「材料」(Journal of the Society of Materials Science, Japan), Vol. 69, No. 9, pp. 650-656, Sep. 2020
解　　説

深層学習を用いた異常検知技術

野村泰稔*

A Review on Anomaly Detection Techniques Using Deep Learning

by

Yasutoshi NOMURA*

Key words: anomalies, outlier, deep learning

1　緒　　言

近年，深層学習に代表される機械学習が様々な分野において大きな発展を遂げており，異常検知分野への実用化・適用研究が活発に進められている．異常検知（Anomaly detection）は，主に品質管理（Quality Control）を目的に発展してきたと言われ，統計学において一世紀近い歴史を有する伝統分野である．一方で，システムや物理現象に対する異常検知が注目されてきたのは，インターネットが社会に浸透し始めた 1990 年以降であり，計算機，ネットワークおよびセンサ技術の発展とともに，現在，オンラインで対象を監視し異常検知を行う需要が高まってきている．そのような中で，機械学習が 2000 年前後から理論・応用面で目覚ましい進歩を遂げ，さらには近年の深層学習の登場以来，工場設備診断，医療診断，コンピュータネットワークへの侵入検知，人の異常行動・不正の検知あるいはノイズ文書の除去等，多種多様な領域での異常検知問題に深層学習が利用されてきている．

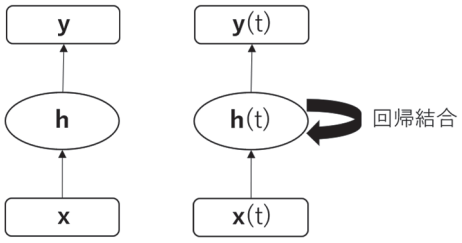
本解説では，深層学習の概要について簡単に触れ，深層学習を用いた異常検知の研究について体系的かつ包括的にレビューしている文献[1]での異常の概念を踏襲し，対象データとして信号・画像・テキストへの展開が可能であり，汎用的に使用できる代表的な教師なし型異常検知技術を紹介する．

2　深層学習の概要

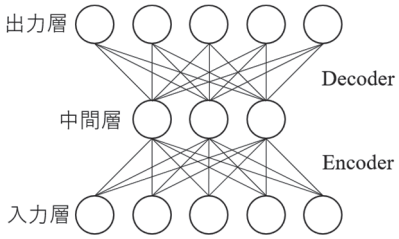
深層学習（Deep learning）は機械学習の一つの手法で，画像認識，音声認識，自然言語処理などの分野で大きな成果を上げている．学習を行うニューラルネットワークにおいて中間層が深く，層が何層にも積み重なっていることから「深層」学習と呼ばれる．ここでは，詳しい計算方法の解説は成書に譲るとして， 基本的な事項についてのみをまとめる．

深層学習のネットワーク構造の基本形態として，深層階層型ニューラルネットワーク（Deep hierarchical neural network），畳み込みニューラルネットワーク（Convolution neural network），再帰型（リカレント）ニューラルネットワーク（Recurrent neural network），自己符号化器（Autoencoder: AE）等がある．階層型ニューラルネットワークや畳み込みニューラルネットワークはデータが入力層から出力層に一方向に流れるため，フィードフォワード型ニューラルネットワーク（Feed-forward neural network）と呼ばれる．一方，リカレントニューラルネットワークは，ある層の出力が遡って入力される回帰結合を持つニューラルネットワークである．図 1 にニューラルネットワークの基本形態を示す．

フィードフォワード型ネットワークの代表例である畳み込みニューラルネットワークは，画像の認識・物体検出・セグメンテーションや画像生成などに用いられる．リカレントニューラルネットワークは，時間とともに次々と入ってくる，長さも決まらないデータを処理するのに適しており，時系列データ・シーケンスデータの分析や翻訳など自然言語処理の分野で盛んに研究されている．自己符号化器は，出力データが入力データをその



(a) フィードフォワード型ネットワーク（左）
(b) リカレントネットワーク（右）



(c) 自己符号化器（Autoencoder: AE）

図 1　ニューラルネットワークの基本構造

＋　原稿受理　令和2年　6月　30日　Received June 30, 2020

*　正 会 員　立命館大学理工学部 〒525-8577 草津市野路東，Dept. of Civil and Env. Eng., Ritsumeikan Univ., Noji-higashi, Kusatsu, 525-8577

まま再構成（復元）するニューラルネットワークである．入力層から中間層への変換器は Encoder（エンコーダ），中間層から出力層への変換器は Decoder（デコーダ）と呼ばれる．中間層のニューロン数を入力層のニューロン数（データの次元数）よりも小さくすることで，次元の縮約が図られる．後述するが，入出力の再構成誤差（復元誤差）を異常検知に利用することも多い．

