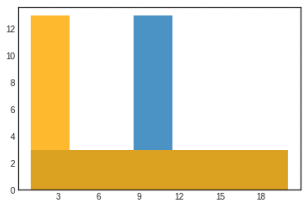
異常検知用の高次元データの可視化に関する研究

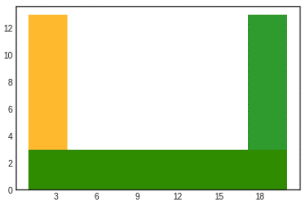
青井 悠佑 †　　岡留 剛 †
Yusuke Aoi　Takeshi Okadome

1. はじめに

現在, 異常検知の分野では高次元のデータを可視化させることによって, 異常データの解釈性を向上させる研究などが行われている. [1] 高次元データを可視化する代表的な手法には高次元データのもつ情報をできるだけ損なわないように低次元空間に縮約する主成分分析や t-SNE(t 分布型確率的近傍埋め込み)[2] などが存在している. 異常検知の際に用いられるノイズの乗った高次元の音声データなどは超平面上に乗っていないため可視化をする際には非線形性を考慮する必要があると一般的に言われている. そのため, 異常検知で用いられる高次元データを可視化する際には主成分分析などの線形的な次元削減手法よりも t-SNE のような非線形的な次元削減手法の方が向いていると考えられている.[3] しかし, t-SNE で分布の類似度を表す尺度として用いられている KL ダイバージェンスには要素の距離を考慮できないなどの問題点があることが知られている. そこで本研究では, t-SNE で距離尺度として用いられている KL ダイバージェンスを別の距離尺度に置き換えることにで, 既存の t-SNE より正確に異常クラスと正常クラスを分離して可視化できる次元削減手法の実現を目指す.



(a)



(b)

図 1: a と b の図の KL ダイバージェンスが同じ値になってしまう例

2. 関連研究

KL ダイバージェンスに代わる分布間の類似度を測る距離尺度として最適輸距離というものが注目されている. 最適輸送距離は KL ダイバージェンスとは違い要素の距離を考慮することができ, さらに距離構造を利用することもできる分布間の類似度を測る距離指標となっている.

2.1 最適輸送距離

最適輸送は分布を比較する際に使われる手法になっている. 比較するヒストグラムを $\mu_0 = \{a_1, a_2, ..., a_m\}, \mu_1 = \{b_1, b_2, ..., b_n\}$ 各点の距離を表す行列を $M \in \mathbb{R}_+^{m \times n}$ とし, 出力である最適輸送距離を次の最適化問題の最適値 γ^* と定義したとき, 以下の式を解くことで最適輸送距離は求められる.

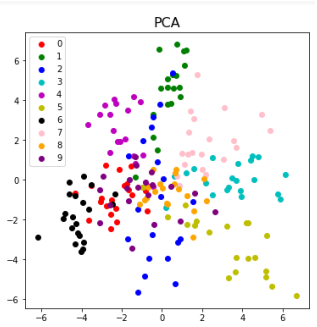
$$\gamma^* = \underset{\gamma \in \mathbb{R}_+^{n \times m}}{arg\min} \sum_{i,j}^{n} \gamma_{i,j} M_{i,j}. \tag{1}$$

$$\text{s.t.} M_{ij} \geq 0, \forall i, j, \sum_{j=1}^{m} \gamma_{ij} = \mu_0, \forall i, \sum_{i=1}^{n} \gamma_{ij} = \mu_1, \forall j$$

