多次元データ可視化のための
重要度・類似度にもとづく散布図の選択と描画の一手法

理学専攻　情報科学コース　1940655　**中林 明日香**　(**指導教員：伊藤 貴之**)

1　はじめに

多次元データは我々の日常生活や専門業務に幅広く存在している．身近な存在である多次元データから重要な知識を得るための方法論として，データの特徴や規則性を観察することがあげられる．その観察手段として多次元データの可視化が有効であると考えられる．

多次元データの可視化の代表的な手法に，散布図行列や平行座標法 (Parallel Coordinate Plots; 以下 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ため，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．また一方で，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．言い換えれば，多次元データを構成する全ての次元を可視化するより，多次元データの中から興味深い特徴や規則性を有する次元だけを事前選択して可視化したほうが，効率よくデータを観察できる場合もある．そこで近年では，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可能性が期待される．

本論文ではこれらの 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，または類似していないいくつかの散布図を，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．

2　関連研究

多次元データの中から重要な部分だけを可視化するためのアプローチとして，可視化する意義の高い低次元部分空間を事前に抽出する手法は従来から数多く提案されている．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などがある．しかしこれらの手法による可視化結果は固定的なものであり，散布図や PCP の表示数を対話的に調節す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 Itoh らは Hidden[1] を発表した．Hidden[1] 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．

3　提案手法

本手法では，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，あるいは類似していない散布図の組み合わせを選出し，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．

本手法では重要度を考慮した散布図選出のための基準として相関係数による基準，エントロピーによる基準，点群領域の細長さによる基準の 3 種類を実装している．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．点群領域の細長さによる基準を適用した際には，点群領域の面積に対し点群領域の外周長が長くなる散布図を，つまり点群が細長く分布する領域を有する散布図を優先的に表示する．

重要度を考慮して散布図を選出すると，似た傾向の見られる散布図ばかり表示してしまうこともある．そこで類似していない散布図の組み合わせを選出するために，前述した 3 つの基準を利用して散布図間の類似度を算出し，これにもとづいて多様な組み合わせの散布図を選出する [2]．具体的には，各散布図の 3 つの基準値を持つ 3 次元ベクトルを生成し，各 3 次元ベクトル間のコサイン類似度を算出する．そして，ユーザ指定の閾値よりもコサイン類似度が高い散布図群を同時に選出しないという制約を設けることで，多様な組み合わせの散布図を選出する．

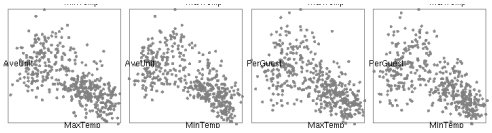
また我々の実装では,「例外点群の抽出」および「例外でない点群の包括領域の生成」に Delaunay 三角分割法を利用している．Delaunay 三角分割法は与えられた点群を連結して三角メッシュを生成する手法であり，三角メッシュを構成する三角形の最小角度が最大になるように三角メッシュを生成するものである．本手法では，各散布図に対して，散布図中の全ての点群を包括する大きな四角形を作成し三角形に分割し，散布図中の点群を 1 つずつ追加して頂点として連結していくことで三角メッシュを逐次的に更新し，全ての点群を追加したら最初に作成した大きな四角形とその頂点に連結される辺を削除する，というインクリメンタル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

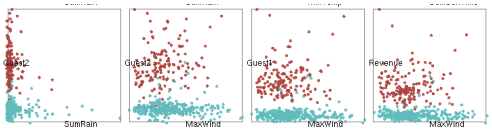
4　実行例

本論文では，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．なお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．

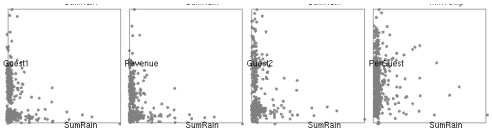
図 1a は本手法の相関係数による基準を適用した可視化の例である．右下がりに点が分布しているような負の相関が見られる散布図が優先的に選出されている．図 1b は本手法のエントロピーによる基準を適用した可視化の例である．カテゴリ変数は平日か否かを用いており，水色は平日，赤色は休日を示している．赤色の点と水色の点が上下によく分離しているような散布図が優先的に選出されている．図 1c は本手法の点群領域の細長さによる基準を適用した可視化の例である．点群領域が細長く分布しているような散布図が優先的に選出されている．図 1d は本手法の類似度による基準を適用した可視化の例である．カテゴリ変数は平日か否かを用いており，水色は平日，赤色は休日を示している．負の相関の強い散布図や点群がラベルごとによく分離している散布図，点群が細長く分布している散布図など，多様な形の散布図が選出されている．



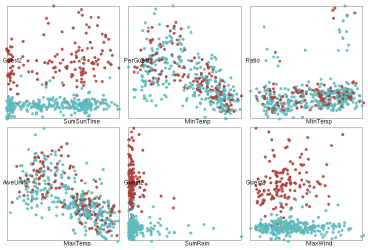
(a) 相関係数による基準を適用した可視化の例



(b) エントロピーによる基準を適用した可視化の例



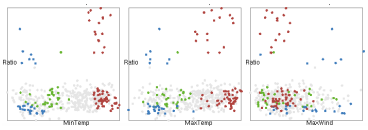
(c) 点群領域の細長さによる基準を適用した可視化の例



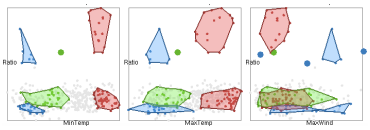
(d) 類似度による基準を適用した可視化の例

図 1: 散布図選出についての実行例

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．他の月は灰色の点で表示されている．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．



(a) 描画処理を行う前



(b) 描画処理を行った後

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，灰色はその他の月)

5　まとめと今後の課題

本論文では，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，あるいは類似していない散布図の組み合わせを選出し，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．

今後の課題として，点群領域の囲み表示と例外点の描画機能において，削除する辺の閾値を全ての散布図で一定にするのではなく，散布図ごとに調節できるようにしたい．現在の実装では散布図ごとに点群の密度の偏りが異なることを考慮していないため，密度の薄い領域にて三角形が削除されないことがあり，これによって必要以上に広い領域が塗りつぶされることがあり，ユーザに誤認を与える可能性がある．そこで，点群の密度の偏りに応じて削除する辺の閾値を散布図ごとに自動的に調節できるようにしたい．

そして，さらに大きな次元を持つ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： 小売店の気象と売上に関する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] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman 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] 伊藤貴之, 中林明日香, 萩田真理子, 散布図選択による多次元データ可視化へのグラフ彩色問題の適用, 第 28 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS), 2020.