InverseVis: 疎な領域へのサンプリング点生成による多次元データ中の「非存在」の可視化

伊藤 貴之 * 田上 湖都 † 矢島 知子 ‡ 李 国政 §

概要. 多次元データの可視化手法は非常にアクティブな話題であり，長年にわたって多くの手法が報告されている．その中でも散布図や平行座標プロットといった幾何学的な手法は，多次元データを構成する各個体を点や線で描くことにより，データ中の密な部位や外れ値が存在する部位を表現する．一方で時として我々は，多次元空間中の「個体が存在しない領域」に着目したいことがある．この要求を解決するために本報告では，多次元空間中の疎な部位に多数のサンプリング点を生成し，その点を可視化することで，個体が存在しない領域の分布を表現する手法を提案する．本手法では多次元空間の生成したサンプリング点に次元削減を適用して散布図で表示する．マウス操作によって特定の点を指定すると，平行座標プロットにより指定された点の各次元の値を表示する．本報告では，含フッ素有機化合物の反応実験データと，学術成績と給与のデータでのケーススタディをもって，提案手法の有効性を検証する．

1 序論

計算機を用いた可視化は一般的に，データ中に特定の事象が「存在する」ことを表現している．例として図 1 に示すように，流体（空気や水）の可視化では流れが存在することを表現する．散布図による多次元データの可視化では点群の中にクラスタや外れ値が存在することを表現する．ネットワークの可視化では 2 点間に関係が存在することを表現する．これらの可視化によって我々は，データ中に特定の事象が存在することを理解することができる．

しかしこのような可視化からは，データ中の事象の「非存在」を発見することが難しい場合がある．計算機を用いた可視化は，特定の事象の存在を点や線などで描画する．可視化のユーザは一般的に，画面上の点や線が描画された箇所に着目する．言い換えれば，点や線が描画されていない箇所にはユーザの注目が集まりにくい．この点に着目して我々は，逆転の発想として，データ中の「存在しない事象」の可視化手法を研究している．本報告はその中でも多次元データに関する手法を提案するものである．

日常社会の情報の多くは多次元データとして記述可能である．これらの多次元データの理解を深める目的で，多次元データの可視化手法が開発されてきた．多次元データの可視化手法は長年にわたって研究されており [10, 12, 22]，特に散布図行列や平行座標プロットに関する論文は非常に多く発表されている．これらの手法は多次元空間中のクラスタや外れ値の「存在」を可視化するのに向いている．

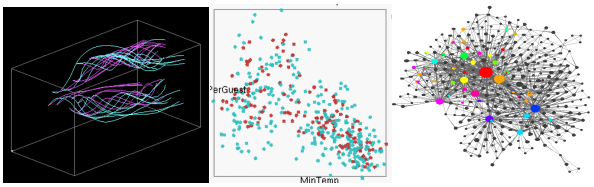

図 1. 計算機を用いた可視化は一般的に，データ中に特定の事象が「存在する」ことを表現している．(左) 流体の可視化．(中) 多次元データの可視化．(右) ネットワークの可視化．

一方で時として我々は，多次元空間中の「個体が存在しない領域」あるいは「個体の密度が疎である領域」に着目したいことがある．例えば，機械学習の訓練データにおいて，どのような数値を有するデータ標本が足りないかを知りたい状況が起こりえる．あるいは，反復的な科学実験や計算機シミュレーションにおいて，どのようなパラメータでの実験や計算が足りないかを知りたい状況が起こりえる．しかし，散布図行列や平行座標プロットから多次元空間中の疎な領域の視認は難しい．

そこで本研究では，逆転の発想として，多次元空間中において個体の密度が低ければ低いほど多数のサンプリング点を発生することで，多次元空間中にて個体が存在しない（あるいは密度が疎な）部位を可視化する手法を提案する．図 2 に本研究の概念を示す．本手法では入力データを構成する多次元空間中の点群 (図 2(a)) の各々がポテンシャルを有すると仮定し，それを補間することで，多次元空間における個体の密度分布 (図 2(b)) を生成する．続いて，その密度分布を参照して，密度が低いほど高い確率

* お茶の水女子大学大学院 理学専攻 情報科学領域
† お茶の水女子大学大学院 理学専攻 化学・生物化学領域
‡ お茶の水女子大学大学院 理学専攻 化学・生物化学領域
§ 北京理工大学

