6R-07

深層学習的手法を用いたクラスタリング手法の提案

増井建斗† 金井祐輔‡ 尾形正泰‡ 宮崎庸平‡ 今井倫太†
慶應義塾大学理工学部† 慶応義塾大学理工学研究科‡

1.はじめに

基本的なクラスタリングアルゴリズムには k-means 法や ward 法,DBSCAN 法が存在するが、書籍の表紙画像等、多数の次元を持った情報には直接適用することができない。高い次元数を持った情報に対するクラスタリングには、次元の呪い[1]と球面集中現象[1]という２つの問題が発生するためである。次元の呪いとは、クラスタリングアルゴリズムの計算量が次元数に応じて大きくなることを言う。球面集中現象は、特徴表現空間の次元数増加に伴って情報同士の距離が大きくなってしまうことを言う。クラスタリングは特徴表現空間上での距離を基準に行われるため、高次元な空間では距離の意味が薄れ、クラスタリングが難しくなる。以上２つの理由により、多数次元の情報は、クラスタリングの前に次元を削減する必要がある。

Stacked Denoising Autoencoder(SDA)[2]は、入力された情報をより少ない次元で表現する様に教師無し学習するアルゴリズムである。SDA は情報から高階な非線形射影関数を学習することに長けており、入力情報をより少ない次元数で表現することが可能である。

本稿では SDA を用いて書籍表紙画像の次元数を削減することで次元の呪いや球面集中現象を回避し、クラスタリングを行う。SDA が書籍表紙画像等の次元削減に有効であることを示すため、同じく次元削減を行うアルゴリズムである主成分分析(PCA)でも比較実験を行った。

2.関連研究

Auto-encoder Based Data Clustering[3]では、StackedAutoencoder の目的関数にクラスタ中心に対する距離を指標として含め、次元削減ではなくクラスタリングを直接行っている。これによって数字の手書き文字データ・セットである MNIST のクラスタリング精度が改善されることを示している。対して本稿では SDA をクラスタリング前の次元削減アルゴリズムとして用いている。クラスタリング対象として数字の手書き文字ではなく書籍の表紙を用いることで、より一般的な画像のクラスタリングに対する SDA の精度を明らかにする。

Dimensionality reduction for clustering : An approach from deep learning

† Kento MASUI, Michita IMAI
Faculty of Science and Technology, Keio University
‡ Yusuke KANAI, Masa OGATA, Yohei MIYAZAKI
Graduate School of Science and Technology, Keio University

3.システム構成

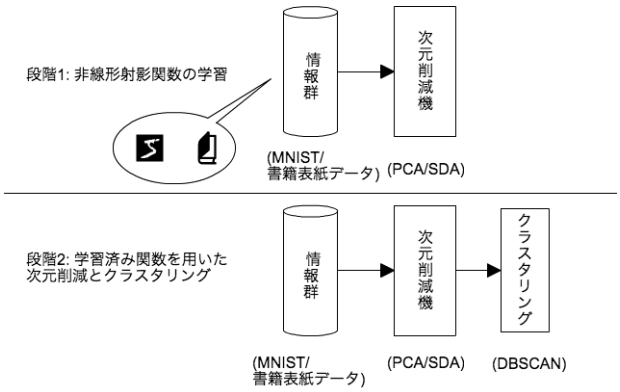

図 1 次元削減機の学習から
次元削減、クラスタリングまでの構成

図 1 に情報の次元削減からクラスタリングまでの流れを示す。システムは２つの段階で構成されている。

１段階目では次元削減機の教師なし学習を行う。次元削減機のモデルとしては PCA か SDA のいずれかを用いる。次元削減機の教師なし学習とは、PCA の場合は情報群の主成分ベクトルを選択させること、SDA の場合は SDA のレイヤーをそれぞれを教師なし学習させることである。

２段階目では１段階目で学習した次元削減機を用いて情報の次元削減を行い、この結果から DBSCAN を用いてクラスタリングを行う。DBSCAN は基本的クラスタリングアルゴリズムの一つで、情報の表現空間における密度を元にクラスタリングを行う。

4.実験

実験は２種類の入力画像セット、２種類の次元削減機を用いて行う。入力画像は画像の機械学習分野で多く用いられている MNIST と、電子書籍サイトで販売されている書籍の表紙画像を用いる。MNIST は 0 から 9 の手書き文字を６万個集めた画像データセットであり、文字一つひとつを 28×28 ピクセルのグレースケール画像として集めている。電子書籍の表紙画像とは、電子

