訓練データ比較のための可視化の一手法

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．本研究では訓練データ群に次元削減を適用して散布図として表示することで，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，散布図上で点群が集中する領域をラベルごとに多角形として表示し，対話的なスライダー操作によってその閾値を設定する．これにより，訓練データ間のラベルの分布の違いなどを観察できる．

キーワード　可視化手法，機械学習，訓練データ群

1 はじめに

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，その一手段として可視化が有効であると考えられる．

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徴量が算出され，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徴量が算出される．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．また，その多角形に含まれない外れ値となる標本群を強調表示する．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ーに提示することを目標とする．

2 関連研究

2.1 転移学習

転移学習 [1] 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 [2] 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) [3] 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 [4] 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 [4] 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．図 1 は Ma ら [4] の可視化画面である．この例では，bike と calculator のように両モデルで同じ精度の高いクラスと，filecabinet と phone のように両モデル間で精度の変わるクラスが存在することが示されている．また図 2 より，2 つのモデルで性能が類似しているクラスでは，2 つの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，2 つのモデルで性能が異なるクラスでは，2 つのドメイン間でパターンが大きく異なっている．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が作れるかをユーザが探索できることを目標にする．

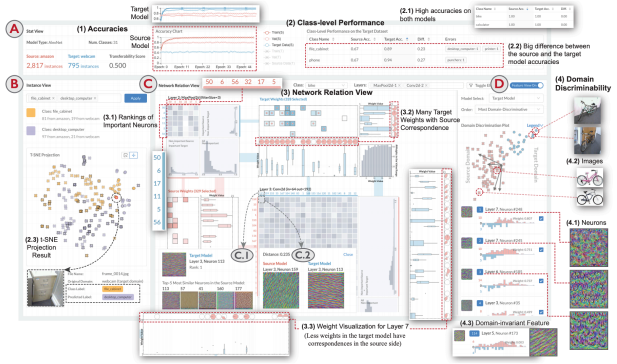

図 1　Ma ら [4] による転移学習の可視化．4 つの可視化コンポーネントで構成されている．

2.2　機械学習のための可視化

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，Swabha らはデータセットの品質を解析し可視化する手法 [5] 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 [6] 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

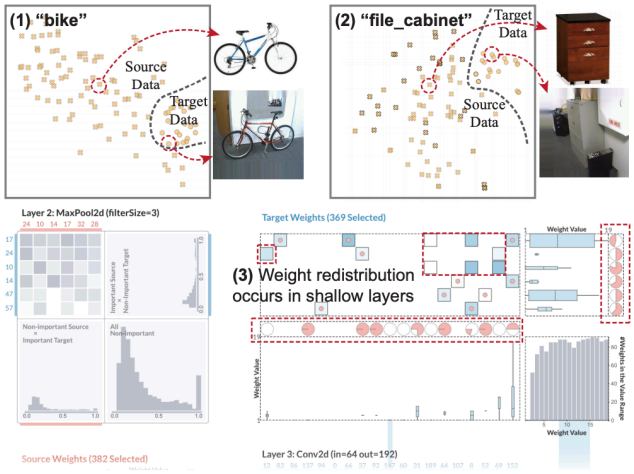

図 2　Ma ら [4] による可視化結果のうち，(1)bike と (2)filecabinet について t-SNE を用いて可視化した結果．

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 [4] 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 [4] 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 [7] 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が，その中でも Hidden [7] 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 [8] 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 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．

本報告の提案手法は，中林らの手法で散布図を選択する代わりに次元削減を適用し,「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するものである．

3 可視化手法

本章では我々の可視化手法について提案する．本手法では 1 章でも論じた通り，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する．現時点での我々の実装では，次元削減手法に t-SNE を採用している．続いて，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する．そして各々の多角形に対して固有の色を割り当てて描画する．以上の描画手法による可視化の例を図 3 に示す．

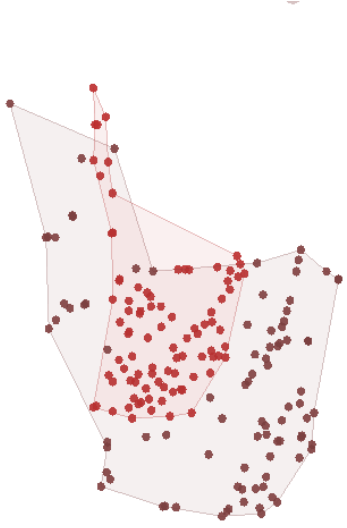

図 3　同一訓練データセットに属して同一のラベルを有する標本群を対象にして，散布図上で点群が集中する領域を多角形で囲んで表示した例.