でサンプリング点を生成する (図 2(c))．この処理により，個体の密度が疎であるほど密度の高いサンプリング点を生成する．

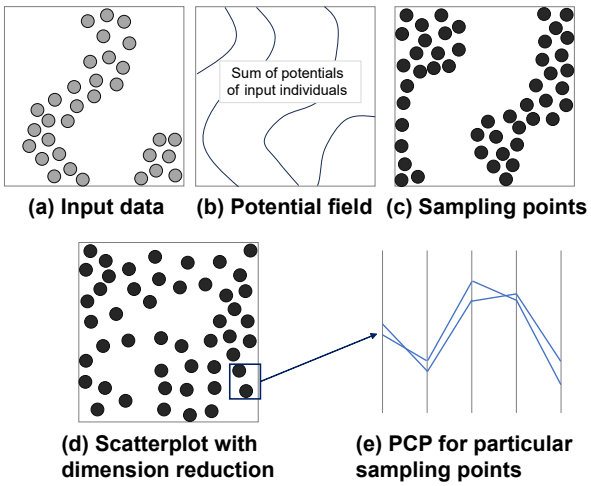

図 2．本研究の概念．本研究では多次元空間中において個体の密度が低い部位にサンプリング点を発生させることで，個体が存在しない（あるいは疎な）部位を可視化する．

サンプリング点の集合を可視化する手段として，提案手法ではまずサンプリング点に次元削減を適用して，2 次元の散布図として可視化する (図 2(d))．そして散布図中の特定の点（または点の集合）をマウス操作で指定すると，それらのサンプリング点が有する各次元の値を平行座標プロットにより可視化する (図 2(e))．この 2 段階の可視化により，まず多次元空間中の疎な領域の全体像を可視化し，続いてユーザが指定する特定のサンプリング点の各次元の値を確認できる．

本報告では，含フッ素有機化合物の反応実験データと，学術成績と給与のデータ [1] でのケーススタディをもって，提案手法の有効性を検証する．

本報告の貢献は以下のとおりである．

- 多次元空間中において個体の密度が低いほど多数のサンプリング点を発生することで，多次元空間中の疎な部分を可視化するというアプローチ．
- 多次元空間中の疎な部分における各次元の値を読み取るために，サンプリング点に次元削減を適用した散布図と，その散布図中の特定の点を表示する平行座標プロットの 2 つの可視化コンポーネントを組み合わせたインタラクションデザイン．
- 含フッ素有機化合物の反応実験データ，学術成績と給与のデータ，を用いた 2 種類のケーススタディによる提案手法の有効性の検証と，ユーザ評価実験結果にもとづく改良案の議論．

2 関連研究

データ空間中の疎領域に着目した手法は科学技術系の可視化手法 (Scientific Visualization) で多用されてきた．Ivan ら [18] は空間中の Sparseness を各部位の重要度に置き換えて特徴的な部位を強調表示するボリュームレンダリング手法を提案している．また流れ場の可視化においても同様に，疎領域に注目することで可視化の視認性を高める議論がなされている [11]．これらの手法はあくまでもデータ中の特徴的な部位の存在を強調表示する手法であり，疎領域そのものの表示を目的としていない．

情報可視化の中でも多次元データを対象とした可視化手法において，疎な領域に存在する Outlier を逃さずに強調表示する手法がいくつか研究されている [3, 4, 5]．これらの手法は疎な領域に存在する Outlier の視認性を高めることを目的としており，疎な領域の全体像を可視化することや，あるいは点群や個体が全く存在しない領域を可視化することを目的としていない．

次元削減と散布図を用いた多次元データ可視化が，多次元空間中の疎な領域の発見に貢献できる，ということを示した研究もいくつか報告されている [8, 23, 25]．また，次元削減によって 2 次元の画面空間に投影された多次元データを，再度 3 次元以上の空間に逆投影することで，多次元データの数値分布の表現性を高めた手法も報告されている [9, 20, 21]．しかしこれらの手法も，多次元空間中の疎な領域を直接的に描画するものではない．

本報告の提案手法で採用しているサンプリング点の生成と同様に，多次元空間中の密度の低い領域に積極的に点を生成するアルゴリズムはいくつかの研究で採用されている [2, 6]．しかしこれらの研究は密度の低い領域の可視化を目的としておらず，本論文とは目標設定が異なる．

本報告で 1 つ目の適用事例として紹介する化学反応の条件設定の探索問題には，実験計画法などの数理的な方法がしばしば適用される．可視化によって実験計画法を支援する手法はいくつか報告されている [7, 16, 24]．本報告で 2 つ目の適用事例では学術成績と給与のデータにおける分布の男女差を可視化している．このような性差の問題を可視化によって分析した事例もいくつか報告されている [13, 17, 19]．これらの用途に適用できる可視化手法を「疎な領域の可視化」という抽象化によって実現した，という点で提案手法はこれらの先行研究と異なる．

