

深層学習による異常検知手法の簡単な比較 (第1報)

廣川 勝久

Anomaly Detection using Deep Learning (I)

HIROKAWA Katsuhisa

教師あり深層学習による異常検知では、学習に正常な教師信号と異常な教師信号を同数程度必要とする。しかし、実際の使用環境では、異常信号の発生は極めて少なく、必要数の異常信号を学習することが困難とされる。一方、深層学習による生成モデルを異常検知に用いることより、正常信号のみの学習から、異常信号を検出することが可能である。本報告では、深層学習の生成モデルを用いた異常検知の評価を行ったので、その結果について述べる。

キーワード：オートエンコーダ、AE、変分オートエンコーダ、VAE、PyTorch、深層学習、Python

1. 緒 言

これまで、多次元のデータが与えられた場合、そのデータ分類のために、より低次元の特徴空間への写像が行われた。低次元への写像は、データの持つ特徴値により決定され、特徴空間の座標が直交する主成分分析や無相関の座標系から成る独立成分分析などが主に解析的な分類手法として用いられた^{1,2)}。

深層学習により主成分分析や独立成分分析を実現しようとする研究も行われたが、一方で多次元データの典型的な特徴を復元する生成モデル、特に画像分野の研究が近年盛んに行われるようになった³⁾。生成モデルでは、多次元データ入力とし、徐々に次元数を減らしながら低次元の潜在変数へと学習を行うことで、各データに共通する典型的な特徴部分が強く学習される。共通しない特異な特徴については学習が行われない。これは、深層学習による生成モデルが、低次元化された潜在変数にデータの特徴を絞り込んだ空間を成しデータ分類を実現していることを意味する。

生物の視覚情報処理も同様に、低次元化が行われている。網膜の受光細胞と比較して、目から脳に情報を伝達する視神経の経路数は著しく少ない。これは、網膜に入力された情報を圧縮することにより脳に送っているからである。我々はこれまで、生物の視覚情報処理を模した学習型画像認識システムの汎化能力について研究を行った⁴⁾。本報告では、同様に情報の低次元化による生成モデルの一種であるオートエンコーダが、その高めた汎化能力により異常検知する能力について検討したので報告する。生成モデルを異常検知に用いる方法では、正常信号のみの学習から、異常信号を検出することができるため、異常信号の教師データの学習が必要ではなく、そのことで想定外の異常信号なども検知できることが期待される。

2. 生成モデルによる異常検知の概要

2.1 生成モデルによる異常検知

現在、深層学習による生成モデルには、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) などが提案されている。深層学習による生成モデルでは、低次元の潜在変数から成る特徴空間の座標から、その座標値に対応した典型的なデータを復元することができる。異常検知では、検知に用いるデータを生成モデルに入力すると、入力データの持つ特徴に対応した空間座標に写像される。この座標値から典型的なデータが復元される。生成モデルによる異常検知は、入力データと復元されたデータを比較し、その差により良否判定される。深層学習による生成モデルでは、正常な部分のみが復元され、異常な部分は復元されない。また、正常なデータに欠損部分があれば、正常なデータとして復元され両データの差分により検出可能となる。

2.2 オートエンコーダ

図1にオートエンコーダのニューラルネットワーク構造を示す。オートエンコーダでは、入力データの低次元化を行うエンコーダ部とデータの生成を行うデコーダ部から成り、エンコーダに入力されたデータは、徐々にノード数を削減した中間層を経由し、最もノード数の少ない潜在変数に接続される。デコーダでは、潜在変数の

