

訓練データの比較可視化のための閾値自動設定

高坂 夏怜[†] 伊藤 貴之[†]

[†] お茶の水女子大学大学院人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: [†]{kosaka.karen, itot}@is.ocha.ac.jp

あらまし 機械学習の用途の多様化に伴い、訓練データの質検証と比較が重要な工程となっている。例えば転移学習において、ソースとターゲットの質の違いを検証することで、モデルの精度低下を防げる場合がある。しかし深層学習の訓練データ群は大規模化しており、その解析は容易ではない。この解決の一手法として我々は訓練データ検証のための可視化に取り組んでいる。本研究では複数の訓練データを同一画面に可視化することによって分布の違いを確かめ、質の違いを発見しやすい視覚的分析を実現する。現段階の実装では、散布図上の点群を連結する三角メッシュを生成し、辺の長い三角形を削除することで、画面上で近い距離に配置された点群を多角形で表現する。本報告では、三角形削除のための辺の長さの閾値をラベルごとに独立に自動設定する手法を提案する。

キーワード 可視化手法、機械学習、訓練データ群

1 研究背景

機械学習を使う目的やデータが多様化していることから、訓練データの比較が重要になっている。例えば転移学習において、ソースデータとターゲットデータの質の違いが訓練後のモデルの精度を下げることが知られている。その他にも例えば、モデルを作成する過程で複数のデータセットの中から訓練データを選定する場合など、データ群の違いを解析することには意義がある。一方で近年、機械学習で使われる訓練データ群は大規模化しており、それにともなってデータ比較の難易度も高まっている。そのため、訓練データを定量的にのみならず、質的に比較することが重要となっており、その手段として可視化が有効であると考えられる。

機械学習のモデルに使われるデータセットに特化した可視化手法として、Xiang ら [10] の研究がある。Xiang らはデータ群を階層的に表示し、さらに正しいラベルへ付け替える手法を提案した。本研究も機械学習のモデルに使われるデータセットである訓練データ群の可視化を行なっているが、Xiang らの研究では画像に付与されているラベルの誤りを正すことに特化しており、1つのデータセットを可視化することが前提となっている。一方で本研究はデータ群の比較に特化しており、複数のデータ群を同時に可視化することを前提としている。そのため本研究では複数のデータセットを比較して、データ群の特徴の分布の違いを把握することや、誤ラベルではないがデータセットで違うラベルがついているデータ群の把握(図 1)を目標としている。



図 1 片方のデータセットには ambulance、もう片方のデータセットには car というラベルがついている例

本研究では、対象とする訓練データ群を以下のように定義する。1 個の訓練データセットは多数の標本で構成される。ここでいう標本とは、画像ファイル、音声ファイル、文書ファイルなどを想定する。現段階の我々の実装では静止画像を対象とする。各標本からは多次元ベクトルとなる特微量が算出され、さらに 1 個以上のラベルが付与される。ただし、現段階の我々の実装では、各標本は排他的に 1 個のラベルを有するものとする。2 個以上の訓練データセットを同一画面に可視化する。このとき全ての訓練データセットにおいて同一の特微量が算出される。各訓練データセットに付与されるラベルは完全に同一でなくてもよい。このような訓練データ群を可視化するための要件として、本研究では以下を掲げる。

要件 1：複数の訓練データ群を同一画面に可視化することで、訓練データ間の分布の違いを表現する。

要件 2：訓練データに付与された各ラベルについて、類似する標本群がどのように分布するか、外れ値となる標本群がどのように分布するか、といった点が理解しやすい表現を実現する。

要件 3：同一のラベルを付与された標本群が、訓練データによってどのように分布の違いを有するかを比較しやすい表現を実現する。以上の要件を満たすために、本研究では以下のようないくつかの可視化手法を提案する。

- 訓練データ群に含まれる全ての標本に対して同一の次元削減手法を適用し、全ての標本を同一の画面空間に写像する。これにより要件 1 を満たす。
- 各々の訓練データで同一のクラスを付与された標本群に対して、画面上で高い密度で分布する標本群を多角形で囲んで表示する。これにより要件 2 を満たす。
- 複数の訓練データに対して、同一のクラスを付与された標本群に同一の色相を与える。これにより要件 3 を満たす。

本研究では上述のような可視化手法を適用することにより、機械学習のモデルの精度を下げる要因をユーザに提示することを目標とする。本報告は提案手法の中でも特に、多角形による囲み表示のための適切な閾値をクラスごとに独立に自動設定す

