1B2-R-3-3

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(1)

A high-dimensional data visualization techinique using a set of low-dimensional plot(1)

鄭 雲珠[*1] 末松 はるか[*2] 伊藤 貴之[*3] 藤巻 遼平[*4] 森永 聡[*5] 河原 吉伸[*6]
Yunzhu Zheng Haruka Suematsu Takayuki Itoh Ryohei Fujimaki Satoshi Morinaga Yoshinobu Kawahara

[*1*2*3]お茶の水女子大学大学院 人間文化創成科学研究科 [*4]NEC ラボラトリーズアメリカ
Graduate School of Humanities and Sciences, Ochanomizu University NEC Laboratories America

[*5]日本電気 情報メディアプロセッシング研究所
NEC Information and Media Processing Research Laboratories

[*6]大阪大学産業科学研究科
The Institute of Scientific and Industrial Research (ISIR), Osaka University.

There have been many visualization techniques for high-dimensional datasets. ScatterPlotMatrix (SPM) is one of the most popular techniques which represents overall features of high-dimensional data spaces. However, looking for interesting patterns by too many Scatterplots is so hard, because they are displayed very small in a limited display space. In this paper, We propose a method for high-dimensional data visualization technique by selecting a set of meaningful Scatterplots and effectively arrange them onto a display space. Our current implementation selects Scatterplots which entropy of classes are lower and therefore particular classes are separated on the display space. It then calculates distances between pairs of Scatterplots base on their similarities of the entropy values, so that the technique can place similar Scatterplots closer. The technique calculates ideal positions of the Scatterplots on the display space from their distances, and arranges their positions by applying a rectangle packing algorithm.

1. はじめに

情報可視化は多種多様なデータの直感的理解を支援するツールとして，またデータからの知識抽出のツールとして有用である．情報可視化が対象とするデータのうち高次元データは，散布図を用いて可視化されるのが最も一般的である．多くの場合において散布図は，直交座標系を構成する各座標軸に，多次元データを構成する任意の次元を割り当て，多次元データの各要素を直交座標系にプロットする．しかしこの表現による可視化結果は，与えられた高次元データのうちわずか 2,3 個の次元で構成される空間を示しているに過ぎない．この問題に対して旧来から，以下のようなアプローチが試されてきた．

- 次元削減手法を適用し，高次元空間全体にわたるプロット間の距離関係や密度分布を保持するように低次元空間で可視化する手法．この手法では各次元の数値を直接読み取ることが難しい．
- 全ての 2 次元ペアに対して散布図を作成し，それをマトリクス状に並べて一覧表示する手法．以下 SPM(ScatterPlot Matrix) と称する．高い次元数を有する多次元データを一画面に全体表示しようとすると，個々の散布図は画面上では非常に小さくなって視認性が低下する，という問題点がある．
- ユーザインタフェースやアニメーションによって高次元空間の理解を支援する手法．例えば Rolling the Dice [Elmqvist08] では，表示対象となる次元を対話操作とアニメーションによってシームレスに切り替えることで，ユーザのメンタルマップを保持しながら高次元空間を可視化する．

本研究では SPM と同様に，任意の 2 次元ペアから生成された複数の散布図を一覧表示することを考える．ここで多くの場合において，2 次元ペアの中には可視化する意義が高いものと低いものがある．そこで本報告では，可視化する意義が高い 2 次元ペアだけを選出して複数の散布図を生成し，これを散布図間の類似性や相関性に基づいて画面上に配置する一手法を提案する．本手法ではまず，高次元データから所定の基準に基づいて複数の 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．ここで所定の基準とは例えば，2 次元間の相関係数の絶対値の高さ，あるいはプロットにクラス情報がある場合にはその散布図上での分離度の高さ，などが有効であると考えられる．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て，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．さらに，この理想位置を参照して，長方形充填アルゴリズム [Itoh06] [Itoh09] を用いて散布図の位置を決定する．

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分析できるようになり，高次元にまたがる数値傾向の理解を支援できると考えられる．

2. 関連研究

2.1 散布図による高次元データの可視化

本章では散布図に関する関連研究について論じる．

高次元空間全体にわたるプロット間の距離関係や密度分布を把握するには，次元削減に基づく座標変換が有効である．高次元データの可視化における次元削減にはその投影手段が大きな

