高次元データ可視化のための次元選択

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．本稿では高次元データの中から特徴的な次元だけを抽出して重点的に可視化するための次元選択手法，およびそれを利用した新しい可視化手法を紹介する．

キーワード：高次元データ可視化，次元選択，平行座標法

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特徴や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特徴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．その特徴や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．

本稿では m 次元ベクトル $a_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ で表現される n 個の個体の集合 $A = \{a_1, a_2, \ldots, a_n\}$ を高次元データと定義する．情報可視化手法の体系の提唱者として知られる Shneiderman は，情報可視化が対象とするデータ構造を 7 種類に分類し，m 次元データ $(m > 3)$ の可視化手法がその一つであると位置づけている．高次元データの可視化手法のサーベイとして Grinstein et al. [1] は 18 種類の手法を紹介している．18 種類の可視化手法は以下のように大別される．

- 2, 3 個の限られた次元だけを選んで可視化する手法．一般的な散布図など．
- 次元削減手法などを利用して高次元データを 2 次元空間に射影する手法．主成分分析，多次元尺度法．自己組織化マップなどを適用する．
- 任意の次元に関する数値分布をスプレッドシートとして網羅的に表現する手法．後述する散布図行列に加えて Table Lens などの手法が該当する．
- 高次元データを構成する各個体を線で表示する手法．後述する平行座標法に加えてレーダーチャートなどが該当する．
- ヒートマップに代表される色ベースの可視化手法．画素単位で数値を表現する手法が多い．
- アイコンやグリフなどの小物体で個体を表示する手法．ここでグリフとは，各部位に数量を割り当てて変色・変形しながら描かれた図形や記号を指す．

これらの手法のうち，高次元データを構成するすべての次元の値を可視化する手段として，散布図行列 (Scatterplot Matrix) や平行座標法 (Parallel Coordinate Plots) が特に知られており，すでに多くの学術論文で議論されている．

散布図行列は，すべての 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．図 1（左）は散布図行列を図解した例である．この例において，左から j 番目で上から i 番目の散布図は，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となっている．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．

平行座標法は図 1（右）に示すとおり，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 $(x_{i1}, x_{i2}, \ldots, x_{im})$ を座標軸上にプロットし，折れ線で結ぶ．平行座標法は，多次元データのすべ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，という点においてほかの手法より優れている．一方で，各個体を表現する折れ線が互いに絡み合うので視認性に問題が生じやすい，

いとう　たかゆき
お茶の水女子大学
〒 112–8610 東京都文京区大塚 2–1–1
itot@is.ocha.ac.jp

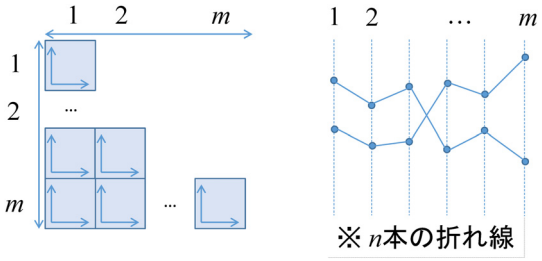

図 1 （左）散布図行列，（右）平行座標法

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すべ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すべての次元に興味深い特徴や規則性が見られるとは限らない．よって必ずしもすべ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いる．本稿では高次元データから可視化する意義のある次元を選択する手法，およびそれを搭載した高次元データ可視化手法をサーベイする．本稿ではこれ以降，高次元データから可視化する意義のある次元を選択することを「次元選択」と呼ぶ．続いて本稿では，筆者が提案した高次元データ可視化手法 Hidden について処理手順と適用事例を紹介し，今後の展望を論じる．

2. 高次元データ可視化と次元選択

高次元データから特徴的な低次元部分空間を抽出して可視化する手法が近年いくつか発表されている．例として，高次元データを限られた数の散布図で表現する手法 [2]，いくつかの低次元な平行座標法で表現する手法 [3]，散布図と平行座標法の組み合わせで表現する手法 [4] などがある．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いなかったため，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 [5]，次元削減を組み合わせて散布図で表示する方法 [6]，平行座標法を用いる方法 [7] などが提案されている．

