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 *

概要. 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，単純かつ対話的なスライダー操作によって選出し，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 種類にわけ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．

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．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．

有名な多次元データ可視化手法に散布図行列や平行座標法 (PCP) があげられる．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．このため近年では，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．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，単純かつ対話的なスライダー操作によって選出する．
- 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．

* お茶の水女子大学

2 関連研究

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として，多次元データ可視化手法 Hidden[1] が挙げられる．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．また Hidden を拡張し PCP と散布図を併用して可視化した手法 [2] 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．本報告も Hidden[1][2] 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．

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[1] 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．

また，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．散布図を構成する全点を連結する三角メッシュを生成し，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 1 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透明度の低い色で描画し，三角形群を透明度の高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．

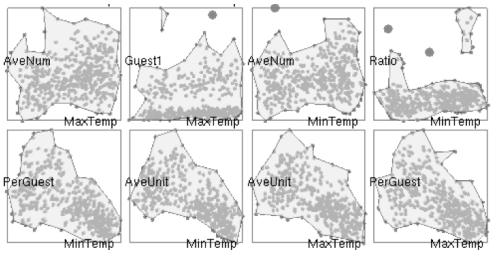

図 1．散布図を「例外点群」と「例外でない点群の包括領域」の 2 種類として描画した例

4　実行結果

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と目的関数の対応表を表 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．

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	最大気温	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 2 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．

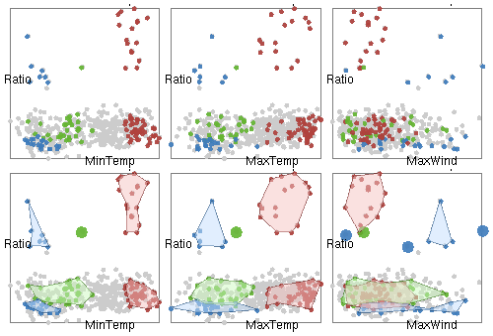

図 2．2 月と 7 月と 11 月の買上率の数値分布 (青色は 2 月，赤色は 7 月，緑色は 11 月)

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単純なスライダー操作によって選出し，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．

今後の課題として，新しい散布図選出基準の実装が必要である．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．また，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] 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*, Vol. 43, pp. 1–13, 2017.

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