る手法を 3.3 節にて新しく提案する。

2 関連研究

2.1 訓練データの比較

本研究は複数の訓練データの比較を目的としている。このような比較が必要となる代表例として、機械学習の一種である転移学習 [4] があげられる。転移学習は、異なる領域やタスクへ情報を転移しながら学習を進める機械学習手法であり、コンピュータビジョンなどの分野で手動ラベリングの負担を軽減するためによく用いられる。転移学習の主な問題点として、異なるドメイン間の分布の不一致がもたらす影響が知られており、この問題点を解決するために多くの研究が発表されている。学習済みモデルを適用する研究 [5] では、同一の構造を有するネットワーク間で相手の学習結果を互いに壊さずに取り込むことを目指している。また、CNN (Convolutional Neural Network) の下層を利用して情報を転移する表現学習という方法もある。代表的な手法としてオートエンコーダー (AutoEncoder) [6] がある。オートエンコーダーはニューラルネットワークの一種で、情報量を小さくした特徴表現を得ることができる。

Ma ら [7] の研究では、転移学習に使うデータの例として Office-31 のデータセットを用いている。Office-31 は転移学習のアルゴリズムを実証するために広く利用されている実世界のデータセットである。Ma ら [7] の研究では、Amazon の商品ページの画像 ("amazon", 合計 2817 枚) をソースドメインデータとして、ウェブカメラの写真 ("webcam", 合計 795 枚) をターゲットドメインデータとして使用している。また 2 つのモデルで性能が類似しているクラスでは、両ドメインの画像は資格や物体の外観などの特徴が共通している。一方、両モデルで性能が異なるクラスでは、両モデル間でパターンが大きく異なる。このようなデータセットによる質の違いの存在を、本研究では可視化によって発見することを目標とする。また先ほどの例のように、質の違うデータ群が存在することで、モデルの精度を下げる可能性がある。

本研究では、質の違いによりモデルの精度を下げる可能性を可視化により発見することで、どのようなデータを使えばより高い精度のモデルを作れるかをユーザが探索できることを目標にする。

2.2 機械学習のための可視化

Bernard ら [8] はデータのラベル付けをするインターフェースを開発した。データのラベリングには、機械学習（特に能動学習）ではモデルを中心としたアプローチ、可視化ではユーザーを中心としたアプローチを採用している。この研究では、これらの異なるラベリング戦略の性能を評価・比較するための実験を実施した。結果として、次元削減によってクラス分布が分離されていれば、視覚的対話型ラベリングは能動学習を上回ることがわかった。さらに、次元削減と、学習モデルの内部の状態を明らかにする可視化を組み合わせることで、視覚的対話型ラベリングの性能が向上することがわかった。Moehrman ら

はユーザーが対話的に画像データを選択し、ラベリングができるようなインターフェースを開発した [9]。これはラベリング作業を簡略化し、ユーザーの負担を軽減することを目的としている。

機械学習のモデルに使われるデータセットに特化した可視化手法として、Xiang らはデータ群を階層的に表示し、さらに正しいラベルへ付け替える手法 [10] を提案した。正しいラベルへの付け替えについては、データ群を t-SNE で次元削減し階層的に表示した後、その中から誤っているラベルがついたデータ群の一部をユーザーが付け直す。DUTI という少ないデータ群からすべてのラベル群を付けかえるアルゴリズム [11] を用いて、ラベル全体を付け直す。データ量が多すぎるときの絞り込み操作を容易にするために階層的な表示を採用している。この絞り込み操作の際に、ランダムに元データからデータを取得してしまうと誤った可能性のついているラベルも一緒に消してしまうことになるので、そのようなデータを削除しないようにデータを削減している。

また Swabha らは、データセットの品質を解析し可視化する手法 [12] を提案した。この手法では、データセットのデータマップを構築し、モデルに関するデータセットを可視化する。具体的には、異なるモデルの学習精度を向上するための貢献度を、データセットごとに分類した。この分類は easy-to-learn, ambiguous, hard-to-learn の 3 領域に区分されており、ユーザはこれを観察することでデータがどのようにモデルの学習に貢献しているかを知ることができる。

また Smilkov ら [13] は、次元削減されたデータをユーザがどのように使いたいかを調査することにより、3 つのタスクを設定してデータセットの可視化を実現した。タスクとしては、1 つ目が局所的な近隣の探索、2 つ目がグローバルジオメトリの表示とクラスタの発見、3 つ目が意味のある「方向性」を見つけるというものである。1 つ目のタスクでは指定された近隣の点が意味的に関連しているか、2 つ目のタスクでは関連するデータのクラスタを見つけること、3 つ目のタスクでは埋め込み空間に意味のある方向性が含まれているかを確かめることを目的とした。

