高次元データ可視化手法 Hidden

伊藤貴之

高次元データの可視化には散布図行列や平行座標法といった手法がよく用いられるが，いずれの手法においても次元数が非常に大きいときに重要な数値分布に着目することが難しくなる．そこで高次元データの中から特徴的な次元だけを自動抽出して低次元データの集合として可視化する手法がいくつか提案されており，著者もこの考え方にもとづく高次元データ可視化手法 Hidden を提案している．本報告では Hidden の処理手順と適用事例を紹介し，さらに学部の講義にて自由課題のツールとして用いている状況を紹介する．

1 はじめに

日常生活や専門業務には高次元データが多数存在しており，そこから発見される規則性はデータ領域における大きな知見となる．決済情報から発見される規則性は顧客満足度向上や売上予測に用いられ，計測情報から発見される規則性は自然現象の理解や予測に用いられる．デジタルコンテンツの特徴量から発見される規則性はコンテンツの認識や推薦に用いられる．その規則性を人間が理解するためには可視化技術による画面表現が有効である．

近年では高次元データからの規則性発見に各種の機械学習が活躍しているが，少なくとも筆者の経験では機械学習のブームと同時に可視化に対する問い合わせが増えている．筆者が耳にする限りでは，機械学習およびその他の分析手法が出した応答に対する説明責任が必要な現場があり，ユーザ自身がそれを理解するために可視化を利用したい，という声が増えているように思われる．このような背景から，高次元データの可視化についても新しい展開が必要であると考えられる．

本稿では m 次元ベクトル $a_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ で表現される n 個の個体の集合 $A = \{a_1, a_2, ..., a_n\}$ を高次元データと定義する．このような高次元データを構成する全ての次元の値を可視化する手段として，散布図行列や平行座標法が知られている．

散布図行列 (Scatterplot Matrix) は，全ての 2 次元ペアを対象として散布図を作成し，それを格子状に並べて一覧表示したものである．言い換えれば，j 番目の変数と i 番目の変数を 2 軸に割り当てた散布図が，左から j 番目，上から i 番目に配置される，というような格子構造が多数の散布図によって形成されたものを散布図行列と呼ぶ．この可視化により，任意の組み合わせの次元間の相関を一画面で視認できる．しかし個々の散布図は画面上では非常に小さくなってしまうため，この可視化結果だけから数値分布を詳細に眺めるのは困難となる．

平行座標法 (Parallel Coordinate Plots) は高次元データを折れ線の集合で可視化する手法である．多次元データを構成する 1 番目から m 番目の次元を表す各座標軸をそれぞれ鉛直な線分で表現し，それを左右方向に並べ，データ中の各個体が有する各変数値 $\{x_{i1}, x_{i2}, ..., x_{im}\}$ を座標軸上にプロットし，折れ線で結ぶ．平行座標法は，多次元データの全ての次元における各個体の値を読み取ることが可能であり，また各次元の数値分布を一画面で一気に視認できる，という点において他の手法より優れている．一方で，

High-Dimensional Data Visualization “Hidden”
Takayuki Itoh, お茶の水女子大学理学部情報科学科, Dept. of Information Sciences, Ochanomizu University.

各個体を表現する折れ線が互いに絡みあうので視認性に問題が生じやすい，隣接していない次元間の相関は読み取りにくい，といった問題点がある．

以上の手法は高次元データを構成する全ての次元を網羅的に可視化するものであるが，一方で高次元データを構成する全ての次元に興味深い規則性が見られるとは限らない．よって必ずしも全ての次元の値を網羅的に可視化するという方針が正しいとも限らない．この点に着目して，可視化するに値する次元だけを選んで表示する可視化手法が近年になって多く発表されており，筆者もこの方針に則った高次元データ可視化手法 Hidden を提案している．本報告では Hidden の処理手順と適用事例を紹介するとともに，筆者の講義科目の自由課題に用いた実例を紹介する．

