深層異常検知モデルの中間表現によるデータ分析手法に関する研究

1X17C001-3　相木将寛
指導教員　後藤正幸

1. 研究背景・目的

多様なデータが活用できるようになった現在，データに基づき異常を検知すると共に，異常データの特徴を発見することは重要な技術となっている．しかし，単にアラートとして異常を検知するだけでなく，異常の特徴を分析可能とする手法やモデルに関する研究は少ない．そこで本研究では，異常検知に影響を与える特徴を発見する手法を構築することを目的とする．

代表的な異常検知手法の 1 つに One-Class Neural Networks(OCNN)[1] がある．OCNN は深層学習モデルである Autoencoder(以下，AE) を用いて縮約した特徴ベクトルをニューラルネットワークに入力する．そのため，ニューラルネットワークの中間層で形成される特徴空間には，入力データの分布情報が集約される．その際，特徴空間の一部の特徴軸を用いて正常データが埋め込まれるように学習されると考えられる．すなわち，異常データは，正常データが埋め込まれる部分空間とは異なる部分空間に埋め込まれ，正常データを表現する特徴軸と異常データを表現する特徴軸が明確に区別されていると考えられる．そのため，ニューラルネットワークの中間層を分析することで，異常検知に寄与している特徴軸を発見できる可能性がある．

以上より，本研究では，OCNN のニューラルネットワークの中間層に得られる中間表現を分析することで異常検知に寄与している重要な特徴軸を特定する手法を提案する．まず，学習した OCNN により，各データに対して正常/異常のラベルを付与する．付与したラベルと中間層の各特徴軸の値を基に，識別器の性能を表す評価指標として用いられる Area Under the Curve(以下，AUC) を特徴軸ごとに算出し，その値を用いて異常検知に寄与している特徴軸を特定する．異常検知に寄与している特徴軸を見つけることで，一部の特徴軸の値だけを利用することでも異常検知が行えるようになり，検知された異常について詳細な分析が可能になる．これにより，OCNN で用いられているニューラルネットワークにおいて，どのようなメカニズムで異常検知が行われているかの解釈への一助になることも期待できる．最後に実データを用いて分析・考察を行い，提案手法の有用性を示す．

2. One-Class Neural Networks(OCNN)

OCNN は AE を用いて抽出した特徴をニューラルネットワークに教師なし学習させることで異常検知を行う手法である．OCNN では，まず AE の事前学習を行う．その上で，事前学習したパラメータを用いて AE のエンコーダ部分と異常検知を行うニューラルネットワークを接続する．このとき，AE のパラメータも再度，ニューラルネットワークのパラメータと合わせて学習することで高い性能を得る．

対象データを $\boldsymbol{x}_n(n \in \{1,2,...,N\})$，ニューラルネットワークの出力層で得られる出力を $\boldsymbol{w}$，データの入力からニューラルネットワークの中間層までの重み行列を $\boldsymbol{V}$ とすると，OCNN の目的関数は式 (1) で与えられる．

$$\min_{\boldsymbol{w},\boldsymbol{V},r} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2}\|\boldsymbol{V}\|_F^2 + \frac{1}{\nu}\cdot\frac{1}{N}\sum_{n=1}^{N}\max(0, r-\boldsymbol{w}\cdot g(\boldsymbol{V}\boldsymbol{x}_n)) - r \tag{1}$$

ここで，ν は異常割合を示すハイパーパラメータ，r は超平面のバイアス，$g(\cdot)$ はニューラルネットワークの活性化関数を表す．式 (1) の $\boldsymbol{w}$,$\boldsymbol{V}$,r の 3 つのパラメータを次に示すアルゴリズムにより逐次的に学習する．

Step1) $\boldsymbol{w}$,$\boldsymbol{V}$ の学習

r を固定し，式 (2) を最小化するように $\boldsymbol{w}$,$\boldsymbol{V}$ を学習する．

$$\min_{\boldsymbol{w},\boldsymbol{V}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2}\|\boldsymbol{V}\|_F^2 + \frac{1}{\nu}\cdot\frac{1}{N}\sum_{n=1}^{N}\max(0, r-\boldsymbol{w}\cdot g(\boldsymbol{V}\boldsymbol{x}_n)) \tag{2}$$