連絡先: 鄭雲珠，お茶の水女子大学大学院人間文化創成科学研究科，東京都文京区大塚 2-1-1, zheng.yunzhu@is.ocha.ac.jp

鍵となる．Leban ら [Leban05] は散布図のための有意義な投影を導くツール VizRank を開発した．しかし依然として，高次元空間を構成する個々の値を読み取りにくいという問題は残っている．

SPM は，高次元データを構成する次元間の関連性を一覧できる手法として，既に広く用いられている．散布図行列は 10 次元程度のデータの分析や観察に適しているが，それよりも次元数が多くなると，個々の散布図は画面上で非常に小さくなってしまい，視認性が低下するという問題がある．そこで本研究のように，多数の散布図から少数の意味ある散布図を選んで可視化するというアプローチが考えられる．Wilkinson ら [Wilkinson05] は SPM の中から少数の意味ある散布図を選ぶ手法を提案している．Sips ら [Sips09] はクラスの一貫性による分布する視点で多数な散布図中から有意義な散布図だけを強調して可視化している．以上のような手段で散布図を選別したとしても，従来の SPM のようなレイアウトでは，類似する散布図間の関連性を見つけにくい．

2.2 長方形の空間充填手法を用いた情報可視化

伊藤ら [Itoh06] は長方形の空間充填に基づく階層型データの情報可視化手法を提案している．この手法は「できるだけ小さな画面空間に大規模な情報を，碁盤状に整理された形式で表現する」というデザイン方針に基づいて階層型データを画面配置する．この手法では図 1 に示すように，階層型データを構成する葉ノードをアイコンで表現し，枝ノードを入れ子状の長方形の枠で表現している．このような画面配置を実現するために，まず最下位階層に属する葉ノードを配置し，その配置結果を囲む枠を生成し…という処理を下位階層から上位階層に向かって反復する．

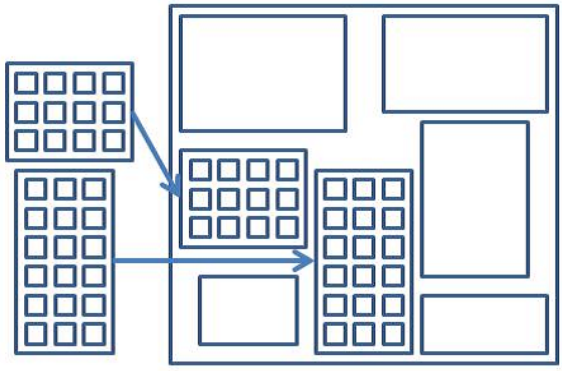

図 1: 長方形充填

この手法において 1 階層を構成するノード群を画面配置するためのキー技術は，ノードを表現する長方形群が以下の条件をできるだけ満たすように各長方形の位置を決定する技術である．

条件 1: 既に配置されている長方形と干渉しない．

条件 2: 配置領域の面積拡大量が最小である．

条件 3: 各長方形に画面上の理想位置が指定されているときには，その理想位置との距離が最小である．

この条件を満たすために，この手法ではまず長方形の配置順を決定する．そして各長方形に対して，[条件 1] を満たすいくつかの位置を画面配置の候補とし，その各々の位置において [条件 2][条件 3] をどの程度満たすかを判定するためのスコアを算出する．そして，このスコアが最良である位置に長方形を配置する．この処理を配置順とおりに反復することで，全ての長方形の位置を決定する．

さらに伊藤らは，この手法を拡張したグラフ可視化手法 [Itoh09] を提案している．この手法では，グラフのエッジにバネを仮想した力学モデルにより，各ノードの画面上の理想位置を決定し，続いて上述の長方形空間充填手法を用いて各ノードを表現する長方形の位置を調整する．

3. 相関性に基づく散布図集合の選択

提案手法の前半では，高次元データの全ての 2 次元ペアを用いて生成した散布図の中から，可視化する意義のある散布図を選択する．現時点での我々の実装では，クラスエントロピーに基づく下記の処理手順で散布図を選択している．

まず全ての 2 次元ペアについて、以下の式で全クラスのクラスエントロピーを算出する。

$$H_{all}(d_1,d_2) = -1/N \sum_{n=1}^{N}\sum_{c=1}^{C} p(y_n = c|x_n^{d_1,d_2}) \log p(y_n = c|x_n^{d_1,d_2}) \quad (1)$$

