DEIM Forum 2022 H34-1

意図的な過学習によるパラメータの変化を用いた外れ値検出

三島　惇也†　　鈴木　　優†

† 岐阜大学　工学部　電気電子・情報工学科　情報コース　〒 501-1193 岐阜市柳戸 1-1
E-mail: †x3033149@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　本研究では，誤りの傾向で分類することが可能な外れ値検出の手法を提案する．誤りの傾向とは，正しいラベルに対し，誤って付与したラベルはどれかという関係のことである．この手法は，誤りの傾向が同じデータ群であれば，学習モデルに入力データ一つを学習させ続けた決定境界はデータ群のどのデータを用いた場合でも類似する境界になる．そして，決定境界を表現するための重みも類似する値になるという仮説に基づいている．仮説に基づき，データセット全体を学習し，最終層以外の勾配計算を止め，データセットから取り出した 1 つのデータを入力し続けた最終層の重みを k-means でクラスタリングすることで外れ値検出を行うことを提案する．教師あり学習が難しいとされる外れ値検出において，提案手法は部分的に教師あり学習を取り入れることに成功した．結果として，提案手法を用いて MNIST で行った実験では，仮説を支持する結果が得られた．具体的には，本当のラベルは 0 で，付与されているラベルは 1 という誤りの傾向ごとにクラスタが分かれた．また，MNIST で行った外れ値の割合が 25%の実験では 2019 年に外れ値検出の SOTA であった $E^3Outlier$ よりも AUROC を約 10%改善することができた．

キーワード　外れ値検出，ニューラルネットワーク，outlier detection，機械学習，クラスタリング，特徴量抽出

1　はじめに

教師ありの機械学習を行うにあたって，教師データとなるデータセットの正確性は極めて重要な要素である．正確なデータセットを使用しない場合，誤った判断をする分類器が構築されてしまう可能性があるためである．

例として，円形のテーブルと円形の椅子を分類したい場合を考える．両者の違いとして円形部分の直径に対して足の長さがどの程度かという部分が考えられる．正しいラベルが椅子であるのにテーブルだと誤ったラベルが付与されたデータが存在する場合には，円形部分の直径に対して足が長いのにもかかわらずテーブルと判断する分類器ができる可能性がある．実際に，ImageNet[1]をはじめとする複数のデータセットにおいて，誤ったラベルがついていたり，ラベルの候補が複数あることが考慮されていなかったりすることが判明した [1].

我々は外れ値検出を用いることにより，誤ったラベルが付与されたデータを特定することが可能となると考えた．外れ値とは観測結果が他の観測結果と大きく離れているものを指す．円形のテーブルと椅子の例で示す．円形部分の直径を 1 とした時，足の長さが 1.5 未満はテーブル，1.5 以上は椅子の可能性が高いとデータの分布から判断されたとする．この場合，足の長さが 3 のテーブルや，足の長さが 0.5 の椅子などが外れ値になる．

本研究ではデータセットに含まれる誤ったラベルが付けられているデータを外れ値であるとして実験を行った．そのため，分布外検出のようなデータセット外のデータに対する検出や，画像自体の異常検出は考えていないことに注意されたい．

外れ値検出に使用する特徴量として，このデータにはこのラベルを付けるという特徴量を抽出することができないかと考えた．そこで，学習済みのモデルに対して，入力データ一つを学習させたときの決定境界の変化を用いることを考えた．学習済みのモデルを使用する理由を以下に示す．学習済みのモデルは各クラスのデータの特徴を学習しているため，クラスごとにデータの分布が固まっていると考えられる．そのため，クラス a の決定境界がクラス b の分布に入り込んだというような特徴が得られると考えたためである．

以上のことから，我々は以下の仮説を立てた．誤りの傾向が同じデータ群であれば，学習モデルに入力データ一つを学習させ続けた決定境界はデータ群のどのデータを用いた場合でも類似する境界になる．そして，決定境界を表現するための重みも類似する値になるというものである．誤りの傾向とは，正しいラベルに対し，誤って付与したラベルはどれかという関係のことである．例えば，真のクラスが 0 の入力データ x_1 と x_2 があるとする．教師データ y_1 と y_2 がともに 1 である場合，0 のデータのラベルを 1 と間違えているため，誤りの傾向が同じである．教師データ y_1 が 1 で，y_2 が 2 の場合，誤りの傾向は異なる．仮説より，外れ値検出に使用する特徴量抽出の手法として，学習済みモデルに一つのデータのみを入力し続け，過学習させたときのモデルの重みを使用することを考えた．この時使用する重みは最終層のみとし，学習済みモデルの更新も最終層のみとした．理由は，ニューラルネットワークの最終層以外は次元削減器であり，学習済みモデルが学習した次元削減の空間を維持するためである．詳しくは 3.1.2 節で述べる．

ここまでの話を前述した円形のテーブルと椅子の例で示す．分類問題の学習が終了したモデルは，正常なデータが多く，外れ値は少ないという特徴により，ある程度正常な分類が可能であると考えられる．学習済みのモデルに対し，正しいラベルが椅子であるのにテーブルだと誤ったラベルが付与されたデータがテーブルと分類されるまで学習させるとする．すると，モデ

1：https://www.image-net.org/

ルの決定境界は歪み，円形部分の直径に対して足の長さが長いものをテーブルと判断するような決定境界になると考えられる．そして，正しいラベルが椅子であるのにテーブルだと誤ったラベルが付与されたデータがもう 1 つある場合，上記のゆがんだ決定境界と類似する決定境界になることが予想される．そして，ニューラルネットワークにおいて決定境界を表現するためのパラメータは重みであるため，重みをゆがんだ決定境界の特徴として使用するというものである．

以上より，本研究で提案する特徴量抽出の手法は，与えられたデータセットをデータセット本来のタスクについて学習したモデルを使用し，データセット内のデータ 1 つを付与されている教師ラベルに分類できるようになるまで学習させたモデルを作成する．そして，学習後の最終層の重みを新たな特徴量として使用するというものである．

