Technical Sheet
No. 22-16

データ解析入門 11 <UMAP による次元削減>

キーワード：UMAP、次元削減、データ可視化、機械学習

はじめに

高次元データの代表的な次元削減手法として主成分分析（PCA）や t-SNE（t-distributed stochastic neighbor embedding）が知られています。本稿では、有用な非線形次元削減手法である UMAP（Uniform manifold approximation and projection）[1]について概説します。

PCA、t-SNE、および UMAP の比較

本稿では、Fashon-MNIST と呼ばれるファッション商品のデータセット [2]を次元削減します。各データは 28×28（784 次元）の グレースケール画像であり、Shirt や Bag などの 10 個のラベルが付与されています。

PCA、t-SNE、および UMAP により、1 万枚の画像を 2 次元に圧縮すると、図 1 に示すプロットが得られます。PCA と比較して、t-SNE および UMAP では異なるラベルデータの重なりが小さいことがわかります。また、UMAP ではラベルごとの凝集性が最も高くなっているように見受けられます。特に、Trouser と Bag はコンパクトに密集しており、その他のデータ群からよく乖離しています。また、PCA、t-SNE、UMAP の実行時間はそれぞれ、約 1 秒、約 222 秒、約 32 秒であり、UMAP の実行時間は t-SNE よりも短くなりました。

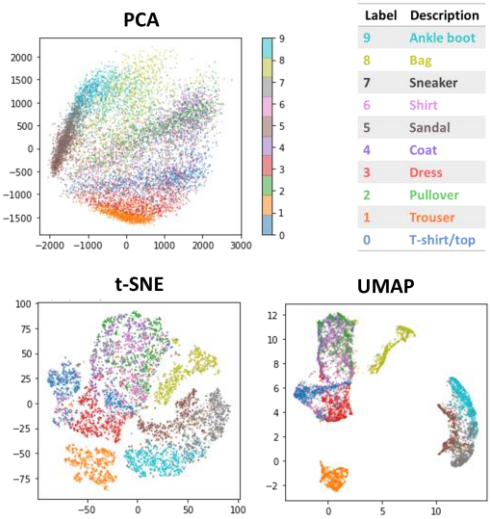

図 1 PCA、t-SNE、および UMAP の実行結果

UMAP の動作原理

前稿 [3]で説明した t-SNE の動作原理と対比しながら、UMAP のアルゴリズムを概説します。

簡潔には、t-SNE と UMAP はいずれも共通した次のステップにより実行されます。

1. 全てのデータについて、高次元データ $\boldsymbol{x_i}$、$\boldsymbol{x_j}$ の類似度に相当する確率（とみなせる値）を算出する。
2. 低次元空間に全データ数 N と同じ数のデータ点を配置する。
3. 高次元空間のデータ点 $\boldsymbol{x_i}$、$\boldsymbol{x_j}$ に対応する低次元空間のデータ点 $\boldsymbol{z_i}$、$\boldsymbol{z_j}$ の確率を算出する。
4. 上記 2 つの類似度が整合するように、低次元空間のデータ点を配置しなおす。
5. 結果が収束するまでステップ 3、4 を繰り返す。

t-SNE と UMAP の相違点

以下に両手法の主な相違点を列挙します。

- ✓ 高次元・低次元空間における確率計算式
- ✓ 局所的なデータ疎密に適応するためのパラメータ（*perplexity* あるいは *n_neighbors*）
- ✓ 低次元データの配置探索のために最小化する目的関数（カルバック・ライブラー情報量あるいは交差エントロピー）

確率計算式の違いから、高次元空間における確率計算量は UMAP の方が小さくなります。また、低次元空間の確率計算に用いる確率分布も異なります。t-SNE では裾の重い t 分布を用いますが、UMAP ではパラメータ *min_dist* により形状が変化する確率分布を用います。*min_dist* を小さくすると、幅の狭い確率分布となり（図 2）、高次元空間で近くに位置するデータは低次元空間で密集することになります。

また、両手法にはそれぞれ *perplexity* と *n_neighbors* という異なるパラメータが存在しますが、いずれも類似した役割を果たします。これらのパラメータは高次元空間の確率計算に用いる確率分布の幅を決定します。

低次元データの配置更新のために最適化する

地方独立行政法人
大阪産業技術研究所 本部・和泉センター
https://orist.jp/
〒594-1157 和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

目的関数も異なります。カルバック・ライブラー情報量を用いる t-SNE では、高次元空間のローカル構造を重視する傾向にあり、グローバル構造の表現が不正確になりやすいと主張されることがあります。例えば、オレンジおよび水色のクラスタ間距離は必ずしも正しい解釈を与えないということです（図 3）。特に、初期配置を乱数で決める場合は念頭に置いておくと無難です [4,5]。

