Technical Sheet
No. 22-18

データ解析入門 12 <ベイズ最適化によるパラメータ調整>

キーワード：ベイズ最適化、t-SNE、次元削減、データ可視化、機械学習

はじめに

高次元データの非線形次元削減手法として、t-SNE（t-distributed stochastic neighbor embedding）[1]や UMAP（Uniform manifold approximation and projection）[2]がよく知られています。これらは非線形なデータ構造の抽出に優れた手法ですが、次元削減結果に影響を及ぼす複数のパラメータが存在します。本稿では、教師なし学習である次元削減の評価指標の一例を紹介し、t-SNE におけるパラメータの自動調整について検討します。

t-SNE のパラメータ探索について

t-SNE は非線形次元削減手法として知られており、Fashon-MNIST と呼ばれるファッション商品のデータセット[3]に適用すると、特徴的な次元削減結果が得られました（図 1）。ただし、図 1 に示した次元削減結果は t-SNE のパラメータの値によって変化します。これまでに *perplexity* という重要なパラメータについて紹介していますが[1]、そのほかにも t-SNE には *learning_rate*（学習率）などのパラメータが存在します。*Perplexity* と同様に、これらのパラメータも次元削減結果に影響を及ぼします。

パラメータの決定にあたり、複数通りの組み合わせを愚直に確認していくことも可能ですが、客観的な指標に基づきパラメータを自動調整できると便利です。そこで、本稿では t-SNE の複数のパラメータを自動調整することを考えます。

次元削減結果の評価指標

パラメータの自動調整には、次元削減結果の評価指標が必要になります。この評価指標としていくつか候補がありますが、ここでは 2 つの指標を考えます。1 つはローカルなデータ構造に関する指標であり、もう 1 つはグローバルなデータ構造に関する評価指標です。

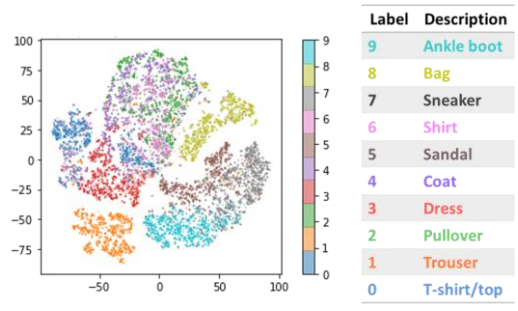

図 1　t-SNE の実行結果

まず、ローカルな指標について考えます。ある高次元データ $\boldsymbol{x}_i$ に注目し、その周辺に存在する k 近傍点のサンプルインデックスを調べます。次に、t-SNE により高次元データ $\boldsymbol{x}_i$ が低次元データ $\boldsymbol{z}_i$ に変換されたとすると、低次元空間においても $\boldsymbol{z}_i$ の周辺には同じデータが存在していることが望まれます（図 2）。そこで、ローカルな指標として k 近傍点のサンプルインデックスが t-SNE 実行後にどれだけ保持されているかという割合を用います[4]。全てのデータについてこの割合を計算し、それらの平均値をローカル構造の指標として用います。

次に、グローバル構造の評価指標について考えます。高次元空間の 2 点間距離の分布が低次元空間でも概ね保持されていることが望まれます。高次元および低次元空間における 2 点間距離のヒストグラムを図 3 に示します。横軸の 2 点間距離はスケーリングしていますが、この分布が近くなるデータ配置を探索します。そのためにスピアマンの順位相関係数を参照することにします[4]。本稿では全データではなく、ランダムに選んだ 1000 点のデー

次元削減後に同じデータが近傍に存在するか？

図 2　k 近傍点を用いた指標

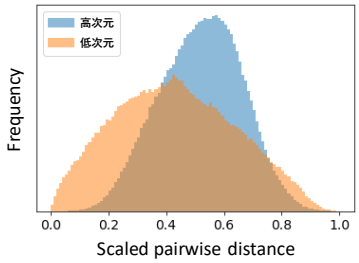

図 3　高次元および低次元空間におけるスケーリングした 2 点間距離のヒストグラム

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）

タから 2 点間距離を計算しました。

以上により、パラメータチューニングのための最適化問題を設定できます。どちらの指標にどの程度重きを置くかという点には任意性がありますが、ここではパラメータ自動調整のデモンストレーションに留め、ローカルあるいはグローバルな指標を最大化するという 2 つの単目的最大化問題に取り組むことにします。

