高次元データ可視化のための低次元プロット表示の改良

酒井えりか　(指導教員：伊藤貴之)

1. はじめに

高次元データの可視化手法としてよく知られた手法に，Scatter Plot Matrix (SPM) や Parallel Coordinate Plots (PCP) があげられる．SPM や PCP などの可視化手法を適用することで，高次元データを構成する次元間の相関を視覚的に表現することが可能である．しかしこれらの可視化手法をもってしても，次元が非常に高いデータにおける次元間の複雑な相関関係を網羅的に表現するのは容易ではない．

その解決の一手段として，高次元空間を低次元空間に分割し，各々の低次元空間を散布図や PCP などのプロット手法で表示する手法[1,2]が知られている．しかし高次元空間を多数の低次元プロットで可視化する際に，個々の低次元プロットが画面上で小さく表示されるために，視認性の点で問題が生じる．

本報告では PCP による低次元プロットの改良手法を提案する．本手法では PCP を構成する折れ線群の束化，また外れ値となる折れ線の強調描画などを採用することで，小さく表示した場合にも重要な点を見逃さないような可視化を実現する．なお我々は提案手法を，画面左側に低次元プロットの集合を，画面右側に次元の散布図を表示する可視化プログラム[3]の上で実装している．

2. 関連研究

低次元プロットの集合による高次元データの可視化手法に，鄭ら[1]や末松ら[2]の手法があげられる．

鄭らの手法[1]では，高次元データから所定の基準に基づき複数の 2 次元ペアを選出し，その各々に対して散布図を生成する．続いて生成した散布図の各々のペアについて，散布図間の類似度距離を算出する．この距離の集合によって構成される行列を用い，グラフ配置または次元削減の各手法を用いて散布図の理想位置を決定する．この可視化手法により各々の 2 次元間の関係を独立に分析するだけではなく，次元ペアと次元ペア間の関係を視覚的に表現し，高次元に跨る数値傾向の理解を支援する．

末松ら[2]は，散布図の代わりに低次元 PCP の集合で高次元データを可視化する手法を提案した．PCP を採用した理由は，散布図の集合による可視化では個々の散布図は 2 次元ずつしか表現できないため，多数の次元にわたって観察される相関関係の理解が難しいと考えたからである．この手法により，高次元データの中の多数の変数間にわたる相関を視覚的に把握することが可能となった．

これらの手法では，低次元空間を表現する散布図や PCP をユーザ操作によって選択することができない．それに対して本研究では，ユーザの対話操作によって PCP を選択的に表示できる可視化環境を前提としている．本研究ではこの可視化環境下において，高次元データ中の興味深い特徴を強調表示することを目標にしている．

3. 低次元 PCP の改良

我々は次元の散布図と低次元 PCP によって高次元データを対話的に可視化する手法[3]を開発している．この手法では画面の左側に低次元 PCP の集合を表示し，画面の右側に次元の散布図を表示する．

本手法では，多次元尺度法 (MDS: Multi-Dimensional Scaling) を用いて次元の散布図を生成する．2 次元間の距離を $d_{ij} = 1.0 - |c_{ij}|$ と定義する．この定義によって計算された距離行列を MDS に適用することで，相関係数の絶対値が高い次元どうしは近くに．低い次元どうしは遠くに配置される．散布図に表示される各々の点は，PCP で表示する各々の座標軸に対応する．この手法では次元の散布図上での選択操作によって，低次元 PCP を選択表示できる．具体的には，次元の散布図の上で特定の点群セットを選択することで，その点群セットに対応する次元群を低次元 PCP として表示する．

PCP による高次元データの可視化では一般的に以下の 2 点が問題になりやすい．

1) 表示する座標軸の増加に伴い，非常に横長な画面空間が必要になり，視認性の低下をもたらす．
2) 折れ線の増加に伴い，折れ線が互いに絡み合い視認性の低下をもたらす．

それに加えて，前に示した操作によって多数の点群セットを選択することで，以下の問題も生じる．

3) 多数の低次元 PCP を一画面に表示することになり，個々の PCP は画面上で小さく表示されることになってしまう．

本手法ではこれらの解決のために PCP の改良手法を適用する．

まず 1)に関する解決を論じる．本手法では次元の散布図を画面右側に配置している．この散布図は相関が高い次元を画面上で近くにプロットする．現時点の我々の実装では，ユーザによるドラッグ操作で次元の散布図上に長方形を描き，その長方形内に入った次元だけを PCP で描画する対話処理機能を有する．このようにして次元の散布図を操作することで，ユーザは主観的に興味深い部分空間を指定し，低次元 PCP を対話的に可視化できる．