2 近年の高次元データ可視化手法

高次元データを可視化する際に，一定の条件を満たす次元のみを選択することで，特徴的な低次元部分空間を可視化することができる．このアイディアにもとづいて，高次元データを限られた数の散布図で表現する手法 [10]，いくつかの低次元な平行座標法で表現する手法 [6]，散布図と平行座標法の組み合わせで表現する手法 [1] などが発表されてきた．しかしこれらの手法では選択される次元の数を対話的に調節する機能を搭載してなかったので，可視化結果を動的に調節することが難しかった．

一方で，対話操作によって選択された少数の次元によって構成される低次元部分空間を，単一の散布図または単一の平行座標法で表現する手法もいくつか発表されてきた．例として，サイコロを転がすメタファを利用して散布図を切り替え表示する方法 [2]，次元削減を組み合わせて散布図で表示する方法 [4]，平行座標法を用いる方法 [5] などが提案されている．しかし単一の散布図や平行座標法で高次元データの全ての特徴を表現できるとは限らない．

最近の高次元データ可視化手法には「次元散布図」を搭載した手法 [7] [8] が提案されている．次元散布図とは，次元の数だけ点を表示した散布図であり，2 点間の距離は次元間の類似度や相関などに対応している．この次元散布図を閲覧しながら対話操作をすることで，ユーザはフレキシブルに次元を選択することができる．また次元をノードとしたグラフを構成し，次元間相関に基づいてノードを配置する手法 [9] も提案されている．

次章で提案する Hidden は，複数の平行座標法による低次元部分空間を対話操作表示し，さらに次元散布図がその操作をガイドするという意味で，上述の各種手法の特徴を組み合わせた手法である．さらに次元間の類似度や相関以外の基準として相関ルールにもとづいた次元選択手法もあわせて搭載しているという点で，従来手法を拡張した手法である．

3 高次元データ可視化手法 Hidden

本章では筆者が提案している高次元データ可視化手法 Hidden [3] を紹介する．Hidden は「HIgh Dimensional Data Exploration and Navigation」の略称であり，高次元データに隠された興味深い知見への探索と誘導を目的とした可視化手法である．

3.1 概要

本章では 1 章で示した高次元データの定義を拡張し，以下のように定式化する．高次元データは n 個の個体を有し，各個体は m 個の変数を有するものとする．変数には m_v 個の実数型変数と m_c 個のカテゴリ型変数が含まれるとする．このとき本稿では高次元データ D を以下のように表記する．

$$D = \{a_1, ..., a_m\}$$
$$a_i = \{v_{i1}, ..., v_{i_n v}, c_{i1}, ..., c_{i_n c}\}$$

ここで v_{ij} は i 番目の個体における j 番目の実数型変数を示し，c_{ij} は i 番目の個体における j 番目のカテゴリ型変数を示す．

本手法の処理手順の概要を図 1 に示す．本手法では実数型変数による各次元を m_v 次元ベクタとして扱い，任意の次元ペア間の距離を算出し，その距離を保持するように散布図を生成する．この散布図を参照しながら，閾値を対話的に調節することで，本手法は可視化する価値のある複数の低次元部分空間を半自動的に抽出し，これらを平行座標法で可視化する．

以下，低次元部分空間抽出のための次元選択手法，および対話操作のための実装について論じる．

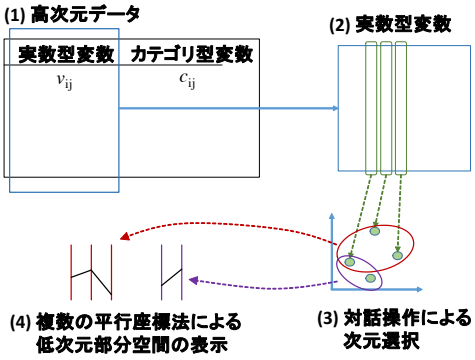

図 1　**Hidden** の処理手順の概要．**(1)** 入力データ．**(2)** 次元間距離の算出．**(3)** 次元間距離を保存した散布図の表示．**(4)** 平行座標法による低次元空間の表示．

