多次元データ可視化のための散布図の選択と描画の一手法

中林 明日香† 伊藤 貴之†

†お茶の水女子大学 〒112-8610 東京都文京区大塚 2-1-1

E-mail: †{g1520533, itot}@is.ocha.ac.jp

あらまし 多次元データの可視化手法として散布図行列や平行座標法などがあるが，これらの手法では膨大な次元数を有するデータにおいて非常に大きな画面空間を必要とする問題点がある．この問題を解決するための一手段として本報告では，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要なものを選出し，さらにその散布図を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する手法を提案する．この可視化手法は，例外点をデータから削除するか否かの判断，例外でない点群のモデル化手法の検討などに有用であると考えられる．本報告では，小売店の気象と売上の関係のデータを題材にして，本手法を用いた可視化の実行例を示す．

キーワード 多次元データ，可視化手法，散布図

1. はじめに

日常生活や専門業務に関するデータの多くは多次元データである．身の回りに存在する多次元データから発見される特徴や規則性は，そのデータを理解し活用するにあたって重要な知識となる．ユーザが理解できる形式で多次元データを可視化することにより，この特徴や規則性を発見することが容易になる．

多次元データの可視化手法としてよく知られるものに散布図行列や平行座標法(Parallel Coordinate Plots; 以下 PCP)があげられる．n次元データを可視化する際に，散布図行列は全ての 2 次元ペアの散布図を作成し$n \times n$の格子状に並べることで表現し，平行座標法はn本の平行な座標軸に変数の値をプロットしそれを折れ線で結ぶことで表現する．これらの手法は多次元データを構成する全ての次元を可視化するものであるが，膨大な次元数を有するデータにおいては非常に大きな画面空間を必要とする問題点がある．また多次元データの全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ため，近年では多次元データから可視化する意義の高い次元だけを選択して表示する手法が多く提案されている．

多次元データを活用する際にはそのモデル化が重要になることもある．多次元データを構成する数値群の中にどのようなノイズや例外値が含まれているかを理解し，適切なスクリーニング処理によってこれらを除去したのちに，どのようなモデルを適用できるかを検討する処理が必要となる場面が多い．例えば機械学習の訓練データに多次元データを利用する際に，このような工程が重要な意味を持つことが多い．このような工程にも多次元データの可視化手法が貢献できることが議論されている．

本報告ではこれらの 2 点に着目した多次元データ可視化の一手法を提案する．本手法は以下の 2 つの処理工程から構成されるものである．

- 多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要ないくつかを選出して表示する．
- 散布図に表示される点群を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する．

本報告の構成は以下の通りである．2 章では関連研究について，3 章では提案手法について述べる．そして 4 章で本手法の実行結果と考察について，5 章で本報告のまとめと今後の課題について述べる．

2. 関連研究

2.1 次元選択を用いた多次元データの可視化

多次元データから可視化する意義の高い低次元部分空間を選択して可視化する手法が近年多く発表されている．例として，多次元データから所定の基準を満たす複数の 2 次元ペアの散布図を生成し，各散布図間の類似度距離に基づいて配置する手法[1]や，所定の基準を満たす次元間の低次元 PCP を生成し，各 PCP 間の次元の共有率から算出される類似度距離に基づいて配置する手法[2]などがある．しかしこれらの手法では PCP や散布図の表示数を対話的に変えることができなかった．この問題点を解決する多次元データ可視化手法として Hidden[3]が発表された．

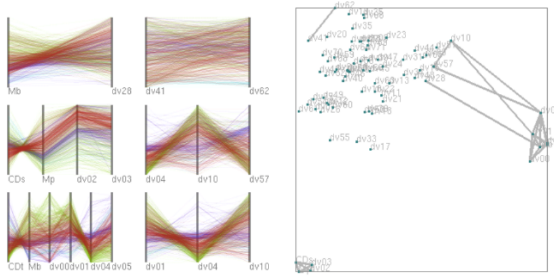

図 1 Hidden による可視化画面[3]