3 多次元空間中の疎な領域の可視化

本章では提案手法の処理手順を示す．具体的には，想定するデータ構造，サンプリング点の発生方法，可視化手法とインタラクションについて述べる．

3.1 データ構造

提案手法では入力情報として N 個の個体 d_i で構成される多次元データ $R = \{r_1, r_2, ..., r_N\}$ を想定する．個体 r_i は M 次元の変数 $r_i = \{x_{i1}, x_{i2}, ..., x_{iM}\}$ を有するものとする．またオプションとして，個体 r_i は目的変数 f_i またはカテゴリ変数 c_i をもちうるものとする．

以上を入力情報として，提案手法はサンプリング点 $S = \{s_1, s_2, ...\}$ を生成する．サンプリング点 s_i は M 次元の変数 $s_i = \{y_{i1}, y_{i2}, ..., y_{iM}\}$ を有するものとする．またオプションとして，目的変数 g_i またはカテゴリ変数 d_i をもちうるものとする．

3.2 サンプリング点の生成

提案手法では入力データとして与えられた点群からポテンシャル場を生成し，それに沿ってサンプリング点を生成する．ここで，単に点群の密度分布を近似するポテンシャル場を生成するのであれば，例えば混合ガウスモデル (Gaussian Mixture Model) を適用することが考えられるが，入力データ中の点の数が少ない場合や点の位置がまばらな場合，あるいは点の密度分布が非連続な場合にうまく動作しないことがある．そこで試行錯誤の結果として，我々は以下のような手法を採用した．

提案手法では各々の個体がポテンシャル場を有するものとする．多次元空間中の位置 X における個体 r_i のポテンシャル場 $p_i(X)$ を以下のガウス関数

$$p_i(X) = a_{pot} \exp(-\frac{r_i - X}{b_{pot}^2}) \qquad (1)$$

で表す．ここで r_i は多次元空間中の個体の位置，a_{pot} および b_{pot} はユーザ定義の定数とする．これを用いることで，位置 X におけるポテンシャル $P(X)$ は，個体 r_i のポテンシャル場の総和 $P(X) = \sum_i p_i(X)$ により求められる．

続いて提案手法では，与えられた多次元空間中のランダムな位置に n_{pot} 個のサンプリング点を生成する．ただしサンプリング点が生成されるか否かの確率を，位置 X から以下の数式 $c_{pot} - P(X)/P_{max}$ によって算出する．ただし P_{max} は $P(X)$ の最大値であるとし，c_{pot} はユーザ定義の定数であり $0 \leq c_{pot} \leq 1$ であるとする．この値が 0 以下となる場合にはサンプリング点は生成されない．

図 3 に提案手法の仕組みを図解する．提案手法では各個体が有するポテンシャルの和によってポテンシャル場を生成し，ポテンシャルが低い位置ほど高い確率でサンプリング点を生成する．

サンプリング点がオプションで目的変数 g_i を有する場合には，サンプリング点の近傍の数個の個体 r_j の目的変数 f_j を補間することで g_i の値を算出する．サンプリング点がオプションでカテゴリ変数 d_i を有する場合には，そのカテゴリ変数がもちうる各々の値について別々にサンプリング点の生成を実行する．例えばカテゴリ変数が値 D_a, D_b, D_c の 3 種類の値をとりうる場合には，カテゴリ変数の値が D_a である個体のみ，D_b である個体のみ，D_c である個体のみを対象として，それぞれのカテゴリ変数値に対してポテンシャル場を生成し，そこからサンプリング点の生成確率を求めて，それぞれのカテゴリ変数値についてサンプリング点を生成する．

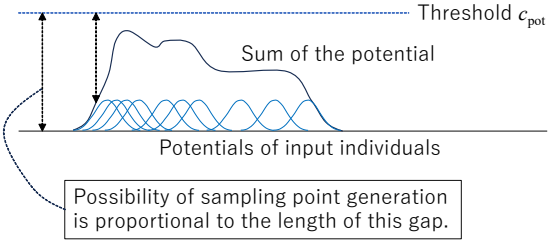

図 3. 入力データ中の個体によるポテンシャル場とサンプリング点の生成確率．

以上の処理により，入力データによって与えられた多次元空間にサンプリング点を生成する．この過程において提案手法では，$a_{pot}, b_{pot}, c_{pot}, n_{pot}$ の 4 つの定数を調節する必要がある．

