# Plotshop：散布図上で２次元データ分布を作成及び編集するための対話的なシステム

浅井　健太郎†　福里　司†　五十嵐　健夫†

**概要**.　本研究では，２次元数値データを散布図上で作成及び，編集するための対話的なシステムを提案する．これまでの２次元数値データを作成する方法は，単純な関数（例：正規分布）を用いるものが主であった．しかし，このような手法で得られるデータは単純な形式のもので，(1)複数の統計的手法の比較や，(2)新たな統計アルゴリズムの適用といった，アルゴリズムの挙動を確認するためのテストデータとして不十分であった．このような問題を解決するために，本研究では，散布図上で２次元数値データ分布を視覚的に作成，編集するための機能（データの追加，削除，変形）を導入している．本システムは，クラスタリングや回帰分析といった統計的なデータ分析手法の特性の理解と比較のための道具としての応用が期待できる．

## 1　はじめに

統計的なデータ分析（例：クラスタリングや回帰分析）を行う場合，新たに開発されたアルゴリズムや既存のアルゴリズムの性質を「理解」することと，「比較」することは非常に重要である．例えば，クラスタリングの手法には，外れ値に対して頑健な手法とそうではない手法が存在する．このような性質を理解するためには，各アルゴリズムにとって得意（or 不得意）な性質を持ったテストデータ（データセット）を事前に用意し，それを入力した際の出力結果を確認する必要がある．テストデータを用いる主な理由として，複数の性質やノイズ成分が含まれる可能性が高い実際のデータとは異なり，テストデータは，単一の性質のみが異なるデータを用意できる点が挙げられる（例：外れ値が多いデータセットと少ないデータセットを用意することで，外れ値に頑強かどうかを検証することができる）．

テストデータを作成する方法として，データ分析用のライブラリに搭載された関数呼び出しによる方法が挙げられる（例：scikit-learn[1] の make_regression や make_classification 機能）．しかし，作成できるテストデータは関数の引数で表現できる形式のデータセットに限られ，より複雑な性質を持ったテストデータを作成することは困難である．また，Data Augmentation 手法を用いることで特定の性質を満たす（深層学習用の）トレーニングデータを作成する方法もあるものの，ユーザが直感的かつ自由にテストデータを作成することは難しい．

このような背景の下，本研究では，２次元のテストデータを自由かつ対話的に作成できるようなシステムを提案する．本システムの狙いは，ユーザが散布図上で描画エディタのような GUI 操作（例：線を描くなど）を行うことで，２次元のデータ分布の作成と既存のデータセットの編集を実現することである．特に，データ分析用アルゴリズムの性質の理解することに特化した編集機能を用いることで，データ分析の視点からも支援することができる．既存のデータ分析ツールでは，CUI 上でのコマンド実行を主に用いているため，データセット自体の作成と編集と可視化は全て個別に行われていた．その一方で，提案システムは，上記の全ての手順を同時に行うことができる．また，Matejka らの研究[8]でも示されるように，たとえ統計値が同じであっても様々なデータの形が考えられるため，単純に統計値を基にデータを生成する方法では，データの形に関する性質の制御には不十分である．一方，本システムでは，データの形と統計値を逐次確認しながら，ユーザがテストデータを作成できるため，データの形の違いによって分析アルゴリズムの挙動がどのように左右されるのかを理解できる点で意義がある．

提案システムを利用する主な利点として，第一に新規（或いは既存）の統計的な分析や複数の分析アルゴリズムを組み合わせた際の挙動を確認できる点が挙げられる．特に，複数の分析アルゴリズムを組み合わせる場合（例：アンサンブル分類器），組み合わせる個々のアルゴリズムの性質のみから，性質の把握することは難しい．そこで，我々のシステムを用いて作成したテストデータを分析することで，全体的な挙動を確認できる．本システムでは，分類器


