

[原著論文]

## Fashion-MNISTのベンチマーク適性の解析

近藤 正紀

キーワード：Fashion-MNIST, MNISTデータ, 機械学習, ニューラルネットワーク,  
ベンチマーク

### Analysis of Fashion-MNIST benchmark suitability

Masanori Kondo

#### Abstract

In the research of neural networks, one data set has played the benchmark role for 20 years, and it is currently used worldwide for such research. In recent years, a second data set has been published with the aim to create a second benchmark and contribute to artificial intelligence. However, the differences between the two data sets and difficulty in learning have not been established.

In this paper, we examine the differences between the two data sets and explore the possibilities of constructing a third data set in the future.

Keywords：Fashion-MNIST, MNIST, machine learning, neural networks, benchmark

#### 要旨

ニューラルネットワークの研究には20年に渡ってベンチマークの役割を担ってきたデータセットが存在し、世界中でニューラルネットワークの研究に使用されている。近年、第二のベンチマークを目指したデータセットが公開され人工知能研究に寄与しているが、両者の違いや学習の困難度等に関する評価が定まっていない。

本稿では第二のデータセットと第一のデータセットの違いを検証し、将来の第三のデータセット構築のための手掛かりを模索する。

#### I 序論

ディープラーニングの基本となるニューラルネットワークにおいては、1998年にYann LeCunらが公開した『MNISTデータ<sup>1)</sup>』と呼ばれる手書き文字画像データセットが画像認識問題のベンチマークとして利用されている。また、LeCunらのウェブサイトではMNISTデータの公開とともに各国の研究者の成果も公開しており、2018年末時点の最高記録は2012年にCireşanらがディープラーニングを用いて達成した、正答率99.77%である。しかしながら、近年MNISTデータに対して「簡単すぎる」という声が聞かれるようになった。

手書き文字データには、MNISTデータ以外にも日本

---

新潟医療福祉大学 医療経営管理学部 医療情報管理学科

[責任著者および連絡先] 近藤 正紀  
新潟医療福祉大学 医療経営管理学部 医療情報管理学科  
〒950-3198 新潟市北区島見町1398番地  
E-mail: masanori-kondo@nuhw.ac.jp

投稿受付日：2019年3月4日

掲載許可日：2019年8月9日

の産業技術総合研究所が作成したETL-8<sup>2)</sup>やNIST Special Database 19<sup>3)</sup>などのデータが存在し、多くの研究者が利用している。また、機械学習用のデータとしてKrizhevskyが公開しているCIFAR-10とCIFAR-100<sup>4)</sup>、National Institutes of Health Chest X-Ray Dataset<sup>5)</sup>など、近年の研究の広まりと共に充実してきているが、ベンチマークとしての地位を確立しているとは言い難い。

このような背景から、新たなベンチマークの候補として開発されたのが、本稿で扱ったFashion-MNIST<sup>6)</sup>という名称のデータである。本稿ではニューラルネットワークの挙動という観点でオリジナルのMNISTデータ同様にFashion-MNISTデータがベンチマークとして機能することを確認し、両データの特性に基づいて将来作られるであろう第三、第四のベンチマークとなるデータの設計指針を示すことを目的とする。本稿の新規性はこの設計指針を示すことにある。

## II オリジナルのMNISTデータとFashion-MNISTデータの概要

オリジナルのMNISTデータ（以下、original MNISTデータと記す）はLeCunらが、NIST（National Institute of Standards and Technology）が作成した米国勢調査局職員の文字からなるSpecial Database 3（SD-3）と高校生の文字からなるSpecial Database 1b（SD-1）を元に構築した、合計70,000枚の画像データである。データは訓練用とテスト用のサブセットに分けられており、訓練用セット60,000枚をSD-3およびSD-1それぞれから30,000枚抽出して構築、同様にテスト用セット10,000枚をそれぞれから5,000枚抽出して構築している。そして以下の優れた特徴を持つ。

- ・ 一つの文字画像はピクセルの8ビットグレースケール画像に統一されている。
- ・ 手書き文字の筆者は250名おり、訓練用セットと

テスト用セットに同じ筆者はいない。

- ・ 二つのデータセットは、ランダムに並べられており、数字や筆者による並びにはなっていない。
- ・ 各画像が表す数字が正しくラベル付けされている。

図1はoriginal MNISTデータの訓練用セットから、最初に現れる0～9の各数字を抽出し、可視化したものである。ただし、通常の8ビットグレースケール画像のピクセル値は最小値0が黒、最大値 $2^8 - 1 = 255$ が白に対応付けられるが、図1は0が白を表すようにグレースケールの反転処理を施している。ニューラルネットワークに与える場合はピクセル値を $0 \leq x \leq 1$ に変換する。

表1はデータセット中のラベルごとの画像の枚数である。ラベルごとに枚数のばらつきはあるが、訓練用とテスト用の構成比率はほぼ同じである。

MNISTデータは1990年代後半におけるコンピューターのハードウェアの能力を考慮すると必要にして十分なデータ量であり、扱いやすさもあって過去20年にわたってニューラルネットワークの標準データとして世界中で使用されてきたが、一方、ベンチマークとしては簡単すぎるのではないかとの疑問も生じてきた。その背景には過去20年間のコンピューター、とりわけハードウェア技術の進歩のため、演算性能が飛躍的に向上し、計算資源のコストが劇的に低下したことがある。例えば、現在一般に使用されているパーソナルコンピューターが搭載しているCPUの中にはその演算性能が当時のスーパーコンピューターを凌駕するものも存在することからも、それは窺える。

Fashion-MNISTデータは、ドイツのZalando Researchがoriginal MNISTデータの次のステップのベンチマークとして公開したものである。紹介文では

“Fashion-MNIST is intended to serve as a direct drop-in replacement of the original MNIST dataset for benchmarking machine learning algorithms.”

