長方形に指定された理想座標値にできるだけ近い位置に、長方形を配置する。

ここで本報告では、入力パラメータ値 x_1 ～x_n の中から特定のパラメータ値 x_i を参照することで、理想座標値の水平方向の座標値を算出することを考える。同様に特定のパラメータ値 x_j を参照することで、理想座標値の垂直方向の座標値を算出することを考える。また「平安京ビュー」を用いて階層型データを可視化する際に、各々のノードの高さ（奥行き方向の座標値）を y 値から算出することを考える。

また本報告では、y 値に最も大きな影響を与える入力パラメータ x_i は、y 値を最も滑らかに遷移させる、という前提に基づいて入力パラメータを特定する。図 4 にこの前提を示す。このようにして特定された入力パラメータ値を可視化結果の横軸および縦軸に対応つけることで、影響の大きい入力パラメータ値と返り値 y の相関性を視覚的に理解しやすくする。次章では、この入力パラメータを自動選択する手法について論述する。

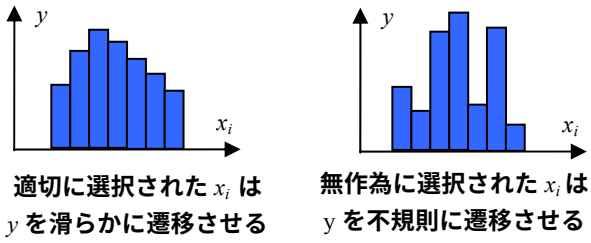

図 4. 選択された入力パラメータ値の変化に対する返り値の変化。

3. 応答曲面法を適用したパラメータ自動選択

各々の入力パラメータ x_i および返り値 y が k 組あるとする。このとき本報告では、「応答曲面法」という手法を用いて、k 組の値を補間する曲面を生成する。応答曲面法は数値解析などの教科書に多数紹介されている、よく知られた手法である。入力パラメータが x_1 ～x_n の n 個であり、曲面が 2 次曲面であるとすると、応答曲面は式(1)で表現される。

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1, j \le i}^{n} \beta_{ij} x_i x_j \quad ..(1)$$

入力パラメータおよび返り値が k 組あり、式(1)の項が全部で p 個あるとすると、式(1)は式(2)のように書き換えることができる。

$$Y = X\beta + \varepsilon \qquad ..(2)$$

ただし

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{k1} & x_{k2} & \cdots & x_{kp} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}.$$

ここで係数 β の不偏推定量 b は式(3)で表現される。

$$b = (X^T X)^{-1} X^T y \qquad ..(3)$$

応答曲面法では、式(3)を用いて係数βを算出した後に、式(4)によって導かれる決定係数 R^2_{ad} を用いることで、曲面の有効性を検証することができる。

$$R^2_{ad} = 1 - \frac{SS_R / (k-p-1)}{S_{yy} / (k-1)} \qquad ..(4)$$

ただし

$$SS_R = \beta_b{}^T X^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},$$

$$S_{yy} = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

決定係数 R^2_{ad} の値が小さすぎる場合には、応答曲面法では、与えられた k 組の入力パラメータと返り値の関係が不規則すぎて、適切な曲面を生成すること自体が困難である、と判断する。筆者らの実装では、このような場合には、入力パラメータの選択自体を断念する。

さもなければ筆者らの実装では、各々のβ値の重要性を、式(5)に示すt検定を用いて判定する。ここでσはβの分散の最尤推定値, C_{jj} は行列$(X^TX)^{-1}$の jj 番目の要素を示す。

$$t = \frac{\beta_j}{\sqrt{\hat{\sigma} C_{jj}}} \qquad ..(5)$$

筆者らの実装では、式(5)を各々のβ値に適用し、t 値の小さいβをゼロにすることで、式(1)に示す応答曲面の項の数を減らす。これを何度か繰り返すことで、返り値に対して影響の大きい項だけが残った、信頼性のある応答曲面を定義する。この応答曲面に対して、最後にもう一度 t 検定を適用し、最も t 値の大きい 2 項に該当する x_i および x_j を入力パラメータに選択する。

4. 実行例

本報告では、x_1 ～x_3 までの 3 つの入力パラメータと、そのパラメータに依存して得られる返り値 y を有する、比較的小規模な多次元パラメータ階層型データを用いて実験を行った。本実験では、応答曲面法を用いて入力パラメータの中から 2 つを選択し、続いてこの 2 つのパラメータを画面空間の横軸および縦軸に対応つけながら、「平安京ビュー」を用いて階層型データを可視化した。図 5 に実行例を示す。ここで可視化結果の色および高さは、以下のような意味をもつものとする。

- ノードの色は、画面空間の横軸と対応つけられた入力パラメータ値x_i より算出されている。この色は単に、x_i の値が適切に各ノードの水平方向の座標値を導出していることを確認するために用いている。
- ノードの高さは、返り値y の関数として算出されている。ノードの高さが位置に対して滑らかに遷移していれば、入力パラメータの選択が適切であったことを裏付けられる。

図 5(左上)は、理想座標値を全く算出せずに、**[条件 1][条件 2]**だけで階層型データを可視化した例である。図 5(右上)は、提案手法によって選択された 2 つのパラメータを用いて理想座標値を算出した例である。図 5(左下)は、無作為に選択された 2 つのパラメータを用いて理想座標値を算出した例である。