3.2　次元選択 (1) 次元間相関に基づく手法

この処理では，高次元データを構成する m_v 個の実数型変数について，各次元ペア間の距離を算出する．現時点での著者らの実装では，j 番目と k 番目の次元の距離を以下のとおり定義する．ただし $f_c(j,k)$ は j 番目と k 番目の間の相関係数である．

$$d_{jk} = |1.0 - f_c(j,k)| \qquad (1)$$

この定義により，正または負の相関が高い次元ペアは距離が小さくなる．このような距離を定義した理由は，平行座標法での可視化では正または負の相関が高い次元を選ぶのが効果的だからである．

本手法ではユーザが設定した閾値 d_{select} よりも距離が小さい次元ペアを連結したグラフを生成し，画面右側に表示する．図 2 においてノードは実数型変数となる各次元を，エッジは距離が d_{select} 以下である次元ペアを表している．そして Bron-Kerboshc のアルゴリズムを適用することで，このグラフからクリーク（部分完全グラフ）を抽出し，これに包括される次元群を平行座標法で表示する．図 2 において (1)~(3) は平行座標法およびそれに対応するクリークを表している．

なお著者らの実装では，図 2 の画面右側におけるノード群の配置に多次元尺度法 (MDS: Multi-Dimensional Scaling) を用いている．また図 2 の画面左側の平行座標法を構成する各次元の並び順を決定するために，クリークを構成する次元群に対して巡回セールスマン問題を適用し，その経路順を次元の並び順としている．

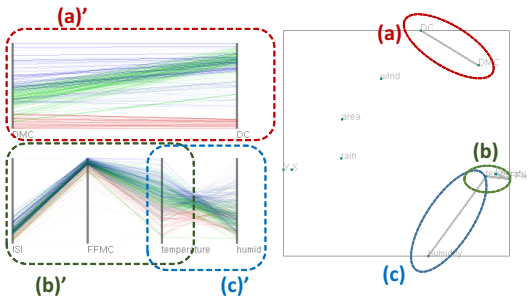

図 2　次元間距離に基づく低次元空間の抽出．(左) 平行座標法によって可視化された低次元空間群．(右) 次元散布図において平行座標法での可視化対象となる次元群が線分で連結されて表示される．

なお，現実の高次元データにはしばしば，あまりにも相関が高すぎて逆に全ての次元を可視化する必要がない，という事例も存在する．そのような場合に備えて著者らの実装では，次元サンプリング処理を実装している．この処理ではユーザが設定した閾値 $d_{remove}(d_{remove} << d_{select})$ よりも距離が小さい次元ペアのうち一方を可視化処理から除外する．

3.3　次元選択 (2) 相関ルールに基づく実装

前節で述べた低次元部分空間の抽出手法は，高次元データ中のカテゴリ型変数を全く参照していない．一方でカテゴリ型変数を用いることで，次元間距離とは別の基準にしたがって興味深い低次元空間を抽出できる．ここではカテゴリ型変数が各個体にラベルを与える役割をもつことを想定して，実数型変数とカテゴリ型変数の間の相関ルールに基づいて低次元空間を抽出する手法を示す．

この手法ではまず，実数型変数となる各次元を等分割する．ここで j 番目の実数型変数の最小値，最大値，分割数をそれぞれ $v_{j_{min}}, v_{j_{max}}, div_j$ とする．このとき j 番目の実数型変数の k 番目の区間 V_{jk} は以下のように定義される．

$$V_{jk} = [(kdiv_j)/(v_{j_{max}} - v_{j_{min}}),$$

$$((k+1)div_j)/(v_{j_{max}} - v_{j_{min}})]$$

このとき本手法では，数値属性相関ルールマイニングを適用して，以下のルールに該当する L と V_{jk} の組み合わせを抽出する．

$$L \rightarrow V_{jk} || V_{jk} \rightarrow L \qquad (2)$$

ここで L はあるカテゴリ型変数における特定のカテゴリ値（特定のラベル）を意味する．よってこのルールは，区間 V_{jk} にある個体はカテゴリ値 L を有する可能性が高い，あるいは逆にカテゴリ値 L を有する個体は区間 V_{jk} にある可能性が高い，ということを意味する．