†東京大学

を複数選択することで，それらを組み合わせたものによる分類結果を表示する機能を有している．また，一部のデータしか手元になく，データ分析を行うことができない場面が存在する．このような状況下において，本システムを用いることで実データを参考しながら仮のデータ（テストデータ）を作成し，データ分析の傾向を事前に検証することができる．つまり，実際のデータが収集できた直後，円滑に分析を始めるための補助ツールとしても期待できる．

## 2 関連研究

散布図中のデータを選択するための手法として，scatterplot brushing[2]が挙げられる．これは，散布図行列内に存在する任意の散布図に対し，ユーザが短形領域を指定することで，指定領域内のデータ点を選択できる手法である．更に，選択されたデータ点が，他の散布図上でどこに位置するかを確認することができる．また，データの可視化方法して，Liu らは Data Illustrator[3]を提案した．この研究では，データの特徴とベクタ図形の対応付けを行うことで，ユーザにとって「理解しやすい」データの提示を実現している．

データの可視化結果に対して直接インタラクションを行う方法として，対話的なデータ分析ツールの Polaris [4]や Tableau[5]が挙げられる．これは，選択されたデータの確認やフィルタリング処理を直接行うことができる．また，新たなアルゴリズムや複数の手法を併用する等，より詳細なデータ分析の手法を試す方法として，python 上のライブラリ matplotlib[6]と plotly[7]が挙げられる．これらのライブラリに事前に用意された関数を用いることで，ユーザはプログラム中からデータを可視化と選択したデータの詳細を確認することができる．しかし，上記のインタラクションは，データの中身を編集するための方法ではなく，あくまでもデータの確認作業に主眼を置いている．

また，Matejka らの研究[8]では，生成したいデータの概形を与えるとその概形に沿った固定された統計値を持つデータを生成することができる．しかし，データ点の値は最適化処理によって決定されるため，データの性質を変更しながらアルゴリズムの挙動を逐次確認するような状況には不向きである．

本システムは，GUI 上でデータを直接作成及び編集できる点が他のシステムと大きく異なり，最も画期的な点である．また新たな試みとして，GUI 上で作成及び編集したテストデータに対し，実験的にデータ分析を行う機能を構築する．この機能は，データの作成（編集）段階で，最適なデータ分析手法を調査できる点で大きな意義があると考えている．更に，従来の CUI 操作とは異なり，本システムではデータの編集，作成と可視化の作業を同時に行うことができるため，作業の効率化も実現している．

## 3 ユーザインタフェース

### 3.1 本システムの目的

本システムは主に（1）既存の 2 次元数値データを編集することと，（2）一からデータを作成することの二つの処理を目的としている．カテゴリや日付といった数値以外のデータは扱わないものとする．本研究で開発したインタフェースを図 1 に示す．