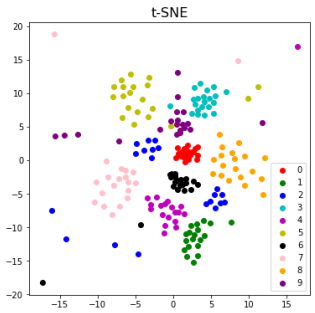
2.2 t-SNE

t-SNE は高次元空間のデータ間の距離とし近いほど確率が高くなり, 遠いほど確率が低くなるような正規分布を仮定している. 高次元空間のデータを低次元の潜在空間（埋め込み空間）に写像（埋め込み）するときには、低次元空間でのデータ間の確率分布を高次元の確率分布と最も似るように選択する. この時, 低次元空間の分布には t 分布を選択している. t 分布を選択する理由としては正規分布より裾野が高い分布を選択することで高次元において中間距離にあるデータ点がより低次元ではより遠くの距離に写像されるようになるためである. このことでクラス間の距離が大きくなる. また, 分布の類似度を測る際には KL ダイバージェンス (相対エントロピー) を利用する. KL ダイバージェンスを誤差関数として確率的勾配降下法で最適化を行う. この時にもクラス内のデータを密集させクラス間の距離を大きくする力が働いている. そのため, t-SNE で高次元データを可視化した際は PCA で可視化した時よりも図 2 のようにクラス間の距離が大きくなる. 確率的勾配降下法で見つかった解が最適解とは限らないので, 初期値をランダムにふりなおし, 繰り返し計算する必要がある. 分散は相対配置エントロピーがすべてのデータ点で一致するように決める.

† 関西学院大学, Kwansei Gakuin University



(a) PCA



(b) t-SNE

図 2: MNIST を PCA, t-SNE で可視化した結果

3. 提案手法

既存の t-SNE で分布の類似度を表す尺度として用いられている, KL ダイバージェンスは要素の距離を考慮できないため, 図 1 のようなオレンジ色のヒストグラムと青色のヒストグラムの KL ダイバージェンスの値とオレンジ色のヒストグラムと緑色のヒストグラムの KL ダイバージェンスの値が同じ値になってしまうというのが問題となっている. そこで, 提案手法では距離尺度に対して, 最適輸送距離を使用することによって, 要素の距離を考慮できるようになるため, 図 1 のようなオレンジ色のヒストグラムと青色のヒストグラムの KL ダイバージェンスの値とオレンジ色のヒストグラムと緑色のヒストグラムの KL ダイバージェンスの値が同じ値になることを防止している.

4. 実験

4.1 予備実験

最適輸送距離が実際に要素の距離を考慮して図 1 のようなオレンジ色のヒストグラムと青色のヒストグラムの KL ダイバージェンスの値とオレンジ色のヒストグラムと緑色のヒストグラムの KL ダイバージェンスの値が異なる値になるのか確かめるため, 図 1 のように 1〜20 の整数が複数回出現するデータでヒストグラムを作成した. オレンジ色のヒストグラムは 1〜3 の整数が 12 個ずつ, それ以外の整数が 3 個ずつ出現するデータになっている. 青色のヒストグラムは 9〜11 の整数が 12 個ずつ, それ以外の整数が 3 個ずつ出現するデータになっている. 緑色のヒストグラムは 18〜20 の整数が 12 個ずつ, それ以外の整数が 3 個ずつ出現するデータになっている. オレンジ色のヒストグラムと青色のヒストグラム, オレンジ色のヒストグラムと緑色のヒストグラムの KL ダイバージェンスの値はどちらも 0.489 で同じ値になった. それぞれ, 最適輸送距離の値を測定するとオレンジ色のヒストグラムと青色のヒストグラムの最適輸送距離は 0.037, オレンジ色のヒストグラムと緑色のヒストグラムの最適輸送距離は 0.118 というように異なる値を出力した. さらに, 今回の結果が高次元データでも同様の結果が得られるか確かめるために図 1 のヒストグラムの 4 次元バージョンを作成し測定を行った. その結果, 上記の結果と同様の結果を得ることができた.