深層学習は，大量のデータで学習することで高い成果を挙げることが知られている．与えられたデータ（情報）の中で，システム自体がデータの特徴を抽出し，自動的にデータから分類や予測ができる最適な特徴を選択する．これが従来の機械学習と異なる点である．従来の機械学習は，分類や予測に寄与するであろう特徴を人手で抽出してきたが，深層学習は人間の経験や勘による特徴抽出を行うことなく，自動的にこれらを行う．そのため，これまで把握できていない新たな特徴を見出す可能性があると言える（図 2）．

3　異常検知問題

3･1　異常の種類

異常検知の問題は，検知する対象により異常の概念が変わってくる．本解説では，文献 [1] の分類を踏襲し，異常の概念を，点異常，文脈依存型異常，集団型異常に分けて説明する．

点異常は，ある確率分布（多峰性分布も含む）から大きく外れたデータであり，代表的な応用例として外れ値検出問題がある．

文脈型依存型異常は，データの変化の傾向から予想できる値から外れたデータ等が該当する．代表的な応用例として，時系列上に現れる急激な変化を，時系列モデルを仮定して検出する変化点検知がある．

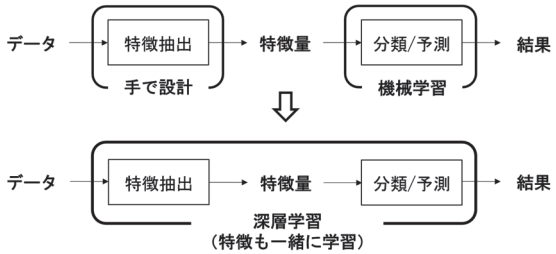


図 2　特徴抽出の変遷
（上：従来技術，下：深層学習）

集団型異常は，個々のデータポイントに注目すれば異常ではないが，ある一連のデータを集団として捉えた場合に，本来，同時に発生することがない状況下で，集中的に発生する状態などを言う．代表的な応用例として，異常行動検出などがある．

異常検知の問題は，時系列信号や画像（輝度値）などの数値データだけでなく，スパムメールの判定などにも定義できる．

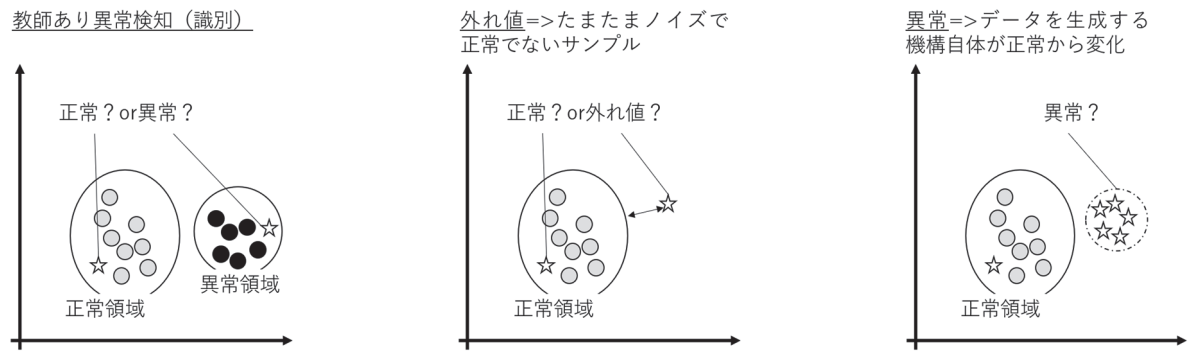
3･2　異常検知における学習形態

機械学習に基づく異常検知は，データの性質に応じて確率分布をどのように学習するかが重要であり，通常，教師あり型と教師なし型の 2 つのタイプに分けられる．

教師あり型は，正常と異常のサンプルがこれまでに十分に獲得されており，教師データとして利用できる場合においては有効である．「正常」や「異常」といったラベル付きデータを教師データとする識別問題として学習ができる（図 3(a)）．また，異常のサンプルが十分になくても，検知したい異常が入出力関係の不一致として定義できる場合，センサデータ等を用いた回帰問題として，従来からある時系列モデルや教師あり学習の各種方法を用いて異常検知ができる．具体的には，正常サンプルのみから，$y = f(\mathbf{x})$を推定し，新しい入力が得られると，その応答（出力）を予測できるシステムを構築する．そして，システムの予測結果と観測データとの逸脱の度合い（予測誤差）から異常を検知するものである．

一方，教師なし型の異常検知は，正常サンプルに比べて異常サンプルが極端に少ない場合，正常クラスの凝集性を学習し，外れ値検出問題に適用される（図 3(b)）．また，この外れ値の蓄積から，データを生成する機構自体がある状態（正常）から変化したと判断するような変化検知問題に展開できる図 3(c)）．なお，教師なし型異常検知は，ほとんどのサンプルを占める正常サンプルを用いることから，半教師あり異常検知と呼ぶ場合もある．