提案手法の特徴量が機能していることを示すため，提案手法の特徴量を用いた誤りの傾向で分類可能な外れ値検出を提案する．提案する手法は，抽出した特徴量を k-means でクラスタリングし，誤りの傾向で分類するという単純なものである．誤りの傾向で分類可能であると示すことができれば，仮説は正しいといえると考えている．また，提案手法は誤りの傾向で分類することができるため，ラベルの訂正に役立つことが期待される．

また，誤りの傾向を考えず，外れ値検出のみを行った場合の精度を測るために，追加でユークリッド距離やマハラノビス距離を用いた外れ値検出も行った．

本研究ではあるデータセットに含まれる誤ったラベルが付けられているデータを外れ値であるとして実験を行った．そのため，分布外検出のようなデータセット外のデータに対する検出や，画像自体の異常検出は考えていないことに注意されたい．

本論文では誤りの傾向毎に分ける実験と，2019 年に画像の外れ値検出で SOTA であった $E^3Outlier$ [2] と提案手法を比較する実験の結果を報告する．実験は MNIST[2]，Fashion-MNIST[3]，SVHN[4]，CIFAR10[5]，CIFAR100[6]を用いて行った．誤りの傾向で分ける実験にでは，MNIST で行った実験は成功したが，他のデータセットは成功とはいえない結果であった．$E^3Outlier$ [2] ととの比較実験では，どのデータセットで行った実験も $E^3Outlier$ を超える結果となった．MNIST の外れ値の割合 (ρ) =0.25 の実験では $E^3Outlier$ の AUROC を約 10%超える結果が得られた．また，提案手法は事前学習モデルの性能に依存する部分があり，精度が高い事前学習モデルを使用する場合は特徴量が機能していると考えられる．

2 関 連 研 究

2.1 外れ値検出

教師なしの外れ値検出は正常値と外れ値のラベルが付けられていないデータを用いて自動的に外れ値を検出する手法である．例えば，オートエンコーダ [3] を用いた手法や，Isolation Forest [4] がある．本研究も教師なしの外れ値検出である．

外れ値は数が少ないという特徴があるため，外れ値の教師データを大量に集めることが困難という問題点がある．さらに，外れ値は正常値以外のものを指すため，全ての外れ値を網羅することが難しいことも問題点として挙げられる．これら 2 点の外れ値が持つ問題点から，教師ありの手法は一般的ではない．また，外れ値検出を目的としたデータ収集を行うことは少なく，他の目的で集めたデータに対して外れ値検出を行うことが多い．そのため，仮に教師ありで外れ値検出を行うとすると，追加で外れ値かどうかのラベル付けを行う作業をしなければならない．データ数によっては作業の依頼などのコストがかかるため，教師データが存在しなくとも検出ができる教師なしの手法が主流となっている．また，正常値のみであれば十分にデータを用意できるため，半教師ありの手法も数多く提案されている．

畳み込み層や多層化といったニューラルネットワークの発展に伴い，画像などの高次元のデータを扱う問題が増加している．そのため，画像の外れ値検出を目的とした手法が提案されている．本研究も画像のデータの外れ値検出を行うため，画像の外れ値検出の手法を以下の 2.1.1 節，2.1.2 節にて紹介する．

2.1.1 画像の教師なしの外れ値検出

教師なしの外れ値検出の例として，本論文で比較実験を行った $E^3Outlier$ [2] という手法がある．$E^3Outlier$ は 2019 年に発表された inlier priority という新しい考え方を取り入れた教師なしの外れ値検出の手法である．正常値と外れ値を学習した際，学習の方向は正常値の方へ偏るという考え方をもとに外れ値検出を行っている．

$E^3Outlier$ の考え方を以下に示す.

未学習の学習器に外れ値検出の対象となるデータに疑似ラベルを付与して学習を行う．学習を行うと全体としては正常値が多いため，更新は正常値の損失を減らすことが優先され，正常値の損失を減らす更新が行われる (inlier priority)．そして，この時更新された方向とは異なる方向へ更新しようとするデータというものは外れ値である可能性が高いという考え方である．

学習の方向を見るという点で本研究と類似する手法であるが，本研究の様に学習済みモデルを利用した手法ではないこと，過学習させるという要素がない点が本研究とは異なる．

他の教師なしの手法として，学習済み EfficientNet [5] を使用した外れ値検出の手法 [6] がある．ImageNet を学習したモデルを使用し，各層の出力を用いて外れ値検出を行うというものである．正常値は出力が同様の値になり，異常なものは出力が異なるものになるという考えに基づいた手法である．この手法は不良品や画像の異常部分などを検出する異常検知の手法である．

学習済みモデルを使用する点，モデルが学習した表現空間を利用する手法である点で本研究と類似している．しかし，学習モデルの出力をそのまま用いるのではなく過学習させるという点と，事前学習に使用するデータセットが ImageNet に固定しないという点で本研究とは異なる．

2：http://yann.lecun.com/exdb/mnist/

3：https://github.com/zalandoresearch/fashion-mnist

4：http://ufldl.stanford.edu/housenumbers/

5：https://www.cs.toronto.edu/~kriz/cifar.html

6：https://www.cs.toronto.edu/~kriz/cifar.html

2.1.2 画像の教師ありの外れ値検出

外れ値検出とは異なるが，類似するタスクである教師ありの異常検知の例として，CutPaste [7] という手法がある．正常な画像の一部を別の場所にコピーすることで異常な画像を生成し，教師ありの異常検知を行う，自己教師あり学習の手法である．正常な画像から外れ値を生成できるため，2.1 節で述べた外れ値（異常値）が少ないという問題を解決している．2021 年に発表され，MV-Tec[7]データセットで SOTA を達成している．

2.1.3 本研究との比較

本研究は，外れ値検出の対象となるデータセットを本来の目的で学習[8]したモデルを使用した手法であるという点が他の研究と異なり，特殊な部分である．外れ値検出を行う場合，通常はすべての過程において教師なしのアルゴリズムを用いる．なぜなら，外れ値は数が少なく，教師データとして使用することが難しいとされているからである．しかし，提案手法ではデータセット全体を本来の目的で学習したモデルを使用することで，部分的に教師ありのアルゴリズムを取り入れることができる．

また，分類問題におけるニューラルネットワークの一つの出力層に着目すると，一つのクラスとそれ以外の分類を学習している．つまり，データセットの分布内の外れ値検出の学習を教師ありで行っていると考えることができる．