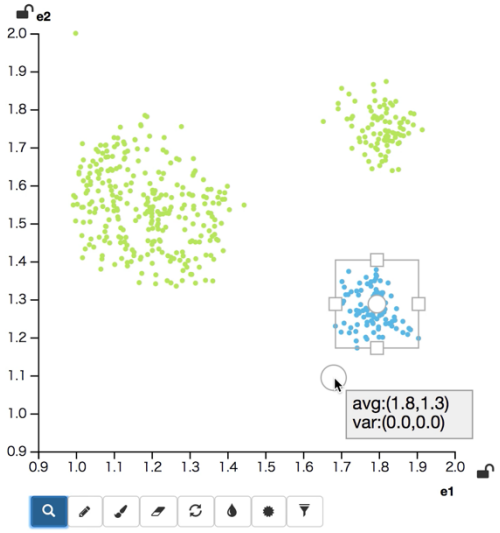



**図 1．提案システムのスクリーンショット．（上部）散布図パネル，（下部）データ操作用のツールボタンが配置されている．ボタン選択によって，選択されたデータ（青点）に対する操作を切り替えることができる．**

### 3.2 インタフェースの説明

インタフェースには，対話的な操作を行うための散布図パネルと，ツールパネルが表示される．図 1 では，ユーザが一部のデータ点を選択した様子を示す．選択されたデータ点は指定された色（青色）で表示され，それらの統計量（平均と分散）が同時に表示される．次に，ユーザはツールパネルを用いることで，選択されたデータ点に編集操作を行うことができる．操作機能として，「統計量の調査」，「ブラシ選択」，「バケツ選択」，「エアブラシ」，「消しゴム」，「拡大・縮小」，「移動」，「回転」，「追加作成」を設計している．また，分析手法を比較するために，手法による分析結果を散布図上で可視化する「分析結果の表示」機能も有している．

### 3.3 散布図上でのデータ点の作成と編集機能

散布図上で行う操作は，「作成」「選択」「編集」の三種類に大別される．本論文では，各機能の詳細の一部を紹介する．なお各操作において，対象のデータ点群の統計量をリアルタイムに可視化することができる．

#### 3.3.1 「作成」機能

**エアブラシ：** ユーザは散布図上に直接ストロークを描くことができる．本システムでは，描いたストロークを中心とする確率分布（ユーザは正規分布と一様分布の二種類を選択可能）を定義し，点群データを追加することができる．また，本システムは，既存のペイントツールと同様に，エアブラシの色（クラスタの種類）や太さ（分布範囲），時間あたりに作成されるデータ点の量を設定する機能を有している．

**データ追加：** 選択されているデータ点から，それらの点を生成するための正規分布モデル（平均及び分散行列）を仮定し，新しいデータ点を追加する．但し，選択されたデータ点（既存のデータ）の分布は考慮せず，モデルのみに基づいて，ランダムにデータ点を生成する．ユーザはスライダー操作によって，追加するデータ点の個数を調整することができる．本機能は python の NumPy ライブラリの関数を用いている[10]．

**データ削減：** データ追加同様選択されているデータ点を基に正規分布を仮定し，そのモデルを基にデータを削減する．選択されている各点に正規分布を元に座標から確率を計算し，削除される確率として点を選択する．削除する点の個数はスライダーを用いて調整することができる．

エアブラシ：先端の円で示される分布を元にデータ点を作成する．

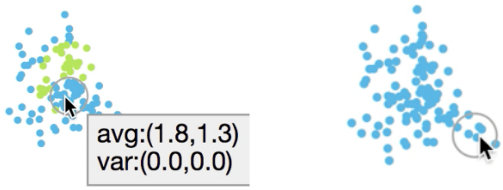
データ追加：既存のデータ分布を元に正規分布を仮定し，データを作成する．

図 2-1. 「作成」機能

#### 3.3.2 「選択」機能

**ブラシ選択：** 「エアブラシ」と同様に，ストロークを描く操作でデータ点を選択することができる．選択されたデータ点は，自動で配色が変更され，どの点群データが選択されているかを視覚的に確認することができる（選択されたデータ点：青点）．

**バケツ選択：** バケツ操作では，散布図中で，データ点が密集している部分をクリックすることで，付近のデータ点をまとめて選択できる．本システムでは，DBSCAN によるクラスタリング結果を参考に，クリックしたデータ点と同じクラスタに属する点群を選択する．この機能を用いることで，データ分析の視点から効率的にデータ点の選択ができる．




ブラシ選択：ドラッグにより，先端の円内部に存在するデータ点を選択する．

バケツ選択：一回のクリックで付近に存在するデータを全て選択する．

図 2-2. 「選択」機能

#### 3.3.3 「編集」機能

本システムでは，選択したデータ点を囲むバウンディングボックス（矩形領域）を構築する．ユーザは，バウンディングボックスを基に，一般的なソフトウェア同様の操作，「平行移動」「拡大縮小」「回転」操作を行うことができる．

