広島県立総合技術研究所東部工業技術センター研究報告　No. 36（2023）
技術ノート

深層学習による異常検知手法の簡単な比較（第 3 報）

1 クラス ニューラルネットワーク

廣川　勝久

Anomaly Detections using Deep Learning (Ⅲ)

One-Class Neural Network

HIROKAWA Katsuhisa

異常検知に深層学習を用いた生成モデルでは，正常信号のみの学習から，異常信号との差異が計算可能となった。本報告では，生成モデルに 1 クラスニューラルネットワーク組みわせることにより，生成モデルの正常信号と異常信号との境界をニューラルネットワークの学習を用いて設定する方法について評価を行った。

キーワード：1 クラスニューラルネットワーク，One-Class Neural Network, オートエンコーダ， AE，VAE，PyTorch

1. 緒　　言

スマートフォンやインターネットのデジタル技術の進歩により，非常に多種多様なデータがクラウドに蓄積されている。これらのデータを解析･利用することにより，新たな価値を生み出すことが試みられている。古くはデータ解析に主成分分析などの数学的手法によるアプローチが用いられたが，近年，クラウドデータの大容量処理には，大容量のパラメータを持つニューラルネットワークに蓄積データを学習させ，学習したデータの確率分布から答えを出力する生成型のモデルの利用が進んでいる。オートエンコーダ（**Auto Encoder: AE**）や敵対的生成ネットワーク（**Generative Adversarial Networks**：**GAN**)などの典型的な生成モデルは [1]，新しい画像生成や，セグメンテーション，自然言語処理などの多くの分野に応用されようとしている。

第 2 報まで，オートエンコーダなどの生成モデルが，正常信号のみの学習から異常信号を検出できることを示した [2,3]。これは特に，生成モデルの潜在変数への低次元クラスタリングと，クラスタから典型的なデータを復元する能力によるものであった。本報告では，生成モデルのオートエンコーダが潜在空間に写像したクラスタデータを入力に用いて学習を行い，1 クラスニューラルネットワーク [4]による正常信号と異常信号の分離を行った。1 クラスニューラルネットワークを用いることにより一定割合で正常信号を含む異常信号を分離可能なクラスタ境界の設定が可能であることを示す。

2. 異常検知の自動化

2.1 学習による異常検知自動化

深層学習による生成モデルでは，低次元化された特徴空間の潜在変数から，典型的なデータを復元することができる。生成モデルによる異常検知は，入力データと潜在変数から復元されたデータを比較し，その差により良否の判定が行われる。生成モデルが成す特徴空間ではその特徴値に合わせたクラスタリングが行われている。1 クラスニューラルネットワークは，生成モデルより行われた正常信号のクラスタリングに対して，分離平面を学習により自動的に設定する手法である。この手法では，学習により設定された分離平面に従って異常信号を正常信号から自動的に分離することが可能となる。

2.2 サポートベクターマシン

サポートベクターマシンは，2 つのクラスタからなるデータを 2 つに分離するための境界を決定する方法である。1 クラスサポートベクターマシンは，サポートベクターマシンの特殊なモデルであり，与えられた全データと原点との分離を行うため，原点と原点に最も近いデータ（サポートベクター）との間の正常値を分離する平面を決定する（**図 1**）。この平面には，原点からの距離（マージン $1/\|\boldsymbol{w}\|$）が最大と成るような $\boldsymbol{w}$ が選ばれる。1 クラスサポートベクターマシンに未知のデータが与えられた場合，線形分離平面より原点に近いデータを異常値とし，平面より遠い場合を正常値とする。ところが，**図 2** に示すように異常データが正常データに囲まれているような実次