本手法では，信頼度の閾値 t_{con} と支持度の閾値 t_{sup} をユーザが対話設定した上で，信頼度および支持度の両方が閾値を超える L と V_{jk} の組み合わせを列挙する．そして L の各々について，相関ルールが 1 個以上存在する実数型変数を列挙し，これらを座標軸とした平行座標法を生成する．図 3 にその例を示す．ここで現時点での実装では，平行座標法を構成する次元群に対して巡回セールスマン問題を適用することで，平行座標法の軸の並び順を特定している．しかし本手法において軸の有効な並び順は他にも考えられる．例えば相関ルールの類似度，あるいは相関ルールの支持度や確信度の高さで軸を並び替えることが考えられる．

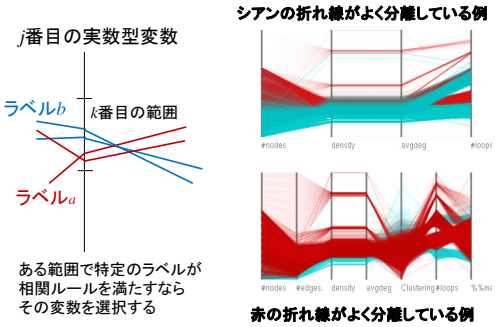

図 3　相関ルールに基づく低次元部分空間の抽出．(左) 平行座標法を構成する j 番目の次元において k 番目の区間を通過する個体に対して相関ルールを適用する．(右) 水色または赤の折れ線が表す個体群について相関ルールが 1 個以上存在する実数型変数群を平行座標法で表現した例．

3.4　対話操作

図 4 に著者らの実装のスナップショットを示す．ウィンドウ左上のラジオボタンはカテゴリ型変数となる次元の一覧となっており，ここから 1 つの次元を選ぶとその次元におけるカテゴリ値で平行座標法の折れ線を色分け表示すると同時に，ウィンドウ左下部にカテゴリ値と色の関係を一覧表示する．画面の右端にはスライダーが表示されている．このスライダーは d_{remove} および d_{select} を調節するため，あるいは t_{con} および t_{sup} を調節するために用いられる．

描画領域の右半分には実数型変数の次元をノードとしたグラフが描画される．スライダーを動かすことでグラフは対話的に更新される．加えて，描画領域上のドラッグ操作によって低次元空間を構成する次元群を手動選択する機能も有する．

描画領域の左半分には平行座標法が表示される．前述のラジオボタンを押したとき，スライダーを動かしたとき，描画領域右側のグラフの一部をドラッグ操作で選択したときに平行座標法が再描画される．

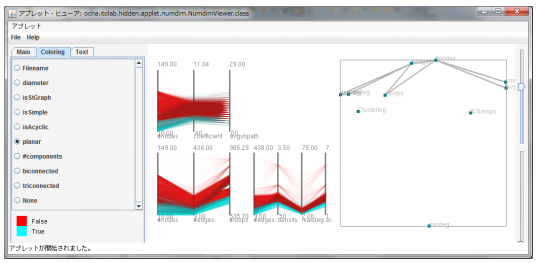

図 4　著者らの実装のスナップショット．

4　適用事例

著者らの論文 [3] では「航空機設計の多目的最適化」と「医療画像の特徴量分析」の 2 種類の適用事例を紹介しているが，本稿では前者について紹介する．

著者らは航空機の翼形状設計の最適化の過程を Hidden で可視化した．この事例では 72 個の設計変数により翼形状を設計し，流体力学シミュレーションにより 4 個の目的関数を算出した．この処理を多目的遺伝的アルゴリズムによって反復することで 776 個

のパレート解を得た．この結果から，776 の個体（= 76 次元ベクトル）を有する高次元データとして可視化した．次元選択には次元間相関を用いた．

本章では設計変数を $dv_{00} \sim dv_{71}$ と記述する．この中でも以下の 6 種類の設計変数は最適解の発見に特に重要な設計変数であることが知られている．

- dv_{00}, dv_{01}: 内翼および外翼のスパン長
- dv_{02}, dv_{03}: 後進角
- dv_{04}, dv_{05}: 翼根の翼弦長

