Technical Sheet
No. 22-14

データ解析入門 10 <t-SNE による次元削減>

キーワード：t-SNE、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）が知られています。PCA は有用な次元削減手法ですが、非線形に分布するデータの可視化には不向きです。本稿では、非線形な次元削減手法として知られる t-SNE（t-distributed stochastic neighbor embedding）を紹介します。

PCA と t-SNE 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 [1]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。

主成分分析（PCA）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。Ankle boot、Sandal、Sneaker といった履物は左上に固まってプロットされていることがわかります。一方で、全体的にプロットが重なっている印象を受けます。

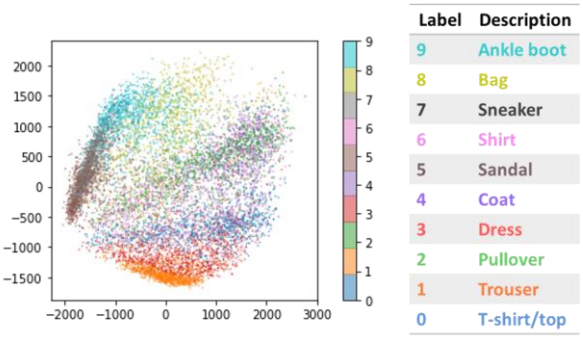

図 1　PCA による次元削減

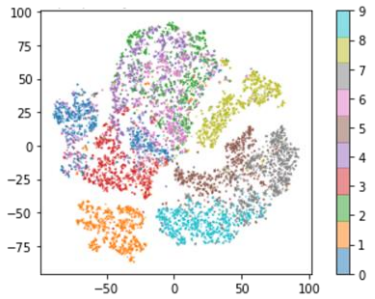

図 2　t-SNE による次元削減

次に t-SNE による次元削減結果を示します（図 2）。PCA の結果とは異なり、t-SNE により次元削減することで、ラベルごとにデータが分布する傾向が強くなりました。t-SNE により非線形な特徴を抽出できたことが示唆されます。

t-SNE の動作原理

PCA ではデータのばらつきに注目し、低次元空間に写像する行列を求めます。一方、t-SNE ではそのような行列を求めることなく、高次元空間で近くのデータは低次元空間でもできるだけ近くに、遠くのデータは低次元空間でも遠くになるようにデータ配置を決定します。

t-SNE のアルゴリズムは以下の通りです。

1. 全データについて、正規分布を仮定し、データ $\boldsymbol{x_i}$、$\boldsymbol{x_j}$ から確率 p_{ij} を算出する[式(1～2)]。
2. 低次元空間に、全データ数 N と同じ数のデータ点をランダムに配置する。
3. 高次元空間におけるデータ点 $\boldsymbol{x_i}$、$\boldsymbol{x_j}$ に対応する低次元空間のデータ点 $\boldsymbol{z_i}$、$\boldsymbol{z_j}$ から t 分布（自由度 1）に基づき確率 q_{ij} を算出する[式(3)]。
4. 確率 p_{ij}、q_{ij} の 2 つの分布が近くなるように、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。

$$p_{j|i} = \frac{\exp(-\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} exp(-\|\boldsymbol{x_i} - \boldsymbol{x_k}\|^2 / 2\sigma_i^2)} \quad (1)$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \quad (2)$$

$$q_{ij} = \frac{(1 + \|\boldsymbol{z_i} - \boldsymbol{z_j}\|^2)^{-1}}{\sum_{k \neq i}(1 + \|\boldsymbol{z_k} - \boldsymbol{z_i}\|^2)^{-1}} \quad (3)$$

ここで、i、j、k はサンプルインデックス、$p_{i|j}$ は計算過程で得られる条件付き確率、σ_i は $\boldsymbol{x_i}$ を中心とした正規分布の標準偏差を表します。また、$p_{ii} = q_{ii} = 0$ とします。

t-SNE はデータ間距離に基づく手法ですが、データ間距離は正規分布および t 分布によりデータの類似度に相当する確率に変換されます。図 3 に

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）

標準偏差が 1 である正規分布および自由度 1 の t 分布を示します。このとき、高次元空間では確率の算出に正規分布を用いますが、低次元空間では t 分布を使用します[式(1～3)]。t 分布は正規分布よりも裾が重いため、高次元空間で近いデータはより近くに、遠くに位置するデータはより遠くに配置されることになります。このような t 分布の特徴を活かし、元のデータ分布の特徴をできるだけ保持します。