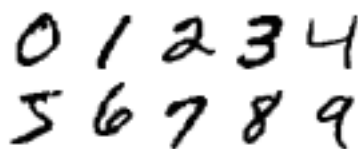












図1 original MNISTデータ内の画像例

表1 original MNISTデータの構成

ラベル	0	1	2	3	4	5	6	7	8	9
訓練用	5923	6742	5958	6131	5842	5421	5918	6265	5851	5949
テスト用	980	1135	1032	1010	982	892	958	1028	974	1009

表2 Fashion-MNISTデータのラベル付けと画像例

ラベル	0	1	2	3	4
名称	T-shirt/top	Trouser	Pullover	Dress	Coat
画像例					
ラベル	5	6	7	8	9
名称	Sandal	Shirt	Sneaker	Bag	Ankle boot
画像例					

と述べられており、また、“MNIST is too easy” “MNIST is overused” “MNIST can not represent modern CV task” という理由から、MNISTデータに対して直接的な差し込み式置き換え（drop-in replacement）を行うことを意図したとある。事実、以下の特徴から original MNISTデータで動作するニューラルネットワークに対して、コードに何ら変更を加えることなく、データファイルの入れ替えのみで動作させることが可能である。

- ・ original MNISTと同じファイル名、同じバイナリー構造である。
- ・ 訓練用60,000枚、テスト用10,000枚の画像からなる。
- ・ 一つの画像は、縦横28×28ピクセル、8ビットグレースケール画像である。
- ・ 各画像は10種類のラベル付けがなされており、ラベルは0～9の値を持っている。

また、以下の点がoriginal MNISTデータと異なる。

- ・ original MNISTデータが線画であるのに対し、Fashion-MNISTデータの各画像はTシャツやスニーカーなど衣類を表しており、ラベルとの対応付けは表2のようにになっている。表中の画像は訓練用データの一部を可視化したもの、図1と異なりグレースケールの反転を行っていない。
- ・ original MNISTデータがラベルごとの画像数が異なるのに対し、Fashion-MNISTデータでは訓練用6,000ずつ合計60,000枚、テスト用1,000ずつ合計10,000枚になっている。

### III 方法

本研究ではoriginal MNISTデータとFashion-MNISTデータの比較を、データそのものの解析とニューラルネットワークを用いた学習実験結果の比較の二つを以て行う。

#### 1 画像の複雑さの解析

本研究では画像の持つ複雑さの指標としてKolmogorov complexityを用いる。Kolmogorov complexityは情報量に基づくバイト列のランダム性の指標であり、『バイト列  $s$  のKolmogorov complexity  $K(s)$  は万能計算機で  $s$  を出力することができる最小サイズのプログラムの長さ』と定義できる<sup>7)</sup>。しかし、任意のバイト列  $s$  から  $K(s)$  を出力する計算可能なプログラムは存在せず、一般に  $K(s)$  は計算不能である。そこで、CilibiasiとVitanyiは  $K(s)$  の概算手法として圧縮を用いることを提案した<sup>8)</sup>。本稿では学習実験で使用するニューラルネットワークが2次元の画像を1次元のバイト列として扱うことに合わせ、Python 3.6のzlib.compress( ) メソッドで圧縮レベルを9として1次元バイト列を可逆圧縮し、圧縮後のバイト数を元画像のバイト数（784バイト）で除して概算した  $K(s)$  を算出する。以下、この概算値を  $\hat{K}(s)$  と表す。バイト列  $s_1$  よりもバイト列  $s_2$  が複雑であれば圧縮後の長さは  $s_2$  の方が長いため、 $\hat{K}(s_1) < \hat{K}(s_2)$  が成り立つ。また、定義から  $\hat{K}(s)$  は  $0 < \hat{K}(s) \leq 1$  を満たす。

#### 2 ニューラルネットワークの数理モデル

データの持つ情報量とニューラルネットワークが見せる挙動との間の関係を測定するために、普遍性定理<sup>9)</sup>を根拠として、図2に示す人工ニューロンが図3のように“入力層 — 隠れ層（1層） — 出力層”の3層に接続されたニューラルネットワークを用いて画像を学習し、画像分類問題の正解率を隠れ層のニューロン数  $h$  の関数として得る。なお、図3のようなモデルでは1層目は単なる入力でニューロンではないとして「2層のニューラルネットワーク」とする文献も多数存在するが、本稿では「3層」と表記し、入力層のノードも「ニューロン」と記述する。

ディープラーニングで多用される畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）

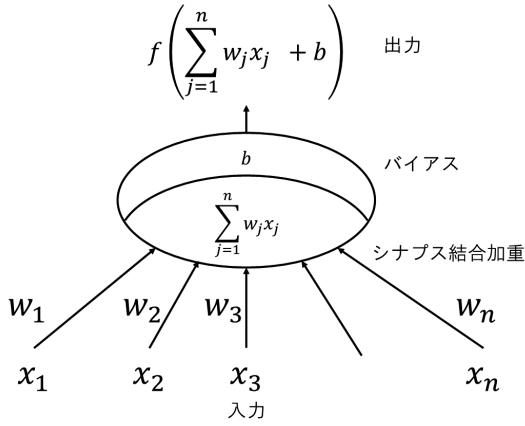


図2 人工ニューロンとパラメータ

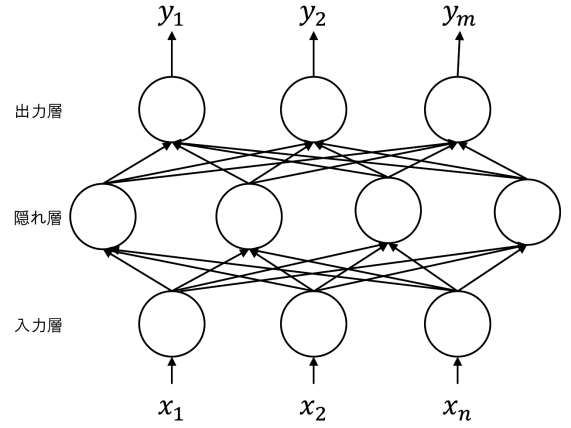


図3 3層ニューラルネットワーク

では入力画像の2次元構造を崩さずに畳み込み演算等を適用するが、図3のネットワークは画像を1次元のバイト列（ベクトル）として扱っている。このモデルはニューラルネットワークの基本形であり、研究され尽くした感も否めないが、データの変更がニューラルネットワークの挙動に与える影響を観察するためのベンチマークプログラムの意味で、本稿では基本形を選択する。

入力層からの入力ベクトルを  $X = [x_1 \ x_2 \ \dots \ x_{784}]$ 、入力層の  $i$  ( $1 \leq i \leq 784$ ) 番目のニューロンと隠れ層の  $j$  ( $1 \leq j \leq h$ ) 番目のニューロンのシナプス結合の重み係数を  $w^{(1)}_{ij}$  とするとき、隠れ層の  $j$  番目のニューロンの出力  $z^{(2)}_j$  を式 (1)、(2) で定義する。

$$a^{(2)}_j = \sum_{i=1}^{784} w^{(1)}_{ij} \times x_i + b^{(1)}_j \quad (1)$$

$$z^{(2)}_j = f(a^{(2)}_j) = \begin{cases} a^{(2)}_j & (\text{if } a^{(2)}_j \geq 0) \\ 0 & (\text{if } a^{(2)}_j < 0) \end{cases} \quad (2)$$