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 [8, 9] 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 [10] も提案されている．

以上を総合すると，次元選択手法を搭載した高次元データ可視化手法の特徴には

- 適切な数の低次元部分空間を構成し，適切な数の散布図や平行座標法で表示する
- 次元選択のために対話操作手法を搭載する
- 次元間の類似度や相関を一覧する散布図やグラフを表示する

といったものがある．これらをすべて満たし，かつ次元間の類似度や相関以外の基準として相関ルールに基づいた次元選択手法もあわせて搭載した手法として，次節では筆者自身による Hidden という可視化手法を紹介する．

3. 高次元データ可視化手法 Hidden

本節では筆者らが提案している高次元データ可視化手法 Hidden [11] を紹介する．Hidden は「HIgh Dimensional Data Exploration and Navigation」の略称であり，まさに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[1,2] である．

3.1 概要

本節では 1 節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．

$$D = \{a_1, \ldots, a_n\}$$
$$a_i = (v_{i1}, \ldots, v_{im_v}, c_{i1}, \ldots, c_{im_c})$$

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 2 に示す．本手法では入力データ（図 2(1)）から実数型変数を抽出する．この実数型変数の各々を n 次元ベクタとして（図 2(2)）

[1] Java による実装 https://github.com/itot0103/hidden
[2] 筆者自身の講義科目の自由課題でのデータ分析の例 http://itolab.is.ocha.ac.jp/~itot/teaching/work/cvis/

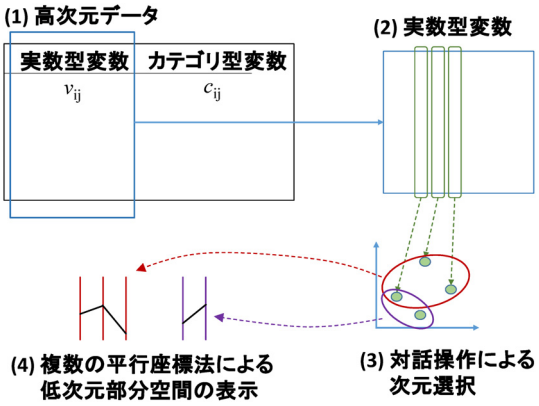

図 2　Hidden の処理手順の概要

その変数ペア間距離を算出し，その距離を保持するように散布図を生成する（図 2(3)）．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（図 2(4)）．

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．

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．現時点での筆者らの実装では，j 番目と k 番目の次元の距離を以下のとおり定義する．ここで $f_c(j,k)$ は j 番目と k 番目の間の相関係数であり，$-1 \leq f_c(j,k) \leq 1$ であるとする．

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

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 3 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．ここから本手法ではエッジで一続きに連結された次元群を抽出し，これらを平行座標法で表示する．現時点での実装では Bron–Kerboshc のアルゴリズムで抽出したクリーク（部分完全グラフ）に包括される次元群を平行座標法で表示している．図 3（左）は平行座標法によって可視化された低次元空間群を示し，図 3（右）では平行座標法を適用された次元群が線分で連結されている．ここで (a)∼(c) は平行座標法

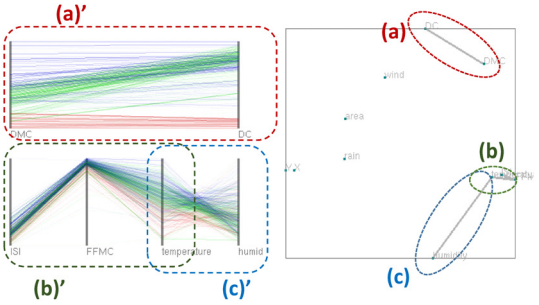

図 3　次元間距離に基づく低次元空間の抽出

での各軸に対応するクリークを表している．

なお筆者らの実装では，図 3 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 3 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することで，隣接次元間の距離 d_{jk} の総和が最小になるような並び順を採用する．