Hidden は画面右部の次元散布図上を対話的に操作することによって選択される低次元部分空間群を，画面左部で複数の PCP によって表示する．図 1 に Hidden による可視化の例を示す．また Hidden を拡張し PCP と散布図を併用して可視化した手法[4]も発表されている．この手法では原則として PCP で多次元データを可視化しつつ，PCP では視認しにくい数値分布を有する 2 軸のみに対して散布図を適用して表示する．

本研究も Hidden[3,4]と同様に重要な次元を選択して可視化する手法であるが，PCP を使わずに散布図のみを適用している．

2.2 散布図による多次元データの可視化

散布図を用いて多次元データを可視化する手法として，Wilkinson らの手法[5]や Dang らの手法[6]が挙げられる．Wilkinson らの手法では，多次元データの各 2 次元ペアから生成される散布図の形状などから Scagnostics と呼ばれる 9 種類の特徴を定量評価し，特定の傾向を持った散布図を生成する 2 変数を推薦する．Dang らの手法では，散布図行列から Scagnostics の基準により各散布図を特徴づけてクラスタリングし，リーダーとなる散布図を選出して類似するものを近くに配置する．これらの手法では，膨大な次元数の多次元データでも非常に大きい画面空間を使うことなく可視化することができる．

現時点での本研究では 3 章にて後述する基準で散布図を選出しているが，Scagnostics を適用することも可能である．

3. 提案手法

本報告では 1 章でも述べた通り，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要なものを選出し，さらにその散布図を構成する点群を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する手法を提案する．

現時点での我々の実装では，散布図の選出基準には Hidden[3]と同じく相関係数またはエントロピーによる基準を採用している．相関係数による基準を適用した際には，各散布図を生成する 2 次元間の相関係数を計算し，その絶対値の大きい散布図を優先的に表示する．エントロピーによる基準を適用した際には，多次元データ中のカテゴリ型変数が各個体のラベルに相当するとみなして，点群がラベルごとによく分離されている散布図を優先的に表示する．

現時点での我々の実装では，「例外点群の抽出」および「例外でない点群の包括領域の生成」に Delaunay 三角分割法を利用している．Delaunay 三角分割法は与えられた点群を連結して三角メッシュを生成する手法であり，三角メッシュを構成する三角形の最小角度が最大になるように三角メッシュを生成するものである．本手法では，各散布図に対して，散布図中の全ての点群を包括する大きな四角形を作成し三角形に分割し，散布図中の点群を 1 つずつ追加して頂点として連結していくことで三角メッシュを逐次的に更新し，全ての点群を追加したら最初に作成した大きな四角形とその頂点に連結される辺を削除する，というインクリメンタルなアルゴリズムを採用している．

このような処理によって生成された三角メッシュから，ユーザ指定の閾値を超える長さの辺を削除することで，図 2 のようにどの点群とも連結されていない点を例外点として抽出する．ユーザによる対話操作で閾値を調節することで，例外点と判定された点の数を調節できる．そして，例外点以外の点で構成される三角形群の領域境界を構成する辺のみを濃い色で描画し，三角メッシュを薄い色で塗りつぶすことによって，点群の包括領域を表示する．

4. 実行結果

我々は本手法を Java Development Kit (JDK) 1.8.0 により実装した．現時点での実装では Hidden[3]の実装を再利用することにより，散布図の選択と対話的表示を実現している．

本報告では，アパレルの小売店における各日の来客数や売上と，その各日の気象値との関係のデータを題材にして，本手法を用いた可視化の実行例を示す．データ中の説明変数(気象の数値・横軸)と目的関数(売上の数値・縦軸)の対応表を表 1 に示す．なお本章で用いるデータは現実のデータに乱数を加算したものであり，現実の数値をそのまま可視化したわけではない点に注意されたい．