ここで  $b^{(1)}_j$  は  $j$  番目のニューロンのバイアス、 $f(\cdot)$  はReLU (Rectified Liner Unit) と呼ばれる活性化関数である。従って、隠れ層の出力ベクトル  $Z^{(2)} = [z^{(2)}_1 \ z^{(2)}_2 \ \dots \ z^{(2)}_h]$  は、 $784 \times h$  行列  $W^{(1)} = [w^{(1)}_{ij}]$ 、バイアスベクトル  $B^{(1)} = [b^{(1)}_1 \ b^{(1)}_2 \ \dots \ b^{(1)}_h]$  を用いて式 (3) で表すことができる。

$$Z^{(2)} = f(X \cdot W^{(1)} + B^{(1)}) \quad (3)$$

同様に、出力層の出力ベクトル  $Y = [y_1 \ y_2 \ \dots \ y_{10}]$  は式 (4)、(5) で定義する。

$$[a^{(3)}_k] = Z^{(2)} \cdot W^{(2)} + B^{(2)} \quad (4)$$

$$y_k = \sigma(a^{(3)}_k) \quad (1 \leq k \leq 10) \quad (5)$$

$W^{(2)}$  は  $h \times 10$  行列、 $\sigma(\cdot)$  はsoftmax function と呼ばれ、式 (6) で定義される。e はNapier数である。

$$\sigma(a^{(3)}_k) = \frac{e^{a^{(3)}_k}}{\sum_{j=1}^{10} e^{a^{(3)}_j}} \quad (6)$$

ただし、実装においては数値演算のオーバーフロー抑制のため、 $a^{(3)}_k$  ( $1 \leq k \leq 10$ ) を  $a'_k = a^{(3)}_k - \max_{1 \leq j \leq 10} \{a^{(3)}_j\}$  ( $1 \leq k \leq 10$ ) と変換する。

softmax functionの性質より出力ベクトル  $Y$  は確率ベクトルとなるため、 $y_k = \frac{\max_{1 \leq j \leq 10} \{y_j\}}$  なる  $k$  を入力画像からニューラルネットワークが予想した解答とする。この解答と正答を比較し、正答との乖離状況を式 (7) の交差エントロピー誤差を用いた損失関数で評価し、損失関数の値が小さくなるようにAmari<sup>9)</sup>が発見した誤差逆拡散法を用いて  $W^{(1)}$ 、 $W^{(2)}$ 、 $B^{(1)}$ 、 $B^{(2)}$  を修正して学習を繰り返すことで正解率の向上を図る。

$$E = - \sum \hat{y} \log y \quad (7)$$

$\hat{y}$  は正解ラベルの値を要素番号とする要素が1、他が0のone-hot表現ベクトル、 $y$  は予測の確率ベクトルである。

なお、本研究ではニューラルネットワークの過学習を抑制する手法であるドロップアウトや荷重減衰などの手法は用いない。また、誤差逆拡散法で用いる学習率も  $\eta = 0.1$  に固定する。なお、誤差逆拡散法には活性化関数が可微分であるという仮定が存在するため、一般に使用されるとおり活性化関数  $f(\cdot)$  の1階導関数を式 (8) に従って実装する。

$$\frac{d}{dx}f(x) = \begin{cases} 1 & (\text{if } x \geq 0) \\ 0 & (\text{if } x < 0) \end{cases} \quad (8)$$

### 3 学習実験の方法

本研究におけるニューラルネットワークの学習は以下に従う。

#### 1) 訓練用セットの分割

学習データの組み合わせによる偏りを抑止するため、訓練用セットを新に訓練用データと検証用データに分割する。その比率はGuyon<sup>10)</sup>によって導かれ、一般的な教科書では5,000枚から12,000枚程度（およそ10%から20%）を検証用データに当てている。本稿では、訓練用データを減らしすぎると学習に支障が出ることと、original MNISTデータを用いたSimard<sup>11)</sup>らの研究が検証用データを10,000枚としたこと、およびテスト用セットが10,000枚であることから、60,000枚の訓練用セットを新たに訓練用データ50,000枚と検証用データ10,000枚に分割する（以下、訓練用データという語は抽出した50,000枚を、訓練用セットという語は抽出前の60,000枚のデータを指す）。この分割は実験を繰り返すごとに無作為にやり直す。この結果、訓練用データの内容は実験ごとに異なることになる。学習の最終評価はニューラルネットワークの汎化性能を見るために、10,000枚のテスト用セットを用いて行う。

#### 2) 成績の決定方法

1回の実験を200エポックとする。ここでエポックとは、訓練用データを学習で全て使い切ることを指す。また、1エポックごとに訓練用データを無作為に並べ替える。この200エポックからなる実験を $h$ を変えながら各数十回繰り返し、各回の誤答率の最小値をその回の成績とする。

#### 3) 学習形態

ミニバッチサイズを100としたミニバッチ学習を採用する。本稿ではランダムに並べ替えられた訓練用データ50,000枚の画像列の先頭から100枚ずつを取り出して1回のミニバッチ学習とするため、500回のミニバッチ学習で1エポックとなる。

#### 4) パラメタの更新

各ミニバッチ学習の終了後、得られた100個の予測ラベルと正解の交差エントロピー誤差を算出し、この結果を基に誤差逆伝播法を用いて加重( $W^{(1)}$ 及び $W^{(2)}$ )とバイアス( $B^{(1)}$ 及び $B^{(2)}$ )を更新する。ミニバッチ学習における交差エントロピー誤差は、正解ラベルのone-hot表現ベクトル群 $\hat{y}_k$  ( $1 \leq k \leq 100$ )と予測の確率ベクトル群 $y_k$  ( $1 \leq k \leq 100$ )を用いて次式で表すことができる。

$$E = - \sum_{k=1}^{100} \sum \hat{y}_k \log y_k$$

なお、 $W^{(1)}$ 、 $W^{(2)}$ は一般に「Heの初期値<sup>12)</sup>」と呼ばれるもので、 $B^{(1)}$ 、 $B^{(2)}$ は0で初期化する。

## IV 結果

### 1 データの特徴分析

original MNISTデータとFashion-MNISTデータの訓練用およびテスト用全140,000枚について、 $\hat{K}(s)$ をまとめたものが表3、表4である。また、図4(a)はoriginal MNISTデータ、図4(b)はFashion-MNISTデータの訓練用セットにおける、各ラベルの $\hat{K}(s)$ の分布の概形である。original MNISTデータではラベル2、4、5、6、9の分布曲線がほぼ重なっている。

### 2 学習実験の結果

図5はoriginal MNISTデータを用いて隠れ層のニューロン数を $h=100$ として得た学習曲線の一つである。学習曲線は誤答率でプロットする。振動はあるものの、200エポックの間に訓練用データの誤答率は0に漸近し、テスト用セットの誤答率は訓練用データの誤答率の変化に添う形で低下し、この例では137エポックで0.0226を記録した後、徐々に上昇している。このケースでは誤答率が上昇に転じた137エポック以降は過学習(overfitting)に陥っており、これ以上の改善は望めない。