他の設計変数には以下が含まれる．

- $dv_{06} \sim dv_{25}$: 翼の反りに関する変数
- $dv_{26} \sim dv_{32}$: 翼の捻りに関する変数
- $dv_{33} \sim dv_{71}$: 翼の厚さに関する変数

4 個の目的関数は以下のとおりである．

- CD_t: 遷音速巡航の抵抗係数
- CD_s: 超音速巡航の抵抗係数
- M_b: 超音速巡航時の翼根にかかる曲げモーメント
- M_p: 翼先端部にかかる捻りモーメント

このデータを可視化した結果を図 5 に示す．画面右側の 2 か所（(a) と (b)）にて強い相関を示す変数群が固まっている．図 5(a) には説明変数 dv_{00}, dv_{01}, dv_{04}, dv_{05} および目的関数 CD_t, M_b の合計 6 個の変数が見られる．図 5(b) には説明変数 dv_{02}, dv_{03} および目的関数 CD_s,M_p の合計 4 個の変数が見られる．

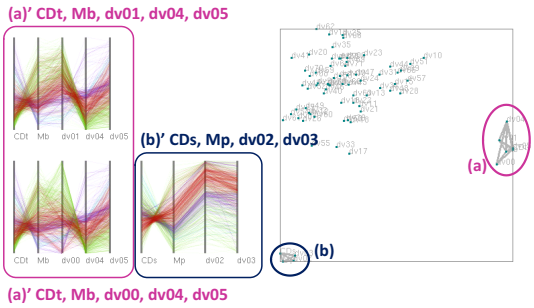

図 5　適用事例 (1)．強い相関を有する 6 個の変数が (a) および (a') で可視化されている．また別の強い相関を有する 4 個の変数が (b) および (b') で可視化されている．

続いて閾値 d_{select} を対話操作によって調節しながら画面左側の平行座標法を観察した．図 5(a') から，CD_t と M_b との間に負の相関があり，目的関数間のトレードオフが示唆される．同様に図 5(b') から，CD_a と M_p の目的関数間にもトレードオフが示唆される．設計変数間の関係に注目すると，図 5(a') から，dv_{00} および dv_{01} の 2 変数は dv_{04} および dv_{05} の 2 変数と負の相関を有することがわかる．また dv_{02} と dv_{03} の間に正の相関が成立するように設計変数を選ぶことがパレート解の発見につながることも示唆される．

一方で，このような強い相関は可視化する前から既知である場合も多い．むしろデータ所有者がいままで気がつかなかった弱い相関を知ることも可視化の意義であると考えられる．その観点からて閾値 d_{select} を調節し，やや弱い相関を可視化した例を図 6 に示す．この結果から著者らは，M_b と dv_{28}，dv_{41} と dv_{62}，dv_{04} と dv_{10}，dv_{10} と dv_{57}，といった組み合わせで相関が見られることを発見した．この結果についてデータ所有者と議論したところ，これらは全て未知の結果であり，航空機設計のあり方および多次元遺伝的アルゴリズムの振る舞いに関する新しい知見につながる可能性がある，とのことであった．

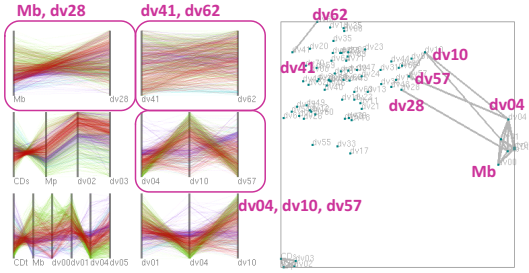

図 6　適用事例 (1)．やや弱い相関まで可視化することでデータ所有者にとって新しい知見をもたらすことができた．

5　講義科目の自由課題として

著者は提案手法をソフトウェアパッケージにして学部講義科目の自由課題にしている．著者は提案手法を Java 言語で実装しており，これとシェルスクリプトファイルを履修者にダウンロードさせて各自の環境で実行してもらっている．このプログラムを用いて，