転移学習に特化した可視化手法として Ma ら [7] の手法がある。この研究は、多くのモデルでは学習データとラベル付されていないデータが同じ分布を構成する、という仮定にもとづいている。しかしこの仮定は現実的には多くの現場において困難である。転移学習はドメイン間の関係をモデル化することでこの仮定を緩和することを意図している。Ma ら [7] は DNN (Deep Neural Network) による学習の過程で、既存モデルから学習した知識が新しい学習タスクにどのように移行されるかを説明するために可視化を適用している。

一方で、モデルを作成する過程で訓練データを選定する状況において、訓練データの品質の違いについて解析し比較するための可視化手法は、まだあまり研究されていない。本研究では、品質が異なる複数の訓練データが手元にある状況を想定して、機械学習に詳しくない人でもデータがどのように異なるのかを理解できるような可視化手法を提案する。また本研究では、訓

練データの質の違いを可視化することで、モデル構築に着手する前の工程で機械学習の質の違いを推察することを目標にする。

2.3 多次元データの可視化

本研究が対象とする訓練データ群は、多次元ベクトルを付与された標本群であることから、これを多次元データとみなして可視化することが可能である。多次元データの可視化は情報可視化の研究の中でも非常に活発に議論されている課題である。

多次元データ可視化の一手法として、Itoh らは Hidden [14] を発表している。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、画面左部で複数の平行座標プロット (PCP: Parallel Coordinate Plots) によって表示する。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をして、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が、その中でも Hidden [14] では PCP や散布図の表示数を対話的に調節することを可能にした。Hidden の考え方を拡張して中林ら [15] は、低次元 PCP の代わりに選択的な散布図集合による多次元データの可視化手法を提案した。この手法の処理手順は以下の 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 变数を 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。

本報告の提案手法に、散布図をプロットした後ユーザーが選んだラベルがついたデータ群を多角形で囲む処理が含まれている。提案手法では、中林らの散布図選択手法の代わりに次元削減を適用した散布図に対して、この多角形囲み処理を用いた「例外点群」と「例外でない点群の包括領域」による描画手法を継承している。具体的には中林らの手法では、「例外点群」および「例外でない点群の包括領域」を描写する処理として Delaunay 三角分割法を用いた手法を採用している。処理手順としてはまず各散布図に対して、散布図中の全ての点群を包括する大きな四角形を生成する。続いて散布図中の点群を 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する。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用いている。

3 提案手法

3.1 処理手順の概要

本手法では訓練データ群の間の分布の違いを確かめるために複数の訓練データ群を同一画面に可視化する。同一画面に複数の訓練データ群を可視化するために複数の訓練データセットに属する全ての標本に対して、その特徴量に次元削減を適用し、2 次元の画面空間に投影している。我々の実装では、次元削減手法に t-SNE を採用している。

訓練データに付与された各ラベルについて、類似する標本群がどのように分布するか、外れ値となる標本群がどのように分布するか、といった点を理解しやすくするために各々の訓練データで同一のクラスを付与された標本群に対して、画面上で高い密度で分布する標本群を多角形で囲んで表示する。また多角形を生成するために中林ら [15] の「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」に使用している手法と同様に、Delaunay 三角分割法を用いた手法を採用している(図 2)。また多角形の生成に際して各々の多角形に対して固有の色を割り当てて描画している。多角形はラベルごとに作成されているのですなわち各々のラベルごとに閾値と固有の色相を設定している。また各データセットに固有の彩度と明度が割り当てている。色相、彩度、明度はプログラムで式を指定して決定しているため、入力されるラベルの数やデータセットの数が異なっていても自動で割り振ってくれるようになっている。

同一ラベルでかつ距離が近いものを基準として我々は、多角形を生成しているが、その距離をどれくらいまで含めるかの境界となる値を閾値と呼んでいる。本手法では各ラベルごとに閾値を自動で計算し、例外点を自動的に判別する。

以上より、本研究では複数の訓練データを同一画面に可視化することによって分布の違いを確かめ、同一ラベルでかつ距離が近いものを基準とし多角形を生成している。またその際に各ラベルに固有の色相を、各データセットに固有の彩度・明度を自動で割り振っている。さらには多角形の閾値をユーザーが指定することもできるが、ラベルごとに自動で計算し、例外点を自動的に判別することも可能である。

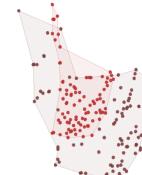


図 2 同一訓練データセットに属して同一のラベルを有する標本群を対象にして、散布図上で点群が集中する領域を多角形で囲んで表示した例。

3.2 点群の密度が高い領域を多角形で囲む処理

点群の密度が高い領域を多角形で囲む処理として、中林ら [15] の「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」に使用している手法と同様に、Delaunay 三角分割法を用いた手法を採用している。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。中林らの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生成し、続いて散布図中の点群を 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適用し