3.3 可視化とインタラクション

提案手法ではサンプリング点の多次元空間中の位置 $s_i = \{y_{i1}, y_{i2}, ..., y_{iM}\}$ に次元削減を適用し，これを散布図で可視化する．現時点での実装では，線形性の高い分布の表現に向いた PCA (Principal component analysis) と，非線形性の高い分布の表現に向いた手法のうち計算時間が比較的短い UMAP (Uniform Manifold Approximation and Projection) を採用している．上述の $a_{pot}, b_{pot}, c_{pot}, n_{pot}$ の 4 つの定数を調節することで，散布図の視認性を調節することができる．なお，サンプリング点が目的変数 g_i またはカテゴリ変数 d_i を有する場合には，散布図を構成する各点の色をもって目的変数またはカテゴリ変数の値を表現する．

提案手法のスナップショットを図 4 に示す．画面左側の散布図にて，ユーザがマウスでホバリングすることで，特定の点を指定することができる．この操作によって特定の点（またはその近傍にある複数の点）を指定すると，画面右側の平行座標プロットを用いてその各点の各次元の値を表示する．これにより，与えられた多次元空間中における疎な領域での数値の例を知ることができる．なお，平行座標プロットの軸の並び順は，ユーザがデータ入力時に指定できる．あるいは最適な並び順の自動設定手法 [14] を適用してもよい．

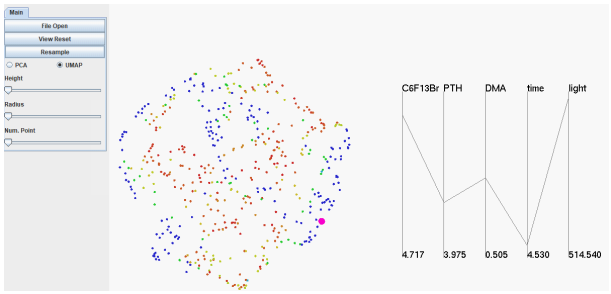

図 4. 提案手法のスナップショット．左側の散布図で特定の点を指定すると，右側の平行座標プロットで各次元の値を表示する，というインタラクションにより疎な領域での数値の例を知ることができる．

4 適用事例

4.1 含フッ素有機化合物の実験データ

化学や生物などの実験科学では，条件設定を変えながら反復的に実験を繰り返すことで，探索的に良好な実験結果を求めるプロセスが必要となることが多い．ここで探索過程となる実験の回数を減らすためには，まだ実施していない実験設定を可視化によりリストアップし，それを実験者が閲覧しながら自らの経験則と照合して，良好な実験結果が得られそうだと予想される実験設定を選ぶことが重要である．本論文ではこの目的に沿って提案手法を適用した事例を紹介する．

含フッ素有機化合物は医薬品や農薬，機能性材料として幅広く利用されており，その新たな合成法の開発は有機化学研究の重要な課題である．本研究では，ハロゲン結合を利用したペルフルオロアルキル化反応を，電子不足な共役オレフィン類，スチレン類及び非共役末端オレフィン類に適用した実験 [15] を題材として，その合成におけるいくつかの説明変数と，目的変数となる収率の関係を可視化した．具体的には，実験時に設定される以下の 4 つの値を 4 次元の説明変数 $r_i = \{x_{i1}, ..., x_{i4}\}$ とし，測定結果となる収率を目的変数 f_i とする 19 組の実験結果を入力データとした．

- ヨウ化ペルフルオロヘキシルの当量
- 酸素の当量
- N, N-ジイソプロピルエチルアミン (DIPEA) の当量
- 光照射時間

そして提案手法により「まだ実験されていない説明変数の例」となるサンプリング点の集合を可視化した．図 5(上) にサンプリング点の集合を散布図として可視化した例を示す．ここで散布図の点は目的変数の値で色付けされており，収率が低い（好ましくない実験結果）と予測されるほど青に近い色で，収率が高い（好ましい実験結果）と予測されるほど赤に近い色で表示されている．

図 5(上) では (a)〜(d) に収率が高いと予測される赤い点が，(e)(f) に収率が低いと予測される青い点が見られる．これらの点における説明変数の値を図 5(下) に示す．このような表示により，収率が高いと予測される説明変数の値を発見し，その中からどの説明変数をもって次の実験を実施するかの議論材料にすることができる．

図 5(下) にて (a)〜(d) の共通点として，DIPEA の当量が 1.0 以上であること，光照射時間が長いことがあげられる．これらの条件を満たしつつ，他の 2 つの説明変数が特定の値を有するときに，実験結果としての収率の高さが期待できる．一方で，図 5(下) にて (e),(f) の共通点として，DIPEA の当量が 1.0 未満であること，光照射時間が短いことがあげられる．これらの条件を満たす際に収率の低さが予測される．この傾向は入力データである 19 回の実験結果とも，またその後の知見とも整合するものであり，提案手法による可視化結果の妥当性を示すものである．