以降では，教師なし型異常検知とハイブリッド型異常検知に関して，具体的な方法論について解説していく．



(a) 教師あり型異常検知　(b) 外れ値検出　(c) 変化検出

図 3　外れ値検出と異常検知

4 教師なし型異常検知

4･1 自己符号化器（Autoencoder）を用いた技術

教師なし型異常検知技術の中核となるのが，前記した次元削減器の一種である自己符号化器（Autoencoder: AE）である．AE は入力と出力が一致するように学習するもので，入力空間を低次の空間に射影する Encoder と，低次の空間から元の空間に戻す Decoder から構成される．例えば，図 1(c)に示したような入力・中間・出力層がそれぞれ 1 層でかつ全結合型ネットワークを内部アーキテクチャとした場合，以下のように表現できる．

$$\begin{aligned}&\text{encoder: } \mathbf{y} = \boldsymbol{f}(\mathbf{W}\mathbf{x} + \mathbf{b})\\&\text{decoder: } \hat{\mathbf{x}} = \hat{\boldsymbol{f}}(\hat{\mathbf{W}}\mathbf{y} + \hat{\mathbf{b}})\\&\text{Autocoder: } \hat{\mathbf{x}} = \hat{\boldsymbol{f}}(\hat{\mathbf{W}}(\boldsymbol{f}(\mathbf{W}\mathbf{x} + \mathbf{b})) + \hat{\mathbf{b}})\end{aligned} \quad (1)$$

ここで，$\mathbf{W}, \mathbf{b}, \boldsymbol{f}(\,)$はそれぞれ入力層から中間層への結合重み，バイアス，活性化関数であり，これらが Encoder となる．一方，$\hat{\mathbf{W}}, \hat{\mathbf{b}}, \hat{\boldsymbol{f}}(\,)$はそれぞれ中間層から出力層への結合重み，バイアス，活性化関数であり，これらが Decoder となる．$\mathbf{y}$ は Encoder により圧縮された情報であり，潜在変数などと言う．AE の学習では，これらのパラメータ$\mathbf{W}, \hat{\mathbf{W}}, \mathbf{b}, \hat{\mathbf{b}}$を決定する．まず，学習データ$\mathbf{x}$に対して，$\hat{\mathbf{x}}$を求める．以下の誤差関数を最小化するように，バックプロパゲーションにより繰り返して更新して最適なパラメータを学習する．誤差関数には，入力が連続値であれば二乗誤差関数が，離散値であれば交差エントロピー誤差がよく用いられる．

二乗誤差：

$$\mathrm{E} = \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2$$

交差エントロピー誤差：

$$\mathrm{E} = -\sum_{n=1}^{N}\sum_{m=1}^{M} \left(x_{n,m} log \hat{x}_{n,m}\right) \quad (2)$$

ここで，$x_{n,m}$，$\hat{x}_{n,m}$は$\mathbf{x}_n$，$\hat{\mathbf{x}}_n$の m 番目の要素である．

AE の中間層を入力次元より少なくしておくことで，入力空間を低次元空間で表現できるようになる．大多数の正常サンプルに対して AE を学習するということは，中間層に正常サンプルの表現空間をつくることに相当する．このようにして作成された AE に対して，異常なサンプルが入力されると，Decoder が入力ベクトルを復元できず，入出力のベクトル（あるいは行列）に大きな誤差が生じる．この再構成誤差から異常を検知することができる（図 4）．内部の Encoder と Decoder のアーキテクチャは入力データが画像の場合，畳み込みニューラルネットワークが利用されることが多く，時系列データなどの系列データの場合，全結合型ニューラルネットワークや回帰型ニューラルネットワークの一種である長短期記憶アルゴリズム [2]（Long short -term memory:LSTM）が利用される．

4･2 敵対的生成ネットワーク（Generative adversarial Network）を用いた技術

生成モデルの代表的な方法である敵対的生成ネットワーク [3]（Generative adversarial network: GAN）と変分自己符号化器 [4]（Variational autoencoder: VAE）を基本とした方法が教師なし異常検知問題に有効であるとして利用されている．生成モデルとは，観測されたデータが何らかの確率モデルから生成されていると仮定し，その生成過程を確率分布によってモデル化し，そのモデルから全く新しいデータ（画像・信号・テキストなど）を作ることを目的としたものである．