図6(a)はoriginal MNISTデータでの、図6(b)はFashion-MNISTデータでの学習曲線を $y$ 軸方向に拡大したものである。いずれもテスト用セットでの最小誤答率をプロットしたもので、隠れ層のニューロン数は図中に示した値を設定している。データの違いによって学習曲線の振動に違いが現れているが、これらは全て急激に低下した後どこかで最小値を記録し、上昇に転じるという共通点がある。また図6(c)、図6(d)に示すようにデータ種別やニューロン数に関わらず訓練用データの誤答率は振動しながらも0に漸近しているため、全ての実験でテスト用セットの誤答率が上昇に転じた時点から過学習を起こしたことになる。

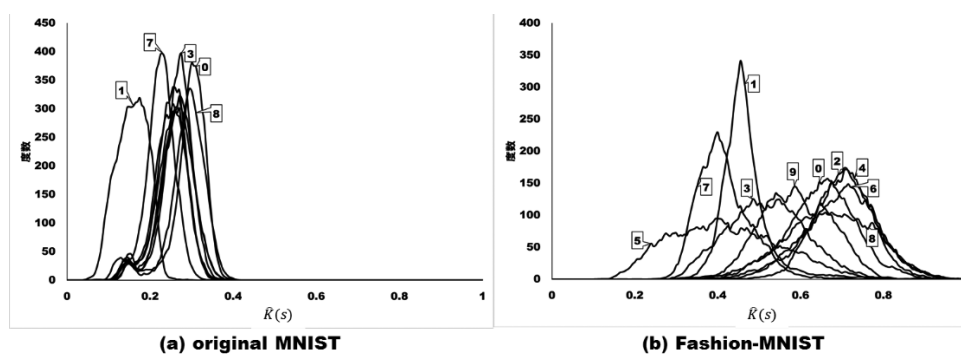
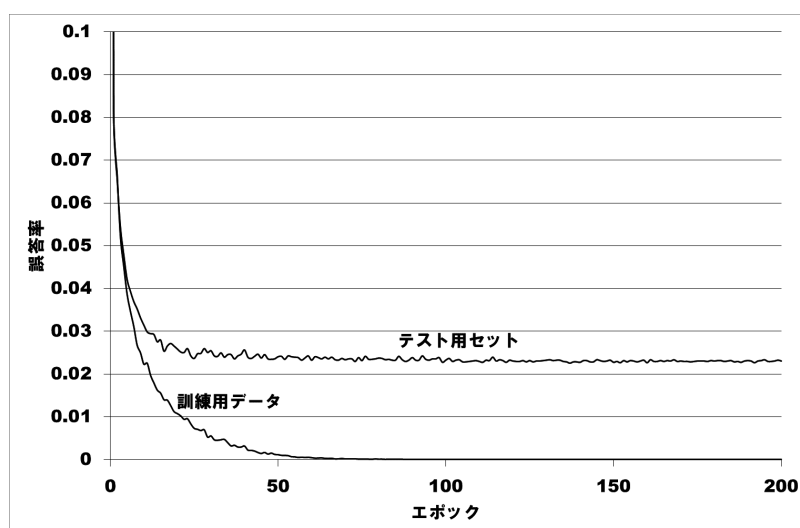
なお、original MNISTデータにおいてニューロン数 $h=10$ と $h=20$ のニューラルネットワークは200エポック経過までに訓練用データの誤答率が0に漸近しなかったため、1,000エポックまで延長して挙動を観察した。その結果、 $h=10$ では1,000エポック経過後でも誤答率が0に漸近しないだけでなく、激しい振動を見せた。徐々に低下する傾向が見られるため10,000エポック以上のどこかで0に達すると予想されたが、実験を打ち切っている。テスト用セットに対する誤答率の最小値は0.055程度であった。また、 $h=20$ のネットワークは500

表 3 訓練用セットのラベルごとの  $\hat{R}(s)$ 

original MNIST			Fashion-MNIST		
ラベル	min/mean/max	S.D.	ラベル	min/mean/max	S.D.
0	.0829/.2877/.3967	.0445	0: T-shirt/top	.3342/.6484/.9617	.0852
1	.0459/.1521/.2959	.0368	1: Trouser	.3023/.4644/.8929	.0610
2	.0702/.2570/.3529	.0442	2: Pullover	.4031/.7081/.9796	.0744
3	.0714/.2566/.3903	.0405	3: Dress	.2551/.5174/.8776	.0945
4	.0816/.2410/.3584	.0394	4: Coat	.4158/.6908/.9770	.0780
5	.0906/.2442/.3635	.0398	5: Sandal	.1122/.4173/.8099	.1287
6	.0778/.2559/.4018	.0431	6: Shirt	.3240/.6931/.9987	.0906
7	.0638/.2175/.3316	.0364	7: Sneaker	.2436/.4136/.7997	.0641
8	.1071/.2832/.4018	.0442	8: Bag	.1658/.6670/.9936	.1088
9	.0804/.2441/.3763	.0384	9: Ankle boot	.3533/.5874/.8457	.0821

表 4  $\hat{R}(s)$  の要約

	original MNIST		Fashion-MNIST	
	min/mean/max	S.D.	min/mean/max	S.D.
訓練用セット	.0459/.2425/.4018	.0554	.1122/.5807/.9987	.1429
テスト用セット	.0523/.2459/.3852	.0558	.1441/.5836/.9847	.1429
全体	.0459/.2430/.4018	.0555	.1122/.5811/.9987	.1429

図 4  $\hat{R}(s)$  の分布状況図 5 学習曲線の例 (original MNIST,  $h=100$ )