図 3 では点群の一属性(平日)が他方の属性(休日)を内包するような形状になっている数値分布の例である．この結果から，客単価や平均買上商品単価は休日よりも平日の方がばらついていることがわかる．

図 4 は秋(9 月～11 月)の平均買上点数と平均買上商品単価の数値分布を示している．冬に近づくにつれて

平均買上商品単価が上昇する傾向にある．これは厚手の服になればなるほど商品の単価が高くなるためと推測される．これに対し平均買上点数は低下する傾向にある．春服・夏服に比べ商品の単価が高いため，各商品の購買に慎重になっている可能性が考えられる．

表 1 データの説明変数と目的関数の対応表

説明変数(気象数値)		目的関数(売上数値)	
MinTemp	最低気温	Revenue	売上
MaxTemp	最高気温	Guest1	購入人数
SumRain	降水量	Guest2	来客人数
SumSnow	降雪量	Ratio	買上率
SumSnowC	積雪量	PerGuest	客単価
SumSunTime	日照時間	AveUnit	平均買上商品単価
MaxWind	最大風速	AveNum	平均買上点数

図 5 は 2 月と 7 月と 11 月の買上率の数値分布を示している．2 月上旬と 7 月下旬のみ突出して買上率が高い期間があり，これは売り尽くしセールなどの特殊なイベントのために，ウィンドウショッピングとして来店した人よりも，最初から商品を購入するつもりで来店する人が多かった可能性が考えられる．また 11 月中に 1 日だけ特に買上率の高い日があることが読み取れる．

5. まとめと今後の課題

本報告では，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要と思われる散布図を選出し，さらにその散布図を構成する点群を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する手法を提案した．本報告で提案した可視化手法は，例外点をデータから削除するか否かの判断，例外でない点群のモデル化手法の検討などに有用であると考えられる．

今後の課題として，まず例外でない点群の包括領域をよりはっきりと描画する必要がある．現時点では 3,4 色以上の領域が重なるとその重なり部分がぼんやりしてしまい視認性が下がってしまう．最低でも 12 色程度の領域が重なってもはっきりと視認できるようにしたい．また例外点(ユーザが選択した点)をデータから削除する機能も実装したい．

また，現時点で実装している相関係数やエントロピーにもとづいた散布図選出手法では，我々が重要であると主観的に判断しているような散布図が選出されないことがある．そこで新しい散布図選出基準を実装することで，このような散布図が選出されるようにしたい．

そしてこれらの機能を実装した後に，より多様なデータセットを本手法に適用し，さらに汎用性に富んだ実装になるように開発を進めたい．

謝辞

データセットを提供して頂いた株式会社 ABEJA 様に感謝いたします．

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, Vol. 18, No. 1, pp. 111-119, 2015.

[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, 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, Vol. 43, pp. 1-13, 2017.

[4] A. Watanabe, T. Itoh, M. Kanazaki, K. Chiba, “A Scatterplots Selection Technique for Multi-Dimensional Data Visualization Combining with Parallel Coordinate Plots”, 21st International Conference on Information Visualisation (IV2017), pp. 78-83, 2017.

[5] L. Wilkinson, A. Anand, R. Grossman, “Graph-Theoretic Scagnostics”, IEEE Symposium on Information Visualization, pp. 157–164, 2005.

[6] Dang Tuan Nhon, Leland Wilkinson, “ScagExplorer: Exploring Scatterplots by Their Scagnostics”, IEEE Pacific Visualization Symposium (PacificVis 2014), pp. 73–80, 2014.