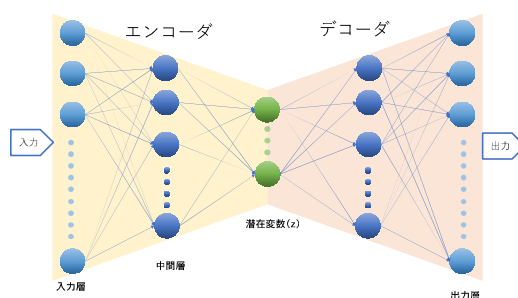


図1 オートエンコーダ

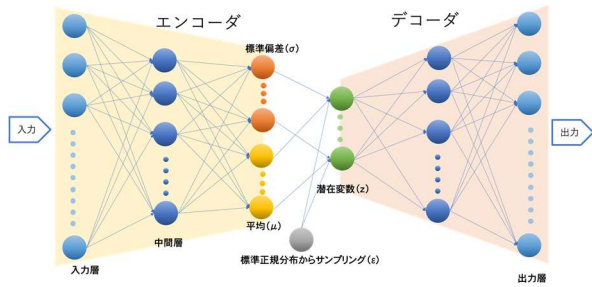


図2 変分オートエンコーダ

層から徐々にノード数を増やした中間層を経由し、最終的に入力データと同じ次元数の出力を行う。学習は入力データと出力データが一致するように、両データの差を損失関数として用いる。オートエンコーダの学習は、一見入力データと出力データを一致させるための学習であるため、無意味な学習であるように感じられるが、入力データの潜在変数への低次元化を行っているため、入力データはその特徴量に従って、潜在変数が成す空間に写像されたクラスタリングが行われる。

2.3 変分オートエンコーダ

一般的なニューラルネットワークでは、学習データに対して少しだけ異なるデータを入力された場合、学習データに対応した出力より少しだけ異なる出力が得られる保証はない。特に、大規模なネットワークが十分な学習が行われない場合や過学習に陥っている場合では、この傾向は顕著となる。深層学習による生成モデルでは、潜在変数の変化に対する汎化能力の高さが必要とされる。これは、潜在変数を変えながら、学習データと似通ったデータを生成するためである。

オートエンコーダの汎化能力を高め、生成モデルとして利用するための深層学習モデルに変分オートエンコーダ (Variational Auto Encoder: VAE) がある。図2にそのニューラルネットワーク構造を示す。変分オートエンコーダでは、汎化能力を高める手段として、潜在変数 z の計算に確率分布を用いている。エンコーダから平均 μ と標準偏差 σ が出力され、その変数から成る正規分布確率からサンプリングを行い、潜在変数 z を計算する。この方法により、学習時には潜在変数にノイズが加えられた状態となり潜在変数の微少の変化に対してもデコーダから同じ出力が得られ、汎化能力が結果的に高められる。しかし、正規分布から単純にサンプリングしたのではサンプリング部分で微分計算がとぎれ、誤差を逆伝搬させることができないため、**reparametrization trick** と呼ばれる手法が用いられる。次式のように標準正規分布からのサンプリングした値 ε を使って、潜在変数 z を計算し、誤差逆伝搬を可能としている。

$$z = \mu + \varepsilon \cdot \sigma \quad (1)$$

変分オートエンコーダの学習は、オートエンコーダと同じく入力データと出力データを一致させるため両データの差を損失関数とする。加えて、潜在変数部分の平均 μ と標準偏差 σ の確率分布についても拘束条件が付加されている。2つの離散確率分布の差異を評価する手法としてKLダイバージェンスがある。離散確率分布 P, Q が与えられた場合、KLダイバージェンス D_{KL} は

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (2)$$

により求められる。変分オートエンコーダでは、エンコーダからの出力平均 μ 、標準偏差 σ の成す潜在変数の確率分布 P と平均 0、標準偏差 1 の正規分布 Q のKLダイバージェンスが最小値となるよう学習を行う。すなわち、2つの確率分布関数を

$$P = N(\mu, \sigma) \quad (3)$$

$$Q = N(0, 1)$$

$\because N(\)$ 正規分布

とすると、損失関数を

$$\text{loss} = D_{KL}[N(\mu, \sigma) \parallel N(0, 1)] \quad (4)$$

定義し、正規分布に出力平均 μ 、標準偏差 σ 代入することにより(4)式を整理すると、KLダイバージェンスは最終的に次式により求めることが可能となる。

$$D_{KL}[N(\mu, \sigma) \parallel N(0, 1)] \approx -\frac{1}{2} \sum (1 - \log(\sigma^2) - \mu^2 - \sigma^2) \quad (5)$$

3. オートエンコーダの異常検知能力

3.1 学習用1次元データ

オートエンコーダの学習には MNIST: Modified National Institute of Standards and Technology⁵⁾ のデータベースによる手書き数字を用いた。手書き数字は 28×28 画素の2次元データから構成されている。今後、センサーなどの1次元時系列データへの適応を考慮し、2次元データから1次元データに展開し728点の学習データとした。フレームワークにはPyTorchを用いて実装を行った。

