卒業論文

意図的な過学習による特徴量抽出と外れ値検出への応用

三島　惇也

2022 年 2 月 9 日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

本論文は岐阜大学工学部に
学士（工学）授与の要件として提出した卒業論文である.

三島　惇也

指導教員：
鈴木 優 准教授

意図的な過学習による特徴量抽出と外れ値検出への応用*

三島　惇也

内容梗概

本論文では，データセットに存在する誤ったラベルがつけられている教師データを検出する外れ値検出の研究結果を報告する．本研究では，外れ値検出と同時に自動的に誤りの傾向で分類することができる手法を提案する．データセットを本来のタスクについて学習した，学習済みの機械学習モデルに対し，１つのデータのみを過学習させた時，誤りのデータであれば決定境界が歪み，他のクラスの分布に入り込んだような決定境界になると考えることができる．そこで我々は，以下の仮説を立てた．学習モデルに同じような誤り方をしたデータを学習させ続けた場合の決定境界は同じような境界になる．そして，決定境界を表現するための重みも同じような値になるという仮説である．以上から，以下に示す手順で外れ値検出を行うことを提案する．データセット全体についてデータセット本来のタスクを学習した事前学習モデルを作成する．そして，事前学習モデルの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によって外れ値検出を行う．提案手法を用いて MNIST で行った実験では，本当のラベルは 0 で，振られているラベルは 1 というような類似したデータごとにクラスタが別れることを確認した．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，機械学習，クラスタリング，特徴量抽出

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1183033149, 2022 年 2 月 9 日.

目次

図目次

表目次

第 1 章　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．

具体的な例として，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の部分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．それをもとに分類を行う学習をした時に，誤ったデータが存在しているとする．すると，テーブルと椅子の足の長さの境界が曖昧になり，正常に分類できなくなることが考えられる．例えば，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ついているデータが存在する場合には，円形の部分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．最近になってImageNet*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にもかかわらずそれが考慮されていなかったりすることが判明した [1].

この問題を解決するために，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で，誤ったラベルがついているデータを見つけ出すことができるのではないかと考えた．このデータにはこのラベルを付けるという特徴を抽出するには，入力データと教師データを組み合わせた特徴量を得る必要がある．

ニューラルネットワークの学習済みモデルに対し，入力データ x を教師データ y であると学習を行う場合について考える．学習を行うと，決定境界は入力データ x を y であると分類できるように更新される．次に，学習済みのモデルに対して何度も誤ったラベルがついている同じデータのみを入力し続け，過学習させた場合について考える．過学習が進むにつれて元の決定境界がゆがみ，入力データを分類できるように無理やり決定境界内に取り込んだような決定境界ができると考えられる．また，学習モデルは通常の分類タスクを行えていることから，入力データの特徴が類似しているものは近くに分布していると考えられる．

以上から，同様の誤り方をしたデータは，一つの入力データのみを学習させ続

*https://www.image-net.org/

け，過学習させたときの決定境界が類似するのではないかと考えた．誤りとは，入力データに本来付与されるべきクラスではない教師ラベルがついている状態のことを指す．例えば，真のクラスが 0 のデータに対して教師ラベルに 1 を振られているデータ同士は同様の誤り方をしていると言える．また，決定境界を表現するのは，学習モデルの重みである．以上のことから，我々は以下の仮説を立てた．同様の誤り方をしたデータ群であれば，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同様の境界になる．そして，決定境界を表現するための重みも同様の値になるというものである．

具体的に示すために，真のクラスが 0 である入力データ x_1 と x_2 について例を示す．

1. x_1 と x_2 の教師データ y_1 と y_2 がともにクラス 1 であった場合，各データのみを用いて過学習させた二つの決定境界は類似したものになる．
2. x_1 の教師データ y_1 がクラス 1，x_2 の教師データ y_2 がクラス 2 であった場合，各データのみを用いて過学習させた二つの決定境界は全く異なったものになる．

という考え方である．

仮説より，入力データと教師ラベルを組み合わせた特徴量を抽出する方法として，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ととらえることができるためである．詳しくは 4.1.2 節で述べる．