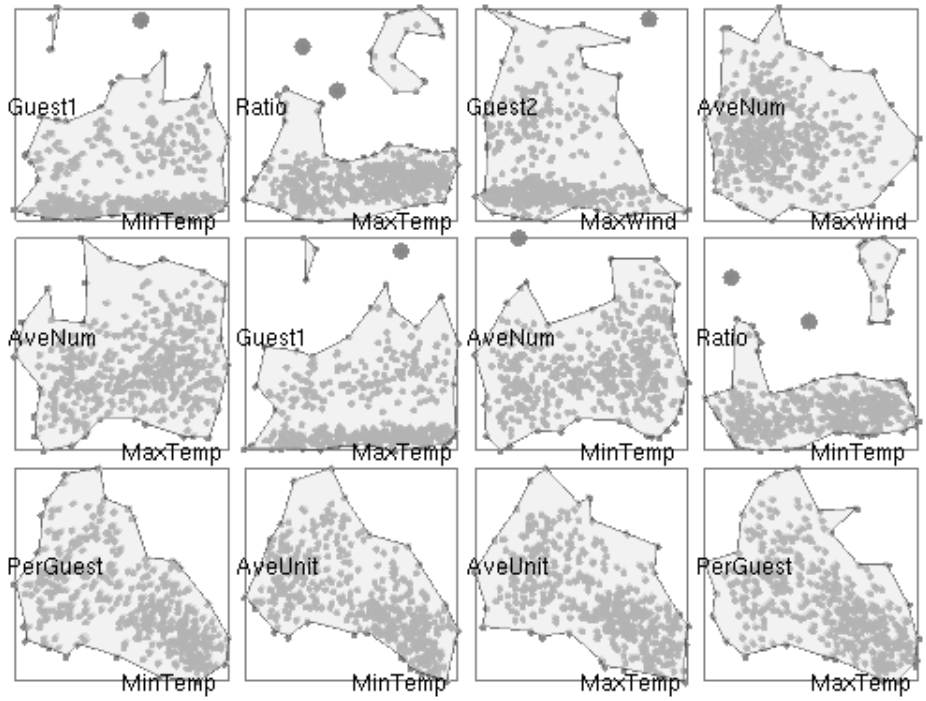

図 2 散布図を「例外点群」と「例外でない点群の包括領域」の 2 種類として描画した例

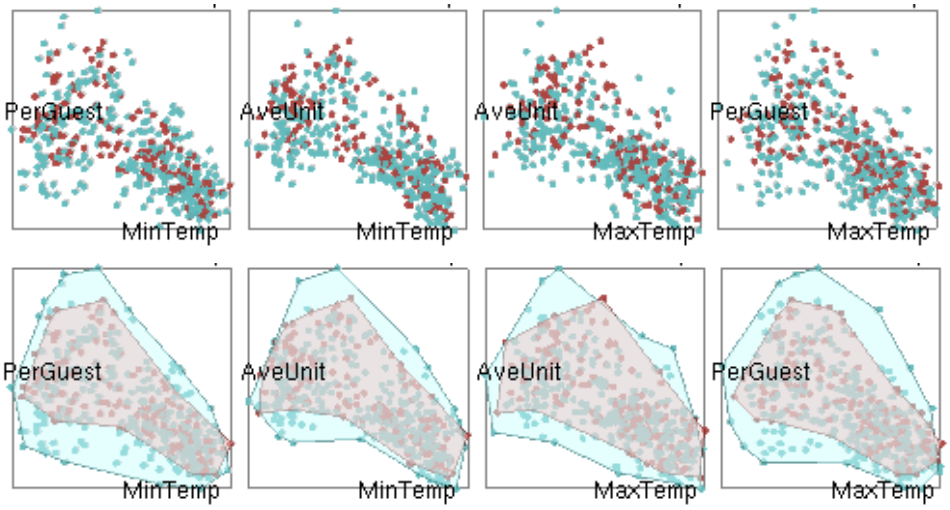

図 3 ある属性が別の属性を内包するような数値分布になっている例
(水色は平日，赤色は休日を示しており，上は包括領域を囲む前，下は囲んだものである)