ブラックボックス関数の最大化

上述の最適化問題はブラックボックス関数の最大化問題と言えます。パラメータを入力すると、上記指標の計算結果が出力されるわけですが、数値が出力される仕組みは不明です。そのような状況下で出力を最大化させなければなりません（図 4）。

最も単純には、大量のパラメータの組み合わせを入力し、最大出力を与えるパラメータを探すというアプローチ（グリッドサーチ）も考えられますが、計算時間などの観点から非現実的であることも少なくありません。そこで、最適化手法の適用を考えます。ただし、今回の目的関数はブラックボックス関数であり、t-SNE のアルゴリズム [1]のような、勾配に基づく解析的な最適化手法の適用は困難です。そこで、入力（パラメータ）と出力（指標）の応答関係を手掛かりに、パラメータを探索します。具体的には、ベイズ最適化と呼ばれる手法を用います。典型的なベイズ最適化では、「データ取得→予測モデル構築・予測→予測値・予測の不確実性を基に次の探索点を決定→データ取得→…」という処理を繰り返すことで、効率的なパラメータ探索を図ります。このとき、予測値の不確実性を探索過程に

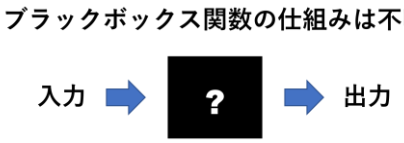

図 4 ブラックボックス関数とは

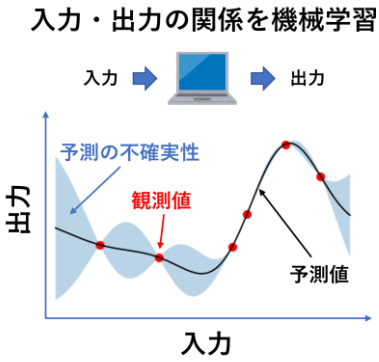

図 5 ガウス過程回帰の概念図

反映するため、ガウス過程回帰などの不確実性を定量できる手法がしばしば用いられます（図 5）。

ベイズ最適化によるパラメータ探索

本稿では、t-SNE のパラメータである *perplexity*、*early_exaggeration*、*learning_rate*、および初期化方法を対象にベイズ最適化を実行しました。今回、80 回という限られた探索回数で、全 20 万通りのパラメータの組み合わせから最適解の探索を試みました。図 6 にベイズ最適化の結果を示します。ベイズ最適化により、ローカル構造、あるいはグローバル構造に重きを置いた次元削減結果が得られました（図 6）。一般的に、ベイズ最適化ではランダムサーチやグリッドサーチよりも効率的な探索が可能になります。ただし、必ずしも大域最適解が得られるわけではなく、探索過程では評価指標の悪い点も選ばれる可能性がある点には留意する必要があります。

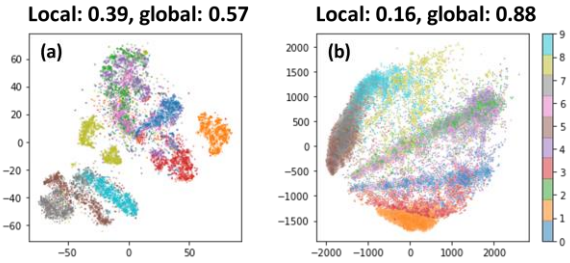

図 6 (a) ローカル構造および(b) グローバル構造に関する指標に基づいたベイズ最適化結果
（各図上部に両指標の値を示す）

おわりに

ベイズ最適化は、機械学習のパラメータ調整に用いられるほか、実験工程の最適化などにも適用可能です。ただし、材料開発に機械学習を適用するためには、個々の材料をコンピュータに認識させる必要があります。次稿では、化合物をコンピュータで演算可能なデータ形式に変換する代表的な方法を紹介します。

参考文献

1) 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)
2) 永廣卓哉：データ解析入門 11 <UMAP による次元削減>、ORIST テクニカルシート、No. 22-16 (2022)
(accessed on March 28th, 2022)
3) https://github.com/zalandoresearch/fashion-mnist/blob/master/README.ja.md
4) D. Kobak and P. Berens, *Nat. Commun.*, **10**, 5416 (2019)

発行日　2022 年 9 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