3.2 オートエンコーダによる低次元化

オートエンコーダは入力データをエンコードするモジュールとエンコードされたデータから入力データを復元するデコーダの2つのモジュールから構成し、次のように各モジュールを接続する単純な構成とした。

```
class AutoEncoder(nn.Module):
    def __init__(self, features):
        super().__init__()
        self.encoder = Encoder(features)
        self.decoder = Decoder(features)

    def forward(self, x):
        y=self.encoder(x)
        x=self.decoder(y)
        return x
```

始めに、各モジュールのニューラルネットワークを全結合型の2層構造とし、出力が0~1の値をとるようにシグモイド関数を用いて、潜在変数の低次元化の能力について検討を行った。6万種類の手書き数字を学習させ、学習後、未学習数字を入力した場合について、データの生成状態の確認を行った。

```
def forward(self, x):
    x = self.fc1(x)
    x = torch.sigmoid(x)
    return x
```

実験の結果、潜在変数のノード数が32~256個の場合については、デコーダから入力データと似たデータが生成されるが、16個以下では完全な生成が行われなかった。次に、クラスタリング能力について調査を行った。2層構

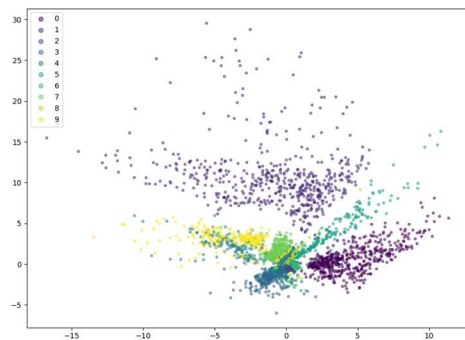


図3 2次元空間へのクラスタリング

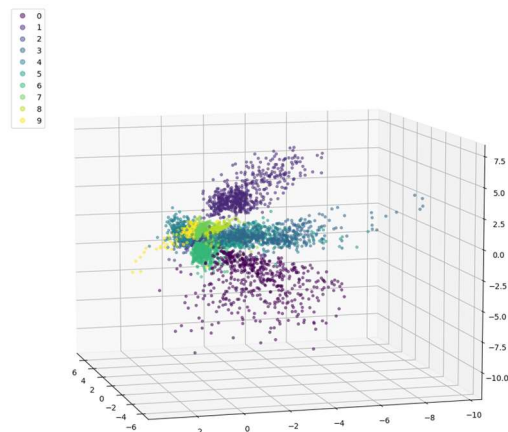


図4 3次元空間へのクラスタリング

造では、潜在変数のノード数が16個以下では能力不足のため、エンコーダ、デコーダとも4層構造とすることで、ニューラルネットワークのクラスタリング能力を高め、潜在変数の低次元化を図った。この場合も、6万種類の手書き数字を学習させ、学習後、未学習数字を入力しクラスタリングの状態を調べた。図3に潜在変数を2次元まで低次元化した場合のクラスタリングの状態を示す。中間層2層のノード数はそれぞれ256個と128個とした。教師なし学習にも関わらず、一部重なりはあるものの手書き数字毎にクラスタリングが行われている。クラスタリングの領域は文字毎に大きく異なる。図4は潜在変数を3次元まで低次元化した場合のクラスタリングの状態を表す。2次元の場合と同様なクラスタリングが行われている。この実験では、変分オートエンコーダとの比較も考慮し、エンコーダ出力段のsigmoid関数は未実装とした。

3.2 変分オートエンコーダによる低次元化

3.2.1 変分オートエンコーダの構成

変分オートエンコーダのエンコーダは、次の示す4層から構成し、出力層に平均と標準偏差をそれぞれ出力する構造とした。標準偏差が正の値のため、softplus関数を出力段に用いた。また、PyTorchのNormal関数により出力値の平均と標準偏差から正規分布の計算を行った。

```
def forward(self, x):
    x = self.fc1(x)
    x = self.relu(x)
    x = self.fc2(x)
    x = self.bn2(x)
    x = self.relu(x)
    mean = self.fc_mean(x)
    div = F.softplus(self.fc_div(x))
    q_z = Normal(mean, div)
    return q_z
```

デコーダについても同様に4層構造とした。

```
def forward(self, x):
    x = self.fc3(x)
    x = self.relu(x)
    x = self.fc2(x)
    x = self.bn2(x)
    x = self.relu(x)
    x = self.fc1(x)
    x = torch.sigmoid(x)
    return x
```