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すべ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筆者らの実装では，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従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 $(0 \leq k < div_j)$ の区間 V_{jk} は以下のように定義される．

$$V_{jk} = [(k(v_{j_{max}} - v_{j_{min}})/div_j + v_{j_{mim}},\\ (k+1)(v_{j_{max}} - v_{j_{min}})/div_j + v_{j_{min}}]$$

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．

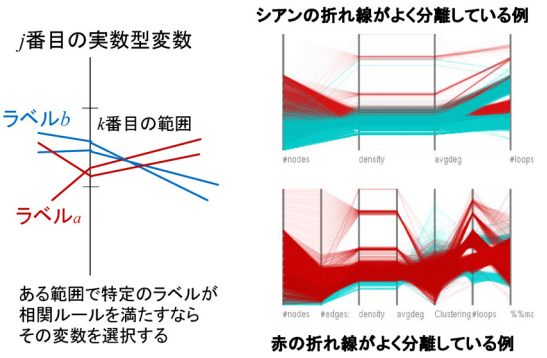

図 4　相関ルールに基づく低次元部分空間の抽出

・ある個体の l 番目のカテゴリ型変数の値が特定のカテゴリ値 L であるとき，その個体の j 番目の実数型変数の値は区間 V_{jk} に属する可能性が高い．
・ある個体の j 番目の実数型変数の値が区間 V_{jk} に属するとき，その個体の l 番目のカテゴリ型変数の値は L である可能性が高い．

本手法では，数値属性相関ルールマイニングにおける信頼度の閾値 t_{con} と支持度の閾値 t_{sup} をユーザが対話設定したうえで，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．ここで図 4（左）に示すように，本手法では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．図 4（右）は，シアン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である．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ほかにも考えられる．たと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．

3.4　対話操作

図 5 に筆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，ここから一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．

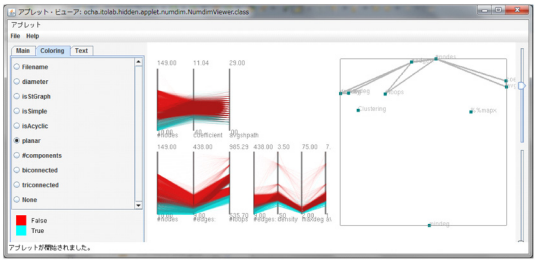

図 5　筆者らの実装のスナップショット

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．またそれに加えて筆者らの実装では，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．

3.5　適用事例 (1)：航空機設計最適化の設計変数・目的関数の相関

筆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．

本節では設計変数を $dv_{00} \sim dv_{71}$ 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