ここで x_n は n 番目のプロット、y_n は n 番目のラベルを表し、N および C はそれぞれプロット数およびクラス数を表す。また $x_n^{d_1,d_2}$ はプロット x_n について d_1 番目の次元と d_2 番目の次元を取り出してできる 2 次元ベクトルを表す。この式によって算出されるクラスエントロピーは、d_1 番目の次元と d_2 番目の次元によって生成される散布図において、クラスラベルが全体的にどの程度分離されているかを示すものである。

続いて以下の式により、c 番目のクラスに関するクラスエントロピーを算出する

$$H_c(d_1,d_2) = -1/N \sum_{n=1}^{N} p(y_n = c|x_n^{d_1,d_2}) \log p(y_n = c|x_n^{d_1,d_2}) + p(y_n \neq c|x_n^{d_1,d_2}) \log p(y_n \neq c|x_n^{d_1,d_2}) \quad (2)$$

これを c の各々の値 $(1 \leq c \leq C)$ について計算する。この式によって算出されるクラスエントロピーは、d_1 番目の次元と d_2 番目の次元によって生成される散布図において、c 番目のクラスがどの程度分離されているかを示すものである。この値に基づいて散布図間の類似度を算出することで、同一のクラスがよく分離されている散布図を画面上で近くに配置することができる。

以上によって算出された H_{all} および H_c の各値について、小さい順に k 位以内となる次元ペアを選出する。この選出結果から重複を取り除いた次元ペア群に対して、散布図を生成する。

4. 散布図集合の画面配置最適化

前章に示した手法で選択した散布図の集合を配置するにあたり，本手法では力学モデルと空間充填モデルを適用した画面配置手法 [Itoh09] を適用する．本手法では画面配置の第一段階として，各散布図の画面上の理想位置を決定する．我々は現時点で，以下の 2 種類の理想位置決定手法を実装している．

- 散布図間の類似度を距離として保持するような次元削減による配置．

- 類似度が高い散布図をエッジで接続して構築されるグラフの配置.

続いて第二段階として，理想位置として与えられた座標値を参照しながら，長方形の空間充填によって位置を適正化する．

以下，各処理について説明する．

4.1 次元削減による理想位置決定

前章で選択した散布図の各々について計算された $H_{all}, H_1, ..., H_C$ の各値を、以後 $C+1$ 次元のベクトルとして扱う。本章ではこれをエントロピーベクトルと称する．ここで散布図の個数を M とし，エントロピーベクトルの次元数を $dim = C+1$ とする．そして j 番目の散布図のエントロピーベクトルを s_j とし，その i 次元目の値を s_{ji} とする．このとき我々の実装では，以下の 3 種類の距離算出式のいずれかにより，エントロピーベクトル s_a,s_b 間の距離を算出する．

$$dist(s_a, s_b) = \sum_{i=1}^{dim} |s_{ai} - s_{bi}|$$

$$dist(s_a, s_b) = (\sum_{i=1}^{dim} |s_{ai} - s_{bi}|^2)^{1/2}$$

$$dist(s_a, s_b) = max_{1<=i<=dim} |s_{ai} - s_{bi}| \quad (3)$$

以上の距離算出手法を，選択された散布図の全てのペアについて適用することで，$M \times M$ の距離行列が生成される．我々の実装では，この距離行列に Isomap を用いて次元削減を行い，その 1,2 次元目を採用することで画面上の理想位置を決定する．

4.2 グラフへの力学モデルによる理想位置決定

前節で示した距離算出結果から，距離行列を生成する代わりに，距離が閾値 $dist$ 以下である散布図ペア間にエッジを生成してできるグラフを考える．このグラフを画面配置するための最も汎用的な手法のひとつに，力学モデルを適用した手法がある．力学モデルに基づく手法では，各エッジにバネを仮想し，反復的解法によってバネの力に関する運動方程式を解くことで，散布図間の距離を適正化する．以上の処理によって得られた各散布図の位置を，画面上の理想位置とする．

4.3 長方形充填

次元削減または力学モデルを用いて算出された理想位置を参照しながら，長方形の空間充填手法 [Itoh06] を用いて，各散布図の画面上の最終位置を決定する．現時点での我々の実装では，散布図は階層化されていないため，各々の散布図に対して 2.2 節で示した長方形配置処理を適用している．一方で，散布図を階層化して，図 1 に示すような入れ子状の階層表現を適用することも，原理的には可能である．