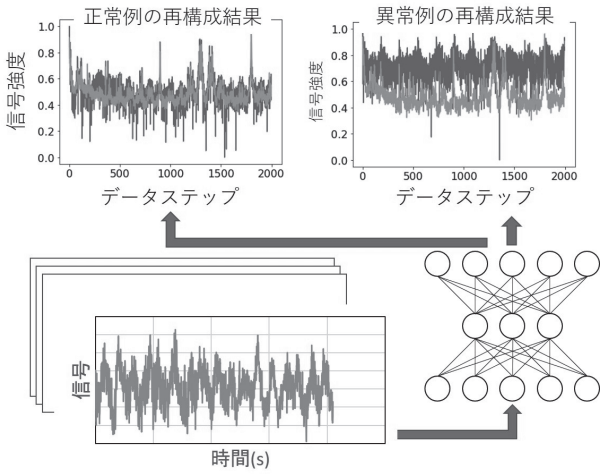


図 4 AE を用いた再構成誤差による異常検知例

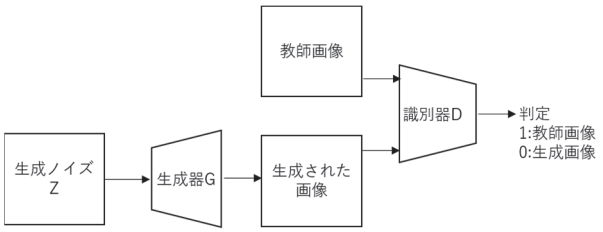


図 5 GAN の基本構造（画像生成の例）

GAN は二種類のニューラルネットワークから構成され，一方は生成器（Generator）と呼ばれ，ランダムノイズを入力としてデータを生成する役割を果たし，もう一方は識別器（Discriminator）と呼ばれ，生成されたデータが，事前に与えた教師データであるのか，あるいは生成器 G から出力されたデータであるのか識別する役割を果たす（図 5）．生成器 G は，識別器 D が教師データであると誤認識するようなデータを生成するように学習を進め，識別器 D は，生成器 G から出力されるデータを誤認識することなく正しく生成器 G からの出力であると認識するように学習を進めていく．

生成ノイズを $\mathbf{z}_i$，生成器 G から生成されるデータを $G(\mathbf{z}_i)$，識別器 D の出力を $y_i = D(G(\mathbf{z}_i))$とし，y_i がニューラルネットワークの活性化処理により 0 から 1 に変換されているものとする．そして，ラベル l に関して，生成器 G が生成したデータのラベルを l=0，教師データのラベルを l=1 とすると，識別器 D の出力が正答かどうかは $y^l(1-y)^{1-l}$で判定される．$y^l(1-y)^{1-l}$は識別器 D の出力が正解ラベルと一致すると 1 となり，誤答すると 0 となる．実際には，識別器 D の出力は，活性化関数（シグモイド関数など）を通じて 0 から 1 の値をとることから，

$y^l(1-y)^{1-l}$も 0 から 1 の値をとる．識別器 D を学習する際の損失関数は，この判定式にミニバッチ分のデータ数があるので，同時確率を考えると，以下のように表される．

$$\prod_{i=1}^{M} y_i^{l_i}(1-y_i)^{1-l_i} \tag{3}$$

これは識別器 D の尤度そのものであり，対数をとると，

$$\sum_{i=1}^{M}(l_i log y_i + (1-l_i) log\,(1-y_i)) \tag{4}$$

となる．識別器 D の損失関数は，この対数尤度にマイナスを掛けた式(5)であり，

$$-\sum_{i=1}^{M}(l_i log y_i + (1-l_i) log\,(1-y_i)), \tag{5}$$

識別器はこれを最小化するように学習される．（最大化問題を解くよりプログラムの実装が容易であるため，通常，ニューラルネットワーク全般の学習には損失関数の最小化が行われる）．

一方，生成器 G の学習時の損失関数は，識別器 D を騙すように学習しなければならないので，式(5)を最大化するようにしたい．ただし，ラベル l_i は生成器 G からの生成データなので常に 0 であることから，第 1 項は消え，さらに$(1\text{-}l_i) = 1$ であり，生成器 G の出力は $y_i = D(G(\mathbf{z}_i))$ となることから，生成器 G の損失関数として，式(5)が以下のように表現できる．

$$\sum_{i=1}^{M} log\,(1-D(G(\mathbf{z}_i)) \tag{6}$$

しかしながら，式(6)では上手く学習が進まないことが原著論文でも指摘されており，結局，$D(G(\mathbf{z}_i))$が 1 を出力してくれればよいと考え，DCGAN[5]など，GAN の内部アーキテクチャに畳み込みニューラルネットワークを用いる方法では，生成器 G の損失関数は，以下となり，これを最小化するように学習を進めている．

$$-\sum_{i=1}^{M} log(D(G(\mathbf{z}_i)) \tag{7}$$

このようなプロセスを応用し，画像データの異常検知問題に適用したものが AnoGAN[6]である．AnoGAN では，まず，教師データとして正常画像を与え，DCGAN などの GAN モデルを学習させる．次に，異常かどうかテストしたい画像（テスト画像）に対して，よく似た画像を生成できるノイズ（生成ノイズ）を求め，それを生成器 G に入力し画像を生成させる．このとき，生成された画像が入力されたテスト画像を的確に再構成していれば正常であると判断し，そうでなければ異常と判断するというのが AnoGAN の処理の流れである．