4.2 学校成績と給与のデータ

機械学習を用いた採用人事システムのバイアスが社会問題になったことがある．具体的には，採用人事のための訓練データに女性のデータが少なかったことから，システムが女性に不利な判定を下す傾向が見られたというものである．このようなバイアスを防ぐ一手段として，データのバランスを事前に観察することが重要である．そこで本論文では，中等教育・高等教育の成績と給与オファーなどをまとめた人材オープンデータを題材として，データ中の男女間の偏りを可視化した事例を報告する．

このデータには以下の 6 つの値が記録されている．

- 中学校での成績
- 高校での成績
- 大学（学部）での成績
- Employment test での成績
- MBA での成績
- 給与

本事例ではこれらの値を 6 次元の説明変数 $r_i = \{x_{i1}, ..., x_{i6}\}$ として扱う．さらに，このデータには各人物の性別が記載されており，これをカテゴリ変数 c_i として扱う．

そして提案手法により「データ中に存在しないタイプの人材の例」となるサンプリング点の集合を可視化した．図 6(上) にサンプリング点の集合を散布図として可視化した例を示す．ここで散布図の点は性別で色分けされており，青は男性，赤は女性を示す．

図 6(上) では (a)(b) に男性と女性の両方が集中しており，(c)〜(f) に女性が集中している．これらの

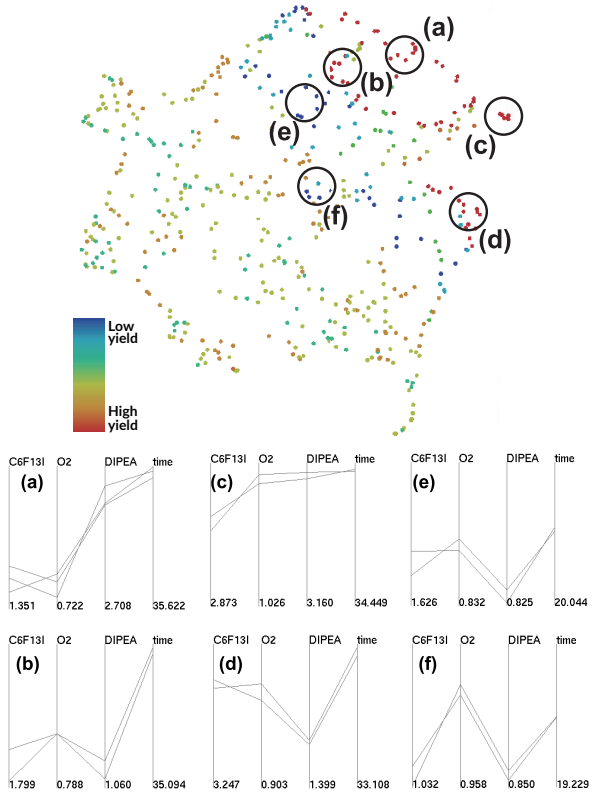

図 5. 含フッ素有機化合物の実験データを用いた実行例．(上) まだ実験されていない説明変数の例となるサンプリング点の集合を散布図で可視化した例．図中の (a)〜(d) に収率が高いと予測される赤い点が，(e)(f) に収率が低いと予測される青い点が見られる．(下)(a)〜(f) の点をマウス操作で指定して説明変数の値を平行座標プロットで表示した例．