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 つを振ら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の手法を提案する．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ことを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．

外れ値は一般的に観測結果が他の観測結果と大きく離れているもののことを指す．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．

外れ値検出の手法は大きく分けて外れ値と正常値のラベルが存在する教師あり，正常値のみのラベルを使用する半教師あり，ラベルを使用しない教師なしの 3 種類がある．しかし，外れ値はデータが少ないため，教師なしで行われることが多い．本研究はデータセットの教師ラベルでの学習は行うもが，外れ値かどうかのラベルは存在しない為，教師なしの外れ値検出である．

画像の外れ値検出では，今回比較した $E^3Outlier$[2] やオートエンコーダなどの手法が提案されている．教師なしの外れ値検出は外れ値や正常値の特徴を用いた検出手法が用いられる．$E^3Outlier$ では inlier priority という考え方をもとに外れ値検出を行っている．inlier priority とは，学習を行う時に正常なデータと外れ値のデータを一緒に学習させると，学習の更新方向は正常なデータに偏るという考え方である．これは正常値の特徴を用いた検出方法であると考えられる．オートエンコーダを用いた手法では外れ値をうまく再現できないという特徴，DBSCAN を用いた手法では外れ値はデータ点が他のデータ点から離れた場所にあるという特徴を用いた検出方法である．

外れ値検出は教師なしの手法で行うことが主流であるが，機械学習全体で見ると教師ありの学習方法の方が性能が高い．外れ値検出の分野においても CutPaste[3] が 2021 年に MV-Tec データセットで異常検知の SOTA を達成しており，この手法は教師あり学習を用いている．このことからも，教師ありの手法を用いることで性能が上がる可能性があることが考えられる．我々の提案手法も教師ありの手法を部分的に取り入れることに成功しているため，性能が向上することが考えられる．

この点については 3.2.3 節にて詳しく述べる

本論文では，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．MNIST[†]，FashionMNIST[‡]，SVHN[§]，CIFAR10[¶]，CIFAR100[‖]を用いて行った実験はどれも $E^3Outlier$ を超える結果となった．また，外れ値検出の際に行うクラスタリングについては，事前学習モデルの性能に依存する部分があり，精度が高いモデルを使用する場合はうまく機能していることがわかった．

本論文の構成は以下の通りである．2 章では基本的事項について述べる．3 章では関連研究について述べる．4 章では提案手法に至った発想，理論，仕組み，提案手法の内容について述べる．5 章では評価実験を行った結果とその考察について述べる．最後に 6 章では本論文のまとめと今後の課題について述べる．

[†]http://yann.lecun.com/exdb/mnist/
[‡]https://github.com/zalandoresearch/fashion-mnist
[§]http://ufldl.stanford.edu/housenumbers/
[¶]https://www.cs.toronto.edu/ kriz/cifar.html
[‖]https://www.cs.toronto.edu/ kriz/cifar.html

第 2 章　基本的事項

2.1　クラスタリング

クラスタリングは教師なし学習の手法の一つである．データの類似度に基づきグループ分けを行う手法のことである．類似度の計算にはユークリッド距離などが用いられ，データ間の距離や，データ点から近い複数のデータから求めたデータの密度を利用してクラスタリングを行う．

クラスタリングには階層的な手法と非階層的な手法がある．階層的な手法では類似度の高いものからまとめ，クラスタを結合していき，最終的に一つのクラスタになるまで継続する手法である．例えば，群平均法と呼ばれる手法では各クラスタのデータ群の平均値のユークリッド距離が近いものから順にクラスタを結合していくことでクラスタリングを行う．この時，各クラスタのデータが一つの場合はそのデータがクラスタの平均となる．最終的にトーナメント表のような，クラスタが結合した結果を示すツリー状のデンドログラムが出力される．例を図 2.1 に示す．このとき，出力されたデンドログラムの足をどこで切るかでクラスタ数が決まる．ま