ている。以上の処理手順により全ての点群を連結する三角メッシュが生成されたら、ユーザ指定の閾値を超える長い辺を有する三角形を削除することで、距離の近い点群だけで構成された三角メッシュを生成する。そして、その外枠を囲む多角形を「例外のない点群の包括領域」として生成するとともに、多角形の外側にある点群を「例外点群」とする。以上の処理に続いて、本手法では以下の3種類の図形を描画する。また図形1、図形2、図形3の例を図3に示す。

図形1：各々の例外点群を小さい円で描写する。

図形2：包括領域の外周となる三角形辺の集合を太い線分で描写する。

図形3：包括領域を構成する三角形群をアルファブレンディングによって半透明描写する。

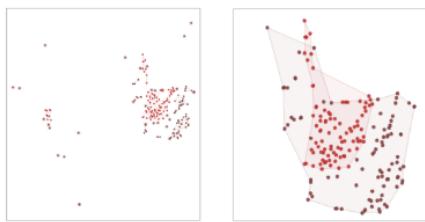


図3 中林ら[15]によるアルゴリズムを使用

3.3 閾値の自動設定処理

前節でも述べたとおり提案手法では、三角メッシュの辺の長さが一定以下である三角形群に含まれる点群を多角形で囲み表示する。以下、三角形を削除する基準となる辺の長さを「閾値」と呼ぶ。閾値はユーザーがスライダー操作で調節できるとともに、各ラベルごとに適切な閾値を自動設定する手法も実装している。自動設定手法の処理手順は以下の通りである(図4)。まずそれぞれの多角形の辺を長い順にソートする。続いてソートされた i 番目と $i+1$ 番目の辺の長さの差($len_i - len_{i+1}$)求める。この差が最大となる i の値を抽出し、値($len_i - 0.5(len_i - len_{i+1})$)を閾値とする。

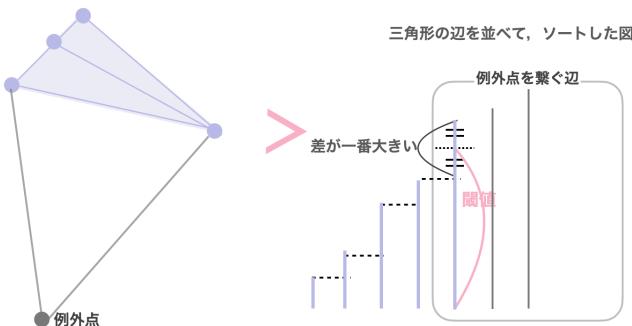


図4 左側の図形の辺をソート。辺の差が一番大きい2つの辺の平均値を閾値としている

3.4 色の指定方法

描画に際して、本手法では各データセット・各ラベルの色を、

HSB 表色系にもとづいて以下の式で指定している。

$$H = 2\pi \frac{i}{N}$$

$$S = B = \alpha \frac{j+1}{M} + (1.0 - \alpha)$$

H は色相、 S は彩度、 B は明度を表している。また色相とは「赤」「黄色」「青」のような色の相違を表し、彩度とは色の鮮やかさを表し、明度とは色の明るさを表す。 N と M ($0 \leq i < N, 0 \leq j < M$) はそれぞれラベルとデータセットの総数であり、 i と j はそれぞれラベルとデータセットの通し番号であり、 α は ($0 \leq \alpha \leq 1$) を満たす実数である。この式により、各データセットに固有の彩度と明度が割り当てられ、各ラベルに固有の色相が割り当てられる。図5はそれぞれのラベルに表示する色を割り振って一覧表示した可視化例である。

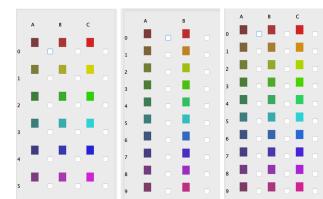


図5 それぞれのラベルに表示する色を割り振って一覧表示した可視化例。行にはデータセット、列にはラベルが書かれている。

3.5 画像の描写処理

提案する可視化システムでは、特徴量とラベルを持った画像群も同時に可視化している。具体的には、散布図に対応する画像を画面右側に表示する。画面に表示する画像の大きさを一定にし、画面をスクロールすることにより全ての画像を閲覧できるようにしている。図6の可視化画面の散布図では、MNIST の0のラベルがついたデータ群が表示されており、全ての点が線で連結されている。そのため、MNIST の0のラベルがついたデータ群の画像が全て表示されている。またラベルを選択する画面で選択したデータと対応し、また閾値のスライダーとも対応している。すなわち、ユーザが選択したラベルと設定した閾値までの画像のみ表示されるようになっている。図7では図6と同様に MNIST の0のラベルがついたデータ群を表示しているが、閾値の値が低いため多角形で囲まれているデータ群も図6に比べて少ない。よって表示されている画像群も図6に比べて少なくなっている。