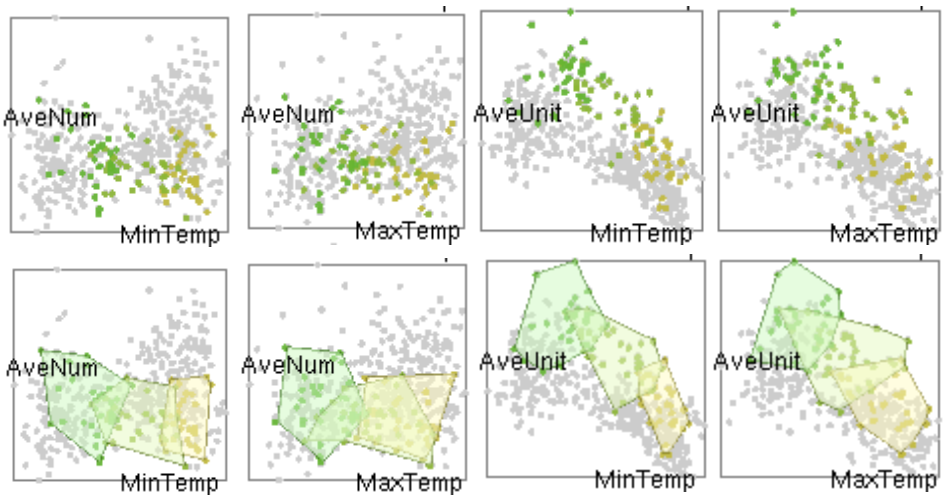

図 4 秋(9 月～11 月)の買上点数と商品単価の数値分布

(黄色は 9 月，黄緑色は 10 月，緑色は 11 月を示しており，上は包括領域を囲む前，下は囲んだもの)

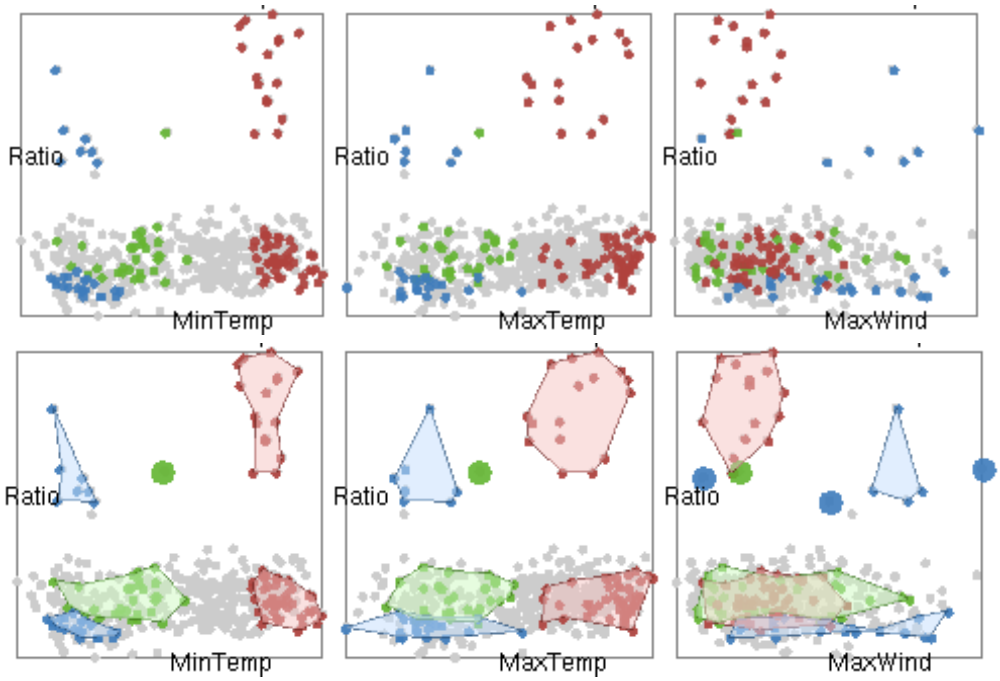

図 5 2 月と 7 月と 11 月の買上率の数値分布

(青色は 2 月，赤色は 7 月，緑色は 11 月を示しており，上は包括領域を囲む前，下は囲んだもの)