一方、交差エントロピーを用いる UMAP は、t-SNE よりもローカルおよびグローバル構造をバランスよく保持できると言われることがあります。ただし、グローバル構造をどれだけ保持できるかは初期配置や *perplexity* などのパラメータにも依存し、近年では発展的な手法も提案されているため、本トピックには議論の余地があるように思われます。

図 4 に初期化方法により UMAP のグローバルな距離関係が変化しうることを示します。左図では、spectral embedding というグラフ行列を用いた次元削減手法で初期データを配置し、右図ではランダムに初期データを配置しました。今回の解析では大きな違いは見られなかったものの、青色、水色、赤色のクラスタに注目すると、クラスタ間距離の大小関係が異なることがわかります。

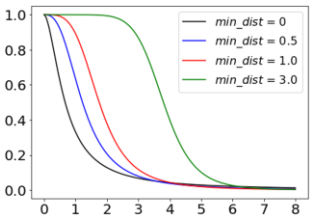

図 2　*min_dist* が異なる確率分布

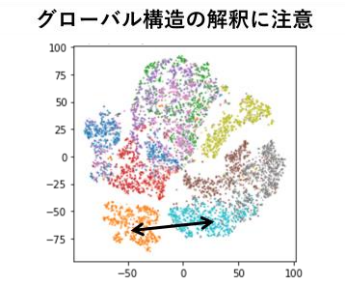

図 3　グローバル構造に関する注意点

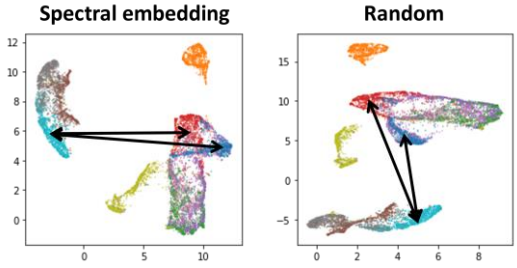

図 4　UMAP における初期化方法の影響

UMAP のパラメータについて

改めて、UMAP におけるパラメータの影響を確認します。ここでは *n_neighbors* と *min_dist* の影響を示します。上述の通り、*n_neighbors* は高次元空間の確率計算に用いる確率分布の幅を決定します。図 5 上段の *min_dist* = 0.1 に注目すると、*n_neighbors* が大きくなるにつれてプロットが広がっている様子がわかります。また、*min_dist* も可視化結果に大きく影響するパラメータであり、*min_dist* を大きくすると近傍点はルーズに埋め込まれます。そのため、コンパクトなクラスタは生成しにくくなります。

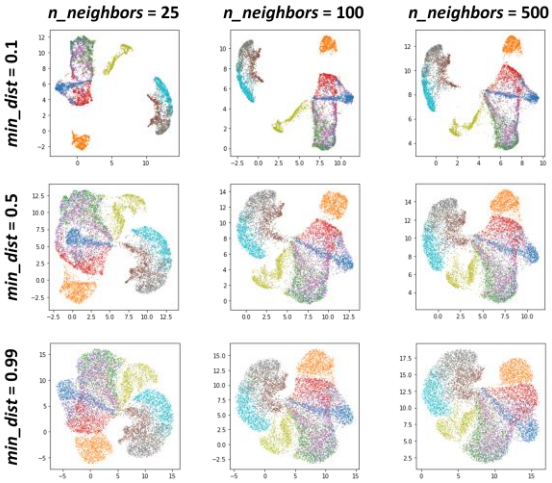

図 5　*n_neighbors* および *min_dist* の影響

おわりに

本稿では UMAP の特徴について確認しました。t-SNE や UMAP には複数のパラメータが存在するため、示唆に富む可視化結果を与えるパラメータを試行錯誤により探索することは重要です。一方で、客観的な評価指標に基づいてパラメータを自動調整することも実践的な手段です。次稿では、可視化手法のパラメータの自動調整について紹介します。

参考文献

1) L. McInnes *et al.*, arXiv:1802.03426v3 (2018)
2) https://github.com/zalandoresearch/fashion-mnist/blob/master/README.ja.md
(accessed on March 28th, 2022)
3) 永廣卓哉：データ解析入門 10 <t-SNE による次元削減>、ORIST テクニカルシート、No. 22-14 (2022)
4) D. Kobak and P. Berens, *Nat. Commun.*, **10**, 5416 (2019)
5) D. Kobak and G. C. Linderman, *bioRxiv* (2019). https://doi.org/10.1101/2019.12.19.877522

発行日　2022 年 8 月 1 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@orist.jp