また本手法では 2)3)の解決のために，形状的に類似する線分群を束ねて表示する「Bundling」という処理を適用する．Bundling は主にネットワークの可視化において広く活用されている[4]. PCP においてもこれを適用することで，折れ線同士が絡み合い視認性が低下する問題を回避できると考える．また本手法では高い相関を有する次元群を選出して PCP で表示することを

前提にすることができる．このとき，正の相関を持つ次元同士の要素を表す折れ線は平行に近く．これらを束にすることで外れ値を発見しやすくなると考えられる．さらにそうして発見された外れ値の折れ線の強調描画などを採用することで，PCP が画面上で小さく表示した場合にも重要な特徴を見逃さないような可視化を実現する．

本手法では PCP を構成する折れ線を色分け表示することを想定している．折れ線で表現される個体にラベルがつけられている場合には，ラベルで折れ線を色分けすることができる．ラベルがつけられていない場合にも，個体群にクラスタリングを適用することで，折れ線をクラスタ分類し，個々のクラスタに固有の色を割り当てることができる．このような色分け表示が適用された時に，本手法では例えば特定の色を割り当てられた折れ線群に対して Bundling を適用する，といった選択的な手段を提供する．

4. 実行例

我々は本手法を Java Development Kit (JDK) 1.7.0 で実装した．

本手法は高次元データに関する汎用的な可視化手法であり，さまざまな分野の高次元データを対象にすることができる．本報告では一例として，飛行機のデータを適用した例を示す．画面左側には 776 本の折れ線を含む PCP が表示され，右側には 76 個の点を含む次元の散布図が表示される．次元の散布図を用いたユーザ操作によって特定の次元だけを選択し，PCP でその次元群のみを可視化することができる．

図 1 には Bundling を適用する前の結果を，図 2 には Bundling を適用した結果を示す．Bundling を適用することで折れ線の絡みが減少し Bundling していない色の折れ線群の特徴が視認しやすいと言える．

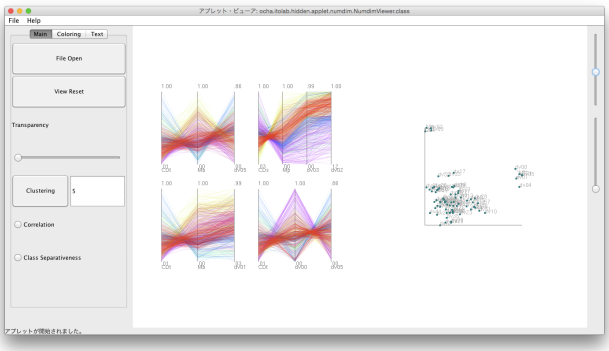

図 1　Bundling 適用前

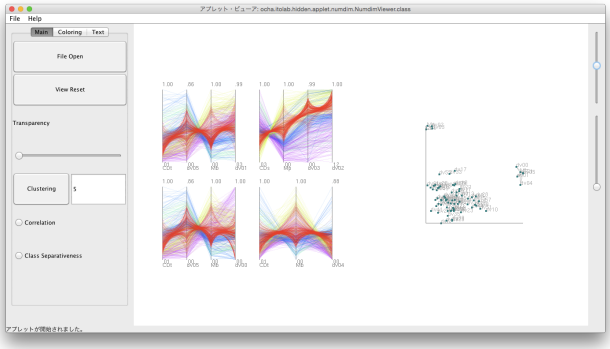

図 2　Bundling 適用結果

5. まとめと今後の課題

本報告では，高次元データを低次元プロットの集合で可視化するための PCP の改良手法を提案した．具体的には PCP を構成する折れ線群の束化や，外れ値となる折れ線の強調描画などを採用することで，各々の PCP が画面上で小さく表示した場合にも重要な特徴を見逃さないように工夫している．

今後の課題として，追加機能として，まだ色分けされていない初期状態でもある閾値を基づいて選択的に Bundling を適用する機能を開発したい．

画面左側の PCP にもユーザ操作を導入したい．また，PCP においては座標軸の並び順が視認性に大きな影響を与えるので，例えばユーザがクリックした順に座標軸を並べ替える，といった操作も考えている．

3 節にて議論した問題点の 2)に対して，Bundling 以外の解決策を考案したい．例えば外れ値となる折れ線の強調描画などを考えている．

これらの機能を開発した後に，さらに多種多様なデータを本手法に適用し，特定のデータに特化した PCP の表示方法についても検討と実装を進めたい．

参　考　文　献

[1] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Scatterplot layout for high-dimensional data visualization, Journal of Visualization, 10.1007/s12650-014-0230-5, 2014.

[2] H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Arrangement of Low Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization, 17th International Conference on Information Visualisation (IV2013), pp. 59-65, 2013.

[3] T. Itoh, et al., High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots and Scatterplots, (国際会議査読中).

[4] D. Holten, Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data, IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No. 5, pp. 741-748, 2006.

.