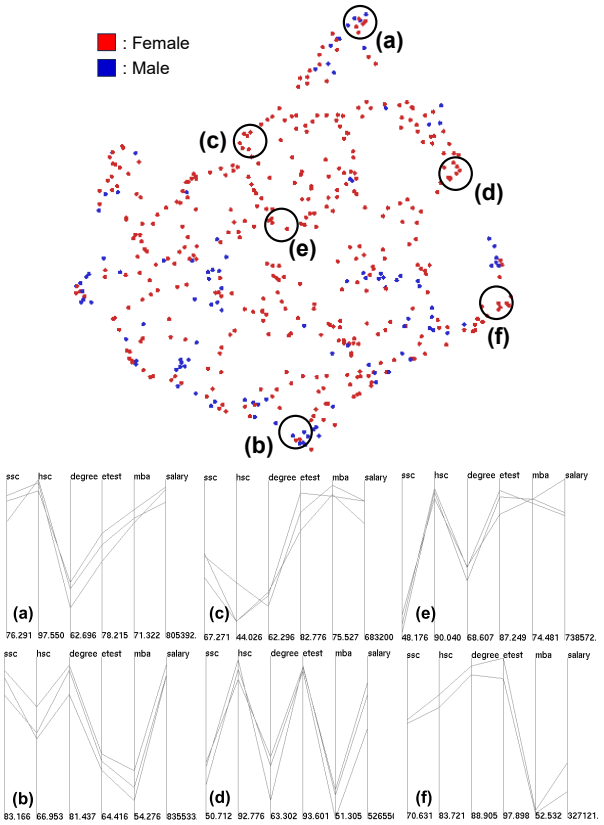

図 6. 学術成績と給与のデータを用いた実行例．(上) データ中に含まれないタイプの人材の例となるサンプリング点の集合を散布図で可視化した例．図中の (a)(b) には男性を示す青い点と女性を示す赤い点の両方が，(c)〜(f) には女性を示す赤い点のみが，それぞれ多数見られる．(下)(a)〜(f) の点をマウス操作で指定して説明変数の値を平行座標プロットで表示した例．

点における説明変数の値を図 6(下) に示す．このような表示により，男性にも女性にも見られないタイプの人材の例，女性 (または男性) にのみ不足しているタイプの人材の例を理解することができる．特に後者のタイプの人材については，散布図中のそれぞれの領域について各変数の値を注意深く観察し，男女間のバランスをとるためのデータの加工や是正を実施するか否かの意思決定を進めることができる．

5 ユーザ実験

4.1 節に示したデータとは別の，5 次元の説明変数を有する含フッ素有機化合物の実験データを用いて，高い収率をもたらす説明変数値のパターンを発見させるユーザ実験を実施した．我々はあらかじめ，自身で当該データを探索し，高い収率をもたらす説明変数値の 9 種類のパターン（図 7 の (a)〜(i)）を発見した．この 9 種類の各々について，ユーザが画面左側のクリック操作によって発見できるかを検証した．ただしユーザにはクリックの回数を最大 10 回に限定した．実験参加者は計算機科学を専攻し可視化やデータサイエンスの授業を履修したことがある 20〜26 歳の大学生・大学院生 22 名であった．

表 1 に 9 種類のパターンを発見できた実験参加者の割合を示す．発見成功率は 80%以上と 70%以下に大きく二分される結果となった．具体的には，図 5 の (b) と (c) を片方だけ発見した人，(e) と (f) を片方だけ発見した人が多く，これらの発見成功率が低い結果となった．この傾向を改善する手段として，散布図の配色が考えられる．図 5 にも示されている通り，現段階の実装では高い収率が予想されるサンプリング点は全て同様に赤色で表示されている．これを改めて，高い収率が予想されるサンプリング点のみに高い彩度を割り当てた上で，説明変数値のパターンによって異なる色相を割り当てる，というように配色を変えれば，多様なパターンの説明変数値

表 1. 高い収率をもたらす説明変数値の 9 種類のパターンに対する提案手法での発見成功者の割合.

パターン	(a)	(b)	(c)	(d)
発見成功率 (単位%)	90.9	59.1	68.2	86.4
(e)	(f)	(g)	(h)	(i)
54.5	68.2	95.5	86.4	90.9

を有するサンプリング点をユーザは積極的に選択できるようになると予想される.

6 制約と展望

本手法には現時点で以下の制約がある. a_{pot}, b_{pot}, c_{pot}, n_{pot} の 4 つのパラメータによって全く異なる可視化結果を生じてしまうため, ユーザがその調節の手間を要するという難点がある. そこで好ましいパラメータ値の自動設定方法などを検討したい. また実装上の問題として, 説明変数の数が増えることで計算時間やメモリ使用量が急増する可能性があるため, これらを抑える実装上の工夫が必要である. またサンプリング点の座標値を単純に乱数で発生させると, 毎回異なる可視化結果を生じてしまうため, これを避けるには擬似乱数を導入する必要がある.

現時点での実装はカーネル密度推定に類似した手法でサンプリング点を生成しているが, 入力データの分布によっては混合ガウス分布などを採用したほうが品質と計算量の両面において有利な可能性もある. また本研究は「疎な領域に満遍なく」サンプリング点を生成して可視化することを目的としているが, 目的に合致した領域のサンプリング点だけを可視化すればいいのであればベイズ最適化などの探索手法を用いたほうが効率的である.

本手法ではサンプリング点の目的変数の値を求める際に, 近傍の数個の個体から局所的な補間によって算出している. 一方で, 目的変数の数値分布によっては, 回帰分析を適用したほうが精度の高い形で目的変数を予測できる場合もある. これについては今後の課題として実装を進めたい.

提案手法が有効に働くと考えられる適用事例として, 以下の条件を満たすデータがあげられる.