点群の密度が高い領域を多角形で囲む処理として，中林らの [8] 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．Delauney 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 t_{len} 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．

図形 1:　各々の例外点群を小さい円で描画する．

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画する．

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画する．

また描画に際して，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する．

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

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

なお N と 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 $(0 \leqq i < N, 0 \leqq j < M)$ であり，a は $(0 \leqq a \leqq 1)$ 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．さらに，図形 1,2,3 には別々のアルファ値（不透明度）$\alpha_1, \alpha_2, \alpha_3$ が割り当てられる．

4 実　行　例

本研究による可視化の実行例を図 4 に示す．この実装ではスライダー操作によっていくつかの値を調節可能である．図 4(1) は閾値 t_{len} を調節するスライダーであり，これを操作することで包括領域の大きさを調節できる．図 4(2)(3) は変数 α_1, α_3 を調節するスライダーであり，これを操作することで例外点および包括領域内部の色の濃さを調節できる．この実行例では，MNIST および USPS という 2 種類の手書き数字画像データセットを入力情報としている．各画像から算出した画像特徴量を次元削減して可視化している．また，0 から 9 までの数字をそのままラベルとして各画像に付与しており，それが可視化結果中の 10 種類の色相として表現されている．

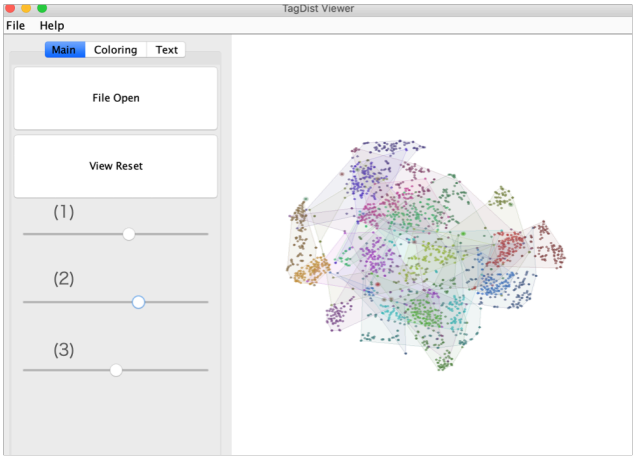

図 4　実行例．この例では MNIST と USPS という 2 種類の訓練データセットを用いている．(1) では包括領域の大きさを調整する．(2) では例外点の色の濃さを調整する．(3) では包括領域内部の色の濃さを調整する．

現在我々は，質の違う複数の訓練データセットを比較することで，モデルの精度を下げる要因を発見する，というタスクに提案手法が有効であるかを検証中である．この検証のための訓練データセットとして，ImageNet のテスト用データセットと，ImageNet と同一ラベルを持ったイラストのデータセットである ImageNet-Sketch を利用している．最終稿ではこの結果についても報告したい．

5　まとめ・今後の課題

本報告では，訓練データ群を比較するための可視化の一手法を提案した．本手法では，特徴量ベクトルとラベルを有する標本の集合によって構成される複数の訓練データセットを仮定し，これらに同一の次元削減を適用して一画面に表示する．本手法を用いることで，訓練データセットの中に潜むモデルの精度を下げる要因を，モデルの中間層などを特徴量として発見することができる．

今後の課題として以下に取り組みたい．まず，多くの点群が集中する包括領域を定義するための閾値について，ふさわしい値を自動設定する，ラベルごとに別々の値を設定できるようにする，といった形で改善したい．また，ラベルやデータセットの数が多い際に，色でこれらを表現する現時点での本手法の視覚表現には限界があるので，これを解決する方法について模索したい．さらに，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．

これらの課題を解決したのちに，多様なデータで本手法の有効性を検証し，さらにユーザ評価実験を通して本手法の有効性を再検証したい．

6　謝　　辞

本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．

文　　献

[1] Sinno Jialin Pan, Qiang Yang, “A Survey on Transfer Learning, Institute of Electrical and Electronics Engineers,” IEEE Pages 1345-1359, 2010.

[2] Chrisantha Fernando, Dylan Banarse, Charles Blun- dell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, Daan Wierstra, “PathNet: Evolution Channels Gradient Descent in Super Neural Networks,” Neural and Evolutionary Computing, arXiv:1701:08734v1, 2017.

[3] Andrew Ng, “Sparse autoencoder,” CS294A Lectire notes, 2011.

[4] Yuxin Ma, Arlen Fan, Jingrui He, Arun Reddy Nelakurthi, Ross Maciejewski, “A Visual Analytics Framework for Explaining and Diagnosing Transfer Learning Processes,” IEEE Transactions on Visualization and Computer Graphics, 2020.

[5] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, Yejin Choi, “Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics,” Proceedings of EMNLP, 2020.

[6] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, Martin Wattenberg, “Embedding Projector: Interactive Visualization and Interpretation of Embeddings,” NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems, 2016.

[7] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jin- man 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.

[8] Asuka Nakabayashi, Takayuki Itoh, “ A Technique for Selection and Drawing of Scatterplots for Multi-Dimensional Data Visualization,”Proceedings of 23rd International Conference on Information Visualisation (IV2019), pp. 62–67, 2019.