1B2-R-3-4

低次元プロットの集合による高次元データ可視化の一手法(2)

末松 はるか[*1] 鄭 雲珠[*1] 伊藤 貴之[*1] 藤巻 遼平[*2] 森永 聡[*3] 河原 吉伸[*4]
Haruka Suematsu Yunzhu Zheng Takayuki Itoh Ryohei Fujimaki Satoshi Morinaga Yoshinobu Kawahara

[*1] お茶の水女子大学大学院人間文化創成科学研究科 Graduate School of Humanities and Sciences, Ochanomizu University
[*2] NEC ラボラトリーズアメリカ NEC Laboratories America, Inc.
[*3] NEC 情報メディアプロセッシング研究所 NEC Information and Media Processing Research Laboratories
[*4] 大阪大学産業科学研究所 The institute of scientific and industrial research, Osaka University

Parallel Coordinates Plot (PCP) is a very popular visualization technique for multi-dimensional datasets. The PCP is useful to explore and analyze the features and the structures in the datasets. It is also usable to find out the correlation or clusters of datasets, but it becomes less effective for datasets which contain large number (i.e. more than 100) of dimensions. This paper proposes a novel PCP-based visualization technique, which is applicable to the datasets containing large number of dimensions. The proposed technique classifies the dimensions of datasets into several groups, and then visualizes the groups as a set of sub-PCPs on a screen. By displaying in this manner, the technique represents the correlations among the dimensions even if the datasets contains large number of dimensions.

1. はじめに

高次元データの可視化には，全ての次元ペアから生成される散布図を格子状に並べて表示する Scatter Plot Matrix(SPM)や，平行な座標軸を次元の数だけ並べて折れ線で高次元数値を表現する Parallel Coordinates Plot(PCP) [Inselberg 1990] がよく使用される．これらの可視化手法は 10 次元程度のデータには有効である．しかし，数十~数百におよぶ高次元データの可視化に適用させる場合，SPM であると各々の散布図が画面上で非常に小さくなり，PCP であると非常に横長な空間が必要となる．双方ともに，視認性の高い状態でディスプレイ上に表示するのが困難である，という問題が生じる．

また別の問題として，多数の次元間にわたる相関関係の表現の問題がある．高次元データではしばしば，任意の次元が多数の他の次元と同時に相関を持つことが想定される．一方で PCP で表現可能な次元間の相関関係は，表示される座標軸の配列に強く依存する．そのため，単一の PCP による高次元データの可視化では，時として多彩な相関の発見を妨げてしまう場合があり，相関関係を視覚的に発見するという重要な目的において十分ではない．

これらの問題に対して鄭ら [Tei 2012] は，所定の基準を満たす次元ペア群から生成された散布図群を画面配置する手法を提案している．しかし散布図の集合による可視化では，個々の散布図は 2 次元ずつしか表現しないため，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場合がある．

本報告では鄭らの手法を拡張して，散布図のかわりに低次元 PCP の集合で高次元データを可視化す手法を提案する．本手法では条件付き独立性の概念を適用して低次元 PCP を生成し，この PCP 群を配置することでデータ全体の一覧表示を実現する．本手法によって，高次元データ中の多数の次元にわたる相関を視覚的に把握することが容易になると考える．

連絡先：末松 はるか，お茶の水女子大学大学院人間文化創成科学研究科，〒112-8610 東京都文京区大塚 2-1-1，haruka@itolab.is.ocha.ac.jp

2. 関連研究

高次元データの可視化には Scatter Plot Matrix(SPM)が旧来から広く用いられてきた．SPM は画面を格子状に分割してできた各領域に 2 次元散布図(Scatter Plot)を配置する方法である．散布図は 2 次元間の相関を把握する時に有効であるが，3 次元以上の相関を同時に見たい場合に適さない．また次元数が高くなると，各散布図の画面上での大きさが非常に小さくなり視認性が低下する，という本質的な限界がある．

Parallel Coordinates Plot(PCP) [Inselberg 1990]は，各次元を鉛直な座標軸として水平方向に並べて一覧表示する高次元データの可視化手法である．隣接する 2 軸間の折れ線が平行に近ければ正の相関を示し，交差している場合は負の相関を示す．各次元を表す座標軸を一覧表示することで，3 次元以上にわたる高い相関を同時に把握することが可能となり，本研究に

適していると考えられる.