ほか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 6 に示す．画面右

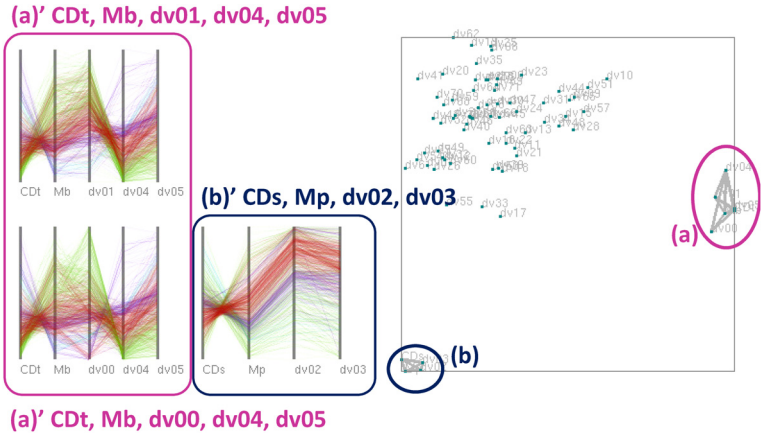

図 6　適用事例 (1) にて強い相関を可視化した結果

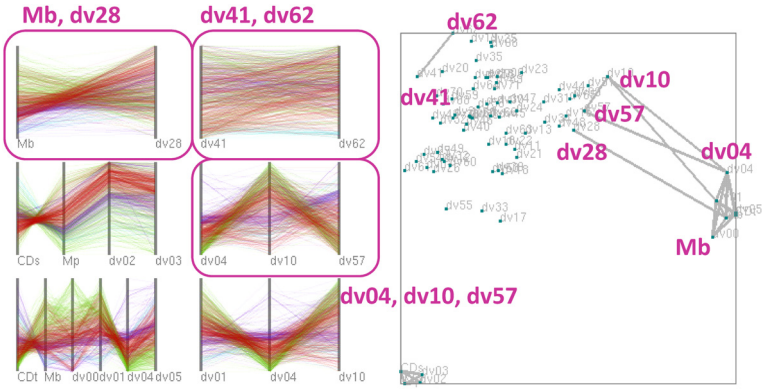

図 7　適用事例 (1) にてやや弱い相関を可視化した結果

側の 2 カ所（(a) と (b)）にて強い相関を示す変数群が集まっている．図 6(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 6(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．

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 6(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 6(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 6(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 7 に示す．この結果から筆者らは，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，これらはすべて未知の結果であり，航空機設計のあり方および多目的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．

3.6　適用事例 (2)：医療撮影画像の特徴量に関する相関ルール

筆者らは腫瘍患者の医療撮影画像を収録した LIDC/IDRI データセット[3] から 933 枚の CT 画像を選び，948 次元の画像特徴量を算出した．それとは別に医療専門家による診断結果と処置結果を各画像に付与した．

[3] http://cancerimagingarchive.net/

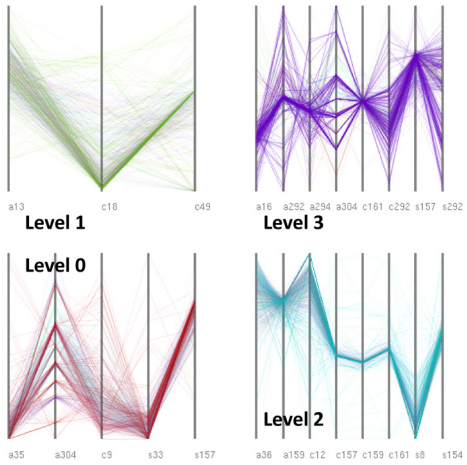

図 8　適用事例 (2) にて診断結果を用いた可視化結果

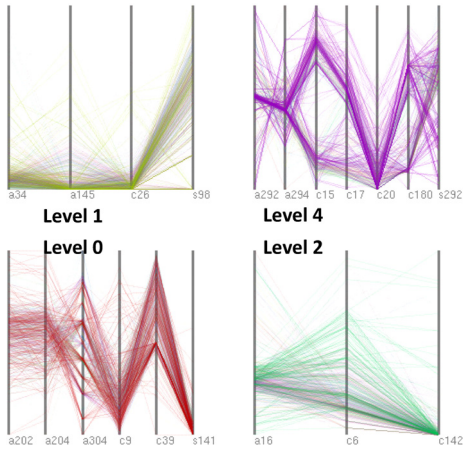

図 9　適用事例 (2) にて処置結果を用いた可視化結果

診断結果は以下の 4 レベルで付与した．

0：不明 1：良性 2：悪性初期 3：悪性かつ転移性

処置結果は以下の 5 レベルで付与した．

0：なし 1：経過観察　2：生検　3：切除　4：進行後治療

図 8 は 4 レベルの診断結果の各々に対して相関ルールを適用した可視化結果である．この結果からいくつかの特徴量が特に診断結果の判別に有用であることが示唆された．特に “coronal GLC texture cluster shade”, “coronal shape solidity”, “sagittal shape roughness” といった特徴量が有用であると考えられる．図 9 は 5 レベルの診断結果の各々に対して相関ルールを適用した可視化結果である．ただし処置結果のうちレベル 3 については相関ルールが見つからなかった．この結果からも，いくつかの特徴量が処置結果の判別に有効であることが示唆された．詳細については筆者らの論文 [11] をご参照されたい．