さらに，教師ありが敬遠される理由の一つである外れ値が少ないという問題点についても，各クラスのデータ数が等しい場合，n クラスの分類を学習した場合 $inlier : outlier = 1 : n-1$ となり， 外れ値のデータ数は正常値以上のデータ数となっている．よって，データセットの分布内という制約はあるものの，外れ値が少ないという問題点も解決していることになる．

以上から，提案手法は他の教師なしの手法とは異なり，実質的に教師ありの外れ値検出を学習したモデルを使用することができるといえる．

2.1.2 節で述べた 2021 年時点での MV-Tec データセットの SOTA である，CutPaste は教師ありの手法である．このように，現時点での機械学習は教師ありの学習の方が精度では優位であるため，教師ありの学習を利用した提案手法も精度が向上するのではないかと考えた．

また，提案手法はどんなニューラルネットワークにも利用することが可能であると考えられ，既存の学習済みモデルを利用することができる手法である点も利点である．

従来の外れ値検出は正常値か外れ値かの判断をするものである．そのため，外れ値がどういった外れ方をしているのかを示すものはない．それに対して，提案手法は誤りの傾向で分ける外れ値検出であるため，外れ値がどういった外れ方をしているのかという点まで考慮できる．学習済みモデルがクラス毎に分類できる能力を利用し，外れ値を誤りの傾向毎に分類可能な手法は我々の知る限り他にない．提案手法は，なぜ外れ値と判断されたか，その外れ値は本来どうあるべきだったかという点を議論できると考えている．これは最近話題となっている説明可能な人工知能に通ずる部分があるのではないかと考えている．

7：https://www.mvtec.com/company/research/datasets/mvtec-ad/

8：MNIST であれば手書き文字認識，SVHN であればカラー画像に写る数字の認識といった，データセットを用いて本来行う分類タスクのことを指す

3 提案手法

本研究で立てた仮説である，誤りの傾向が同じデータ群であれば，学習モデルに入力データ一つを学習させ続けた決定境界はデータ群のどのデータを用いた場合でも類似する境界になる．そして，決定境界を表現するための重みも類似する値になるという仮説についての詳細は 3.1.1 節で述べる．立てた仮説をもとに，本研究では以下の手法を提案する．

（1） データセットを本来の目的で学習し，学習後のモデルを保存しておく．3.2 節で述べる．

（2） 保存したモデルを読み込み，最終層以外の勾配計算を停止する．データセットのデータを一つだけ取り出し，何度も学習させ，学習後のモデルの最終層の重みをそのデータの特徴量とする．3.3 節で述べる．

（3） 抽出した特徴量を用いて k-means でクラスタリングを行う．そして，できたクラスタや抽出した特徴量のベクトルを用いてスコアを算出し，外れ値検出を行う．3.4 節で述べる．

3.1 理論と仕組み

本節では立てた仮説の発想と理論，仕組みについて述べる．

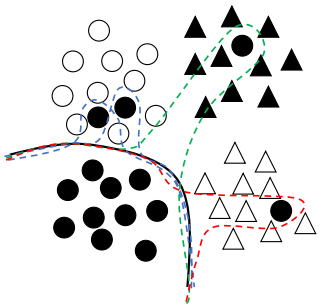

図 1 決定境界の変化の例

3.1.1 本研究の発想

例として，4 クラス分類の問題を考える．学習済みモデルの決定境界が図 1 の実線，外れ値を入力し続け，過学習させた場合の決定境界が図 1 の破線ようになるとする．正常値を分類可能になるまで過学習した場合，もともと分類可能であるため，決定境界は動かない．しかし，外れ値を分類可能になるまで過学習した場合，決定境界は歪み，図 1 の破線のようになると考えられる．外れ値が入り込んだクラスの分布ごとに破線の色を変えてある．このとき，過学習後の決定境界に注目すると外れ値が入り込んだクラス，つまり誤りの傾向ごとに決定境界の形が大きく変わることが予想される．図 1 の色分けした破線を見ると，青色の破線は同じような決定境界であると判断することができる．これが仮説を立てるに至った発想である．

誤りの傾向で分けるためには，決定境界の形を表す特徴量が必要である．決定境界を決めるパラメータはモデルが持つ重みである．そのため，外れ値を学習し続け，過学習させたモデルが持つ重みを新たな特徴量とし，外れ値の検出に使うことがで

きるのではないかと考えた．

3.1.2 理　　論

本研究で提案する手法は以下の理論に基づいたものである．ニューラルネットワークを用いた機械学習を最終層以外の処理と，最終層の処理に分けると以下のように説明できる．

最終層以外　次元削減を行いながら，入力を変換し，決定境界を引きやすいような分布へと変換する処理．

最終層　変換された特徴量の分布を用いて決定境界を引く処理．

つまり，ニューラルネットワークの本質は最終層以外は次元削減器を学習し，最終層はそれに決定境界を引くことを学習しているといえる．そのように考えた理由を以下に示す．

どれだけ複雑な学習器を定義して学習を行ったとしても，そのモデルの最終層への入力を再現するような次元削減の関数 $f(x)$ さえ定義できれば，関数 $f(x)$ と最終層のみで分類器を作成することができる．このことは今まで行われてきた数多くの研究が証明している．

例えば，VGG [8] や ResNet [9] など，ニューラルネットワークの多層化に成功し，精度の向上を図ってきた研究がある．しかし，これらのネットワークも最終層に入力する特徴量を関数 $f(x)$ で置き換えることさえできてしまえば不必要となる．なぜなら，ニューラルネットワークの全てのモデルは最終層の入力と，最終層の重みを用いて最終的な出力を決めているからである．

以上から，最終層以外の部分は次元削減器であり，最終層によって決定境界が引かれ，出力を得ていると考えることができる．本研究の理論は，オートエンコーダに近い考え方であると考えている．オートエンコーダは出力が入力になるように学習するが，本研究では出力は教師ラベルになるように学習する点が異なる．ニューラルネットワークの学習は，教師ありのエンコーダを作成し，最終層でどのクラスに分類されるかを判断していると考えることができる．