4.2 比較実験

この比較実験では, 提案手法と既存の t-SNE で高次元データを可視化した際のクラスタリング精度を比較するための実験を行った. クラスタリング精度を比較するために用いる指標は調整ランド指数 (Adjusted Rand Index) と正規化相互情報量（Normalized Mutual Information）[4] というクラスタリング結果を評価する際に使われる指標を用いる. この二つの指標は正解ラベルを用いて評価する指標となっている. 調整ランド指数は, 同じクラスタに属すべきデータ同士が正しく同じクラスタに属しているかを表している指標となっていて, 1〜0 の間で数値を出力する. 正しくクラスタリングされているほど 1 に近い数値を返す. 正規化相互情報量も [0,1] の値を取り, 値が高いほど優れているということを表す指標となっている. 最初に提案手法がクラスタリング手法として使用できることを確かめるために, 簡単なデータを作成し実験に用いた. 作成したデータはクラス数が 2 で 6 次元の教師ありデータで図 3 のように分布している.

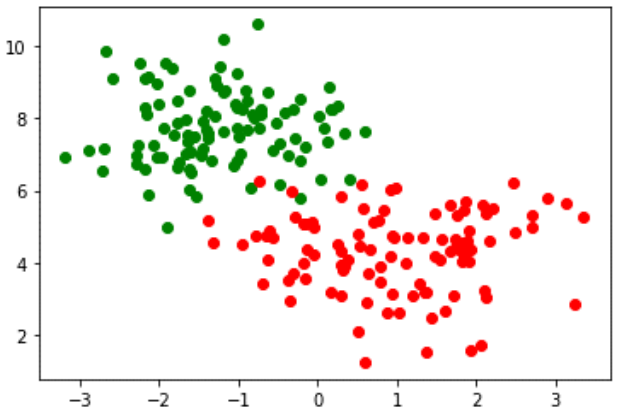
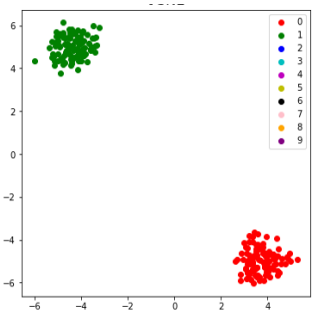
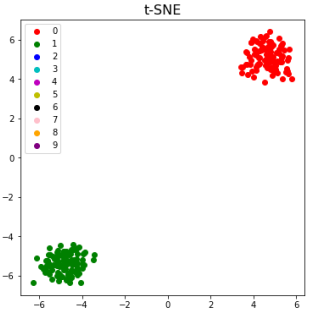


図 3: 作成したクラス数が 2 で 6 次元のデータセット

このデータを実際に提案手法で可視化した結果が図 4 の (a) である. また, 既存の t-SNE で可視化した結果が図 4 の (b) となっている.



(a) 提案手法



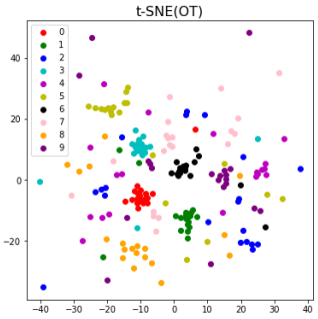
(b) 既存の t-SNE

図 4: 作成したクラス数が 2 で 6 次元のデータセットの可視化結果

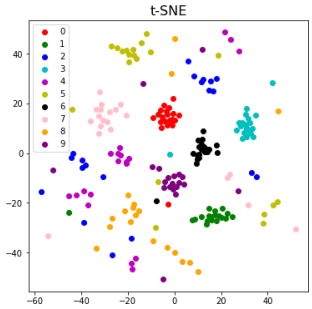
表 1: 作成したデータセットでの実験結果

	提案手法	既存の t-SNE
ARI	1.0	1.0
NMI	1.0	1.0

二つの手法ともに同じクラスタに属すべきデータ同士が正しく同じクラスタに属するように可視化できていることがわかる. また, 異なるクラスタ間の距離が大きくなることによりデータが違うクラスに所属していることが視覚的にもわかりやすくなっている. さらに, 両手法とも ARI, NMI の数値がともに 1.0 となっており, 高い精度のクラスタリングができているということがわかる. このことから, 提案手法は複雑ではないデータであれば既存手法と同等のクラスタリング精度を示すことができるとわかる. 次に, 主成分分析で可視化した際にクラスタ同士が重なり合ってしまうような高次元のデータを用いて, 提案手法と既存の t-SNE で高次元データを可視化した際のクラスタリング精度を比較する. この実験では手書き数字の画像データセット MNIST を用いる. 実際に MNIST を提案手法と既存手法で可視化したものが図 5 である.