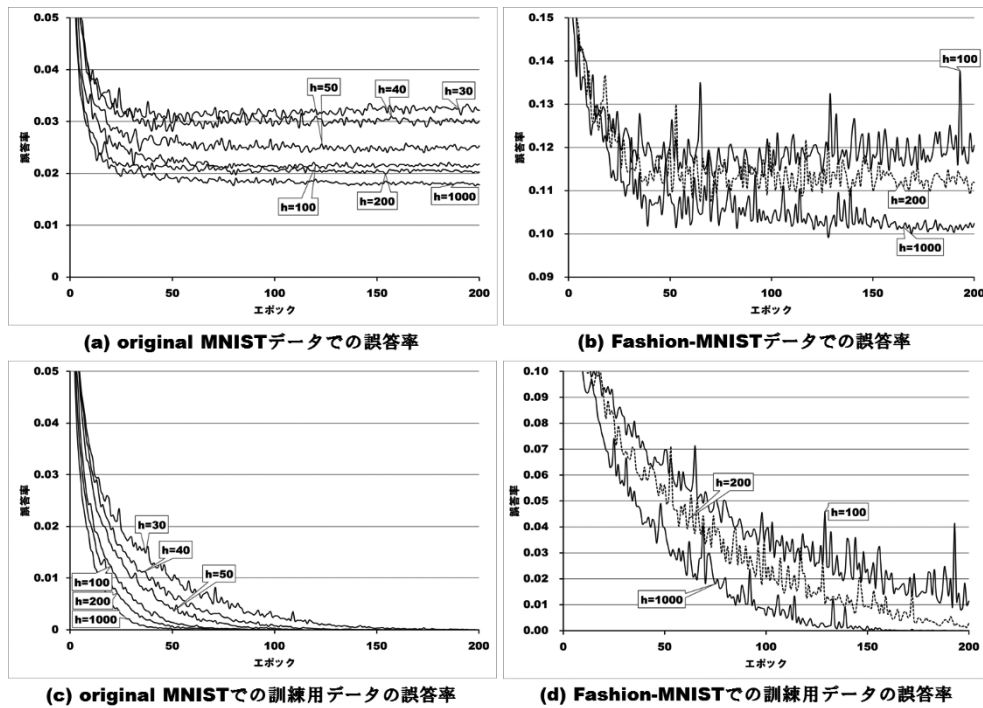


図6 典型的な学習曲線

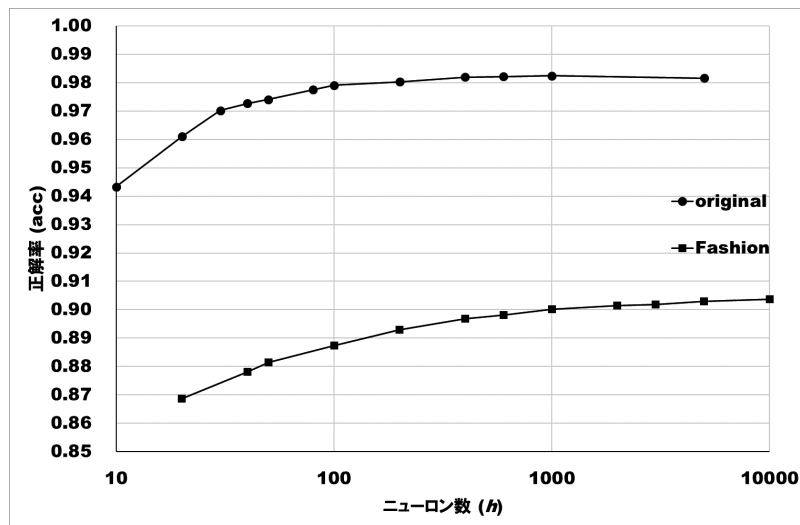


図7 ニューロン数に対する正解率

エポックまでに誤答率は0に漸近したが、テスト用セットに対する誤答率の最小値は0.04程度であった。

このようにして隠れ層のニューロン数  $h$  に対して同一の学習実験を100回程度繰り返し、正解率の最大値の平均値を  $h$  の関数としてプロットしたものが図7である。この2本の曲線それぞれに対して以下の近似式を二乗誤差  $10^{-5}$  以下の精度で得た。

$$\begin{aligned} \text{Acc}(h) &= -0.0805 \times h^{-0.5} + 0.9864 & (\text{original MNIST}) \\ \text{Acc}(h) &= -0.1665 \times h^{-0.5} + 0.9050 & (\text{Fashion-MNIST}) \end{aligned} \quad (9)$$

$h = 1000$ での正答率がoriginal MNISTデータで0.98以上、Fashion-MNISTデータで0.90程度であることから、original MNISTデータよりもFashion-MNISTデータは正解しにくい、即ち学習しにくいということがわかる。

### 3 誤答の状況

表5はoriginal MNISTデータで、表6はFashion-MNISTデータでそれぞれ最大正答率を示したときの混同行列である。これらの混同行列から得た多クラス分類アルゴリズムの品質メトリックス（適合率、再現率、Fスコア）をまとめたものが表7である。品質メトリック

表5 original MNISTデータで最大正解率を記録したときの混同行列  
h=1000, 正解率=0.9824

真のラベル	予想したラベル										
	0	1	2	3	4	5	6	7	8	9	
	0	971	1	1	1	1	0	2	1	2	0
	1	0	1124	3	1	0	1	2	2	2	0
	2	4	2	1012	1	2	0	1	4	6	0
	3	1	0	4	994	0	3	0	1	3	4
	4	1	1	3	1	962	0	2	1	1	10
	5	2	0	0	8	1	871	5	0	4	1
	6	4	3	1	1	2	4	943	0	0	0
	7	1	2	6	2	0	0	0	1008	4	5
	8	5	0	2	3	3	1	2	3	952	3
9	2	3	1	3	5	2	2	2	2	987	

表6 Fashion-MNISTデータで最大正解率を記録したときの混同行列  
h=5000, 正解率=0.9042

真のラベル	予想したラベル										
	0	1	2	3	4	5	6	7	8	9	
	0	874	1	15	16	5	1	77	0	11	0
	1	3	978	1	14	2	0	2	0	0	0
	2	19	1	846	9	69	1	53	0	2	0
	3	22	4	7	910	31	0	19	0	7	0
	4	0	1	76	25	850	0	46	0	2	0
	5	0	0	0	1	0	971	0	16	2	10
	6	115	0	82	28	56	0	711	0	8	0
	7	0	0	0	0	0	12	0	970	0	18
8	5	0	5	6	2	3	5	4	970	0	
9	0	0	0	0	0	7	1	30	0	962	

表7 品質メトリックス

	ラベル	0	1	2	3	4	5	6	7	8	9
original h=1000 Acc=0.9824 ep=198	適合率	.9798	.9894	.9796	.9793	.9857	.9875	.9833	.9863	.9754	.9772
	再現率	.9908	.9903	.9806	.9842	.9796	.9765	.9843	.9805	.9774	.9782
	F <sub>1</sub>	.9852	.9899	.9801	.9817	.9826	.9820	.9838	.9834	.9764	.9777
Fashion h=1000 Acc=0.8980 ep=146	適合率	.8395	.9868	.8162	.9008	.8171	.9777	.7526	.9482	.9719	.9678
	再現率	.8580	.9750	.8170	.9080	.8400	.9650	.7180	.9700	.9670	.9620
	F <sub>1</sub>	.8487	.9809	.8166	.9044	.8284	.9713	.7349	.9590	.9694	.9649
Fashion h=5000 Acc=0.9042 ep=180	適合率	.8420	.9929	.8198	.9019	.8374	.9759	.7779	.9510	.9681	.9717
	再現率	.8740	.9780	.8460	.9100	.8500	.9710	.7110	.9700	.9700	.9620
	F <sub>1</sub>	.8577	.9854	.8327	.9059	.8437	.9734	.7429	.9604	.9690	.9668

h : 隠れ層のニューロン数、Acc : 正解率、ep : エポック



スの算出はSokolova<sup>13)</sup>の方法に従い、ラベル  $i$  ( $0 \leq i \leq 9$ ) の適合率  $P_i$ 、再現率  $R_i$  を、混同行列  $C = [c_{ij}]$  を用いて次のように定義される。

$$P_i = \frac{tp_i}{tp_i + fp_i}, \quad R_i = \frac{tp_i}{tp_i + fn_i}$$

ただし、

$$tp_i = c_{ii}, \quad fp_i = \sum_{0 \leq k \leq 9, k \neq i} c_{ki}, \quad fn_i = \sum_{0 \leq k \leq 9, k \neq i} c_{ik}$$