AnoGAN では，図 6 に示すようにテスト画像とよく似たデータを生成する生成ノイズを求める際に，テスト画像と生成画像のピクセルごとの絶対値誤差（residual loss と呼ばれる）と識別器 D の中間層（出力層の一つ手前の層）で得られるテスト画像と生成画像の特徴量の絶対値誤差（discrimination loss と呼ばれる）を合わせてバックプロパゲーションして，生成ノイズ $\mathbf{z}$ を調整する．具体的には，全結合層を更新学習することで生成ノイズ $\mathbf{z}$ が調整される．この全結合層を更新する際の損失関数 Loss は以下の通りである．

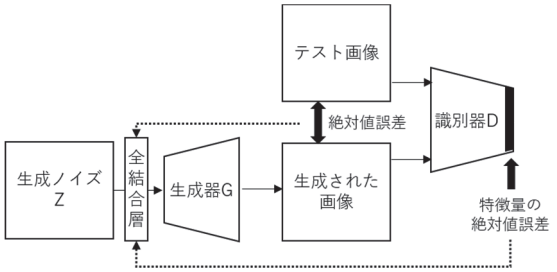


図 6　AnoGAN の学習プロセス

$$\begin{aligned}\text{Loss} = (1-\lambda)\cdot\text{residual loss} \\ +\lambda\cdot\text{discrimination loss}\end{aligned} \tag{8}$$

λはこれらの二種類の loss のバランスを調整する変数で，原著論文では 0.1 が使用されている．この損失関数を最小化して求められる最終的な生成ノイズ $\mathbf{z}^*$を生成器 G に入力し画像を作成する．そして，テスト画像に対する異常判定は，生成ノイズ $\mathbf{z}^*$を用いて得られる画像とテスト画像の Loss（式(8)）を計算し，これを異常度として閾値処理することになる．原著論文では医療画像に適用し，有効性が示されている．一方で，AnoGAN による異常検知は，事前に，DCGAN などの GAN モデルを構築しておき，さらに，テスト画像に対して正確に再現するような生成ノイズ $\mathbf{z}$ を，数 epoch 繰り返して更新学習する必要があり，異常検知に時間が掛かるという問題がある．そのため，現在，この方法を解決するための方法の一つとして Efficient GAN[7]という方法等が提案されている．この詳細は文献 [7]を参照されたい．

4･3　変分自己符号化器（Variational Autoencoder）を用いた技術

一方，もう一つの代表的な生成モデルに VAE がある．これは AE と同様に入力データと出力データが一致するように学習するものであるが，潜在変数の取り扱いが異なる．

VAE の Encoder は入力 $\mathbf{X}$ が得られると，その $\mathbf{X}$ を生み出した潜在変数 $\mathbf{z}$ を推論するが，Encoder はその潜在変数 $\mathbf{z}$ の変分事後分布$q_\phi(\mathbf{z}|\mathbf{X})$における平均ベクトル$\boldsymbol{\mu}_\mathbf{z}$と標準偏差ベクトル$\boldsymbol{\sigma}_\mathbf{z}$の二つの要素を出力する．そして，これらをパラメータとして使った正規分布により，潜在変数 $\mathbf{z}$ が確率的にサンプリングされ，その潜在変数 $\mathbf{z}$ から Decoder は，$\mathbf{X}$ の条件付き確率$p_\theta(\mathbf{X}|\mathbf{z})$における平均ベクトル$\boldsymbol{\mu}_\mathbf{X}$と標準偏差ベクトル$\boldsymbol{\sigma}_\mathbf{X}$の二つの要素を出力し，元のデータ $\mathbf{X}$ を復元する．なお，これらの変分事後分布$q_\phi(\mathbf{z}|\mathbf{X})$と条件付き確率$p_\theta(\mathbf{X}|\mathbf{z})$は多変量正規分布としてモデル化される．

VAE では，潜在変数 $\mathbf{z}$ を確率分布としてモデル化することで，同じ入力を与えても，毎回異なる潜在変数 $\mathbf{z}$ が得られることになり，Decoder の出力を連続的に変化さ