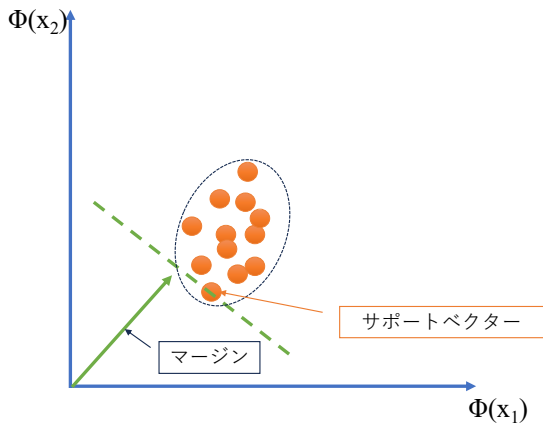

図 1 1 クラスサポートベクターマシン

2023. 7. 12　加工技術研究部　受理

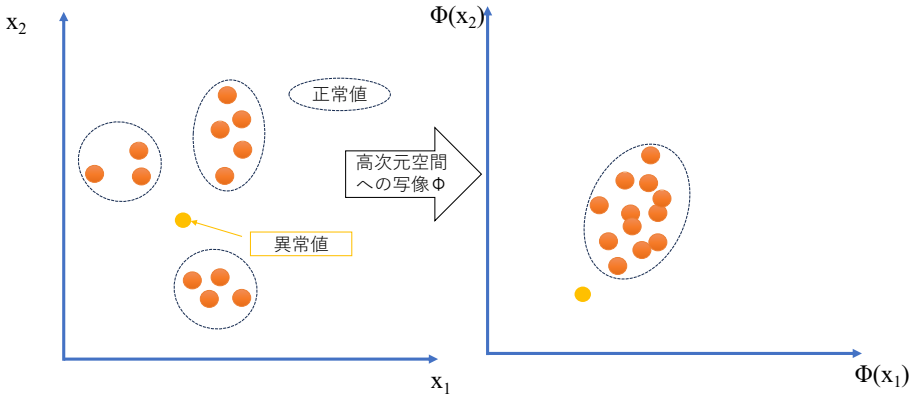

図 2 カーネルトリックによる分離

元の空間では線形分離が不可能な場合がある。サポートベクターマシンでは，入力データを高次元空間に写像することにより線形分離を行うことが出来る。この方法はカーネルトリックと呼ばれ，例えば$\Phi(\cdot)$を高次元空間への写像関数とすると，2 次元の空間に写像された場合の分離直線は，次のように表される。

$$f(\Phi(x_1),\Phi(x_2)) = w_1\Phi(x_1) + w_2\Phi(x_2) - r = \boldsymbol{w}\Phi(X) - r \quad (1)$$

多次元空間における 1 クラスサポートベクターマシンでは，次式より最適なパラメータ $\boldsymbol{w}, r$ を求める。

$$\underset{\boldsymbol{w},r}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + \frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r - (\boldsymbol{w}, \Phi(X_n))\right) - r \quad (2)$$

$$\because \max(s,x) = \begin{cases} 0 & x < s \\ x & x \geq s \end{cases}$$

ここで，ν は正常値が分離平面から原点側にはみ出し，異常値となることを許容する割合（ソフトマージン）を示す。

2.3 1 クラス ニューラルネットワーク

1クラスニューラルネットワークは1クラスサポートベクターマシンの計算をニューラルネットワークにより実現したものと考えることができる。1クラスニューラルネットワークの構成を**図 3** に示す。オートエンコーダにより低次元化された特徴値を 1 クラスニューラルネットワークでは入力とする。入力層，中間層，出力層の 3 層構造となっており，サポートベクターマシンの写像関数$\Phi(\cdot)$は中間層ノードの非線形関数$g(\cdot)$により実現している。また，入力データを重み付けされたニューロンを伝搬させ，非線形関数に入力する構造となっている。学習は次式により行われる。

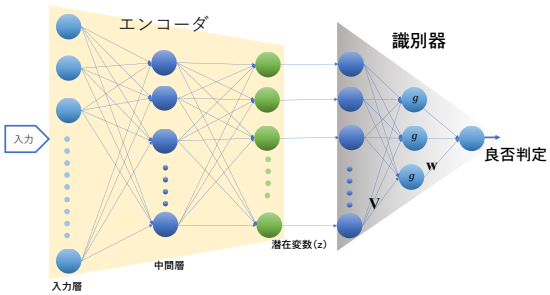

図 3 1 クラスニューラルネットワークの構成

$$\underset{w,V,r}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|V\|_F^2 + \frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r - \left(w, g(VX_n)\right)\right) - r \quad (3)$$

ここで，F は高次元空間の次数を示す。$\boldsymbol{w},\boldsymbol{V}$，$r$ のパラメータを同時に学習できないため，まず，1 クラスニューラルネットワークにオートエンコーダを伝搬させたデータを入力し，r の値を定数とし，勾配計算を行い重み $\boldsymbol{w},\boldsymbol{V}$ を更新する。

次に，更新後の重み $\boldsymbol{w},\boldsymbol{V}$ を使い，再度 1 クラスニューラルネットワークにオートエンコーダを伝搬させたデータを入力し次式が最大となる r を求める。

$$f(r) = \frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r - \left(w, g(VX_n)\right)\right) - r \quad (4)$$