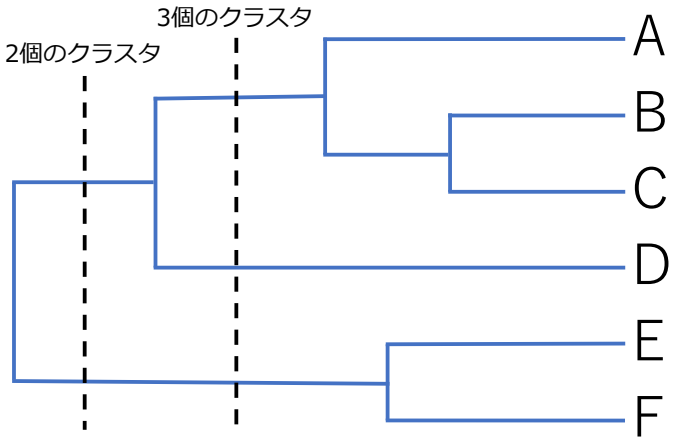

図 2.1　階層的クラスタリング行った時のデンドログラムの例

た，逆の手順を辿り，一つのクラスタから細分化していく手法もある．

非階層的な手法ではあらかじめクラスタの分割数を決めておく手法 (k-means など) や, 同じクラスタと判断する条件を設定しクラスタリングを行う手法 (DBSCAN など) がある．非階層的な手法としてよく使用されており，本研究でも使用した k-means について詳しく説明する．k-means で k 個のクラスタを作る場合，以下の操作を行うことでクラスタリングを行う．

(1) k 個のクラスタの中心点を無作為に定める．
(2) それぞれのデータ点を一番近い中心点のクラスタに属するものとし，k 個のクラスタを作成する．
(3) それぞれのクラスタの重心を求め，クラスタの中心点をデータ群の重心に更新する

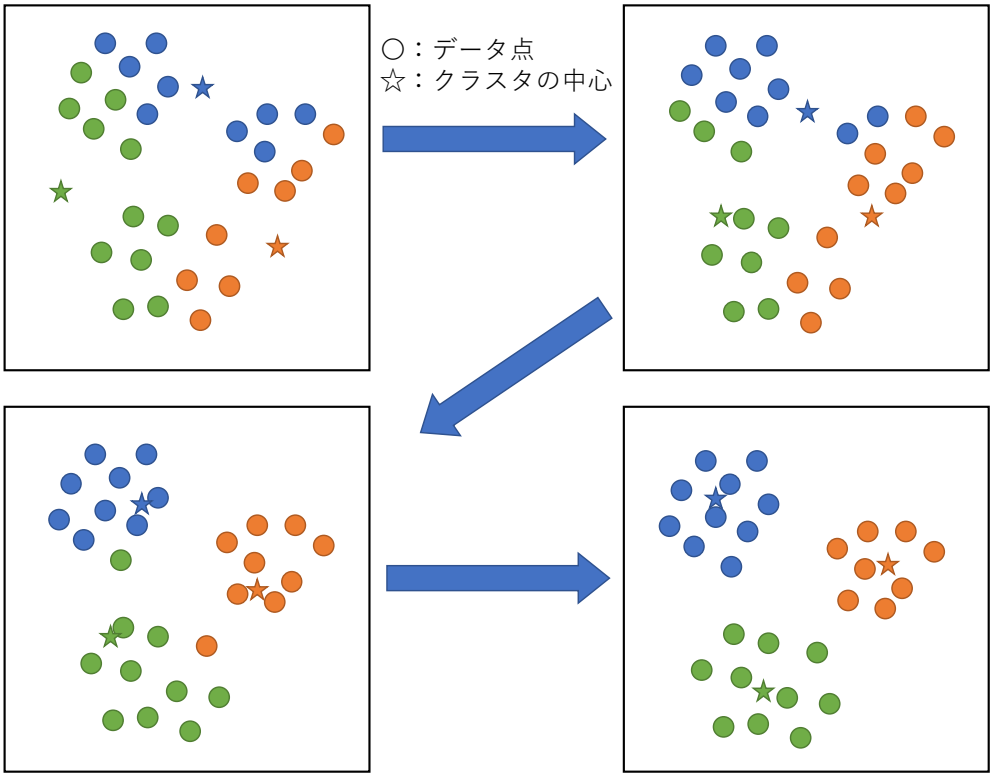

図 2.2　k-means の例

(2) と (3) をクラスタの変動が起きなくなる，または定めた回数分繰り返す．例として，4 回で収束した場合を図 2.2 に示す．以上のような手順でクラスタリングを行う手法が k-means である．