Step2) r の学習

Step1 で算出した $\boldsymbol{w}$，$\boldsymbol{V}$ を用いて，式 (3) を最小化するように r を学習する．

$$\min_{r} \frac{1}{\nu}\cdot\frac{1}{N}\sum_{n=1}^{N}\max(0, r-\boldsymbol{w}\cdot g(\boldsymbol{V}\boldsymbol{x}_n)) - r \tag{3}$$

Step1 と Step2 を繰り返し，r と $\boldsymbol{w}$，$\boldsymbol{V}$ を交互に学習する．学習終了時に得られるパラメータを $\hat{r}, \hat{\boldsymbol{w}}, \hat{\boldsymbol{V}}$ とする．

3. 提案手法

3.1. 概要

本研究では，OCNN のニューラルネットワークの中間表現を分析することで異常検知に寄与する重要な特徴軸を特定する手法を提案する．提案手法では，まず OCNN の出力 $\hat{\boldsymbol{w}}\cdot g(\hat{\boldsymbol{V}}\boldsymbol{x}_n)$ によりデータの正常/異常を判定する．

また，データ $\boldsymbol{x}_n$ を入力して得られる中間層の特徴軸 $k \in \{1,2,\ldots,K\}$ の値を $g_k(\hat{\boldsymbol{V}}\boldsymbol{x}_n)$ と表す．

このとき，全データに対する OCNN の判定結果を正解とし，特徴軸 k のみの値 $g_k(\hat{\boldsymbol{V}}\boldsymbol{x}_n)$ で分類したときの，AUC_k を算出する．この AUC_k より，異常検知に寄与している特徴軸を明らかにする．

3.2. アルゴリズム

Step1) $\hat{\boldsymbol{w}}\cdot g(\hat{\boldsymbol{V}}\boldsymbol{x}_n)$ より $\hat{r}$ を閾値として得られる正常/異常の判定結果を，各データに対するラベルとする．

Step2) Step1 で得たラベルを用いて，全入力データから出力される N 個の値 $g_k(\hat{\boldsymbol{V}}\boldsymbol{x}_n)$ の大小により，閾値で分類する場合の AUC を算出し，AUC_k とする．

Step3) 閾値 α よりも AUC_k が高い特徴軸を抽出する．

4. 実データ分析

4.1. 分析条件

本研究では，“0”–“9” までの手書き数字画像データセットである MNIST[2] と文書のデータセットである読売新聞記事データ [3] を用いる．

MNIST は，“0” を正常データとして 4,936 件，“1”–“9” を異常データとして計 50 件を用いた．ニューラルネットワークの入力層は 32 次元，中間層の次元数は 16，32，64 の 3 種類で実験を行った．

読売新聞の記事データは，出現頻度が高い名詞・動詞の出現回数が 30 以上の 7,432 単語を特徴量として使用した．「犯罪・事件」のラベルが付与された文書を正常データとして 1,500 件，「政治，経済，社会，スポーツ，文化，生活，科学」の 7 種類がラベルとして付与された文書を異常データとして計 70 件を用いた．ニューラルネットワークの入力層は 300 次元，中間層の次元数は 150，300，600 の 3 種類で実験を行った．提案手法の閾値は MNIST，読売新聞記事どちらも $\alpha = 0.6$ に設定した．

4.2. 分析結果

MNIST，及び記事データにおいて，閾値以上の特徴軸の数及び異常検知手法としての精度 (AUC) を表 1,2 に示す．

表 1: MNIST に対する実験結果

中間層の次元数	16	32	64
抽出された次元数	5	5	9
異常検知精度 (AUC)	0.555	0.631	**0.728**

表 2: 読売新聞に対する実験結果

中間層の次元数	150	300	600
抽出された次元数	6	11	275
異常検知精度 (AUC)	**0.705**	0.649	0.683

表 1,2 より MNIST では中間層の次元数が 64 次元の場合において最も精度が高くなり，読売新聞では 150 次元の場合において最も高くなった．また，中間層の特徴軸の中でも AUC_k が高くなる特徴軸が一定数存在することが MNIST，読売新聞の双方で確認できた．これにより，OCNN で用いられるニューラルネットワークでは，異常検知に寄与する特徴軸が特定可能であることが示された．