である。また、Fスコア  $F_1$  は適合率と再現率の調和平均として定義される。なお、これらの式では慣例に反し、行列の行番号と列番号は0から始まるものとする。

original MNISTデータにおいて10,000枚のテスト用セットに対する最大正解率は  $h = 1000$  のときに98.24%を記録し、この時点で訓練用データに対する正解率は100%、検証用データに対しては98.02%を示した。このときのラベルごとの適合率は最大値98.94%、最小値97.54%、再現率は最大値99.08%、最小値97.65%であった。これに対し、Fashion-MNISTデータを用いて同じ  $h = 1000$  という条件での最大正解率89.80%を得たときの混同行列からは適合率は最大値98.68%、最小値75.26%、再現率は最大値97.50%、最小値71.80%を得た。適合率の最大値はほぼ同一であるが最小値が極端に小さくなっている。また、これらの値はFashion-MNISTデータで最高正解率を示した  $h = 5000$  のニューラルネットワークにおいてもほぼ同様である。 $h = 1000$  のネットワークに対してほとんどのラベルにおいて適合率と再現率が上昇したため正解率が90.42%になっているが、一部僅かに低下したラベルが存在する。また、ラベル6の再現率は70%台前半と変化せず、また最大の低下量を示した。

## V 考察

### 1 画像情報の複雑さ

original MNISTデータとFashion-MNISTデータのバイト列の圧縮によって近似した  $\hat{K}(s)$  は、original MNISTデータで0.24、Fashion-MNISTデータで0.58という平均値を得ている。一般に圧縮アルゴリズムはバイト列中の何らかのパターンの繰り返しを手掛かりに圧縮を行うので、 $\hat{K}(s)$  が小さい、即ち圧縮後のバイト列が短くなっている場合はパターンの繰り返しが多いと考えてよい。図1、表2を見てもわかるとおりoriginal MNISTデータの各画像は数字の周囲の余白、即ちピクセル値0が連続する領域が非常に大きい。一方Fashion-MNISTデータの画像は衣類が大きい関係で余白が小さい。“0の連続”というパターンが高い圧縮率に貢献しているこ

とが明らかなので、これが  $\hat{K}(s)$  が小さくなった理由の一つと考えられる。一方、余白が小さくとも描かれた衣類の模様が単調では圧縮しやすくなり  $\hat{K}(s)$  は大きくなる。Fashion-MNISTデータの中には  $\hat{K}(s) = 0.9987$  とほとんど圧縮できていない画像も存在し、ランダム性の高い模様を持つ衣類が存在していることがわかる。

画像の複雑さという点では、図4の分布状況からもoriginal MNISTデータは  $\hat{K}(s)$  が小さい、即ち変化の乏しい単調な画像、Fashion-MNISTデータは  $\hat{K}(s)$  も中程度以上で、それなりの複雑さを持った画像と判断でき、これは視覚的直感的に得た印象と一致する。

参考として、National Institutes of Health Chest X-Ray Datasetから無作為に1枚画像を取り出したところ、この1024×1024ピクセル、8ビットグレースケール画像のKolmogorov Complexityは0.600であったことから、Fashion-MNISTデータの平均的な画像は胸部X線画像と同程度の複雑さを持つことが推測できる。

original MNISTデータ、Fashion-MNISTデータともデータ設計の詳細は明らかにされていないが、original MNISTデータでは基になったNIST SD-1とSD-3で各数字の構成比率に偏りがあり、LeCunらはそれを踏襲したものと推測できる。ただし、NISTのSpecial Database Catalog<sup>14)</sup>によると両データベースとも現在非公開となっており、推測の域を出ない。手書き数字の場合はラベルが同一であれば字形も似ているため、 $\hat{K}(s)$  も特定の値の近傍に集中することが予想される。そのため構成比率が似ているならば必然的に分布状況は似ると考えられる。一方、Fashion-MNISTデータは人工的な画像データであり、ラベルごとに衣類の形状に類似性はあっても模様は千差万別であるため、 $\hat{K}(s)$  が広範囲に分布したと考えられる。

表3、表4、図5より、original MNISTデータの  $\hat{K}(s)$  は比較的狭い範囲に集中していること、Fashion-MNISTデータの  $\hat{K}(s)$  はoriginalの2.5倍程度の範囲に広がって分布していることがわかる。また個々のラベルの分布状況を見ると、original MNISTデータのラベル2、4、5、6、9の分布曲線がほぼ重なっている。一方Fashion-MNISTデータでは、ラベル2と4は近い分布状況だがそれら以外のラベルは分布の広がり方もピークの位置も異なっている。

本稿ではデータ圧縮によって  $\hat{K}(s)$  を定義しており、前述のとおり圧縮後のサイズは元のバイト列中のパターンの大きさや出現頻度に依存する。 $\hat{K}(s)$  が大きい場合は同一パターンの出現頻度が小さい、パターンの大きさが小さい、あるいはその両方が成り立っていることを意味しており、 $\hat{K}(s)$  が小さい場合はその逆である。このように同一ラベルでバイト列中のパターンの含まれ方が

大きく異なるデータ群をニューラルネットワークに順次与えるとき、式(1)から始まる一連の計算とシナプス結合加重等を更新する誤差逆拡散法はパターンを検出してデータ群を分類する問題を解決するように動作するため、過剰な更新が起こって収束が遅くなるだけではなく、収束せずに振動が起こる場合がある。これが学習の困難さとなって現れるので、original MNISTデータは比較的容易に学習可能なデータ、Fashion-MNISTデータはoriginalに比べて学習困難なデータとしてそれぞれベンチマークになりえる能力を持っていると考える。

## 2 ニューラルネットワークの挙動とFashion-MNISTのベンチマーク適性

図6の学習曲線を比較すると、original MNISTデータよりもFashion-MNISTデータのほうが振動の振幅が大きいことが見て取れる。大きな振幅はシナプス結合加重やバイアスの更新がうまくいっていないことを示しており、学習率 $\eta$ 等のハイパーパラメータや損失関数、更新手法の選択に原因がある。即ち、データの違いがハイパーパラメータ等の選択に影響を与えることが示されている。