高次元空間におけるデータ点 $\boldsymbol{x_i}$、$\boldsymbol{x_j}$ の確率p_{ij}の分布と近くなるように、低次元空間における確率q_{ij}の分布を求めます。このために、カルバック･ライブラー情報量という確率分布間の距離のような尺度を最小化します[式(4)]。

$$C = \sum_i \sum_j p_{ji} log \frac{p_{ji}}{q_{ji}} \qquad (4)$$

カルバック･ライブラー情報量は、低次元空間におけるデータ点 $\boldsymbol{z_i}$ で微分可能であるため、勾配情報を利用した手法により $\boldsymbol{z_i}$ を最適化できます。例えば、確率的勾配降下法などを用いることができます。

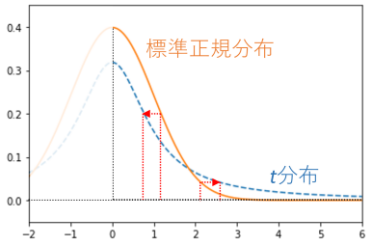

図 3　正規分布および t 分布

t-SNE のパラメータについて

以上が t-SNE の動作原理ですが、式(1)の標準偏差 σ_i に依存して得られる結果は変化します。データ点 $\boldsymbol{x_i}$ の周りにデータが密集しているときはσ_iを小さくし、疎らであれば大きくするべきであり、適切なσ_iを探索するために *perplexity* というパラメータを用います。*Perplexity* が大きければσ_iも大きくなり、遠くに位置する $\boldsymbol{x_i}$ の近傍点の学習への寄与が大きくなります。図 4 に示すように *perplexity* の値によって得られる結果は変わります。*Perplexity* の検討範囲として、5～50 が目安とされることが多いようですが、データ数が多い場合は、より大きな *perplexity* も検討すると良いかもしれません。

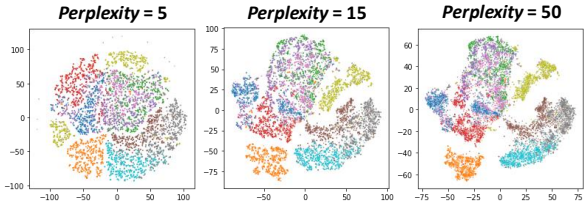

図 4　*Perplexity* の影響

PCA と t-SNE の組み合わせ

非線形な特徴の可視化に優れる t-SNE ですが、データが高次元になると実行時間が長くなります。実行時間短縮のためには、t-SNE の実行前に PCA により次元削減しておくことも有効です。

図 5 に、t-SNE、PCA と t-SNE("With PCA"と付記)、PCA による次元削減結果を示します。PCA を前処理として用いる場合、各図上部の丸括弧に記載した累積寄与率 [2]を基準に主成分を選択しました。予め PCA を実行することで t-SNE の計算時間が短縮されましたが、一般的に過度な次元削減は PCA における情報損失を増加させてしまいます。一方、次元削減が適度であれば、ノイズの影響が軽減される可能性があります。

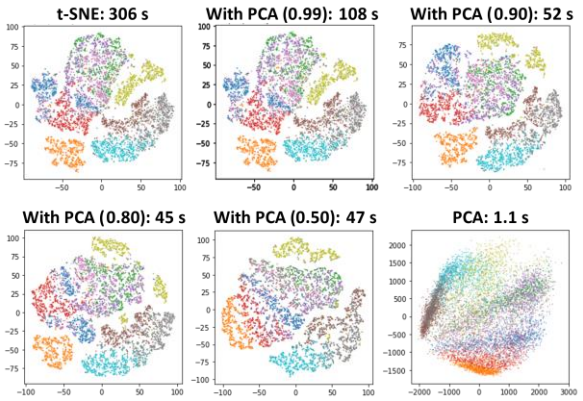

図 5　t-SNE、PCA を前処理とした t-SNE("With PCA"と付記)、および PCA の実行結果(各図上部に累積寄与率および実行時間を示す)

おわりに

t-SNE は非線形データを可視化できる手法ですが、PCA のように写像する行列を求めているわけではありません。したがって、t-SNE により得られた低次元空間上に新たなデータを埋め込みたい場合は工夫が必要になります [3]。次稿では、t-SNE とよく対比される非線形次元削減手法を紹介します。

参考文献

1) https://github.com/zalandoresearch/fashion-mnist/blob/master/README.ja.md
(accessed on March 28th, 2022)
2) 永廣卓哉:データ解析入門 3 <主成分分析によるデータマイニング>、ORIST テクニカルシート、No. 21-25 (2021)
3) P. G. Poličar *et al.*, *Mach. Learn.* (2021). https://doi.org/10.1007/s10994-021-06043-1

発行日　2022 年 7 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