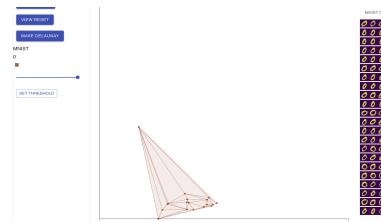


図6 画面の右側に多角形で囲まれている画像が表示してある。

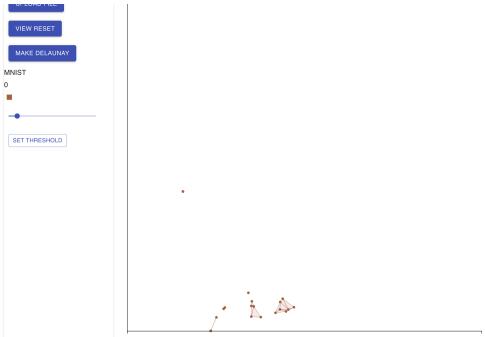


図 7 図 6 の閾値を変えて表示した例. 多角形で囲まれている部分の画像だけが表示されているのがわかる.

3.6 ユーザーインターフェース

本手法を操作するためのユーザーインターフェースを図 8 と図 9 に示す. 画面左側の操作パネルを操作することで, 画面右側のパネルの表示内容をインタラクティブに変更できる. 画面左側の操作パネルには main タブと color タブがある. 図 8 の操作パネルは main タブを, 図 9 の操作パネルは color タブを表示している. main タブでは上にボタンが 3 つあり, ファイルのアップロードができる UPLOAD FILE ボタン, 操作のリセットができる VIEW RESET ボタン, Delaunay の三角形の追加と散布図の描写ができる MAKE DELAUNAY ボタンがある. UPLOAD の際に選択するファイルには, 画像の 2 次元の座標値, ラベル, 画像パスが入っている必要がある. またファイルをアップロード後, MAKE DELAUNAY のボタンを押すと, それぞれのラベルの閾値を変更できるスライダーが表示される. color タブではラベルとそれに付与した色相, チェックボックスの表が表示されている. チェックボックスにマークし SUBMIT ボタンを押すと, ユーザーが選択したラベルの画像群のみ表示される. ラベルが多い時や特定のラベルがついた画像群のみ表示指定ときに利用できる.

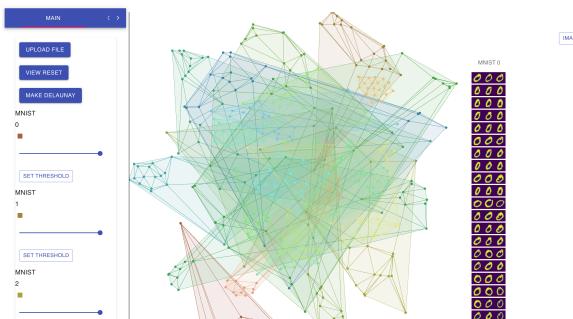


図 8 画面左側の操作パネルに main タブを表示した画面

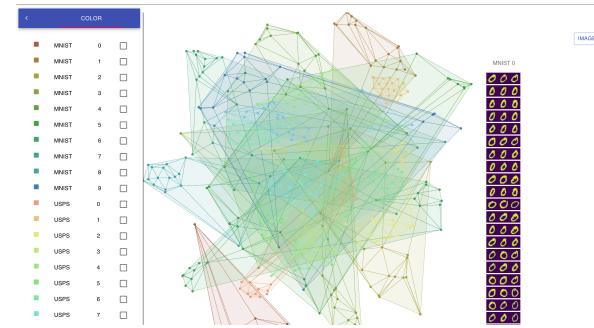


図 9 画面左側の操作パネルで color タブを表示した画面

インターフェースの操作手順としては, UPLOAD FILE ボタンを選択し (図 10), 表示したいデータの情報が入ったファイルを選択する (図 11).



図 10 ファイルが何も選択されていない初期画面

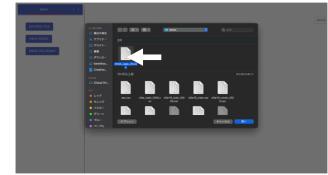


図 11 UPLOAD FILE ボタンを押した後の画面. ユーザーが表示したいファイルを選択する.