4.3. AUC_k が高い特徴軸の可視化分析

AUC_k が閾値を超えた特徴軸を抽出し，可視化を行うことで，異常検知が可能であることを確認する．精度が最も高くなった MNIST の 64 次元，読売新聞の 150 次元の中間層から閾値を超えた特徴軸を抽出し，主成分分析により 3 次元に集約して可視化する．図 1 に MNIST，図 2 に読売新聞に対する可視化を示す．ここで，図 1,2 において，紺色が正常データを表し，その他の色が異常データを表す．

図 1,2 より MNIST，読売新聞それぞれにおいて，抽出した特徴軸だけで正常データと異常データが判別できていることが分かる．次に，主成分分析の 3 次元空間上で中央値から最も距離の離れているデータを分析する．図 3 に MNIST において最も中央値から外れた 3 点の異常データを示す．

図 3 より，正常データ (“0”) とはその造形が大きく異なる “3”，“8”，“5” のような数字が中央値から外れていることが

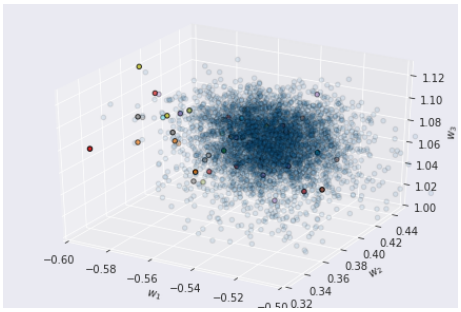

図 1: 特徴軸の可視化 (MNIST)

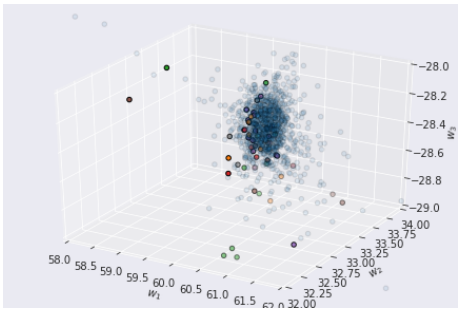

図 2: 特徴軸の可視化 (読売新聞)

図 3: 最も中央値から外れた 3 点の異常データ

確認できた．また，読売新聞では中央値から外れた異常データのうち，密集していた 3 点の記事データを分析したところ，すべて「生活」というカテゴリで分類される記事であった．そのため，位置が中央値から大きく外れていても，カテゴリが同一の記事同士の距離が近くなることが確認できた．これらの主成分分析の可視化と異常データの分析により，本研究における異常検知に寄与する特徴軸の発見が有効であると考えられる．このように，OCNN で用いられているニューラルネットワークにおいてどのように異常検知が行われているのかについて，その解釈の一助になると考えられる．

5. まとめと今後の課題

本研究では，OCNN の中間層から異常検知に大きく寄与する特徴軸を明らかにする手法を提案した．そして，提案手法を実データに適用し，その有効性を確認した．この特徴軸の発見により，OCNN による異常検出メカニズムの解釈の一助になると考えられる．今後の課題としては，中間層を分析することで正常データに複数のクラスのデータが存在している場合のクラス分類に対する知見の獲得が挙げられる．

参考文献

[1] Chalapathy, R., *et al.*, “Anomaly detection using one-class neural networks,” *arXiv preprint arXiv:1802.06360*, 2018.

[2] Yann LeCun, *et al.*, “Mnist handwritten digit database,” *ATT Labs (Online). Available: http://yann. lecun. com/exdb/mnist*, 2010.

[3] 読売新聞社, “2015 年　読売新聞記事データ集 (邦文),” 日外アソシエーツ, 2016.