エンコーダからの正規分布からのサンプリングには、rsample関数により、逆伝搬可能とした。

```
class AutoEncoder(nn.Module):
    def __init__(self, features):
        super().__init__()
        self.encoder = Encoder(features)
        self.decoder = Decoder(features)

    def forward(self, x):
        q_z = self.encoder(x)
        x=self.decoder(q_z.rsample())
        return x,q_z
```

学習時は入力と出力の差分の値とKLダイバージェンスの

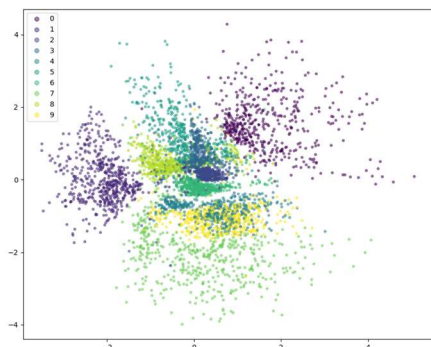


図5 2次元空間へのクラスタリング

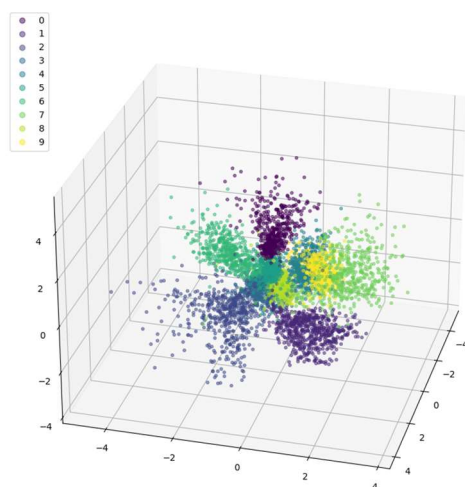


図6 3次元空間へのクラスタリング

値を足し合わせた損失関数の逆伝搬を行い重みの更新を行った。ここで、KL ダイバージェンスは、予め平均 0、標準偏差 1 の正規分布を計算し、エンコーダからの正規分布出力から `kl_divergence` 関数により求めた。

```
outputs, q_z = model(images)
loss1 = criterion(outputs, images).sum(dim=1).mean()
loss2 = torch.distributions.kl_divergence(q_z, p_z).sum(dim=1).mean()
loss = alpha * loss1 + loss2
loss.backward()
optimizer.step()
```

3.2.2 変分オートエンコーダのクラスタリング

図5は変分オートエンコーダによる2次元空間へのクラスタリング状態を表す。また、図6は3次元空間でのクラスタリングを表している。両図ともオートエンコーダと比較して、KL ダイバージェンスを用いたことにより、クラスタリング領域の広がりが大幅に抑制されている。変分オートエンコーダでは、学習時に、正規分布によるサンプリングを行い、学習データの潜在変数周辺の領域についても学習が行われているため、クラスタリング領域の抑制は潜在変数の未学習の領域を減少させ、汎化能力を高めることが可能となる。