3.1.3 提案手法の仕組み

ニューラルネットワークの最終層以外は入力の次元を削減しつつ，決定境界を引きやすいような分布へと変換するために存在している．そのため，最終層への入力は各クラスごとに偏った分布になっていると考えることができる．そのため，3.1.1 節で述べた考え方を適用することができる．

各クラスごとに偏った分布になっている状態で正常値を分類できるまで過学習しても決定境界は変化しない，または少しだけ変化した決定境界になるものの，他のクラスの分布に入り込んだ決定境界にはならないはずである．

一方，外れ値を分類できるようになるまで過学習すると，決定境界は他のクラスの分布に入り込んだ決定境界になる．

この決定境界の違いの特徴を得ることで，誤りの傾向の特徴を得ることができる．決定境界の特徴を得る手段として，最終層の重みを取得する．3.1.2 節で述べたように，決定境界は最終層によって引かれるからである．そして，最終層の重みを用いてクラスタリングを行うことで図 1 で示したような決定境界が入り込んだクラス毎のクラスタが生成される．

以上が本研究の仕組みである．

3.2 事 前 学 習

本節では提案手法の（1）である事前学習について述べる．本研究において事前学習は与えられたデータセットの本来のタスクについて学習することを指す．例えば，MNIST で行う場合，手書き文字認識の学習を行うことになる．事前学習が終了したモデルは保存する．保存したモデルが提案手法の（2）で行う個別の学習に使用するモデルとなる．

また，事前学習に用いる学習モデルについては層の数や，畳み込み層の有無はそれぞれのデータセットに合わせて変更しても問題ない．むしろ学習する問題に応じて変更すべきであると考えている．なぜなら，学習モデルの精度が高いほどモデルの表現空間はクラスごとに分かれていると考えられるためである．

3.3 個別の学習

本節では提案手法の（2）である個別の学習について述べる．手順は以下のとおりである．

手順 1　検出対象となるクラスの教師ラベルが付与されているデータセットのデータを一つずつ取り出す．

手順 2　3.2 節で保存しておいたモデルを読み込み，追加の学習を行う．学習時には最終層以外の勾配計算は止める．これは，データセット全体を学習して得たモデルの表現空間を失わないための処理である．取り出したデータのみを用いて Accuracy が 1 になるまで，または loss の値が十分に下がるまで学習を行う．

手順 3　学習が終了したら，最終層の重みを保存しておく．

以上の手順をデータセットのデータ数分繰り返す．手順 1 を検出対象となるクラスで分けて行う理由は 3.4 節で行うクラスタリングのクラスタ数を抑えるためである．そのため，個別の学習自体はデータセット全体をそのまま行っても問題はない．

3.4 重みのクラスタリングと外れ値検出

本節では，提案手法の（3）である最終層の重みを用いたクラスタリングと外れ値検出について述べる．

3.4.1 クラスタリング

本節では，最終層の重みを用いたクラスタリングについて説明する．

まず，保存しておいた最終層の重みを新しい特徴量としたデータセットを作成する．

例　元のデータセットのデータ数が N，最終層のユニット数が 10（バイアスを含めて 11），出力層のユニット数が 10 の場合，できるデータセットは 110 次元のベクトルが N 個並んだデータセットとなる．

次に，作成したデータセットを k-means でクラスタリングする．この時のクラスタの数は使用するデータセットのクラス数とする．この作業を行うことで正しいクラスのクラスタと誤りの傾向のパターンで別れたクラスタの作成を行う．

例　10 クラスのデータセットを用いる場合，クラスタ数は 10 となる．教師ラベルにクラス 0 が付与されているデータのクラスタリングを行う場合，本当にクラス 0 のデータ群，本当

はクラス 1 のデータ群，本当はクラス 2 のデータ群，... のように誤りの傾向毎のクラスタに分かれる．

3.4.2 外れ値検出のスコア算出方法

本節では，外れ値検出に使用するスコアの算出方法について述べる．

算出方法 1　3.4.1 節で作成したクラスタの最大クラスタのデータ数を N_{max}，各クラスタ x のデータ数を N_x として，$1 - N_x/N_{max}$ で求められる値をスコアとする．

算出方法 2　最大クラスタの中心座標と各データ x の座標間のユークリッド距離 D_x を求める．D_x のうち最大のものを D_{max} とし，$1 - D_x/D_{max}$ で求められる値をスコアとする．

算出方法 3　作成したクラスタは関係なく，マハラノビス距離を用いてスコアを算出する．各データ x のマハラノビス距離を M_x とし，最大のものを M_{max} とする．$1 - M_x/M_{max}$ で求められる値をスコアとする．

以上の 3 種類のいずれかを用いて外れ値検出を行う．どの算出方法がより優れているのか，という議論は 4 章にて行う．

4 実　　験

本章では実験の手順，結果，考察について述べる．

提案手法を用いて誤りの傾向で分かれる外れ値検出が可能であることを確認するために実験を行った．実験は 2 種類行う．提案手法を用いることで誤りの傾向で分ける実験と，提案手法と $E^3Outlier$ の精度を AUROC の値で比較する実験の二つである．データセットには MNIST，FashionMNIST，SVHN，CIFAR10 と，CIFAR100 の上位クラスの 5 種類を用いた．

実験の目的は二つある．一つ目は，提案手法で抽出した特徴量を用いることで誤りの傾向で分けることができるかどうかを調査すること，二つ目は，提案手法で抽出した特徴量を用いて外れ値検出を行うとどの程度の精度が出るのかを調査することである．以上二つの観点から実験の考察を行う．

実験手順について，一部を除いて 3 章で述べた提案手法と変わらない為，異なる点のみ 4.1 節にて述べる．また，実験条件の詳細は 4.2 節で述べる．

4.1 提案手法と異なる点

提案手法の説明と異なる点は，外れ値検出用データの作成方法である．本来であれば，誤りを含むデータセットで事前学習を行い，付与されたラベル毎に外れ値検出を行う．しかし，比較対象とした $E^3Outlier$ で行われた実験と同様の実験を行うため，事前学習に使用するデータセットはある 1 つのクラスのみにしか外れ値がないデータセットになっているという点が提案手法の説明と異なっている．