但し，実際の計算は，1 クラスニューラルネットワークからの出力値を並び替えた小さい値から，$\nu \cdot N$番目の分位数を選択すれば良い。

$$\hat{y}_n = w \cdot g(VX_n) = \{y_1, y_2 \cdots y_n\} \quad (5)$$

$$r = y_{\nu \cdot N}$$

3. ニューラルネットワークへの実装

3.1 学習用 1 次元データ

1 クラスニューラルネットワークに入力するための低次元化された潜在変数は，畳み込みオートエンコーダにより生成した。MNIST : Modified National Institute of Standards and Technology[5]のデータベースによる 6 万字のグレスケール手書き数字を畳み込みオートエンコーダに学習させることにより，各文字に対応したクラスタリングが潜在空間では行われる。28×28 画素の 2 次元手書き数字を畳み込みオートエンコーダにより，徐々に次元数を減しながら最終的に 1 次元の潜在変数を生成し，1 クラスニューラルネットワークの入力としている。クラスタリングが行われた潜在空間に対して 1 クラスニューラルネットワークの学習により最適な分離平面を設定する。ニューラルネットワークのフレームワークには PyTorch を用いた。

3.2 学習用 1 次元データ

1 クラスニューラルネットワークは 2 層の全結合型のネットワークとなっている。高次元への写像関数には sigmoid を用いた。ネットワーク伝搬後の出力を x，学習に必要な各層の重みを v,w とした。

```
def forward(self, x):
    x = self.fc2(x)
    v = self.fc2.weight
    x = torch.sigmoid(x)
    x = self.fc1(x)
    w = self.fc1.weight
    return x, v, w
```

入力データは，オートエンコーダのエンコーダからの出力を 1 クラスニューラルネットワークに渡すことによって出力される。1 クラスニューラルネットワークの学習時に損失がエンコーダへ伝搬することを防ぎ，エンコーダの学習が行われない方式とした。

```
def forward(self, x):
    _,y,_ = self.encoder(x)
    x, v, w = self.suportvector(y.detach())
    return x, v, w
```

また，3 式の損失関数は次のとおり実装し，r は定数として学習を行った。

```
y_hat, v, w = modelOC(images)
r = new_r(y_hat, nu).item()
loss1 = 0.5 * LA.vector_norm(v,ord=fsize)**2  + 0.5 * LA.norm(w)**2
loss2 = (1 / nu) * torch.mean(F.relu(r - y_hat))-r
loss = loss1 + loss2
loss.backward()
optimizerOC.step()
```

定数 r は各層の重み更新後，再計算を行っている。

```
def new_r(y_hat, nu):
    r = torch.quantile(input=y_hat, q=nu)
    return r
```

4. クラスタデータの識別能力

4.1 クラスタ特徴値の高次元空間へのマッピング

実験では，まず，全ての手書き文字を変分オートエンコーダに学習させ，低次元化された特徴値を潜在変数に出力できるようにエンコーダ部分の重みを学習する。次に，この変分オートエンコーダのエンコーダ部分と 1 クラスニューラルネットワークを接続する。学習を終えたエンコーダからの出力を 1 クラスニューラルネットワークに入力し，1 クラスニューラルネットワーク部分の学習を行った。1 クラスニューラルネットワークの高次空間次数は 256 を設定した。25 回の学習後，学習に使用した 6 万文字を入力し，得られた原点までの距離に従って画像の並び替えを行った。**図 4** に原点に最も近いと判別された画像を左上から順番に右下に並べた。これらの画像は全て分離平面より原点側の異常値と判定されたものである。**図 5** に原点から最も遠い画像を左上から右下に 64 枚を並べた。原点に近い画像は，線が細く乱れた文字が多く見られる。一方で原点から遠い画像では文字は太く，傾