PCP の問題点には以下のようなものがあげられる.

1) プロット数が多い場合に, 散布図と比べて可視化結果が煩雑になり視認性が低下しやすい上に, 描画速度も低下しやすい. この問題については, プロットのクラスタリングやサンプリングに関する研究が多数報告されている.

2) 次元間の相関関係は複雑であり, その全ての単一の PCP で表現するのは限界がある. 一方で, 観察可能な次元間の相関関係は, 表示する座標軸の整列順に強く依存する. そのため座標軸を効果的に並べ替えることで, 少しでも多くの相関関係を単一の PCP で表現する研究も報告されている. しかしそれでも, 単一の PCP での表現には限界があることには変わりない.

3) 次元数の増加に伴い, 非常に横長な画面空間を必要とする.

本報告の提案手法は, 上記の問題点のうち 2)3)の解決を試みるものである. 低次元 PCP 集合の生成と画面配置によってこれらの解決を試みる, という手法は我々の調査した限りではまだ見当たらない.

3. PCP 集合の生成

本章では低次元 PCP を生成するための各処理について論じる.

3.1 条件付き独立の判定

低次元 PCP には以下の条件を満たすことが求められる.

(1) 同一の低次元 PCP に属する次元間には高い相関性を有する.

(2) 異なる低次元 PCP に属する次元間には低い相関性を有する.

ただし、単純に相関のみを用いてグループ分けを行おうとすると, 見かけ上の相関による単純な(同じような挙動を示す)グループのみが得られてしまう可能性がある. 従ってここでは, 条件付き独立性に基づいて, 変数のグループ分けを行う. 条件付き独立性の判定は, n 個の変数において互いに重ならない任意の 3 つの部分集合 XA ,XB ,XC を与えた時, 次のように行われる.

1. それらを用いて計算される条件付き相互情報量(式(i))を計算する.

$$I(X_A, X_B \mid X_C) \quad \text{(i)}$$

2. もしその値が極めて 0 に近い場合, 変数集合 A と B は, C を与えた時に条件付き独立となると判断される.

3.2 クリーク選択による PCP 集合の生成

条件付き独立性を判定し, 任意の変数集合を与えた時, 互いに独立にならない 2 変数が異なるグループに入るように, [Chechetka 2008]の手法を用いてグループ分けを行い, 低次元 PCP を生成する.

例として 4 変数{ X1 ,X2 ,X3 ,X4}の場合を考える. 条件とする変数集合 XC={X3}として条件付き相互情報量を計算した時, XA={X1 ,X2}, XB={X4}となる条件付き相互情報量が 0 になった場合, 4 変数は{X1 ,X2 ,X3}と{X3 ,X4}に分割される. {X1 ,X2 ,X3}と{X3 ,X4}のいずれも他の条件のもとでこれ以上分解されなくなった時, 図 1 のようにこの 4 変数はこの 2 つのクリークに分けられ, (1), (2)の条件を満たす低次元 PCP 生成を実現する.

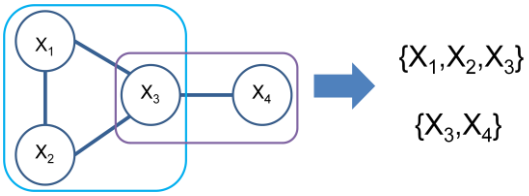

図 1 低次元 PCP 集合の生成

3.3 各 PCP における軸順番の決定

現時点での我々の実装では貪欲的に座標軸の順番を決定している. PCP の特徴として, 隣接する二軸間の相関が最も視認しやすく, 相関のある軸が隣接して配置されることが望ましいため, まず各 PCP の軸中で, 最も相関が高い二軸を算出し, 配置を行う. 決定した二軸の右から順にそれ以外の軸を一つずつ配置し, 軸の配置を決定する. この処理については既存の軸順番選択手法によって改善が可能である. 一例として, 次元をノードとするグラフに巡回セールスマン問題を適用して軸順番を決定する手法 [Zhang 2012] の適用が考えられる.