書籍販売サイトで公開されている表紙画像を無作為に4万件選んだものである。今回の実験では選択された表紙画像をすべて、色情報を保ったまま 28×28 の解像度へ縮小して用いている。

表 1:SDA の構成

	書籍表紙画像	MNIST
レイヤーサイズ	[5000, 3000, 10]	[1000, 1000, 1000, 10]
学習率	0.01	0.01
学習回数	15	15

書籍表紙画像、MNIST に対する SDA はそれぞれ違う構成で実験を行った。それぞれのデータに対する SDA の構成を表 1 に示す。

PCA は書籍表紙画像と MNIST のそれぞれに対して同じパラメータを用いた。PCA によって得られる主成分から 10 成分を使用している。

MNIST と書籍表紙画像、それぞれに対して SDA, PCA のそれぞれで次元削減を行った。得られた4通りの次元削減結果に対して DBSCAN でクラスタリングを行い、その結果を確認した。また、それぞれのクラスタリング精度についてシルエット係数を用いて評価した。シルエット係数は、ラベル付けがされていないデータを対象とするクラスタリングの精度を評価する係数である。シルエット係数は-1 から 1 の値を取り、数値が大きいほどクラスター内部の密度が高いことを意味する。0 に近い場合は、クラスタ同士が重なっていることを表現している。

5.結果

前節で説明した 4 通りの実験結果について、それぞれのシルエット係数を表 1 に示す。書籍表紙においては、SDA のシルエット係数が PCA のシルエット係数を上回っているのに対し、MNIST ではこの関係が逆転している。これはクラスタリング対象の性質が異なるために起きていると言える。MNIST は手書きの数字で構成されているためその構成パターンも少なく、SDA が得意とする高次な特徴表現の学習が生かされていないと考えられる。書籍表紙画像のクラスタリングにおいては SDA がより高いシルエット係数を示しており、複雑なデータに対して SDA が PCA に対して比較的有効な特徴表現を獲得できていることが確認できる。

図 2 に、Multi-dimensional Scaling(MDS)を用いて計算したクラスタリング結果の分布を示す 。MDS は多次元情報を相対距離に応じて任意の次元に配置する。今回の実験では入力画像を 10 次元まで次元削減したが、更にこの MDS を用いて 2 次元まで圧縮し、グラフにプロットした。

表 2:クラスタリング結果のシルエット係数

	SDA	PCA
書籍表紙画像	−0.129	−0.235
MNIST	−0.322	−0.152

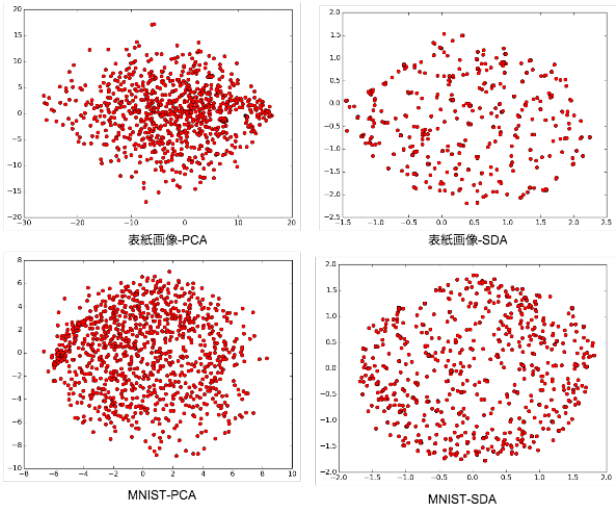

図 2:クラスタリング結果の可視化

書籍の表紙画像と MNIST, どちらの画像情報に対する次元削減結果も、SDA による次元削減がより疎な分布を作り出していることが確認できる。

6.結論

本稿ではクラスタリングに必要とされる次元削減を PCA と SDA で行い、その性能差を評価した。SDA は PCA と比較してより高階な非線形射影関数を学習するとされている。今回の実験結果として、SDA はより高階な情報を持った入力に対してクラスタリングに適した次元削減を行うことがわかった。

7.参考文献

[1] 石井健一郎、前田英作、上田修功、村瀬洋 “わかりやすいパターン認識” オーム社 1998

[2] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *The Journal of Machine Learning Research* 11 (2010): 3371-3408.

[3] Chunfeng Song, et. al, “Auto-encoder Based Clustering”, 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I