各自で探してきたスプレッドシート形式の数値データを CSV 形式ファイルに変換して可視化し，そこから見つけた知見を説明するレポートを作成することを課題としている．授業での説明はほとんどなく，プログラムの操作方法などについては説明書を各自で読んで自習する形式の課題となっている．データから特徴や規則を発見する目を養うこと，データの入手手段や背景知識を文献として紹介すること，などの経験を兼ねた課題となっている．

なお著者の勤務先学科では Java 言語プログラミングは必修科目であるため，履修者は必要に応じてプログラムを拡張することもできる．当プログラムは 50 個程度のクラスで構成されており，Eclipse などの統合開発環境を用いた一定規模のプログラムを拡張する練習にもなる．

2015 年度からこの課題を出しているが，「プログラムを起動できない」「プログラムの操作方法がわからない」「データを表示しても何もわからない」といった質問や相談は 1 件も届いていない．このことから本プログラムは学部生の自由課題に使える程度にはわかりやすいデータ観察ソフトウェアであると考えている．

6 まとめ・今後の課題

本報告では著者が開発した高次元データ Hidden について紹介した．今後の課題として以下を考えている．

まず現在の実装に関する機能上の改良がいくつか考えられる．現時点では以下が課題としてあげられる．

- 画面左側の平行座標法の実装の改良．軸の並び順を決定するアルゴリズムの再検討，レンダリングの改良などが考えられる．
- 平行座標法に関する画面上での対話操作の実装．例えばデータの絞り込み，軸の順番の入れ替えなどの機能が考えられる．
- 画面右側のグラフの改良．例えば画面配置アルゴリズムの再検討，次元間距離の算出式の再検討などが考えられる．

また，提案手法に関する実験や検証を進める必要がある．例として，個体数や次元数が非常に大きいデータにてどこまで視認性や操作性を確保できるか，被験者がデータからどの程度適切に知識を発見できるか，といった点を実験したい．

もう一つの課題は本ソフトウェアのシステム化があげられる．例えば本稿の 4 章であげた適用事例では多目的最適化処理が終了したあとにデータをまとめて可視化しているが，最適化処理に提案手法を組み込むことで，最適化処理を対話的に調節することが可能になる．また現在進行中の研究課題では，判別分析や回帰分析といった機械学習の各処理と統合することで，機械学習の精度をあげるための対話的データ処理ツールとして提案手法を活用しようとしている．

参 考 文 献

[1] J. H. T. Claessen, J. J. van Wijk, Flexible Linked Axes for Multivariate Data Visualization, *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2310–2316, 2011.

[2] N. Elmqvist, P. Dragicevic, J. Fekete, Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation, *IEEE transactions on Visualization and Computer Graphics*, 14(6), 1141–1148, 2008.

[3] T. Itoh, A. Kumar, K. Klein, J. Kim, High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots, *Journal of Visual Languages and Computing*, 2017.

[4] S. Liu, B. Wang, P.-T. Bremer, V. Pascucci, Distortion-guided structure-driven interactive exploration of high-dimensional data, *Computer Graphics Forum*, 33(3), 101–110, 2014.

[5] K. Nohno, H.-Y. Wu, K. Watanabe, S. Takahashi, I. Fujishiro, Spectral-based contractible parallel coordinates. *18th International Conference on Information Visualisation*, 7–12, 2014.

[6] H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Arrangement of Low-dimenional Parallel Coordinate Plots for High-dimensional Data Visualization, *17th International Conference on Information Visualisation*, 59–65, 2013.

[7] C. Turkay, A. Lundervoid, H. Hauser, Representative factor generation for the interactive visual analysis of high-dimensional data *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2621–2630, 2012.

[8] X. Yuan, D. Ren, Z. Wang, C. Guo, Dimension Projection Matrix/Tree: Interactive Subspace Visual Exploration and Analysis of High Dimensional Data, *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2625–2633, 2013.

[9] Z. Zhang, K. T. McDonnel, E. Zadak, K. Muller,

Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map, *IEEE Transactions on Visualization and Computer Graphics* , 21(2), 289–303, 2015.

[10] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Scatterplot Layout for High-Dimensional Data Visualization, *Journal of Visualization*, 18(1), 111–119, 2015.