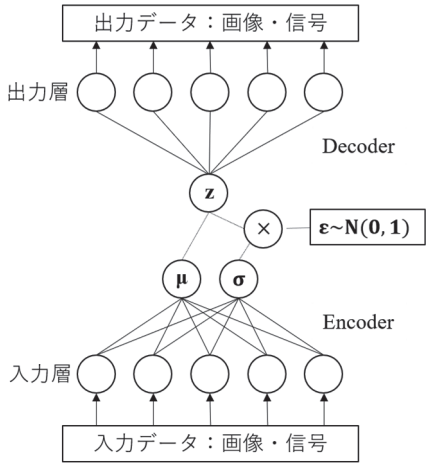


図 7　VAE の構造と Reparametrization Trick

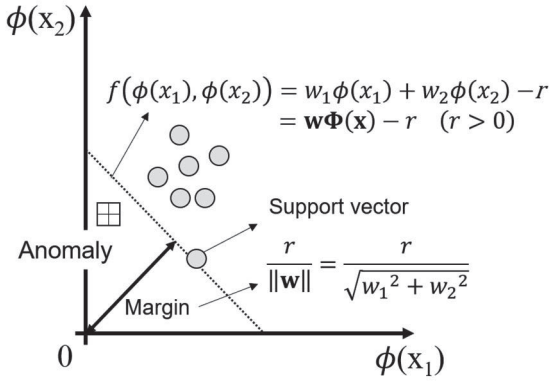


図 8　one-class SVM の概念図

せることが可能となる．また，通常**μ**，**σ**，**z** はベクトルあるいはバッチを考慮した場合は行列の構造となる．

なお，サンプリングは Reparameterization Trick という方法 [4]が使用され，図 7 のように，平均 0，分散 1 あるいは I の正規分布からサンプリングされる**ε**を用いて潜在変数を表現する．

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}\boldsymbol{\sigma} \tag{9}$$

これは Encoder や Decoder のアーキテクチャを学習するときのバックプロパゲーションを可能にするための工夫である．

そして，内部アーキテクチャを最適化するときの具体的な損失関数$L_{VAE}(\mathbf{X})$は以下の通りである．

$$L_{VAE}(\mathbf{X}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{X})||p(\mathbf{z})) - logp_\theta(\mathbf{X}|\boldsymbol{\mu}_\mathbf{z}) \tag{10}$$

第 1 項は正則化項を表しており，カルバックライブラーダイバージェンスにより，潜在変数 **z** の分布が標準偏差 1（あるいは I），平均 0 の正規分布 N(0,1)にどれだけ離れているかを評価している．第 2 項は Encoder で推論した潜在変数 **z** の分布を得た状態での，Decoder の出力の対数尤度であり，入力と VAE の出力の再構成誤差を表す．結局，損失関数は以下のように分解できる．

$$L_{VAE}(\mathbf{X}) = D_{VAE}(\mathbf{X}) + A_{VAE}(\boldsymbol{X}) + M_{VAE}(\mathbf{X}) \tag{11}$$

ただし，$D_{VAE}(\mathbf{X})$，$A_{VAE}(\mathbf{X})$，$M_{VAE}(\mathbf{X})$ はそれぞれ，

$$\begin{aligned} D_{VAE}(\mathbf{X}) &= \textstyle\sum_{j=1}^{N_\mathbf{z}} \frac{1}{2}\left(-log\sigma_{z_j}^2 - 1 + \sigma_{z_j}^2 + \mu_{z_j}^2\right), \\ A_{VAE}(\mathbf{X}) &= \textstyle\sum_{i=1}^{N_\mathbf{x}} \frac{1}{2} log 2\pi\sigma_{x_i}^2\Big|_{z=\mu_z}, \\ M_{VAE}(\mathbf{X}) &= \textstyle\sum_{i=1}^{N_x} \frac{1}{2}\frac{\left(\mu_{x_i} - x_i\right)^2}{\sigma_{x_i}^2}\Big|_{z=\mu_z}, \end{aligned} \tag{12}$$

である．*i*，*j* はそれぞれ入力データ **X** と潜在変数 **z** の要素の番号を表し，$N_\mathbf{x}$, $N_\mathbf{z}$ は入力データ **X** と潜在変数 **z** の要素数（次元数）である．ここで，入力データの要素数とは，入力データが画像の場合，その縦ピクセル数×横ピクセル数×（カラー:3，白黒:1）を意味する．$D_{VAE}(\mathbf{X})$は正則化項を表しており，$A_{VAE}(\mathbf{X})$+$M_{VAE}(\mathbf{X})$は再構成誤差を表している．なお，この $A_{VAE}(\mathbf{X})$+$M_{VAE}(\mathbf{X})$の式は，Decoder における条件付き確率$p_\theta(\mathbf{X}|\mathbf{z})$が多変量正規分布に従うと仮定した場合に得られるものである．$A_{VAE}(\mathbf{X})$は多変量正規分布の確率密度関数の積分を 1 にするものであり，$M_{VAE}(\mathbf{X})$はマハラノビス距離やユークリッド距離を正規化したものに近いことが分かる．なお，この条件付確率$p_\theta(\mathrm{X}|\mathrm{z})$が多変量ベルヌーイ分布に従うと仮定した場合は，式(2)で示した交差エントロピー誤差が評価されることになる．

以上が画像や信号を生成する際の設定であるが，この損失関数$L_{VAE}(\mathbf{X})$を異常度として定義し，教師なし型異常検知を行う研究が報告されている [8]．さらに，近年，異常度として $M_{VAE}(\mathbf{X})$のみを用いる方法が提案されている [9]．この文献 [9]では，この正則化項を取り除いた非正則化異常度 $M_{VAE}(\mathbf{X})$を評価することで，正常状態・異常状態の出現頻度に依存せず，複雑な構造を有する工業製品の画像データによる教師なし型異常検知が高精度に行えることが報告されている．