(a) 提案手法



(b) 既存の t-SNE

図 5: MNIST データセットの可視化結果

図 5 から, ラベルが 0,1,3,5,6,9 のデータに関しては既存の t-SNE と同様に正確にクラスタリングできていることがわかる. しかし, 2 や 7 などのラベルを持つデータは既存手法よりもクラスタリング精度が低くデータが分散してしまっている. 既存の t-SNE の ARI は 0.44 で提案手法の ARI は 0.37 となり, 既存の t-SNE の方が ARI の数値が少し高くなった. また, NMI に関しても既存の t-SNE で可視化した際の値が 0.549 となり, 提案手法で可視化した際の値は, 0.500 となった. これらのことから, 既存の t-SNE より少しクラスタリング精度が落ちてしまうことがわかる.

表 2: MNIST での実験結果

	提案手法	既存の t-SNE
ARI	0.37	0.44
NMI	0.500	0.549

5. 考察

初めに, 今回の MNIST を可視化した際に提案手法が既存手法よりも ARI と NMI の数値が少し低くなっていた理由について考察する. これは, 実験の際に学習のステップ数を提案手法と既存手法で合わせたため, 最適輸送距離を使用している提案手法では, 既存手法よりも計算が複雑になってしまい, MNIST データセットのようなクラス数が多く高次元空間上でデータ集合が複雑な形

をしているデータセットでは, 計算が収束するまでの学習回数の差が大きくなってしまっているためだと考えられる. そのため, 実験のために作成したクラス数が 2 で 6 次元の教師ありデータのような, クラスの数が少なく高次元空間上でのデータ集合の形が単純なデータセットの場合, 既存手法と提案手法の計算が収束するまでの時間の差が小さく同じステップ数で ARI と NMI の値が等しくなったと考えられる.

6. まとめと今後の展望

本研究では, 高次元データの可視化の際に使われる t-SNE の中で使用されている距離尺度を別の距離尺度に変更した新たな手法を提案し, 既存手法と提案手法のクラスタリング精度を比較した. そこで, MNIST データセットのような計算に時間がかかるデータセットに対しては, 提案手法の方が既存手法に比べて, 同じ学習回数におけるクラスタリング精度が少し低くなるということが分かった.

今後の課題としては, クラスタリング精度の向上と処理速度の向上, 提案手法を異常検知用のデータセットに使用した際の既存手法とのクラスタリング精度の比較などが挙げられる. 処理速度の向上に対しては, 最適輸送距離の計算をシンプルかつ高速に計算できるようにしたシンクホーンアルゴリズムを用いることで処理速度を向上できると考えている. 異常検知用のデータセットでの実験に関しては ToyADMOS や MIMII などの異常検知用に用意された高次元データセットを利用していく予定である.

参考文献

[1] Shi, Lei Liao, Qi He, Yuan Li, Rui Striegel, Aaron Su, Zhong. (2011). SAVE: Sensor anomaly visualization engine. *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011*, Proceedings. 201-210. 10.1109/VAST.2011.6102458.

[2] van der Maaten, L. Hinton, G. (2008). *Visualizing Data using t-SNE . Journal of Machine Learning Research*, 9, 2579–2605.

[3] Tagawa, Yuki, Rytis Maskeliūnas, and Robertas Damaševičius. 2021. *"Acoustic Anomaly Detection of Mechanical Failures in Noisy Real-Life Factory Environments" Electronics* 10, no. 19: 2329. https://doi.org/10.3390/electronics10192329

[4] Strehl, Alexander and Ghosh, Joydeep. *"Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions.."* J. Mach. Learn. Res. 3 (2002): 583-617.

[5] LeCun, Yann and Cortes, Corinna. *"MNIST handwritten digit database."* (2010): .