#### 3.3.4 「分析結果の表示」機能

クラスタリングや回帰を用いたデータの分析結果を散布図上で可視化する機能を有している．可視化方法として，クラスタリングの場合はクラスタごとに異なる色付け処理を行い，回帰では回帰直線を表示するものとなっている．本システムでは分析手法として，四種類のクラスタリング手法「k-means 法」「DBSCAN 法」「Ward 法」「MeanShift 法」と二種類の回帰分析手法「線形回帰」「RANSAC 法」を用意しており，ユーザはチェックボックスをクリックすることで，分析手法を切り替えることができる．本論文では，上記の手法は scikit-learn に用意されている関数を用いて実装した．今後は，用意されている手法だけでなく，ユーザの書いた任意の外部コードを適用できるようにしていきたい．

## 4 結果

提案システムで作成したデータ分布の一例を図 3.1 に示す．これらの分布は，ベンチマーク[1][9]として挙げられていたデータ分布例を参考に作成したものであり，いずれも従来手法（関数の呼び出し）では作成自体が困難でかつ，統計的な分析アルゴリ

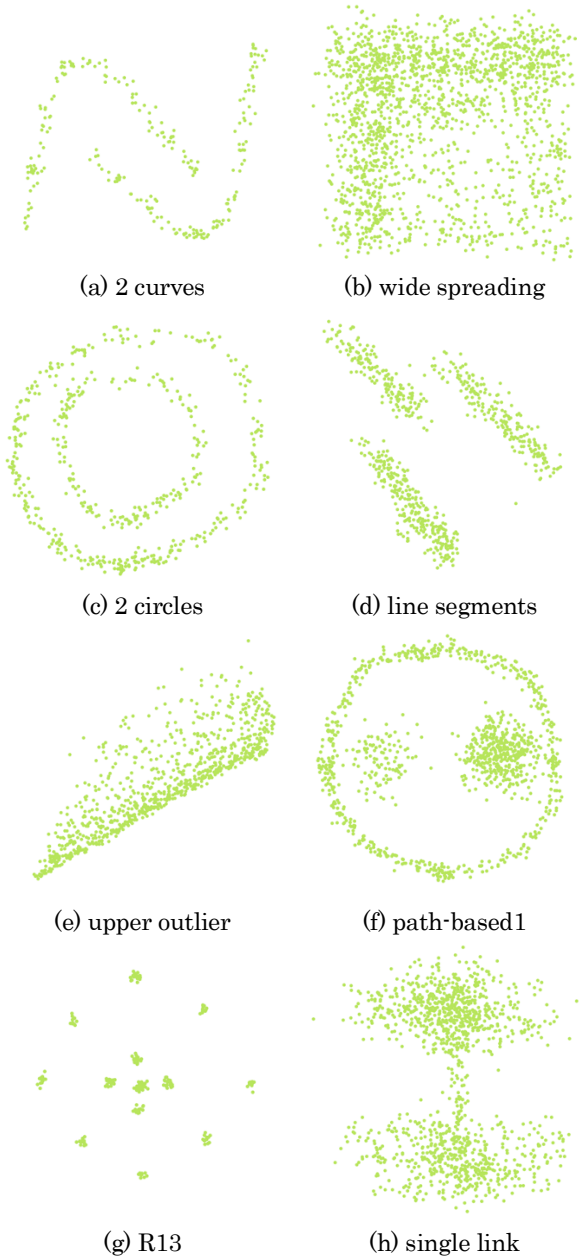



図 3-1. テストデータの作成例

ズム手法の比較及び調査を行うにあたり用いられるデータ分布である．

## 5 利用例

### 5.1 クラスタリング手法の挙動比較

GUI 上でのテストデータ作成及び分析が，有効になる場面の一例として，クラスタ手法の比較検証が挙げられる．例えば，クラスタ間の距離が十分に離れたデータ分布（図 4(1)(2)）の場合，多くのクラスタリング手法が有効である．しかし，クラスタ間の距離が近いデータ分布（図 4(3)(4)）の場合，適切なクラスタリングができない可能性がある．つまり，入力となるデータ分布に対し，どのようなクラスタリング手法が適切かといった手法自体に関する性質の調査が重要となる．各クラスタリング手法の結果が異なる主な要因としては，(1)クラスタ間の距離，(2)クラスタ間の密度の違い，(3)外れ値などが挙げられる．