その後, MAKE DELAUNAY ボタンを押すと Delaunay 三角分割法を用いて三角メッシュが作成され表示される (図 12). またその際に右側にデータの画像も表示されている. color タブに遷移するとそれぞれにどのような色が割り振られているか表示されている. また見やすさのために自分が見たいラベルのついたデータを選ぶことも可能である (図 13, 図 14).



図 12

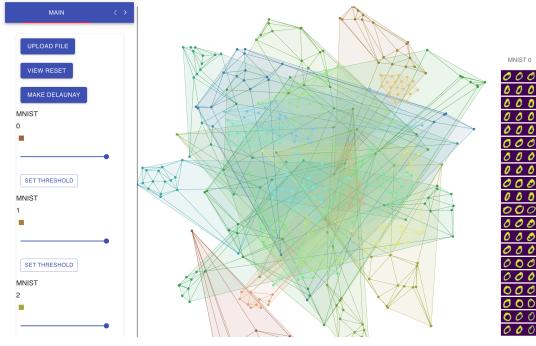


図 13 すべてのラベルが表示されている状態の画面

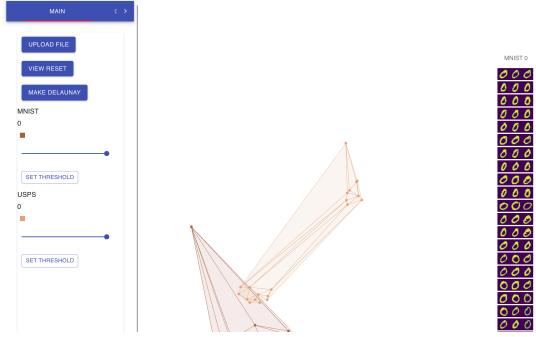


図 14 図 13 から 2 つのラベルを選び表示した画面

3.7 実装方法

我々は本手法の実行環境として, MacBook Air, macOS 11.6.1 を利用した. またフロントエンドの実装はフレームワークは React を利用し, 言語は JavaScript を選択した. バックエンドの実装はフレームワークは Flask を利用し, 言語は Python を選択した.

4 可視化例

4.1 手書き数字

MNIST と USPS のデータ群を同時に t-SNE にかけて 2 次元に次元削減し可視化した. 類似したデータ群を本報告のユーザーインターフェースで可視化した時, どのような結果になるのかを確かめるために 6 と 9 のラベルがついた画像をそれぞれのデータセットから選んだ. MNIST の 6 が茶色, 9 が緑色, USPS の 6 が赤, 9 が水色で表されている(図 15).

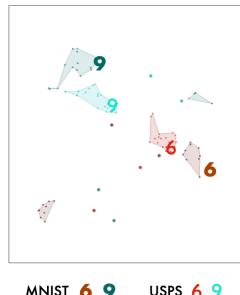


図 15 MNIST の 6 が茶色, 9 が緑色, USPS の 6 が赤, 9 が水色で表されている

同じラベルを持っており, 異なるデータセットに属する点群は近づいた. 6 のラベルがついた点群と 9 のラベルがついた点群は離れた位置にあるため, 特徴は異なることがわかった(図 16). より数が多く集まっている部分を 9 の正規のクラスタと考

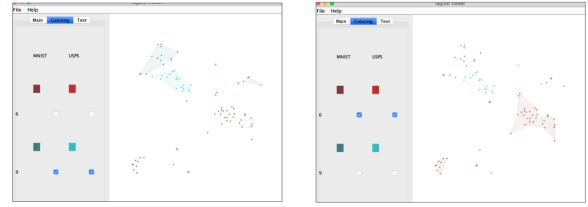


図 16 左側は各々のデータセットの 9 のついた画像のみを表示しており, 右側は 6 のついた画像のみを表示している.

え, それ以外の場所にある点を例外点と考えた. 外れ値となっているデータ群を確認してみたところ, 誤ったラベルがついていることはなく普通に見れば 9 と感じるデータ群であった(図 17).