外れ値検出用データの作成は $E^3Outlier$ の論文と同様に，あるクラスの全データに他のクラスのデータが ρ% 含むようにランダムにデータを入れるという方法を取った．ただし，誤りの傾向でデータを分けるという実験も行うため，他のクラスのデータはそれぞれ均等に入るように調整した．外れ値である点は変わらない為，同条件での実験であると考えている．

4.2 実 験 条 件

MNIST，FashionMNIST はテストデータを含む 70,000 件のデータ，SVHN はテストデータを含む 99,289 件のデータ，CIFAR10，CIFAR100 はテストデータを含む 60,000 件のデータを使用して実験を行った．あるクラスに外れ値が含まれる割合は 5%から 25%まで 5%刻みで実験を行う．実験はそれぞれ 5 回ずつ行い，その平均値を報告する．

また，学習に使用するモデルを図 2 に示す．MNIST，FashionMNIST には Model1，SVHN と CIFAR10 には Model2，CIFAR100 には Model3 を使用し，実験を行った．損失関数は CrossEntropyLoss, 最適化手法には Adam を使用し，入力データは-1 から 1 に正規化した．

個別の学習における最終層の重みを保存する条件は，Accu-

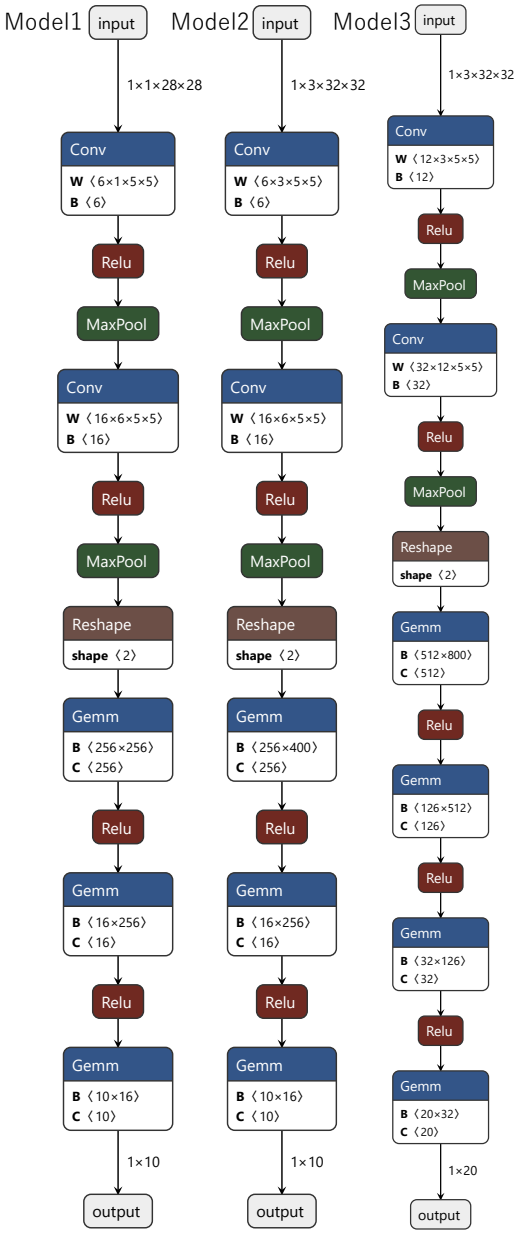

図 2　実験に使用したニューラルネットワークの図

racy が 1 になる，または loss の値が 0.0001 未満になる，の 2 パターンで実験を行った．外れ値検出のためのスコアは 3.4.2 節の算出方法 1 から 3 の 3 パターンで実験を行った．したがって，報告する結果は 6 パターンとなり，それぞれと $E^3Outlier$ の比較を行った．

4.3 実験結果

4.3.1 誤りの傾向で分ける実験の結果

k-means でクラスタリングを行った後，クラスタ毎にそのクラスタのラベルを決める．クラスタのラベルの決め方を以下に示す．

（1） クラスタ内に含まれるデータに対して本来付与されるべきラベルを参照し，集計を行う．

（2） 集計した結果，最大となったラベルをクラスタのラベルとする．

個別の学習を Accuracy が 1 になった時点で止める実験と loss が下がりきった時点で止める実験の両方を行ったが，Accuracy が 1 になった時点で止める実験の方がどのデータセットにおいても性能が高かった．そのため，Accuracy が 1 になった時点で止める実験の結果を報告する．

MNIST，FashionMNIST，SVHN を使用した実験でクラスタリングを行い，クラスタのラベルが時のクラスタのラベルが 0〜9 に分かれた時の Accuracy の平均を求めたものを表 1 に示す．太字表記になっている部分が正常値 (inlier) のクラスである．表の値は，MNIST/FashionMNIST/SVHN の順でそれぞれの結果を示している．表の一番下の項目 (can't_classified) はクラスタのラベルが 0〜9 に分かれなかった回数を示している．CIFAR10，CIFAR100 は全ての条件においてクラス毎にクラスタが分かれなかったため表は記載しない．

また，同じ数字の文字認識データセットであり，分類難易度の異なる MNIST，SVHN について，クラスタのラベルが 6 のクラスタから 25 枚ずつランダムで表示した画像を図 3 と図 4 に示す．画像は各データセットの ρ=0.25，inlier がクラス 0，個別の学習を Accuracy=1 で止める実験の結果のうち，1 回目の実験のクラスタのラベルが 6 のものである．MNIST では，6 のみが出力されているのに対し，SVHN では 0 や 5 といった形が類似するものが混ざっていることが確認できる．

4.3.2 外れ値検出の実験の結果

5 個のデータセット MNIST，FashionMNIST，SVHN，CIFAR10，CIFAR100 で実験を行った結果を表 2 に示す．

ρ はあるクラスに外れ値データが含まれる割合，それぞれの AUROC の値は，5 回実験を行った平均値 (μ) ±振れ幅（$||\mu-\mu$ から最も離れた値 $||$）となっている．また，比較した中で 1 番 AUROC の値が高いものを太字表示してある．