2.2　ニューラルネットワーク

ニューラルネットワークとは，人間の神経細胞のニューロンとそのつながりを模してつくられた機械学習の手法である．人間のニューロン一つをパーセプトロンと呼ばれる人工的なニューロンに置き換え，人間の神経回路と同様に複数個つなげてネットワークを構築するというものである．

ニューラルネットワークはパーセプトロンを多数結合したモデルであり，複雑な関数も近似することが可能であるとされている．従来の機械学習の手法 (クラスタリングやロジスティック回帰など) よりも分類問題や回帰問題を解くための関数を複雑にできるため，従来では解けなかった問題も解くことができる．

レコメンデーションや自動運転などの分野において，ディープラーニングを用いた手法がよく利用されている．ディープラーニングとは，ニューラルネットワークのうち，隠れ層が多層構造になっているモデルで学習することである．ディープラーニングを用いることで，隠れ層が 1 層の時よりもさらに複雑な問題にもうまく対応できるとされており，画像などの高次元データを扱う問題であっても利用できる．そのため，様々な分野においてディープラーニングを用いた手法が主流になりつつある．

ニューラルネットワークの学習について簡単に説明する． ニューラルネットワークは入力層，隠れ層 (中間層)，出力層に分かれており，各層の間にはニューロン同士の関係の強さを示す重みが存在する．人間の神経回路でいうシナプスの結合強度のことであり，重みはそれを人工的に再現したものである．この重みにより，一つ前の層のパーセプトロンから入ってきた情報をどの程度重要視するかという部分が決まる．図 2.3 に示したように入力に対してそれぞれの入力の重要度を示す重み w を与え，重みをかけて合計した値から出力を得ている．図で示した例は，入力として特徴量が 3 種類で True か False の 2 値分類のタスクを行う場合の例である．