4. PCP の画面配置最適化

続いて, 前章に示した手法で低次元 PCP 群を画面配置する. 本手法では鄭らの手法 [Tei 2012] と同じく, 力学モデルと空間充填モデルを併用した画面配置手法 [Itoh 2009]を適用する.

画面配置の処理手順は以下のとおりである.

[理想座標値算出 a] 任意の低次元 PCP ペアに対して, 次元の共有率に基づいて類似度距離を算出する. 例として, 任意の 2 つの PCP, A, B それぞれの軸を Ai(1≤ i ≤n), Bj(1≤ j ≤m)と与えた時, A, B の類似度距離(式(ii))を Ai, Bj の相関係数(式(iii))を用いて算出する.

$$(A,B) = 1 - \frac{1}{|A||B|\sum_{A_i \in AB_j \in B} \sum |R_{AB}|} \quad \text{(ii)}$$

$$R_{A_iB_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_i-\bar{A})(B_j-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\sqrt{\sum_{j=1}^{m}(B_J-\bar{B})^2}} \quad \text{(iii)}$$

この類似度距離を全ての低次元 PCP ペアについて算出することで距離行列を生成する. そしてこの距離行列に対して Isomap などを適用して次元削減を実行し, その結果から 1,2 次元目を座標値とすることで, 各 PCP の理想座標値を算出する.

[理想座標値算出 b] 任意の低次元 PCP ペアに対して, 次元の類似度が一定以上であるものをエッジで連結することで, 低次元 PCP のグラフを生成する. このグラフに対して, エッジにバネを仮想した力学モデルを適用することで, 各 PCP のり総座標値を算出する.

[空間充填] 上記のいずれかの手法によって算出された理想座標値を参照しながら, 低次元 PCP が互いに干渉しないように, また画面占有面積の拡大を抑えるように, 空間充填モデルにより各 PCP の配置を調整する.

[理想座標値算出 b]を適用した場合の処理手順を図 2 に示す.

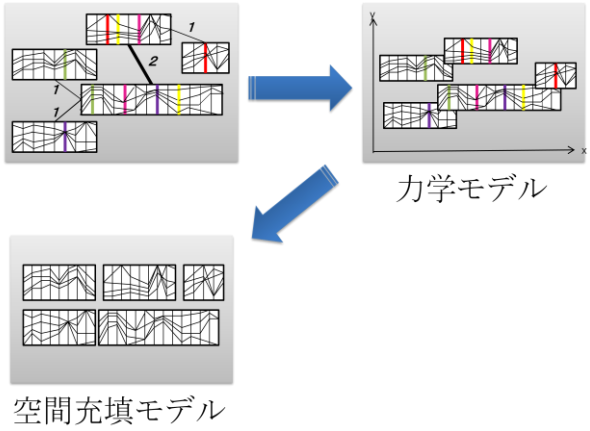

図 2 配置過程

5. 実行例

我々は本手法のうち PCP 選択および次元削減を MATLAB および Python 2.7 で実装し, 力学モデルと空間充填モデルを Java Development Lit (JDK) 1.6.0 で実装した. そして, UCI Machine Learning Repository で公開されている Segmentation Data [Data] をサンプルデータとして可視化を試みた. このデータは, 7 枚の画像を 3×3 ピクセルの画像領域で分割して, 各領域について全 19 次元の特徴量(画像の領域重心, RGB 値, 彩度など)を算出したものである. 結果として本手法では, 次元数 19, プロット数 210 の高次元データとして扱っている.