図7の曲線は最大正答率の平均値を隠れ層のニューロン数 $h$ の関数とみなすことができるが、仮に“学習の困難度”という第3の次元を仮定すると2本の曲線は3次元空間内の曲面を $\text{Acc}-h$ 座標平面に平行な2枚の平面で切った切り口を投影したものと捉えることができる。例えば本稿で測定した $\hat{K}(s)$ の平均値を第3の次元とすると $\hat{K}(\text{original})=0.243$ と $\hat{K}(\text{Fashion})=0.581$ の平面である。ただし、70,000枚からなる画像集合を特徴付ける指標はこれ以外にも存在すると考えられる。図7の曲線を与える曲面は複数の指標によって4次元以上の空間の曲面として張られているものと予想される。

図7および式(9)から $h$ の全区間でこの近似が成り立つとすると、どちらの曲線も $h \rightarrow \infty$ の極限において正解率は1に達することはないため、これが3層ニューラルネットワークの限界を示す可能性があり、今後の解析を必要とする。また本稿で用いたニューラルネットワークは、784ニューロンの入力層と10ニューロンの出力層を $h$ ニューロンの隠れ層で全結合したものであり、結合加重行列 $W^{(1)}$ 、 $W^{(2)}$ の大きさはそれぞれ $784 \times h$ 、 $h \times 10$ になるため、学習に必要な空間計算量は定数 $A$ を用いて $(781+1) \times h + (h+1) \times 10 + A$ と表すことができる。Landauの記号を用いると $O(h)$ と評価でき、同様に学習に必要な時間計算量も $O(h)$ となる。一方式(9)は $O(h^{-0.5})$ と見積もることができ、十分大きな $h$ に対して $h > h^{-0.5}$ であるから、闇雲に隠れ層のニューロン数 $h$ を増やしても $h$ の増加に見合った正解率の向上が見られなくなることが示されている。

本稿ではoriginal MNISTデータに対する最大正解率として、隠れ層のニューロン数 $h=1000$ のときに98.24%を、Fashion-MNISTデータに対しては $h=5000$ で90.42%得ている。LeCunらのウェブサイトにはoriginal MNISTデータを用いた研究を多数紹介している。3層ニューラルネットワークを扱ったものは多くないが、LeCunらが1998年に著した論文<sup>15)</sup>で、画像を無加工で使用し $h=1000$ で得た正解率95.5%とアファイン変換を施した画像で $h=300$ のネットワークによる正解率98.4%を得たこと、最高記録としてSimardらが2003年に $h=800$ のネットワークを用いて無加工画像を用いて得た正解率98.4%と弾性歪みを与えた場合の正解率99.3%を得た論文<sup>11)</sup>などが紹介されている。本稿と同様に画像を無加工で用いたケースと比較すると、LeCunらは損失関数として二乗誤差を、本研究は交差エントロピーを用いたことが最大正解率の違いに現れたものと考えられる。また、Simardらは学習率 $\eta$ を100エポックごとに小さくするという手法を適用したこと以外では本稿との違いは隠れ層のニューロン数のみであり、学習率の調整方法と非公開となっている活性化関数や処理系、乱数の違いが結果の違いに現れたものと考えられる。Simardら得た98.4%は式(9)において $h=800$ としたときの値98.3554%と非常に近いことから、実験に使用したニューラルネットワークそのものの動作には特別な問題はなく、式(9)の妥当性も高いと考える。

一方Fashion-MNISTデータでは本稿の一連の実験ではテスト用セットに対する最大正解率は隠れ層のニューロン数 $h=5000$ のときに90.42%を得ている。LeCunら同様Zalando Researchのウェブサイト<sup>16)</sup>でも実験結果の集積を行っており、全て1年以上前の実験結果ではあるが2018年末の時点での最高正解率はSupport Vector Machineを用いた89.7%、本稿と同じ3層ニューラルネットワークでは $h=100$ の87.7%が記載されている。Fashion-MNISTデータの作者であるXiaoらはデータの紹介論文<sup>17)</sup>の中で $h=100$ の3層ニューラルネットワークで本稿と同じ活性化関数ReLUを用いた場合、original MNISTデータで97.2%を得られるニューラルネットワークでFashion-MNISTデータでは87.1%を得たことを報告している。損失関数については触れていないため同一の実験を再現することはできないが、これと比較すると本稿の一連の実験で得た $h=5000$ での90.42%は先行論文と比較して高い正解率を得ているが、 $h$ を大きくしたことによる計算コストの増加に見合う正解率の向上になっているとは言い難い。

これらを踏まえると、Fashion-MNISTデータを3層の全結合ニューラルネットワークで学習したときの、効率的な学習という観点での限界点は隠れ層のニューロン

数が1000程度、同様にoriginal MNISTデータの場合はその限界点が100程度であることが考えられる。そして式(9)のようにoriginal、Fashion両データで同型の式が得られたことは、Fashion-MNISTデータがoriginal MNISTデータと同様の挙動を引き起こしたと考えられるため、表現力が隠れ層のニューロン数に依存することとも合わせるとFashion-MNISTデータがベンチマークとして機能できることを示唆していると考ええる。

### 3 データの複雑さと誤答の状況

表6から、本稿の実験では2: Pullover、4: Coat、6: Shirtを相互に取り違える確率が高く、また0: T-shirt/topを2: Pulloverや4: Coatに誤ることは少ないが6: Shirtに誤ることが多い。さらに、7: Sneakerと9: Ankle bootの間に若干の取り違えが見られ、8: Bagは1と9以外の全てとの間に僅かに取り違えが見られる。

誤りの原因として、まず、Pullover-Coat-Shirt相互、T-shirt/top-Shirt相互、Sneaker-Ankle boot相互の誤りの原因は、形状や模様にしたものが多いためということが挙げられる。記述者にもPulloverとCoat、T-shirt/topとShirtの間で誤判定が少なからず生じた。物理的な物体であれば大きさ、材質等からラベル付けの誤りは生じ難いはずだが、本データのように大きさの情報を消去し形状とグレースケールの模様のみから判定することは人間にも困難であると考ええる。

一方、8: Bagが他の多くのラベルとの間で取り違えが起こった原因は、少々事情が異なると考える。BagがCoatやShirtとして予想されたものは、バッグのハンドルやショルダーテープと呼ばれる部分が衣類の襟の部分に当たるとして処理されたことが原因と考えられ、SandalやSneakerとして予想されたものは靴のような形状のバッグを学習した結果だと考えられる。バッグに四角形の形状のものが多いこと、四角形でない場合は比較的自由度が高いことが取り違えの原因と考えられる。