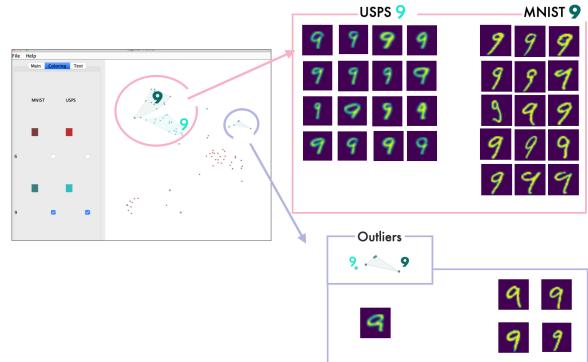


図 17 各々のデータセットの 9 のついた画像のみを表示

6 のラベルがついたデータ群でも同様の結果が得られた(図 18). この可視化例では 1 つのラベル・データセットごとに 20 枚の画像を表示した. また特徴量は t-SNE をデータ群にそのまま適用し, データの分布を表示した. そのためより多くの画像を適用すると結果が変わることもある. また, 特徴量の取り方によっても結果が変わることもある.

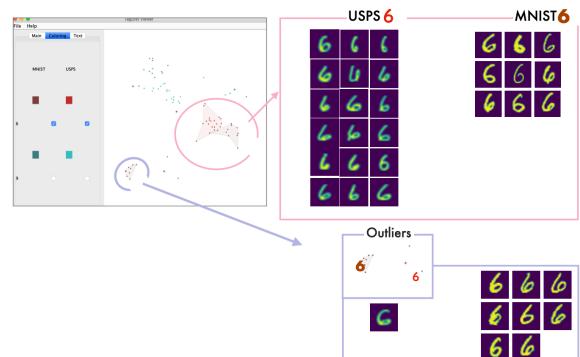


図 18 各々のデータセットの 6 のついた画像のみを表示

4.2 CIFAR-10

CIFAR-10 のデータ群は、CNN を利用して特微量を抽出したのち t-SNE にかけて 2 次元に次元削減した。誤ったラベルのついたデータ群を本報告のユーザーインターフェースで可視化した時、どのような結果になるのかを確かめるために 8 の ship がついたデータ群の一部を 9 の truck のラベルに付け替えて可視化した(図 19)。



図 19 8 の ship がついたデータ群の一部を 9 の truck のラベルに付け替えた

図 20 では誤ったラベルがついた画像を含むデータ群と正しいラベルのみがついたデータ群をそれぞれ可視化した。

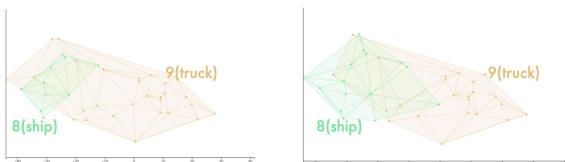


図 20 左側は ship を truck のラベルに変更したものを含むデータ群の可視化。右側にある図は正しいラベルがついたデータ群の可視化。

次に誤ったラベルがついたデータ群に閾値を自動設定した。図 21 では本来ならば緑色に表示されるはずの 8 の ship が付いた画像群の一部に 9 の truck のラベルがついている。灰色の丸で囲んだ点が、本来緑色に表示されるはずだった 8 の ship のラベルが付いた画像である。ここで閾値を自動で設定すると全て独立した点となり緑には多角形がなくなった。本報告では閾値の自動設定は辺の長さを利用している。この可視化結果では、ship の画像群に対応する三角メッシュの辺の長さがどれも近かったために、閾値がうまく設定できなかったと思われる。この例に限らず現状の実装では、閾値の自動設定が全て成功するとは限らない。外れ値をより見つけやすくするために、閾値をスライダー操作をして直して、ユーザー自身が外れ値を探す必要がある場合もあり得る。また現在の実装では、可視化結果の各々の点にどの画像が対応するかすぐにはわからないので、マウスオーバーして表示させるなどの実装の必要性がこの例でわかった。

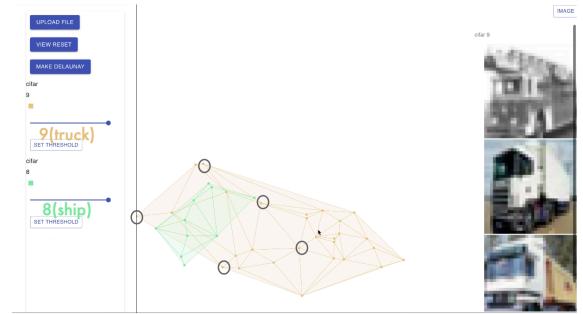


図 21 灰色の丸で囲んだ点が、本来緑色に表示されるはずだった 8 の ship のラベルが付いた画像

4.3 考 察

以上の結果から本手法の長所として、どんなデータ群の特徴が類似しているのか、またそれによってラベルの付与が適切かどうかを視認しやすくなる。また特微量の算出手法がそのデータセットに適しているかどうかがわかる。またユーザーが観察したいラベルを任意に選択できるので、あらかじめ比較したいラベルがわからなくても、画面上で操作しながら確認できることがメリットとしてあげられる。一方で、画像と点の対応を示すためのマウスオーバー機能が必要であることがわかった。また、辺の長さにもとづいて閾値を自動設定しているので、辺の長さが均等であると全てが独立した点となり多角形が作成されないといった問題点が見つかった。