図 3 に可視化結果の例を示す. ここでは 7 枚の画像から算出された特徴量が 7 色に色分け表示されている. 各低次元 PCP は 2~12 次元で構成されていて, エッジの総数は 20 本である. 我々の実装では, 拡大縮小, 平行移動, 各々の低次元 PCP の次元表示や図 4, 5 のような折れ線のラベル別表示, 透過などの各機能を搭載している.

数十次元以上の高次元データを可視化するにあたり, 従来の PCP 手法では非常に横長なグラフが生成され, 現在普及しているディスプレイでの表示が困難な場合があった. それに対して本手法では, 高次元データを低次元のサブグループに分割し各々を PCP として表示して, それをディスプレイサイズに合わせて配置する. そのため, 視認性の高い状態でのデータ全体の可視化が容易となる.

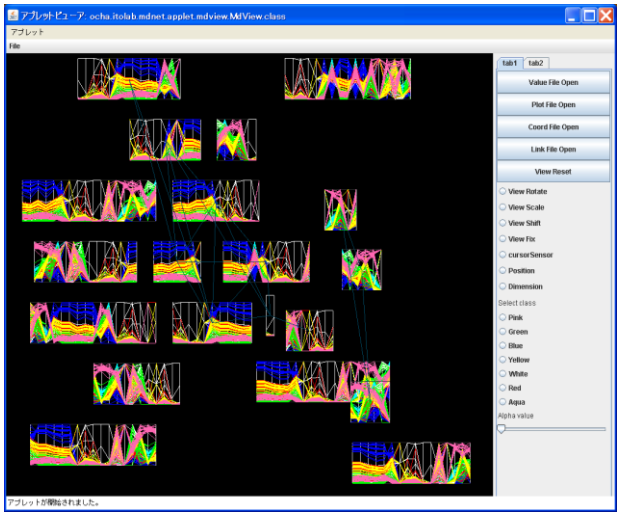

図 3 実行結果

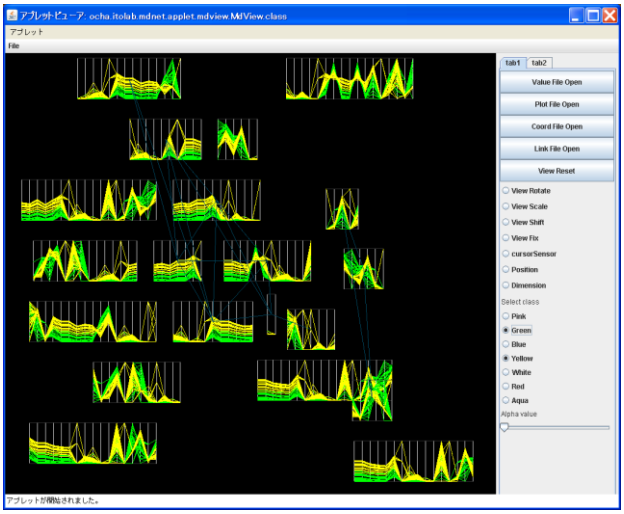

図 4 ラベル別選択（黄色と緑色の折れ線を選択表示）

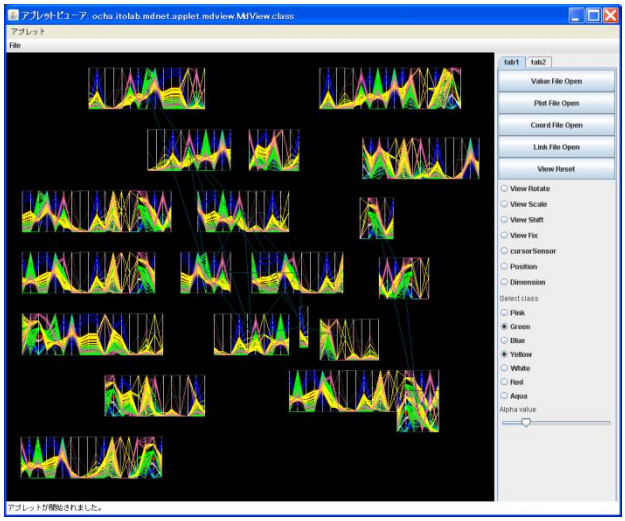

図 5 透過（黄色と緑色の折れ線以外を透過表示）

また，次元の共有度が高い低次元 PCP 同士が近くに配置されるため，低次元 PCP 同士の比較も容易となる．例えば，図 6 は黄色と緑色の折れ線で描画されるラベルを選択した結果である． 上の低次元 PCP と下の低次元 PCP どちらにも存在する次元 9 は輝度を示す．上の低次元 PCP では明度を示す次元 16 が次元 9 との間で平行となることから正の相関を持ち，下の低次元 PCP でも同様に RGB 値の G 値と R 値を示す次元 12, 10 が次元 9 と正の相関を持つ．G 値が高ければ R 値も高いということから，輝度が高ければ黄色の割合も高いことが分かる．また，輝度が高ければ明度も高いことが上の低次元 PCP から示されるので，2 枚の画像から全体的に輝度が高ければ明度と黄色の割合が高いことが分かる．このように一つの PCP からは判断しづらい，枝分かれを伴う相関の発見を本手法ではより視認出来る形でユーザーに提供する．

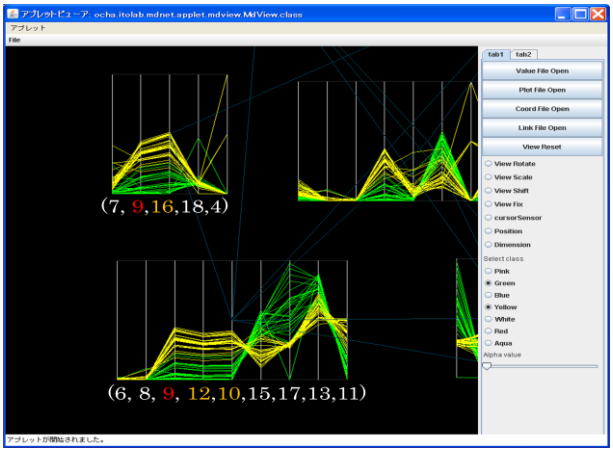

図 6 低次元 PCP 間の考察

6. まとめと今後の課題