図 4-1 は，クラスタ間の距離の違いが，k-means 法や Ward 法を用いた階層的クラスタリングの結果にどのような影響を与えるかを確認したものである．この結果から，クラスタ間の距離が互いに離れている場合は適切なクラスタリングができるものの，クラスタ間の距離を近づけた場合，k-means 法では適切なクラスタリングができないことが視覚的に確認できる（図 4-1(3)）．このような検証は，単独のクラスタリング手法の比較に限らず，実際の現場でも利用される複数の手法を組み合わせた分析手法の検証にも応用できる（例：複数のクラスタリング手法で得られたそれぞれの結果を基に，多数決方式で最終的なクラスタを決定する手法）．

図 4-2 は，二つのクラスタが存在するテストデータの一方のクラスタにノイズデータを追加し，DBSCAN によるクラスタリング結果を可視化したものである．この結果から，ノイズデータがある程度増えた時，当初二つあったクラスタが一つのクラスタとして分類されたことが確認できた．

以上をまとめると，本システムを用いることで，ユーザは性質の異なるテストデータを容易に作成することができる．また，作成したテストデータの分

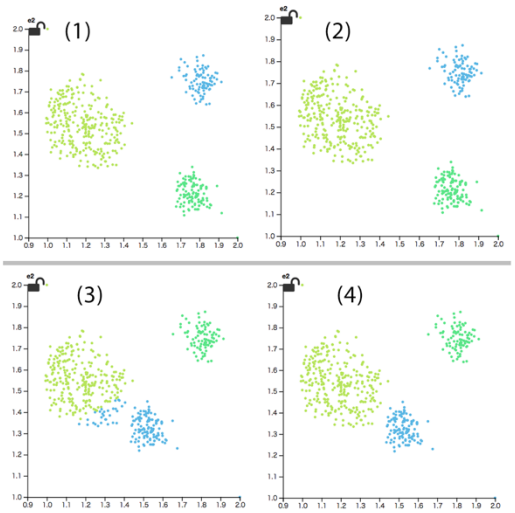



図 4-1. クラスタ間の距離の違い（(1) (2)：距離が十分に離れている場合，(3) (4)：距離が近い場合）に基づくクラスタリング結果．(1) (3) は k-means 法，(2) (4) は Ward 法を用いた階層的なクラスタリング結果（提案システムの「分析結果の表示機能」を用いて可視化及び比較）．

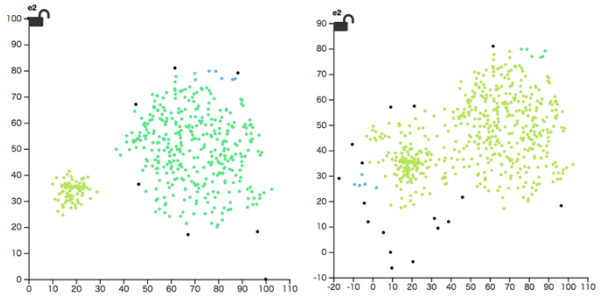


**図 4-2. ノイズデータを追加した場合の DBSCAN 法によるクラスタリング結果．ノイズデータをある程度追加すると，二つのクラスタとして判定されたテストデータが，一つのクラスタとして判定されてしまったことが分かる（黒点：DBSCAN 法によって外れ値と検出されたデータ点）．**