表 2 より，提案手法を用いることで，$E^3Outlier$ よりも高い AUROC を出すことができることが示された．

4.4 考　　察

4.4.1 誤りの傾向で分ける手法についての考察

実験の結果から，誤りの傾向毎に分かれた割合を表 4 にまとめた．クラスタリングで誤りの傾向で分けることは，MNIST については成功と言えるが，FashionMNIST，SVHN については成功とまでは言えず，場合によっては誤りの傾向で分かれるという結果となった．CIFAR10 ，CIFAR100 は誤りの傾向で分けることができなかった．

実験自体はうまくいかなかったが，仮説は正しいといえると考えている．そう考えた理由は，図 4 を見ると，出力に入り込んだデータは形が類似するものであったからである．具体的には 6 のクラスタに対し，他の文字に比べて 6 と形が似ている 0 や 5 が入り込んでいる．本当に正しいかどうか，現時点では判断しかねるため，さらなる検証が必要であると考えている．

また，表 1 と，表 3 の事前学習終了後の Accuracy の値を見ると，データセットの難易度が上がるにつれて徐々に誤りの傾向で分類できなくなっていることがわかる．MNIST の実験において概算ではあるが，25%外れ値が存在する場合で約 90%の外れ値を誤りの傾向で分類可能であることを確認した．このことから，提案手法は誤りの傾向で分けることが可能であるが，事前学習のモデルの性能に依存する部分があることが考えられる．また，SVHN，CIFAR10，CIFAR100 については，カラー画像である点や，背景の存在により特徴量の散らばりが大きくなっていることも原因として考えられる．

外れ値の割合が高いほど，誤りの傾向毎に分かれやすいという傾向がある．クラスタのラベルの決め方によるものだと考えられる．単純に外れ値の数が増えることによってクラスタ内の外れ値のデータ数が増え，外れ値が優勢になり，誤りの傾向で分かれやすくなっていると考えられる．

しかし，一般的に外れ値の割合は低い．そのため，外れ値の割合が高いほど検出されやすい傾向にある手法を用いるのは好ましくない．提案手法では誤りの傾向毎のクラスタを作成するためにクラスタ数を指定できる k-means を用いた．しかし，inlier となるデータ数が多いため，クラスタの中心とユークリッド距離を用いてクラスタリングを行う k-means は不適切な手法であった可能性がある．この問題については，他のクラスタリング手法を用いてクラスタリングをする，または外れ値検出をした後に外れ値をクラスタリングを行うことで，提案手法よりも改善される可能性があると考えている．

4.4.2 外れ値検出を行うことについての考察

表 2 の結果より，すべてのデータセットにおいて，Accuracy が 1 になった時点で個別の学習を止め，マハラノビス距離を利用したスコアを用いる手法が最も AUROC の値が高くなることを確認した．事前学習を止める条件，外れ値検出に使用するスコアの二つの条件について，上記の条件が最も性能が良い理由を以下で考察する．

個別の学習を Accuracy が 1 になった時点で止める手法の方が性能が良い点については，元のモデルからの決定境界の形状の変化の程度が異なることが原因であると考えられる．Accuracy が 1 になった時点で止めると，決定境界の形状の変化は最小限となる．しかし，loss が下がりきるほど学習すると決定境界の形状が大きく変化する．データごとに特化した境界になり，誤りの傾向が同じデータの境界の形が類似したものにならないと

表 1　各実験で作成したクラスタの Accuracy(inlier:0) (MNIST/FashionMNIST/SVHN)

ρ	ρ=0.05	ρ=0.10	ρ=0.15	ρ=0.20	ρ=0.25
True_label:0	**0.99/0.00/0.00**	**0.99/0.99/0.00**	**0.98/0.98/0.96**	**0.98/0.97/0.95**	**0.97/0.96/0.92**
True_label:1	0.99/0.00/0.00	1.00/0.98/0.00	1.00/0.98/0.70	1.00/0.98/0.77	0.99/0.98/0.78
True_label:2	0.98/0.00/0.00	0.99/0.56/0.00	0.99/0.63/0.83	1.00/0.72/0.87	0.99/0.73/0.93
True_label:3	0.99/0.00/0.00	0.99/0.52/0.00	1.00/0.66/0.76	1.00/0.76/0.87	0.99/0.79/0.81
True_label:4	0.99/0.00/0.00	0.99/0.74/0.00	1.00/0.75/0.83	1.00/0.76/0.84	1.00/0.80/0.87
True_label:5	0.99/0.00/0.00	0.97/0.89/0.00	1.00/0.95/0.80	0.99/0.96/0.72	0.99/0.97/0.84
True_label:6	0.97/0.00/0.00	0.99/0.59/0.00	0.99/0.54/0.79	1.00/0.60/0.80	0.99/0.57/0.85
True_label:7	0.99/0.00/0.00	1.00/0.94/0.00	1.00/0.94/0.90	0.99/0.94/0.86	0.99/0.94/0.85
True_label:8	0.98/0.00/0.00	0.99/0.87/0.00	1.00/0.93/0.83	1.00/0.91/0.86	0.99/0.92/0.79
True_label:9	0.99/0.00/0.00	0.99/0.95/0.00	0.99/0.90/0.75	0.99/0.96/0.79	0.99/0.96/0.76
can't_classified	0/5/5	0/4/5	0/4/4	0/2/4	0/0/4

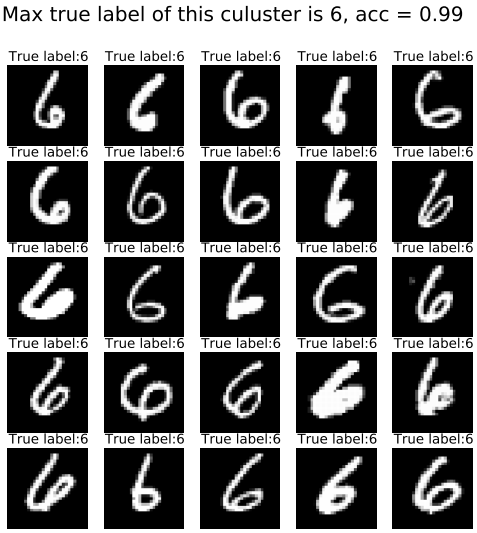

図 3　クラスタリングの結果 (MNIST，cluster:0)