4･4　ハイブリッド型技術

教師なし型異常検知には，ハイブリッド型のものもある．例えば，事前に訓練された転移学習モデルや AE を特徴抽出器として利用し，集約された情報を one-class SVM などの従来の機械学習アルゴリズムに入力し，外れ値検出を行う方法も広く利用されている．ただし，転移学習モデルや AE で獲得される特徴ベクトルは異常検知を的確に行えるように表現されているわけではなく，つまり，異常検知用に特化されていないことから，外れ値の検出に失敗してしまうことがある．この問題を克服するため，近年，one-class SVM での考え方を踏襲した one-class ニューラルネットワーク [10]（one-class NN）が提案されている．

one-class NN を説明する前に，one-class SVM に簡単に触れると，この方法は教師なし学習に属し，正常データのみを用いて異常検知することができる．図 8 に one-class SVM の概念図を示す．この方法は，原点と正常データのマージンを最大化するように境界を決定する．SVM と同様に，カーネルトリックを使用して非線形分離を行うことやソフトマージンを採用してマージンの内側にデータが入り込むことを許容するなどの工夫が

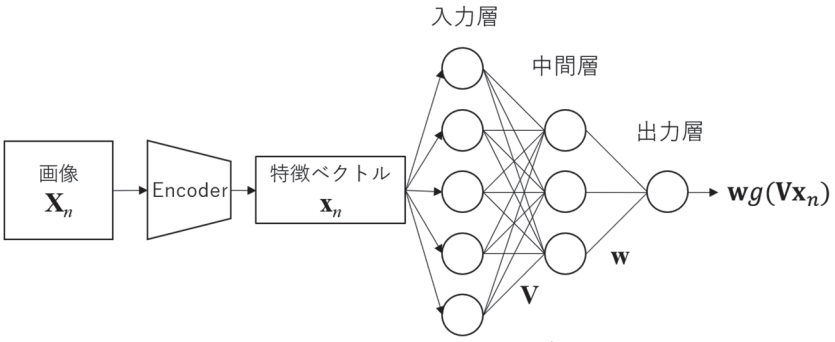


図 9　one-class NN の構造

可能である．マージンを最大化するのは，結局は，$\|\mathbf{w}\|^2/2-r$を最小化することと同値であるので最小化問題に置き換える．さらに，one-class 分類に失敗することもある程度許容するために，以下の項を付け加える．

$$\frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r-\mathbf{w}\boldsymbol{\Phi}(\mathbf{x}_n)\right) \tag{13}$$

∑の内部は正常な点が境界より原点側に入り込む場合にのみペナルティを付与するものである．νは分類の失敗を許容するハイパーパラメータであり，$0<\nu\leq 1$ の範囲で設定される．詳細は割愛するが，この最適化問題は不等式制約付きの最小化問題となるので，ラグランジュ未定定数を導入し．双対問題へ変換してから解く．One-class SVM 等，SVM 全般の解法は文献 [11]が詳しい．

one-class NN は one-class SVM におけるカーネルトリックを使用して高次元空間に写像する部分を全結合型ニューラルネットワークの一層目の重み行列に置き換え，他の層の重みも含めバックプロパゲーションにより最適化する．そして，これと交互に，原点と教師データ（正常データ）のマージンを最大化するように境界を決定することで，異常検知に適した特徴表現を獲得できるとしている．図 9 にネットワークの構造を示す．図中の $\mathbf{V}, \mathbf{w}$ はそれぞれ，入力層から中間層の重み行列，中間層から出力層の重みベクトルである．$g(\ )$は中間層の活性化関数であり，原著論文 [10]ではシグモイド関数が使用されている．one-class NN の評価関数をまとめると以下の通りである．

$$\begin{aligned}&\mathrm{arg}min_{\mathbf{w},\mathbf{V},r}\frac{1}{2}\|\mathbf{w}\|_2^2+\frac{1}{2}\|\mathbf{V}\|_F^2\\&+\frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r-\mathbf{w}\mathbf{g}(\mathbf{V}\mathbf{x}_n)\right)-r\end{aligned} \tag{14}$$

ここで，r は正常か異常であるかを判定する際の境界に対するバイアスであり，これも最適化対象となる．

この評価関数を最小化する際，$\mathbf{w}, \mathbf{V}, r$ が初期化されている状態から，まずは，r は初期値のまま固定した状態で，以下の評価関数を用いてニューラルネットワークの重み $\mathbf{w}$, $\mathbf{V}$ をバックプロパゲーションにより更新する．

$$\begin{aligned}&\mathrm{arg}min_{\mathbf{w},\mathbf{V}}\frac{1}{2}\|\mathbf{w}\|_2^2+\frac{1}{2}\|\mathbf{V}\|_F^2\\&+\frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r-\mathbf{w}\mathbf{g}(\mathbf{V}\mathbf{x}_n)\right)\end{aligned} \tag{15}$$