布の違いから，クラスタリング結果にどのような影響があるのかを事前に確認することができる．

### 5.2 回帰分析手法の挙動比較

回帰分析手法の中には，外れ値をはじめとする様々な要因で，推定結果が大きく異なる手法が存在する．図 4-3 は，平均二乗誤差最小化による推定と二乗誤差刈り込み平均最小化による推定結果に対し，ノイズデータがどの程度の影響を与えるかを確認したものである．この結果から，平均二乗誤差最小法は外れ値の影響を大きく受けることが分かる．また，二乗誤差刈り込み平均最小法は平均二乗平均誤差最小法と比べ，より適切な回帰分析ができることが理解できる．このような検証方法を通して，ユーザは回帰分析手法の頑健性を確認することができる．

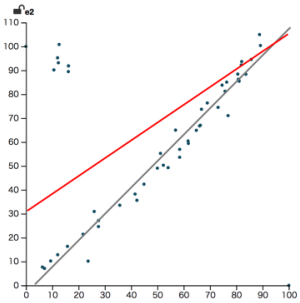
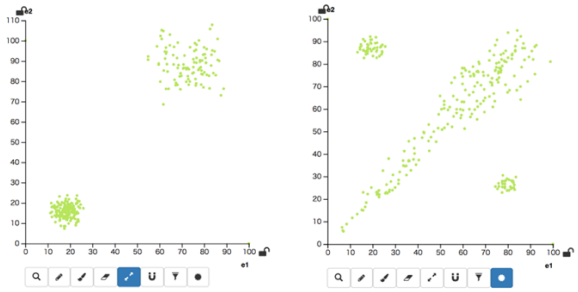


**図 4－3，線形回帰モデルの推定（赤線：単純な平均二乗誤差最小化による推定，黒線：二乗誤差刈り込み平均最小化による推定）．**

### 5.3 複雑なデータ分布の作成

従来のサンプルデータを作成する主なライブラリとして，scikit-learn が挙げられる．作成する手順としては，scikit-learn に事前に用意されている関数「make_classification」を使うものである．make_classification では，いくつかの引数が用意されている．しかし，この関数では，全てのクラスタが均一であることを仮定しているため，より複雑な構造（例：複数の異なる正規分布で定義されるクラスタ群）のパラメータ（例：分散や平均）を自由に変更することはできない．つまり，密度の異なるクラスタが混在するようなテストデータ(図 5-1(1))を用意することができない．

また，回帰直線（曲線）とクラスタが混在しているデータ分布(図 5-1(2))を作成する場合，従来は「make_regression」関数を実行する必要がある．その後，複雑なコマンド操作でデータの可視化と平行移動を行わなければならず，非常に難しいと言える．その一方，我々のシステムでは，上記のような複雑なテストデータを容易に作成及び編集することができる．



**(1) 分散の異なる二つの正規分布**　**(2) 回帰直線とクラスタが混在する分布**

**図 5-1. 複雑なテストデータの一例**

## 6 今後の課題

今後の課題として，より直観的な編集機能の確立を目指し，研究に従事する予定である．

**コーディング初学者によるデータ処理の支援：** 近年，ビジネス領域において機械学習技術が注目されつつある．それに伴い，コーディング初心者が機械学習を使ってデータ分析を試みる機会が今後増えていくことが予想される．しかし，実際のビジネスで使用される分析ツールは，対話的な操作を使用できないツールが主であり，可視化と分析，データの編集が完全に分離されている．これは，初学者によるデータ分析にとって，大きな妨げとなることが予想される．そこで，提案システムの散布図上のデータを直接選択と編集，及び可視化機能を用いることで，コーディングによる操作に慣れていない人材にも，より直感的なデータ分析の方法を提供できると考えている．今後，現実のデータ分析でより使いやすい形に拡張する予定である．

**教育用システムとしての実装：** 近年，機械学習に

関する知識を習得するための初心者向けのサービスも登場している．その中でも特に，外れ値やクラスタの距離，回帰といったノイズ成分による影響などの特性を知ることができれば，分析手法の理解が進むことが予想される．そこで，本システムを基に，各手法とその特性を確認できるオープン教材を作成することで，初学者が機械学習を学ぶ上での大きな助けとなる．また，手法の理解を深めるための編集機能を開発する予定である．