- 多次元の各変数が連続値であるデータ. 離散値をとる変数, 順列変数, カテゴリ変数などは望ましくない.
- 色付けによってデータの分布を読み取れるデータ. 分散の大きい目的変数や, 目安として 10 種類以下の値をとるカテゴリ変数が該当する.

以上の条件を踏まえた上で, さらに多くの事例に提案手法を適用したい. 本報告で紹介した有機化合物の実験以外にも, 科学技術, ビジネス, 社会現象に関する幅広い予測問題に本手法を適用したい. また, 本報告で紹介した学術成績のデータ以外にも, 機械

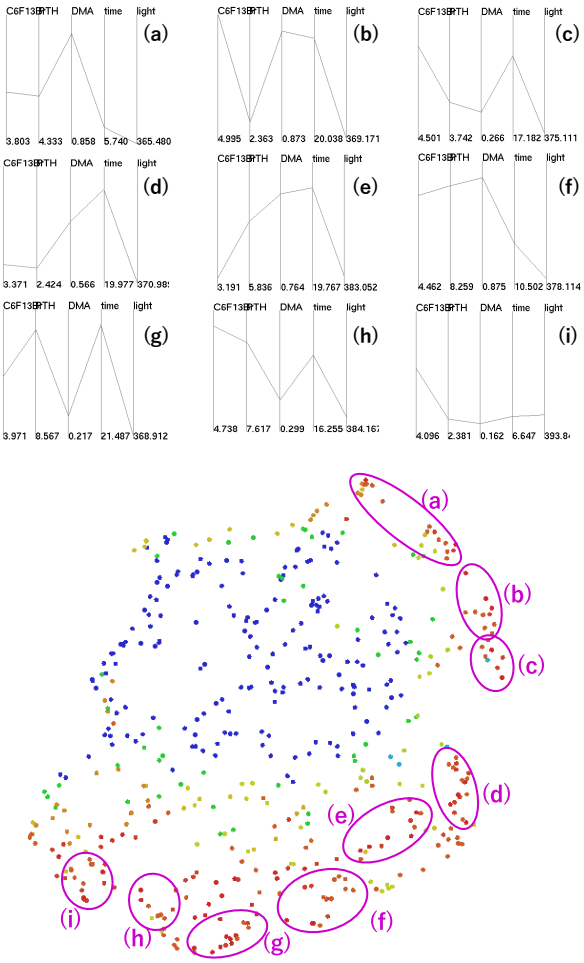

図 7. 高い収率をもたらす説明変数値の 9 種類のパターンとその散布図上での位置.

学習に用いる訓練データ全般にわたってそのデータ内のバランスを観察する目的で本手法を適用したい. さらに他の目的として, 特徴量を算出可能なデジタル作品群（例えば音楽や絵画など）に本手法を適用し, データ内に存在しない特徴量を有する作品としてどのようなものがあげられるかを観察したい.

7 まとめ・今後の課題

本報告では, 多次元空間中の疎な部位に多数のサンプリング点を生成し, その点を可視化することで, 個体が存在しない領域の分布を表現する手法を提案した. また, 含フッ素有機化合物の反応実験データと, 学術成績と給与のデータでのケーススタディをもって, 提案手法の有効性を検証した.

今後の課題として, 前章で述べた制約を解消する改良手法開発, さらなる適用事例の開拓, より実用シナリオに近いユーザ評価実験に取り組みたい.

参考文献

[1] Campus Recruitment Data, `https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement`.

[2] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah. MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. *IEEE Transactions on Software Engineering*, 44(6):534–550, 2017.

[3] L. Blouvshtein and D. Cohen-Or. Outlier Detection for Robust Multi-Dimensional Scaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2273–2279, 2018.

[4] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer. Clusters, Trends, and Outliers: How Immersive Technologies Can Facilitate the Collaborative Analysis of Multidimensional Data. In *Proceedings of the CHI conference on human factors in computing systems*, pp. 1–12, 2018.

[5] N. Cao, Y.-R. Lin, D. Gotz, and F. Du. Z-Glyph: Visualizing outliers in multivariate data. *Information Visualization*, 17(1):22–40, 2017.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[7] F. Cheung. A Figure One Web Tool for Visualization of Experimental Designs. *Journal of Open Research Software*, 8(6), 2020.

[8] E. P. dos S. Amorim, E. V. Brazil, J. Daniels, P. Joia, L. G. Nonato, and M. C. Sousa. LAMP: Exploring high-dimensional spacing through backward multidimensional projection. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 53–62, 2012.