表6及び表7から、再現率が極端に低い6: Shirtの適合率と再現率を向上させることができれば、全体の正解率は大きく向上することがわかる。original MNISTデータでも特定のラベルの再現率が低くなるケースは存在したが、現在は解消できることが知られている。Fashion-MNISTデータの開発者がoriginal MNISTデータを“too easy”とする理由はここにもあると考えられる。Fashion-MNISTデータでの正答率向上のためにはさらに異なるブレイクスルーを必要としている可能性もあり、その意味でも第二のベンチマークの役割を果たしていると考ええる。

### 4 第三のMNISTデータ

ここまでの議論でoriginal MNISTデータとFashion-MNISTデータには以下の性質があることがわかった。

- ・ 図5に見るとおり、Fashion-MNISTデータの各画像の $\hat{K}(s)$ の分布はoriginal MNISTデータの2.5倍程度の広がりを持っている。
- ・ Fashion-MNISTデータの学習の困難さは $\hat{K}(s)$ の分布の広さに由来すると考えられる。
- ・ 式(9)に見るとおり、3層のニューラルネットワークにおいて隠れ層のニューロン数と正答率の間には同型の関係が存在する。

ここから、第三のMNISTデータが満たすべき条件として、本稿では次を提案する。

- ・ 各画像は $\hat{K}(s) > 0.5$ とする。
- ・  $\hat{K}(s)$ を大きくするためにも、画像内の余白を小さくするか、余白の無い画像とする。

$\hat{K}(s)$ の分布については、original MNISTデータが $0.04 < \hat{K}(s) < 0.41$ であるのに対し、Fashion-MNISTデータは $0.11 < \hat{K}(s) < 1.00$ であることから、original MNISTデータの分布と重ならないという条件である。また、画像の余白とそれによって形作られる物体の輪郭はパターンを生み出すことになり、結果的に学習を容易にする効果がある。パターンを抽出しにくいデータという形でベンチマークになる $k$ とができると考える。

28×28ピクセルの8ビットグレースケール画像は全部でおよそ $10^{1889}$ 枚存在し、単純計算で70,000枚の画像セットを $1.4 \times 10^{1884}$ セット取り出すことができる。意味のある画像の数は今後の解析を待つ必要があるが十分な数の画像が眠っているものと予想する。

## VI 結論

本稿ではFashion-MNISTデータの存在によってKolmogorov complexityがデータセットの学習の困難さの指標として機能する可能性を示すことができた。特に圧縮を基にして近似した $\hat{K}(s)$ は画像のサイズという要素を除外できるという点で異なるデータセット間での比較が可能となるメリットがある。この比較が可能であったのはFashion-MNISTデータの開発者の言う“drop-in replacement”が可能なデータ、すなわち入力画像の物理的特性が同一であるために、学習の結果の違いが画像データの論理的特性の違いのみに由来することによる。データの物理的特性が異なる場合は同一のニューラルネットワークを適用することが不可能であるので、Fashion-MNISTデータの存在によってデータセットの違いを論ずるための指標を作成できたことになる。

また、Fashion-MNISTデータは学習が困難であるだけでなく、極端に再現率が低いラベルがあるなどoriginal MNISTでは現れにくい現象も得られている。近い将来Fashion-MNISTデータでの誤答率が0に近づき、同時に第三のMNISTデータが作られることがある

ならば、Fashion-MNISTで得た知見が役立つことは疑いない。Fashion-MNISTデータのベンチマークとしての地位は揺ぎ無いものであろう。

## 謝辞

Fashion-MNISTデータを開発したZalando Research他、各種データセットを開発・公開している皆様に感謝申し上げます。

## 利益相反

本研究には開示すべき利益相反は無い。

## 文献

- 1) LeCun Y, MNIST handwritten digit database, <http://yann.lecun.com/exdb/mnist/index.html>, 2018年1月25日.
- 2) 国立研究開発法人産業技術総合研究所, Specification of ETL-8-etlcdb, [http://etlcdb.db.aist.go.jp/?page\\_id=2461&lang=ja](http://etlcdb.db.aist.go.jp/?page_id=2461&lang=ja), 2018年12月25日.
- 3) National Institute of Standards and Technology, NIST Special Database 19 NIST, <https://www.nist.gov/srd/nist-special-database-19>, 2018年12月25日.
- 4) Krizhevsky A, Nair V, Hinton G, CIFAR-10 and CIFAR-100 datasets, <https://www.cs.toronto.edu/~kriz/cifar.html>, 2018年12月25日.
- 5) Kaggle, NIH Chest X-rays Kaggle, <https://www.kaggle.com/nih-chest-xrays/data>, 2018年12月25日.
- 6) Zalando Research, GitHub-zalandoresearch/fashion-mnist: A MNIST-like fashion product database. Benchmark, <https://github.com/zalandoresearch/fashion-mnist>, 2018年1月25日.
- 7) Bennett CH, Gács P, Li M, et al.: Information distance, IEEE Trans. Information theory, 44 (4) : 1407-1423, 1998. doi: 10.1109/18.681318.
- 8) Cilibrasi R, Vitanyi PMB: Clustering by compression, IEEE Transactions on information theory, 51 (4) : 1523-1545, 2005. doi: 10.1109/TIT.2005.844059.
- 9) Amari S: A theory of adaptive pattern classifiers, IEEE Transactions on electronic computers, EC-16 (3) : 299-307, 1967. doi: 10.1109/PGEC.1967.264666.
- 10) Guyon I: A scaling law for the validation-set training-set size ratio, AT&T Bell Laboratories, 1997. doi: 10.1.1.33.1337.
- 11) Simard PY, Steinkraus D, Platt JC: Best practices for convolutional neural networks applied to visual document analysis, Proceedings of the seventh international conference on document analysis and recognition (ICDAR 2003) : 958-963, 2003. doi: 10.1109/ICDAR.2003.1227801.
- 12) He K, Zhang X, Ren S, et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015 IEEE International conference on computer vision (ICCV) : 1026-1034, 2015. doi: 10.1109/ICCV.2015.123.
- 13) Sokolova M, Lapalme G: A systematic analysis of performance measures for classification tasks, Information processing & management, 45 (4) : 427-437, 2009. doi: 10.1016/j.ipm.2009.03.002.
- 14) National institute of standards and technology, Special database catalog NIST, <https://www.nist.gov/srd/shop/special-database-catalog>, 2018年1月25日.
- 15) LeCun Y, Bottou L, Bengio Y et. al.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (11) : 2278-2324, 1998.
- 16) Zalando research, Benchmark dashboard, <http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/>, 2018年2月25日.
- 17) Xiao H, Rasul K, Vollgraf R, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms: arxiv: 1708.07747, 2017, <https://arxiv.org/abs/1708.07747>, 2018年1月25日.