ニューラルネットワークの学習はバックプロパゲーション（誤差逆伝搬）によっ

て行われる．学習方法は大きく分けると以下の手順となる．

(1) ある入力データをニューラルネットワークを用いて予測を行う．

(2) 教師データと比べて出力がどうだったか確認する．

(3) 欲しい出力が得られていない場合，それを修正するために重みに修正を加える．

この時の (3) で重みに加える修正の計算がバックプロパゲーションである．(3) で得たい出力と実際の予測との誤差を用いて予測とは逆の順番（出力層→隠れ層→

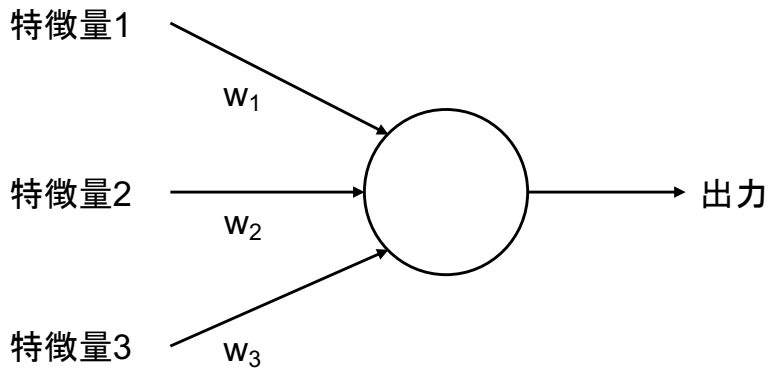

図 2.3　人工ニューロン（パーセプトロン）の例

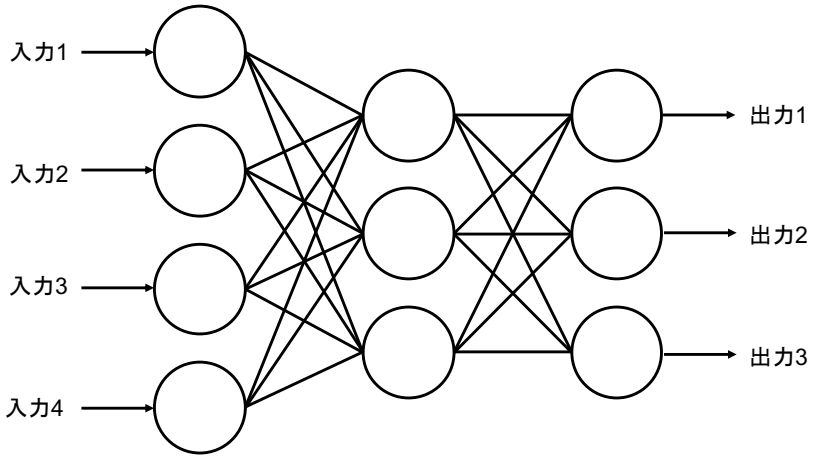

図 2.4　ニューラルネットワークの例

入力層）に重みを修正するというものである．(1) から (3) の操作を繰り返すことで学習が進み，問題を解くことができる．

入力が 4 個の特徴量で 3 クラス分類のニューラルネットワークの例を図 2.4 に示す．このようなネットワークを構築し，学習を行うことで複雑な問題に対しても精度の高い予測，分類が可能となる．ResNet など，ニューロンをさらに結合し，多層化して精度の高いモデルを作成している研究が数多く行われている．

2.3 外れ値検出

1 章でも述べたように，外れ値とは一般的に観測結果が他の観測結果と大きく離れているもののことを指す．この外れ値を検出するタスクが外れ値検出である．

類似するタスクに異常検知や，分布外検出というものがある．これらのタスクは，異常検知の一部として外れ値検出や分布外検出があるという関係にある．異常検知では不良品の検出など，異常なものの検出すべてを指す．これに対し，外れ値検出は統計的におかしいものの検出を指し，分布外検出は機械学習において，学習したデータセットに存在しないカテゴリのデータを検出するタスクを指す．詳しくは 3.2 節にて述べる．

正常な分布から外れた値のことを指すため，外れ値は数が少ないとされている．そして，正常な分布以外の値を取るものすべてが外れ値であるため，外れ値が取る値の範囲は正常値に比べはるかに広いとされている．このように外れ値検出は 2 値分類ではあるものの，データの偏り，値の範囲などの問題がある．そのため，通常の 2 値分類の問題とは区別されている．

外れ値検出では，正常値はデータの分布が近く，外れ値の分布は離れていると考えることができる．そのため，外れ値検出の手法として，データの分布が正規分布に従うとして外れ値検出を行う手法がよく使用される．データの分布が密集している部分は正規分布の値も高く，外れた値であれば正規分布の値は低くなるからである．データの分布を正規分布に近似し，ある値以下を取るデータはすべて外れ値であるとする手法である．データの分布が正規分布に従わない場合は，入力 x を $\log(x)$ や $\sqrt{x}$ に変換し，正規分布の形にすることを試みたり，カーネル密度推定という手法を用いて分布を求めたりして外れ値検出を行う．

他にはマハラノビス距離を用いた手法やクラスタリング（DBSCAN）を用いた手法などがある．

マハラノビス距離をもちいた手法では，データの分布に基づいた距離を持つため，ある一定距離以上のものを外れ値として判定する手法が取られる．マハラノビス距離とは，平均，共分散行列を用いて，データの分布に基づいた等高線を引くようなイメージの距離関数である．計算式は，$\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ で，この時の x がデータ点, μ がデータ全体の平均，Σ が共分散行列である．データの分布が楕円形であったりした場合にも外れ値の距離のみが大きくなるという特徴がある．例を図 2.5 に示す．次のような分布の場合は示した箇所が外れ値であると判断される．