[9] M. Espadoto, G. Appleby, A. Suh, and D. Cashman. Unprojection: Leveraging Inverse-projections for Visual Analytics of High-dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1559–1572, 2021.

[10] G. Grinstein, M. Trutschl, and U. Cvek. High-Dimensional Visualizations. In *Proceedings of the Visual Data Mining Workshop (KDD)*, Vol. 2, p. 120, 2001.

[11] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf. The State of the Art in Flow Visualization: Dense and Texture-Based Techniques. *Computer Graphics Forum*, 23(2):203–221, 2004.

[12] R. Mazza. Introduction to Information Visualization. *Springer*, 2009.

[13] Y. Nakai, T. Itoh, H. Takahashi, S. Nakashima, and T. Yamamoto. Hierarchical Data Visualization of Gender Difference: Application to Feeling of Temperature. In *27th International Conference Information Visualisation (IV2023)*, pp. 178–183, 2023.

[14] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering. In *IEEE Symposium on Information Visualization*, pp. 89–96, 2004.

[15] K. Tagami and T. Yajima. Halogen-bond-promoted hydroxyperfluoroalkylation of olefins with molecular oxygen under visible-light irradiation. *Asian Journal of Organic Chemistry*, 12(8):e202300273, 2023.

[16] F. Taverna, J. Goveia, T. K. Karakach, S. Khan, K. Rohlenova, L. Treps, A. Subramanian, L. Schoonjans, M. Dewerchin, G. Eelen, and P. Carmeliet. BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. *Nucleic Acids Research*, 48(W1):W385–W394, 2020.

[17] N. Tovanich, P. Dragicevic, and P. Isenberg. Gender in 30 Years of IEEE Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):497–507, 2022.

[18] I. Viola, A. Kanitsar, and M. E. Groller. Importance-driven feature enhancement in volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):408–418, 2005.

[19] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):966–975, 2022.

[20] Y. Wang, A. Machado, and A. Telea. Quantitative and Qualitative Comparison of Decision-Map Techniques for Explaining Classification Models. *Algorithms*, 16(9):1–26, 2023.

[21] Y. Wang and A. Telea. Fundamental Limitations of Inverse Projections and Decision Maps. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2024.

[22] P. C. Wong and R. D. Bergeron. 30 Years of Multidimensional Multivariate Visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, Vol. 2, pp. 3–33, 1994.

[23] S. Wu and T. W. S. Chow. PRSOM: a new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks*, 16(6):1362–1380, 2005.

[24] J. Xia, H. L. N, M. L. Mayer, O. M. Pena, and R. E. W. Hancock. INVEX-a web-based tool for integrative visualization of expression data. *Bioinformatics*, 29(24):3232–3234, 2013.

[25] J. Zhao, M. Fan, and M. Feng. ChartSeer: Interactive Steering Exploratory Visual Analysis With Machine Intelligence. *IEEE Transactions on Visualization and Computer Graphics*, 28(3):1500–1513, 2022.

未来ビジョン

本研究の主な未来ビジョンは以下の 2 点である．

1 点目は「非存在の可視化」という概念を多次元データ以外の多様なデータに適用するという点である．本報告では多次元データに限定してその空間中の非存在の可視化を試みた．一方で，可視化が対象とするデータは多次元データの他にも，ネットワーク，時系列データ，地理データ，テキストデータなど多岐にわたる．あるいは，主に科学技術を対象とした物理空間でのデータもある．将来的なビジョンのひとつとして，これらの多様なデータを対象として可視化技術全般において「非存在の可視化」という概念を実現したい．あくまでも例として，毎年見られるはずなのに今年見られない現象を可視化する，ネットワークの中で連結されて然るべき関係を有するのにエッジで連結されていない部位を可視化する，といった技術の研究開発に臨みたい．

2 点目は「データ中の非存在の発見」という工程を価値創出につなげるという点である．非存在の可視化によって「何が存在しないか」を知ることができれば，存在しないものを新しく生成してみようという動機付けが生まれる可能性がある．例えば，多次元の音楽特徴量空間に多数の楽曲を配置して，そこから非存在な音楽特徴量を導くことができれば，非存在な音楽特徴量を有する楽曲を生成 AI に作らせてみる，という試行錯誤的な創造につなげることができる．あるいは，多次元の味覚特徴量空間に多数の食品を配置して，そこから非存在な味覚特徴量を導くことができれば，非存在な味覚特徴量を有する架空の食品を生成して試しに味見してみることが可能になる．従来の「何が存在するか」を知る可視化と違って，「何が存在しないか」を知るための可視化を確立することで，これまで存在しなかった新しい作品や商品の創出を促進できるのではないか…という議論を展開したい．