図 5(右上)と図 5(左下)では、画面空間の横軸にしたがって、ノードの色相が滑らかに青から赤に変化しているのが観察される。図 5(左上)では、このような結果は観察されない。以上の観察結果から、x_i の値を用いて各ノードの理想座標値を算出することにより、x_i が適切に各ノードの水平方向の座標値を導出していることを裏付けている。筆者らは同様に、画面空間の縦軸と対応つけられた入力パラメータ値 x_j が、適切に各ノードの垂直方向の座標値を導出していることも確認した。

また、図 5(右上)では、ノードの高さが画面空間の横軸に沿って滑らかに変化しているのが観察される。図 5(左上)および図 5(左下)では、ノードの高さの滑らかな変化は観察されない。以上の観察結果から、提案手法によって選択された入力パラメータが、ノードの高さの滑らかな遷移を実現していることがわかる。これは提案手法が、適切な入力パラメータ選択を実現していることを裏付けている。

筆者らは、他のいくつかの階層型データに対しても、提案手法が適切な入力パラメータ選択を実現していることを確認している[5]。

5. まとめ

本報告では、入力パラメータ値の中から、返り値に最も大きな影響を与える入力パラメータを選択する手法を提案した。提案手法では、応答曲面法を用いて入力パラメータ値および返り値を補間する曲面を生成し、その各項に対して t 検定を適用することで、最も影響の大きい入力パラメータを特定することで、入力パラメータを自動選択する。さらに本報告では、多次元パラメータ階層型データを構成する多次元パラメータの中から 2 つの入力パラメータを選択し、それらを画面空間の横軸および縦軸に対応つけながら、「平安京ビュー」を用いて階層型データを可視化することで、提案手法による入力パラメータの自動選択が適切であることを検証した。

今後の課題として、図 5 に代表される実行結果を数値的に評価することで、提案手法による入力パラメータ選択の妥当性を再検証することがあげられる。

筆者らは現在、提案手法を心臓細胞シミュレーション[6]のための入力パラメータ自動選択に適用することを検討している。このシミュレーションでは、細胞内外を往来する各イオンの濃度に代表される、非常に多数のパラメータを入力して、その活動電位の時間変化などを算出している。この算出結果の実測値との誤差を最小にする入力パラメータ値を導出するため、またこの誤差値に対して特に影響の大きい入力パラメータを特定し、その誤差値との相関性を分析するために、提案手法は有用であると考えられる。今後のもうひとつの課題として、心臓細胞シミュレーション以外にも、提案手法を有効利用できるアプリケーションを探すことがあげられる。

参考文献

[1] Obayashi S., Sasaki D., Visualization and Data Mining of Pareto Solutions Using Self-Organizing Map, Second International Conference of Evolutionary Multi-Criterion Optimization, pp. 796-809, 2003.

[2] Koyamada K., Sakai K., Itoh T., Parameter Optimization Technique Using the Response Surface Methodology, IEEE Engineering in Medicine and Biology Society, 2004.

[3] Itoh T., Koyamada K., HeiankyoView: Orthogonal Representation of Large-scale Hierarchical Data, International Symposium on Towards Peta-Bit Ultra Networks (PBit 2003), pp. 125-130, 2003.

[4] Itoh T., Yamaguchi Y., Ikehata Y., and Kajinaga Y., Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, IEEE Transactions on Visualization and Computer Graphics, Vol. 10, No. 3, pp. 302-313, 2004.

[5] Kiyoshi T., Itoh T., Koyamada K., Sakai K., Iwashita T., Kanazawa M., Visualization of Multi Parameter Hierarchical Data Using Automatic Dominant Parameter Determination Technique, NICOGRAPH International 2005.

[6] Sarai N., Noma A., simBio: A Java Package for Biological Simulation, 5th International Conference on System Biology, p373, 2004.

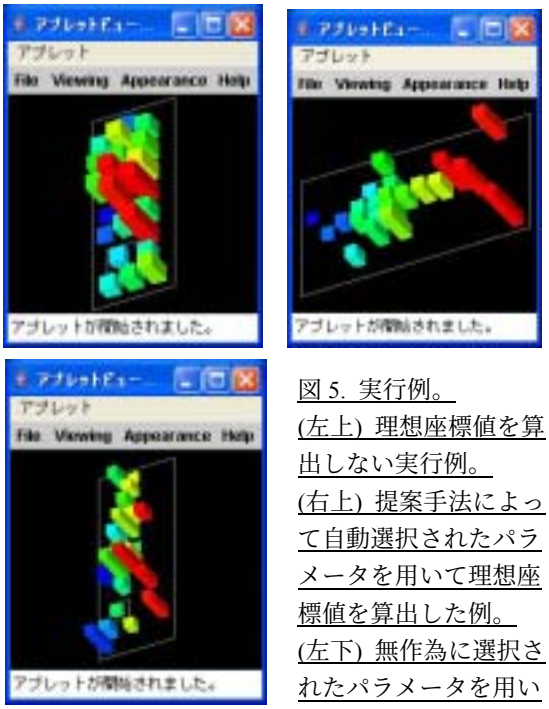

図 5. 実行例。
(左上) 理想座標値を算出しない実行例。
(右上) 提案手法によって自動選択されたパラメータを用いて理想座標値を算出した例。
(左下) 無作為に選択されたパラメータを用いて理想座標値を算出した例。