訓練データ比較のための可視化の一手法

理学専攻　情報科学コース　2040642　高坂夏怜　(指導教員：伊藤 貴之)

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．

訓練データ群に特化した可視化手法として，Xiang ら [1] 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を目標としている．

本研究では，訓練データは静止画像を対象としている．現段階の我々の実装では，1 つの画像に対し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した際，全ての訓練データセットにおいて同一の特徴量を算出する．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，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．

2　処理手順

2.1　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 [2] の手法と同様に，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．全ての点群を連結する三角メッシュが生成されたら，ユーザー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．

2.2　閾値の自動設定

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同時に，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．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．

2.3　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．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．

2.4 ユーザーインターフェース

画面左側の操作パネルを操作することで，画面右側の表示パネルに表示するもの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 1）．main タブでは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 DELAYNAY ボタンがある．またファイルをアップロード後，MAKE DELAUNAY 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．

図 1: 左側が main タブを表示した画面，右側が color タブを表示した画面

3 可視化事例

3.1 手書き数字

MNIST と USPS のデータ群を同時に t-SNE にかけて 2 次元し可視化した．類似したデータ群を本手法で可視化した時の結果を確かめるために 6 と 9 のラベルがついた画像を選んだ．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．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．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え，それ以外の点を例外点と考えた．外れ値となって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 2)．

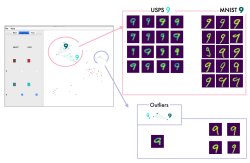

図 2: 各々のデータセットの 9 のついた画像のみを表示

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 3)．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可能性がある．また，特徴量の取り方によっても結果が変わる可能性がある．

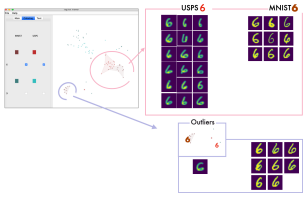

図 3: 各々のデータセットの 6 のついた画像のみを表示

3.2 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正確かどうかがわかりやすくなる．また特徴量の取り方がそのデータセットに適しているかどうかがわかる．ユーザーが見たいラベルのデータ群を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，画像と点の対応がマウスオーバーでわからないことや，閾値の自動設定を辺の長さにもとづいていることから，距離が均等の場合に自動設定を実行すると全てが独立した点となり多角形が作成されない，といったデメリットも明らかになった．

4 まとめと今後の課題

本論文では，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を提案した．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する．

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．

参考文献

[1] 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.

[2] 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.