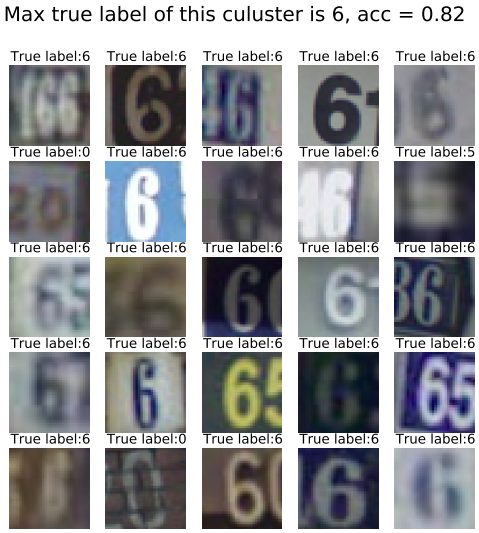

図 4　クラスタリングの結果 (SVHN，cluster:0)

表 2　実験結果 AUROC（%）（$E^3Outlier$ の値は論文より引用）

Dataset	ρ	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
MNIST	0.05	93.37 ± 2.67	99.43 ± 0.52	**99.56 ± 0.34**	99.00 ± 1.25	98.33 ± 4.23	99.13 ± 0.33	95.16 ± 0.15
	0.1	94.37 ± 2.70	97.41 ± 5.83	**99.58 ± 0.34**	99.11 ± 1.03	97.72 ± 4.21	98.81 ± 0.49	94.09 ± 0.13
	0.15	95.02 ± 1.87	97.65 ± 6.23	**99.62 ± 0.24**	99.05 ± 0.92	96.69 ± 7.94	98.63 ± 0.45	92.85 ± 0.15
	0.2	95.28 ± 2.25	97.97 ± 5.24	**99.64 ± 0.22**	98.85 ± 1.15	95.68 ± 6.61	98.52 ± 0.63	91.31 ± 0.16
	0.25	95.14 ± 2.09	96.98 ± 6.09	**99.61 ± 0.16**	98.39 ± 2.97	97.65 ± 4.35	98.30 ± 0.83	89.77 ± 0.25
FashionMNIST	0.05	93.69 ± 2.27	96.86 ± 4.89	**98.21 ± 0.37**	89.32 ± 5.24	88.95 ± 11.57	93.82 ± 1.63	94.05 ± 0.13
	0.1	93.19 ± 1.99	94.80 ± 6.03	**97.89 ± 0.46**	89.59 ± 4.52	90.05 ± 9.07	92.48 ± 1.39	93.27 ± 0.14
	0.15	93.96 ± 1.98	92.54 ± 12.80	**97.78 ± 0.29**	90.46 ± 4.31	88.52 ± 9.60	91.81 ± 1.89	92.3 ± 0.1
	0.2	93.91 ± 1.66	93.12 ± 10.80	**97.78 ± 0.21**	90.72 ± 4.30	90.03 ± 8.41	91.70 ± 1.38	91.18 ± 0.15
	0.25	93.46 ± 1.62	92.40 ± 13.59	**97.61 ± 0.24**	90.40 ± 4.57	86.39 ± 9.97	90.86 ± 1.65	89.61 ± 0.55
SVHN	0.05	88.68 ± 2.51	93.12 ± 13.35	**96.91 ± 0.51**	86.06 ± 1.86	81.76 ± 19.35	90.03 ± 1.31	88.9 ± 0.24
	0.1	87.77 ± 2.43	91.67 ± 13.63	**96.69 ± 0.37**	85.41 ± 2.86	81.00 ± 14.87	88.23 ± 1.60	86.01 ± 0.18
	0.15	87.73 ± 1.65	90.65 ± 10.93	**96.63 ± 0.36**	84.66 ± 2.27	73.80 ± 28.49	87.16 ± 1.47	83.32 ± 0.46
	0.2	87.48 ± 2.68	88.51 ± 19.06	**96.48 ± 0.39**	83.28 ± 3.63	74.80 ± 22.64	85.78 ± 2.15	80.97 ± 0.25
	0.25	87.73 ± 1.87	89.91 ± 20.38	**96.34 ± 0.40**	82.67 ± 3.49	77.48 ± 22.60	84.46 ± 1.59	78.84 ± 0.26
CIFAR10	0.05	86.41 ± 2.01	83.68 ± 18.45	**90.91 ± 1.20**	78.65 ± 2.45	73.85 ± 15.03	81.62 ± 2.55	85.65 ± 0.42
	0.1	85.23 ± 2.16	85.35 ± 15.39	**90.26 ± 1.00**	78.10 ± 2.85	67.88 ± 25.59	79.97 ± 2.75	83.53 ± 0.2
	0.15	84.77 ± 1.86	80.52 ± 21.58	**89.94 ± 0.79**	77.76 ± 2.57	68.64 ± 16.34	78.47 ± 3.08	81.33 ± 0.27
	0.2	83.83 ± 1.76	82.30 ± 23.35	**89.45 ± 0.92**	77.29 ± 2.61	65.50 ± 25.00	77.47 ± 3.44	79.32 ± 0.16
	0.25	83.31 ± 2.01	78.54 ± 26.26	**89.13 ± 0.79**	76.51 ± 3.24	66.78 ± 21.36	76.54 ± 3.17	77.37 ± 0.2
CIFAR100	0.05	81.71 ± 2.68	78.56 ± 16.28	**85.66 ± 1.98**	74.83 ± 3.25	67.19 ± 21.59	79.48 ± 2.71	85.65 ± 0.42
	0.1	82.09 ± 2.14	77.55 ± 15.61	**85.60 ± 1.65**	74.34 ± 2.54	68.72 ± 18.34	77.51 ± 2.18	83.53 ± 0.2
	0.15	82.18 ± 1.69	74.37 ± 18.99	**85.41 ± 1.31**	73.41 ± 2.65	66.31 ± 19.27	75.65 ± 2.12	81.33 ± 0.27
	0.2	81.99 ± 1.45	75.33 ± 14.30	**84.90 ± 1.19**	73.11 ± 2.92	67.09 ± 18.82	74.34 ± 2.39	79.32 ± 0.16
	0.25	81.86 ± 1.36	78.89 ± 11.20	**84.63 ± 1.16**	72.43 ± 2.73	65.51 ± 15.34	73.19 ± 2.76	77.37 ± 0.2