5. 実行例

我々は本手法のうち散布図選択および次元削減を Python 2.7 で実装し，力学モデルと空間充填モデルを Java Development Kit (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 枚の画像を各々30 個のブロックに分割して，各ブロックにおいて 18 次元の特徴量を算出したものである．つまり，このデータは次元数 18，要素数 210 の高次元データである．

図 2 および図 3 に，複数の散布図でこのデータを可視化した例を示す．いずれの例にも，類似する可視化結果を得た散布図が画面上で近接していることが見て取れる．例えば図 2 では，画面右上部に青いクラスのプロットが明確に分離した散布図が多く見られ，画面右下部に緑のクラスのプロットが明確に分離した散布図が多く見られる．また画面左側には正または負の強い相関性を有する 2 次元が選ばれた散布図が多く見られる．図 3 でも同様に，同一のクラスが明確に分離した散布図どうしが近接したり，同様な相関性を有する散布図どうしが近接しているのが随所に見られる．

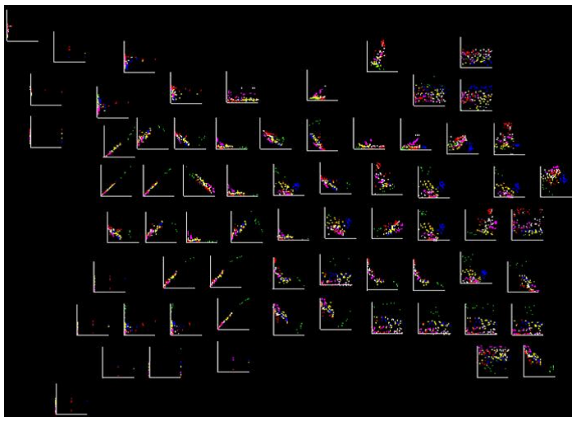

図 2: 68 個の散布図を選択して可視化した結果.

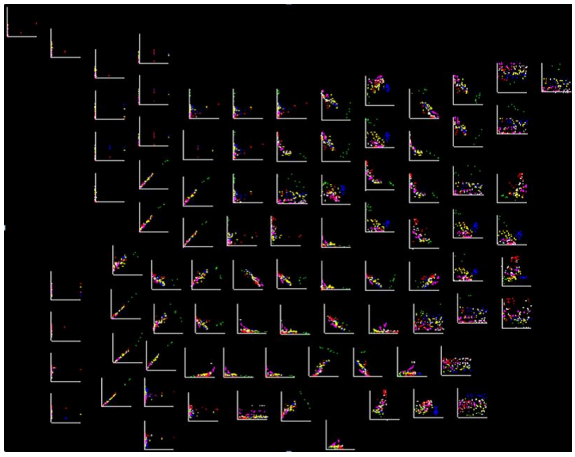

図 3: 87 個の散布図を選択して可視化した結果.

6. まとめ・今後の課題

本報告では，高次元データを構成する全ての 2 次元ペアから生成される散布図のうち，特定の基準を満たす散布図を選択し，類似する散布図が画面上で近隣するように一覧表示する可視化手法を提案した．今後の課題として，この散布図の配置結果に対する定量評価および主観評価を進めていきたい．また現実のデータに適用して，どのような知見が視覚的に発見できるかを検証したい．

参考文献

[Elmqvist08] 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.

[Itoh06] T. Itoh, H. Takakura, A. Sawada, K. Koyamada, Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, *IEEE Computer Graphics and Applications*, 26(2), 40-47, 2006.

[Itoh09] T. Itoh, C. Muelder, K. Ma, J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, *IEEE Pacific Visualization Symposium*, 121-128, 2009.

[Leban05] G. Leban, I. Bratko, U. Petrovics, T. Curk, B. Zupan, VizRank: Finding Informative Data Projections in Functional Genomics by Machine Learning, *Bioinformatics Applications Note*, 21(3), 413-414, 2005.

[Wilkinson05] L. Wilkinson, A. Anand, R. Grossman, Graph-Theoretic Scagnostics, *IEEE Symposium on Information Visualization*, 157-164, 2005.

[Sips09] M.Sips: Selecting Good Views of High-Dimensional Data Using Class Consistency, *Eurographics/IEEE-VGTC Symposium on Visualization*, 831-838, 2009.

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