本研究では，高次元データの条件付き独立性に着目して低次元 PCP 群を生成し，次元共有度の高い低次元 PCP が近くになるように一画面に配置する，という考えに基づいた高次元データ可視化手法を提案した．

今後の課題として，低次元 PCP 生成結果の客観的検証， 配置結果の客観的検証， 低次元 PCP の視認性に関する主観評価などがあげられる．また，既存の PCP 手法で既に採用されている諸機能，例えば折れ線のサンプリングや，適応的な座標軸の反転などを実装したい．

参考文献

[Inselberg 1990] A. Inselberg, B. Dimsdale: Parallel Coordinate: A Tool for Visualizing Multi-Dimensional Geometry, IEEE Visualization, 361-370, 1990.

[Tei 2012] 鄭, 末松, 伊藤, 藤巻, 森永, 河原, 低次元プロットの集合による高次元データの可視化(1), 2012 年度人工知能学会全国大会, 1B2-R-3-3.

[Chechetka 2008] A. Chechetka, C. Guestrin: Efficient Principled Learning of Thin Junction Trees, Advances in Neural Information Processing Systems 20, 273-280, 2008.

[Zhang 2012] Z. Zhang, K. T. McDonnell, K. Mueller, A Network-Based Interface for the Exploration of High-Dimensional Data Spaces, IEEE Pacific Visualization Symposium, 2012.

[Itoh 2009] T. Itoh, C. Muelder, K.-L. Ma, J. Sese: A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, IEEE Pacific Visualization Symposium, 121-128, 2009.

[Image Segmentation Data Set] UCI Machine Learning Repository, Image Segmentation Data Set, http://archive.ics.uci.edu/ml/datasets/Image+Segmentation