次に，更新された $\mathbf{w}, \mathbf{V}$ を用いて，以下の評価関数により r を更新する．

$$\mathrm{arg}min_r\frac{1}{\nu}\frac{1}{N}\sum_{n=1}^{N}\max\left(0, r-\mathbf{w}\mathbf{g}(\mathbf{V}\mathbf{x}_n)\right)-r \tag{16}$$

ただし，この解 r は単純に$\hat{y}_n=\mathbf{w}\mathbf{g}(\mathbf{V}\mathbf{x}_n), \{n=1,\ldots,N\}$の内，$N\cdot\nu$番目に小さい値が更新された r となる．この証明は原著論文 [10]を参照されたい．なお，νは one-class SVM で登場した分類の失敗を許容するハイパーパラメータである．そして，更新された r を式(15)に代入し，順次 $\mathbf{w}, \mathbf{V}$ を更新していく．以上の手順を収束するまで繰り返すことで，$\mathbf{w}, \mathbf{V}, r$ が獲得される．これらのパラメータが獲得された後，テストデータに対して，以下のようにスコアを計算し，$S_n \geq 0$ ならば正常と判断し，$S_n < 0$ ならば異常と判断する．

$$Decision\ score: S_n=\hat{y}_n-r \tag{17}$$

なお，one-class NN の原著論文 [10]では，事前に Convolutional AE（畳み込み型 AE : CAE）を用いて，データセットの再構成学習をしており，Encoder で特徴ベクトル（中間層で得られる情報）を得ておく必要がある．この再構成学習にはデータの構造に応じて AE や LSTM などの利用が考えられる．原著論文 [10]では，本手法の工業・産業分野への応用はされていないものの，いくつかのベンチマーク問題において，有効性が示されていることから，注目すべき方法論として考えられる．

5　結　　言

本解説では，近年活発に研究が進められている深層学習を用いた異常検知問題について，特に代表的な教師なし型異常検知技術について紹介した．深層学習の各手法を異常検知問題に適用しようとすると，世界中で新しい方法論が活発に開発されていることから，どの方法論が最も適切かを事前に把握することは難しい．一方で，現在，Python 言語で記述できる深層学習用の無償のフレームワーク Tensorflow や Pytorch で実装された異常検知技術がインターネット上に数多くオープンソースで公開されている．個人/商用利用問わず，誰でも簡単に利用することができることから，問題に応じて適切な方法論を試行錯誤的に調査することが重要であろう．

また，本解説では，異常検知問題における従来の機械学習技術については触れなかったが，文献 [1]でも述べら

れているように，深層学習は従来の機械学習技術と比較して，データの規模が大きくなるにつれて性能が向上する傾向がある．従来の機械学習技術を用いた異常検知に関しては，文献[12),13)]に詳細かつ丁寧にまとめられているので，そちらを参照されたい．

参考文献

1) Chalapathy, R. and Chawla, S.: Deep Learning for Anomaly Detection: A Survey, airXiv preprint airXiv:1901.03407v2, (2019).
2) Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, Neural Computation, Vol.9, No.8, pp.1735-1780, (1997).
3) Goodfellow, I. et al.: Generative Adversarial Nets, Electronic proc. of Neural information processing systems conference (NIPS2014), Advances in Neural information processing systems 27, (2014).
4) Kingma, D.P. and Welling, M.: Auto-Encoding Variational Bayes, airXiv preprint airXiv:1312.6114v10, (2014).
5) Radford, A., Metz, L., and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, arXiv preprint arXiv:1511.06434, (2015).
6) Schlegl, T. et al.: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, Proc. of International conference on information processing in medical imaging, pp.146-157, (2017).
7) Zenati, H. et al.: Efficient Gan-based Anomaly Detection, arXiv preprint arXiv:1802.06222v2, (2019).
8) Ribeiro, M., Lazzaretti, A.E., and Lopes, H.S.: A study of Deep Convolutional Autoencoders for Anomaly Detection in Videos, Pattern Recognition Letters, Vol. 105, pp.13-22, (2017).
9) Tachibana, R., Matsubara, T. and Uehara, K.: Anomaly Manufacturing Product Detection using Unregularized Anomaly Score on Deep Generative Models, Proc. of the 32th annual conference of the Japanese society for Artificial Intelligence, 2A1-03, (2018).
10) Chalapathy, R., Menon, A.K. and Chawla, S.: Anomaly Detection using One-class Neural Networks, arXiv preprint arXiv:1802.06360v2, (2019).
11) 平井有三：はじめてのパターン認識，森北出版，(2015).
12) 井手剛 : 入門　機械学習による異常検知，コロナ社，(2015).
13) 井手剛，杉山将：異常検知と変化検知，講談社，(2017).