**多次元データの分析のための拡張：** 多次元データ作成のために，我々は多次元データ中の任意の２次元の組み合わせでの散布図を，ユーザが我々のシステムを用いて繰り返し定義することで実現することを構想している．但し，多次元データの可視化自体が困難であるため，洗練された方法と我々の研究の組み合わせを模索する必要がある．

**外部ツール（ソフトウェア）との連携：** 実際の研究や開発の導入を支援するために，python や R といった言語や，それを用いたソフトウェアとの連携を考えている．これらツールとの連携により，現実のデータ分析において，本システムの使用が進むと期待される．

**将来システムに実装したい機能：** 現時点で「作成」，「選択」，「編集」操作を，より効率的に行うために，以下の機能案を考えている．

1) 図形描画： 図形を散布図上に描くことで，その内部にデータ点を生成する機能.
2) 3D データ読込: 既存の３D モデリングツールで作成した３D メッシュデータを基に， ３D 形状でモデリングしやすいデータ分布を作成する機能.
3) 短形領域に対する拡大縮小だけではなく，自由形状での拡大縮小を行う機能.
4) ブラシ機能について：正規分布と一様分布以外の確率分布のブラシ形状.
5) データ選択について：選択されたデータの生成モデルを正規分布以外で仮定する機能.（例：ガウス混合分布）
6) スタンプツール：一般的なデータ分布を１回のクリックで生成するためのツール.
7) 編集方法について：統計グラフ（例：ヒストグラム）を直接操作することで，データ分布を編集する機能.

## 7 むすび

本研究では GUI 上での直感的な操作で２次元データを編集及び作成するための対話的なツールを開発した．このツールによって，従来はコマンド操作で行なっていた諸々のデータに対する編集や，テストデータの作成，可視化の過程を分離せず，一貫して行うことが可能となった．また，データ編集や作成の過程にデータ分析の知見を交えることでより効率的な操作を実現した．更に，本システムを用いることで，実際のデータを想定した上での事前準備することができる点で，意義が大きいと考えている．これらの成果を発展させて，研究の場においては新しいアルゴリズムの開発，実用の場では，持っている実際のデータセットに適合するように複数の手法を比較する場面において，本システムの導入を目指していきたいと考えている．

## 謝辞

本研究は JST CREST grant JPMJCR17A1 の助成を受けたものである．

## 参考文献

[1] Scikit-learn. http://scikit-learn.org/stable/ (閲覧日：2018/05/29)
[2] RA Becker and WS Cleveland. “Brushing scatterplots.” *Technometrics*, 29(2), pp.127-142, 1987.
[3] Z.Liu, J.Thompson, A.Wilson, M.Dontcheva, J.Delorey, S.Grigg, and J.Stasko. “Data Illustrator: Augmenting Vector Design Tools with Lazy Data Binding for Expressive Visualization Authoring.” *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.123:1-123:13, 2018.
[4] C.Stolte, D.Tang, and P.Hanrahan. “Polaris: A system for Query, Analysis, and Visualization of Multidimensional Relational Databases.” *IEEE Transactions on Visualization and Computer Graphics,* 8(1), pp.52-65, 2002.
[5] Tableau. https://www.tableau.com/ja-jp (閲覧日：2018/05/29)
[6] Matplotlib. https://matplotlib.org/ (閲覧日：2018/05/29)
[7] Plotly. https://plot.ly/ (閲覧日：2018/05/29)
[8] Matejka and Fitzmaurice, "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1290-1294, 2017
[9] Clustering basic benchmark. https://cs.joensuu.fi/sipu/datasets/ （閲覧日：2018/05/29）
[10] NumPy. https://docs.scipy.org/doc/numpy/