考えられる．そのため，Accuracy が 1 になった時点で止める手法の方が AUROC が高かったのではないかと考えている．

外れ値検出を行う際のスコアはマハラノビス距離を用いたスコアが最も性能が良い点について考察する．クラスタを用いたスコアはクラスタごとにスコアが付くため，角ばった AUROC を描く．そのため，高い値を出せないことが考えられる．ユークリッド距離を用いたスコアは最大クラスタの中心点から円形に広がっていく距離を用いる．そのため，正常値の分布が円形

表 3　事前学習後のモデルの Accuracy の平均値

ρ	MNIST	FashionMNIST	SVHN	CIFAR10	CIFAR100
0.05	0.99	0.93	0.93	0.76	0.66
0.10	0.99	0.93	0.93	0.76	0.66
0.15	0.98	0.92	0.92	0.75	0.65
0.20	0.97	0.91	0.91	0.74	0.66
0.25	0.96	0.91	0.90	0.74	0.65

表 4　誤りの傾向毎に分かれた割合

	MNIST	FashionMNIST	SVHN	CIFAR10,CIFAR100
ρ=0.05	5/5	0/5	0/5	0/5
ρ=0.1	5/5	1/5	0/5	0/5
ρ=0.15	5/5	1/5	1/5	0/5
ρ=0.2	5/5	3/5	1/5	0/5
ρ=0.25	5/5	5/5	1/5	0/5

でない限り外れ値のスコアが高くなる可能性がある．上記二つのスコアに比べ，マハラノビス距離を用いるスコアは，全体の分布をもとに距離を算出するため，外れ値の距離が大きくなる可能性が非常に高い．異常より，マハラノビス距離を利用したスコアを用いる方が高い AUROC になると考えられる．

また，表 2 の結果と表 3 の事前学習終了後の Accuracy の値は相関があるように見える．表に示した 25 個のデータを用いて算出した相関係数は 0.99 であった．外れ値検出においても，事前学習に使用するモデルの性能に依存する部分があると考えられる．CIFAR10 や CIFAR100 も学習モデルを変更し，Accuracy が高いモデルを使用することでさらに高い AUROC を出すことができると考えられる．

5 おわりに

本研究では，公開されているデータセットに誤ったラベルが存在することを問題視し，問題解決のための研究を行った．

誤りの傾向が同じデータ群であれば，学習モデルに入力データ一つを学習させ続けた決定境界はデータ群のどのデータを用いた場合でも類似する境界になる．そして，決定境界を表現するための重みも類似する値になるという仮説を立てた．そして，学習済みモデルに一つのデータのみを学習させ続けた後の最終層の重みを用いて誤りの傾向で分ける外れ値検出を提案した．

実験の結果から，学習済みモデルに一つのデータのみを過学習させた最終層の重みという特徴量は，誤りの傾向で分けることが可能な特徴量であることがわかった．しかし，学習済みモデルの性能に依存する部分があるという問題点も見つかった．

本研究の仮説について，MNIST の実験では正しいといえる結果が得られた．この結果から，事前学習のモデルに依存する部分があるが，誤りの傾向が同じデータ群は一つのデータのみを過学習させた最終層の重みが類似する値になるといえる．

ただし，本当に仮説が正しいのかという点については，さらなる検証が必要であると考えている．理由は，MNIST 以外の実験では，誤りの傾向毎に分かれるとは言い切れない結果が確認されたためである．今後以下に示す 2 種類の実験を行い，仮説が正しいかどうか検証したいと考えている．

（1） 他のデータセットを使用し，同様の結果が得られるか．

（2） ResNet [9] や EfficientNet [5] といった精度が高いモデルを用いた場合，精度の向上が見られるか．

また，外れ値検出においては，マハラノビス距離を用いたスコアを用いることによって $E^3Outlier$ よりも精度が高い検出器を構築することができた．

今後の展望としては，どのようなデータセットであっても誤りの傾向毎に分けることができるようにしたいと考えている．実現することができれば，データセットの作成をサポートできると考えている．具体的には，データセット作成時に提案手法を用いることで外れ値検出だけでなく，ラベルの修正のアシストを行うことができると考えている．これにより，誤ったラベルが付与されたデータの修正の手間が少なくなり，ラベルの修正作業に手を付けやすくなるのではないかと考えている．

外れ値検出においては，今後の研究次第で報告した精度以上のモデルが作成できる可能性がある．使用する機械学習モデルを ResNet や EfficientNet といった制度が良いとされるモデルに変更するなどして精度の向上を図り，本研究と同様に $E^3Outlier$ [2] と比較を行っている SLA^{2}P [10] との比較や，現時点の SOTA 手法との比較も行ってみたいと考えている．

また，本研究で提案した特徴量を外れ値検出以外に役立てることができないか模索したいと考えている．

謝辞　本研究の一部は JSPS 科研費 18H03342，19H04221，19H04218 の助成を受けたものです．

文　献

[1] Curtis G. Northcutt, et al. Pervasive label errors in test sets destabilize machine learning benchmarks. *CoRR*, Vol. abs/2103.14749, , 2021.

[2] Siqi Wang, et al. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Advances in NIPS*, Vol. 32, 2019.

[3] G E Hinton, et al. Reducing the dimensionality of data with neural networks. *Science*, Vol. 313, No. 5786, pp. 504–507, July 2006.

[4] Fei Tony Liu, et al. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

[5] Mingxing Tan, et al. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th ICML*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, 09–15 Jun 2019.

[6] Oliver Rippel, et al. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. *CoRR*, Vol. abs/2005.14140, , 2020.

[7] Chun-Liang Li, et al. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on CVPR*, pp. 9664–9674, June 2021.

[8] Karen Simonyan, et al. Very deep convolutional networks for large-scale image recognition. *CoRR*, Vol. abs/1409.1556, , 2014.

[9] Kaiming He, et al. Deep residual learning for image recognition. *CoRR*, Vol. abs/1512.03385, , 2015.

[10] Yizhou Wang, et al. Sla^{2}p: Self-supervised anomaly detection with adversarial perturbation. *CoRR*, Vol. abs/2111.12896, , 2021.