4.4 ま と め

本報告では、訓練データ群の間の分布の違いを確かめるための可視化手法に関する統報を報告した。本手法ではラベルを有する画像群を訓練データセットに仮定して、これらに同一の次元削減を適用して一画面に表示する。本手法では散布図上の点群を連結する三角メッシュを生成し、辺の長い三角形を削除することで、画面上で近い距離に配置された点群を多角形で表現する。本報告では、三角形削除のための辺の長さの閾値をラベルごとに独立に自動設定する手法を提案した。本手法を用いることでデータセットの構成の理解を支援し、また複数のデータセットの比較も容易になると期待される。

今後の課題として以下に取り組みたい。まず、各標本が 2 つ以上のラベルを有するときの視覚表現についても検討したい。また点に対応する画像をマウスオーバーで表示する実装を加えたい。

文 献

- [1] M. Liu, S. Liu, H. Su, K. Cao, J. Zhu, “Analyzing the noise robustness of deep neural networks,” In Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST), 2018.
- [2] S. Liu, C. Chen, Y. Lu, F. Ouyang, B. Wang, “An interactive method to improve crowdsourced annotations,” IEEE Transactions on Visualization and Computer Graphics, 2018.
- [3] D. Tanaka, D. Ikami, T. Yamasaki, K. Aizawa, “Joint optimization framework for learning with noisy labels,” In

- CVPR, pages 5552–5560, 2018.
- [4] S. Pan, Q. Yang, “A Survey on Transfer Learning,” Institute of Electrical and Electronics Engineers, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, Issue 10, pages 1345-1359, 2010.
 - [5] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. Rusu, A. Pritzel, D. Wierstra, “PathNet: Evolution Channels Gradient Descent in Super Neural Networks,” Neural and Evolutionary Computing, arXiv: 1701:08734v1, 2017.
 - [6] A. Ng, “Sparse autoencoder,” CS294A Lecture notes, 2011.
 - [7] Y. Ma, A. Fan, J. He, A. Nelakurthi, R. Maciejewski, “A Visual Analytics Framework for Explaining and Diagnosing Transfer Learning Processes,” IEEE Transactions on Visualization and Computer Graphics, Vol. 27, pages 1385-1395, 2021.
 - [8] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, M. Sedlmaier, “Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study,” in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pages 298-308, Jan. 2018.
 - [9] J. Moehrmann, S. Bernstein, T. Schlegel, G. Werner, G. Heidemann, “Improving the usability of hierarchical representations for interactively labeling large image data sets,” In Proceedings of the 14th International Conference on Human computer Interaction: Design and Development Approaches Volume Part I, HCII ’11, pages 618–627, 2011.
 - [10] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, S. Liu, “Interactive Correction of Mislabeled Training Data,” 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 57-68, 2019.
 - [11] X. Zhang, X. Zhu, S. Wright, “Training set debugging using trusted items,” In Proceedings of The Thirty Second AAAI Conference on Artificial Intelligence (AAAI), 2018.
 - [12] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. Smith, Y. Choi, “Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics,” Proceedings of EMNLP, 2020.
 - [13] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. Viegas, M. Wattenberg, “Embedding Pro jector: Interactive Visualization and Interpretation of Embeddings,” NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems, 2016.
 - [14] 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, pages 1–13, 2017.
 - [15] A. Nakabayashi, T. Itoh, “A Technique for Selection and Drawing of Scatterplots for Multi-Dimensional Data Visualization,” Proceedings of 23rd International Conference on Information Visualisation (IV2019), pages 62–67, 2019.
 - [16] B. Bederson, B. Shneiderman, M. Wattenberg. 2002, “Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies,” ACM Trans. Graph. 21, pages 833–854.
 - [17] B. Bederson, “PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps,” In Proceedings of the 14th annual ACM symposium on User interface software and technology (UIST ’01). Association for Computing Machinery, New York, NY, USA, pages 71–80, 2001.
 - [18] A. Gomi, R. Miyazaki, T. Itoh and J. Li, “CAT: A Hierarchical Image Browser Using a Rectangle Packing Technique,” 2008 12th International Conference Information Visualisation, 2008, pages 82-87, doi: 10.1109/IV.2008.8.
 - [19] Mnist handwritten digit database. <http://yann.lecun.com/exdb/mnist/>. Accessed: 07-October-2018.