図 4 原点に近い異常画像

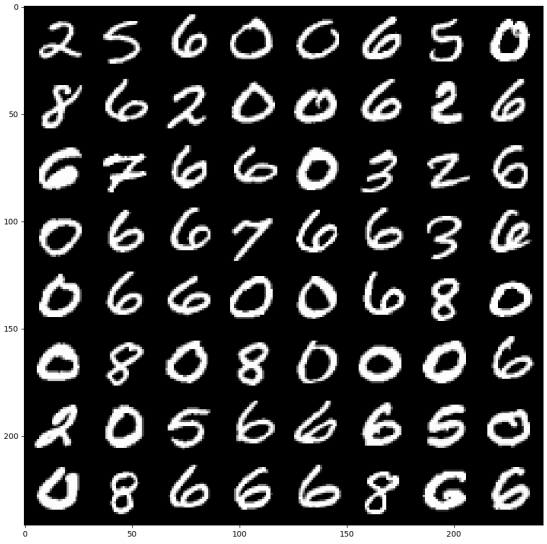

図 5 原点から遠い正常画像

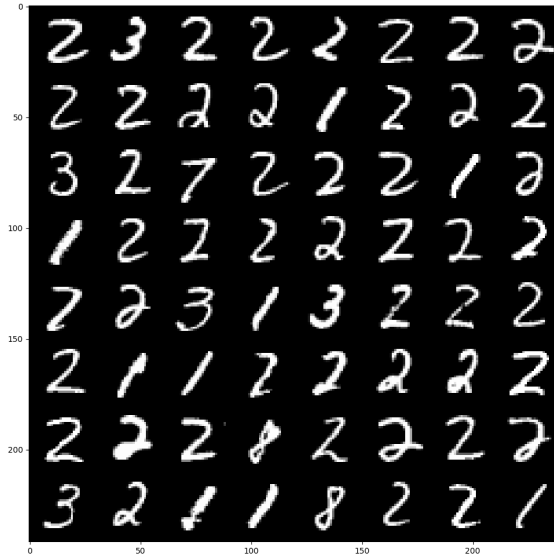

図 6 原点に近い異常画像

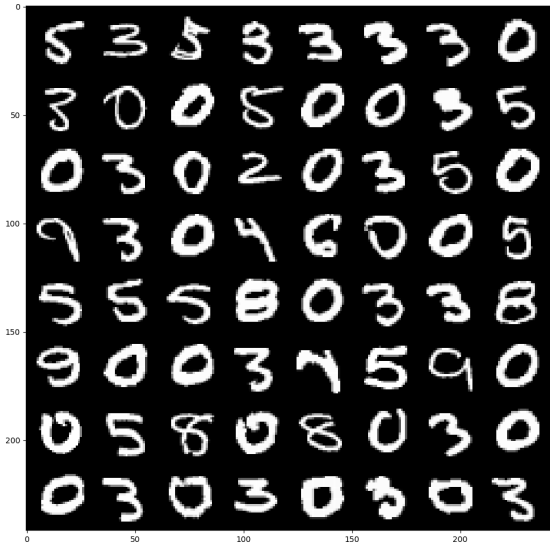

図 7 原点から遠い正常画像

向が同じ文字となっている。このことから, 1 クラスニューラルネットワークにより，典型的な特徴値（正常値）から異常値が学習により自動的に分離されたと考えることが出来る。

また，**図 4**, **図 5** の画像の違いは，1 クラスニューラルネットワークの特徴値の高次元空間へのマッピング傾向を示すものと考えられる。しかし, 更に学習を進めると，原点から最も近い画像と原点との距離が長くはなるが，各画像が近づいた座標にマッピングされる傾向がある。高次元空間での広がりは縮小され，典型的な特徴値と異常値の混在が始まる。

4.2 未学習特徴値の識別能力

全文字を学習したオートエンコーダのエンコーダ部分を使い，1 クラスニューラルネットワークには 0, 6, 9 の手書き数字 3 種類のみを学習させ，評価を行った。学習後全ての手書き数字 6 万字を入力し，その識別能力を調べた。**図 6** に原点に最も近い異常画像を左上から並べ，**図 7** に原点から最も遠い正常画像を左上から並べた。異常値を示す画像のほぼ全てが未学習の文字であり，数回の試行では，ソフトマージンνに応じた割合で学習した文字が含まれる場合もあった。一方，正常画像には，学習・未学習の両方の文字が含まれる結果となった。この結果より，たとえオートエンコーダによって特徴値によるクラスタリングが行われたとしても, 1 クラスニューラルネットワークは学習に用いた特徴値のみの異常検知制御が可能であり，他の特徴値は高次元空間ではクラスタリングされず，識別することは出来ない。従って 1 クラスニューラルネットワークを異常検知に用いる場合，オートエンコーダが出力する特徴値の中で，必要とする特徴値全てを学習する必要がある。

5. 結　　言

正常信号のみの学習により, 異常信号の特異な部分や欠損部分検出できる生成モデルは，今後応用が期待される深層学習分野である。情報の低次元化が可能な一種の生成モデルであるオートエンコーダに 1 クラスニューラルネットワークを組み合わせ，異常信号と正常信号との学習による分離能力について評価を行った。1 クラスニューラルネットワークを組み合わせることにより，予め設定した一定割合の異常値を含む分離がニューラルネットワークの学習により可能であることを示した。

但し，今後，1 クラスニューラルネットワーク異常信号の分離に用いるためには, 1 クラスニューラルネットワークによって写像される高次元空間での広がりの設定がパラメータとして重要となると考えられる。

文　　献

1) 毛利 拓也 他: GAN ディープラーニング実装ハンドブック, 秀和システム（2021）.
2) 廣川 勝久：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による異常検知手法の簡単な比較（第 1 報）35（2022）.
3) 廣川 勝久：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による異常検知手法の簡単な比較（第 2 報）36（2023）.
4) R. Chalapathy, A. K. Menon, S.Chawla: Anomaly detection using one-class neural networks. arXiv: 1802. 06360v2, (2019).
5) THE MNIST DATABASE of handwritten digits：http://yann.lecun.com/exdb/mnist/