4.　今後の展望

高次元データを可視化してユーザがその分布を目視確認する意義，またそのために次元選択という工程を採用する展望について述べたい．

可視化の業界では Visual Analytics [12] という新しいフレームワークが欧米や中国をはじめとする主要国で採用されている点を紹介したい．Visual Analytics は可視化と分析を反復することによって大規模かつ複雑なデータから知識を発見しようというフレームワークである．Visual Analytics における可視化側の主な役割の中には「データ中の重要な部位を画面から発見して対話操作により指定し，次に適用する分析の手法を特定し，データ中の対象範囲を絞り込む」という役割がある．この「重要な部位を絞り込む」という操作には「次元を絞り込む」という操作も含まれる．このような場合に次元選択という工程が重要となる．

1 節で論じた決済情報や計測情報を扱う現場では，判別分析・回帰分析といった分析処理が適用されることが多い．このような処理において，たとえば

- 分析処理からの異常値の除外の有無，および異常値の検出基準
- 寄与度の低い次元の除外の有無，およびその判断基準

は時として分析精度を上げるための重要なファクタとなる．一方で異常値や次元を除外する判断には往々にして専門家の判断が必要な場合も多く，そのためには専門家がデータを眺める必要がある場合も多い．学術研究としては既に判別分析・回帰分析の精度向上を支援するための Visual Analytics のシステムがいくつか発表されている．

このような分析処理には近年では機械学習が適用される機会が増えているが，一方で機械学習の処理は高度化とともにブラックボックス化される傾向もあり，「機械学習がなぜこの分析結果を出したのか説明できない」という理由で採用を見送る現場も現れ始めている．そこで機械学習の振る舞いを理解するための可視化の試みも増えており，その目的で高次元データ可視化手法が適用される機会も今後は増えると考えられる．この用途においても，可視化の対象を絞り込むための次元選択は重要な技術要素となるであろう．

参考文献

[1] G. Grinstein, M. Trutschl and U. Cvek, “High-dimensional visualizations,” In *Proceedings of KDD*

Workshop on Visual Data Mining, 2001.
[2] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga and Y. Kawahara, "Scatterplot layout for high-dimensional data visualization," *Journal of Visualization*, **18**, pp. 111–119, 2015.
[3] H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga and Y. Kawahara, "Arrangement of low-dimenional parallel coordinate plots for high-dimensional data visualization," In *Proceedings of 17th International Conference on Information Visualisation*, pp. 59–65, 2013.
[4] J. H. T. Claessen and J. J. van Wijk, "Flexible linked axes for multivariate data visualization," *IEEE Transactions on Visualization and Computer Graphics*, **17**, pp. 2310–2316, 2011.
[5] N. Elmqvist, P. Dragicevic and J. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Transactions on Visualization and Computer Graphics*, **14**, pp. 1141–1148, 2008.
[6] S. Liu, B. Wang, P.-T. Bremer and V. Pascucci, "Distortion-guided structure-driven interactive exploration of high-dimensional data," *Computer Graphics Forum*, **33**, pp. 101–110, 2014.
[7] K. Nohno, H.-Y. Wu, K. Watanabe, S. Takahashi and I. Fujishiro, "Spectral-based contractible parallel coordinates," In *Proceedings of 18th International Conference on Information Visualisation*, pp. 7–12, 2014.
[8] X. Yuan, D. Ren, Z. Wang and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, **19**, pp. 2625–2633, 2013.
[9] C. Turkay, A. Lundervoid and H. Hauser, "Representative factor generation for the interactive visual analysis of high-dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, **18**, pp. 2621–2630, 2012.
[10] Z. Zhang, K. T. McDonnel, E. Zadak and K. Muller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Transactions on Visualization and Computer Graphics*, **21**, pp. 289–303, 2015.
[11] T. Itoh, A. Kumar, K. Klein and J. Kim, "High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots," *Journal of Visual Languages and Computing*, 2017. https://doi.org/10.1016/j.jvlc.2017.09.005
[12] D. A. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer and G. Melancon, "Visual analytics: Definition, process, and challenges," *Visual Data Mining* (*LNCS 4950*), pp. 154–175, 2008.