3.3 変分オートエンコーダによる異常検知

変分オートエンコーダの異常検知能力を調べるため、6万種類の手書き数字を学習させ、未学習数字を入力した



図7 入力用未学習データ

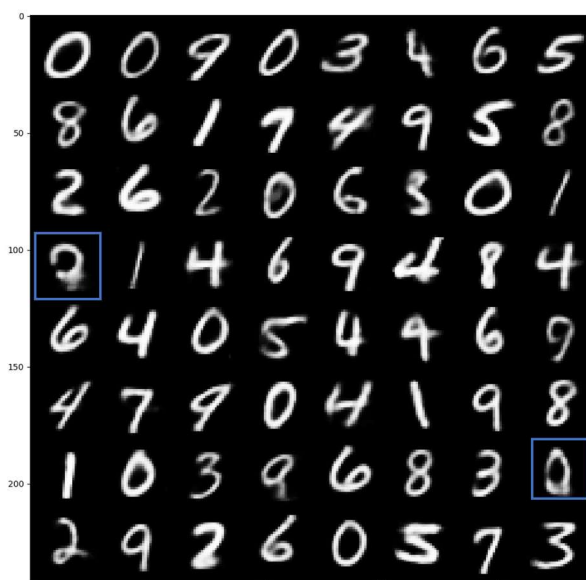


図8 変分オートエンコーダの生成データ

場合の生成データから異常検知を行った。異常検知用の学習では、64ノード数の潜在変数を用いて実験を行った。図7に用いた未学習の入力データを示す。図8は入力データに対する変分オートエンコーダから生成された1次元出力データを2次元画像に再構成したものである。両画像を比較すると、入力画像の赤丸部分が、変分オートエンコーダにより修正されている。入力データに特異な部分の信号があれば、生成データからは削除され、入力データに欠損部分があれば、データが補われ典型的な信号として復元されている。このことから、変分オートエンコーダを用いて、入力データと出力データの差分により、特異または欠損の異常信号の検知の可能性が示された。但し、変分オートエンコーダであっても出力データの青色の四角で囲んだデータなどのように典型的なデータに復元できない場合もあることが分かった。

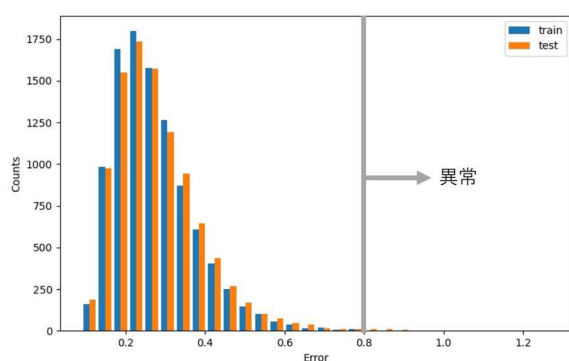


図9 学習データと未学習データの異常検知

図9に学習データと未学習データに対する、入出力間の差分の合計を示した。青の棒グラフは学習データ、オレンジ色の棒グラフは未学習データによる差分値を示す。未学習データは、学習データと比較して差分が大きくなる傾向がある。変分オートエンコーダを異常検知として用いる場合、差分に閾値を設定し、閾値以上を異常検知とすることも可能である。学習データの差分の最大値を閾値として、未学習データを入力した場合に異常検知された手書き文字を図10に示す。図10は入力データ(a)に対する生成データ(b)を示す。右の2組の数字は数字以外のノイズ信号により異常検知されたものと考えられる。また、左3組は学習されたクラスターでは、生成不可能な入力データであったことが想定される。

4. 結 言

情報の低次元化による生成モデルである変分オートエンコーダの高めた汎化能力による異常検知について検討を行った。生成モデルを異常検知に用いる方法では、正常信号のみの学習から、異常信号の検出能力を有し、異常信号の特異な部分や、信号の欠損部分検出の可能性を示した。また、異常信号が教師データの作る学習クラスターの範囲外の場合にも異常検知の可能性を示した。今後、異常信号の発生が極めて少なく必要数の異常信号を学習

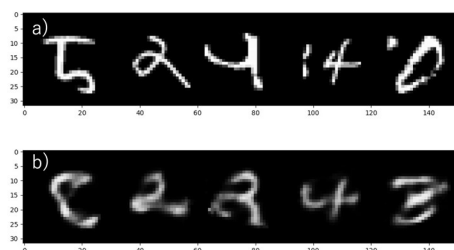


図10 閾値を使った異常検知

a) 入力データ, b) 生成データ

することが困難とされる実際の使用環境における異常検知に応用可能な有効な手段である。本報告で利用した手書き数字は、背景の値が0、文字の値が1の2値識別程度の簡単な検知能力で十分であったが、センサーなどの1次元時系列信号の異常検知への応用では、僅かなアナログ値の変動を検知する必要があり、実用化のためには、具体的な事例による検知能力向上の開発が必要であると思われる。

文 献

- 1) T Yamaguchi, K Hirokawa, K Itoh: Independent component analysis by transforming a scatter diagram of mixtures of signals. Opt. Comm. **173**(1-6), 107-114 (2000).
- 2) K. Hirokawa, T. Yamaguchi, K. Itoh: Independent component analysis based on scatter diagram. Proc. SPIE **3740**, 142-145 (1999) .
- 3) 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) .
- 4) K Hirokawa, K Itoh, Y Ichioka: Invariant pattern recognition by neural networks combined with optical wavelet preprocessor. Opt. Rev. **7**(4), 284-293 (2000).
- 5) THE MNIST DATABASE of handwritten digits : <http://yann.lecun.com/exdb/mnist/>