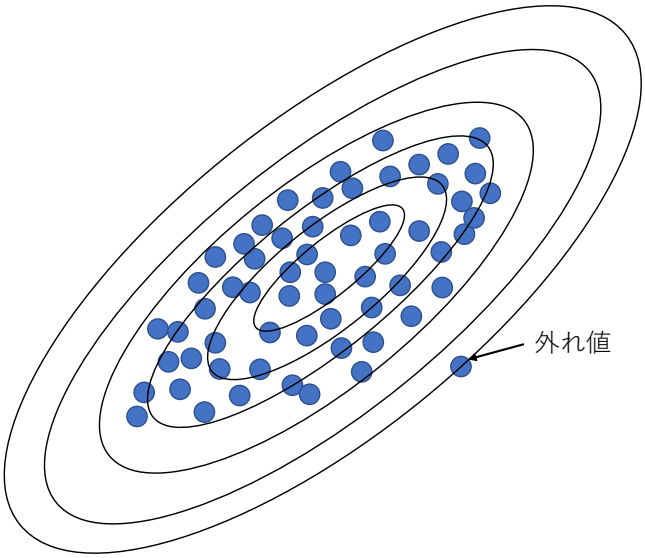

図 2.5　マハラノビス距離を用いた外れ値検出の例

DBSCAN を用いた手法では，あるデータ点から一定の距離以内にあるものを探索した結果，同じクラスタに属さないとされたものを外れ値とする手法である．DBSCAN とは，以下の手順でクラスタを作成する手法である．

(1) あるデータ点から設定した半径だけ離れた範囲を探索し，他のデータが存在したらそのデータを同じクラスタとする．
(2)（1）と同じ処理を同じクラスタと判断されたデータ点から行う
(3)（2）を繰り返し，同じクラスタと判断されたデータ点がなくなったら，(1)

に戻り，未探索のデータで同様の処理を行う．

以上を繰り返していくと，クラスタに属するデータと，どのクラスタにも属さないデータが出てくる．これらのデータのうちどのクラスタにも属さないデータを外れ値であると判断する手法が DBSCAN を用いた外れ値検出の手法である．図 2.6 に例を示す．

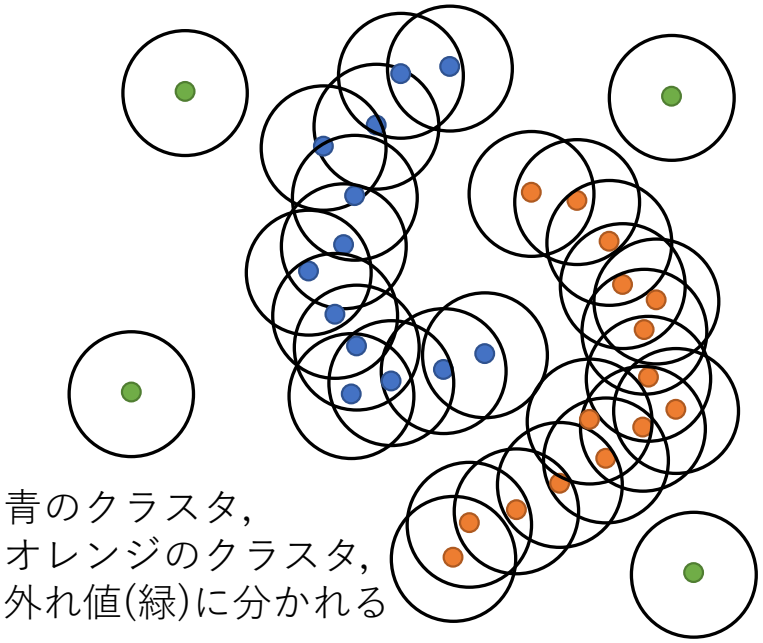

図 2.6　DBSCAN を用いた外れ値検出の例

2.4　特徴量抽出 (次元削減)

特徴量抽出とは，与えられた入力の特徴を維持したまま入力よりも小さな次元のデータへと変換し，新たな特徴量を得る手法のことである．主成分分析や特異値分解による次元削減や，3.1 節で述べるオートエンコーダ [4] を用いた次元削減などが特徴量抽出の手法としてよく利用される．

これらの手法に共通するのは，入力よりも小さな次元のデータになること，データの特徴を出来る限り損なわないように変換を行っていることである．このような

データの変換を特徴量抽出という．

2.5 評価指標

本研究で用いる評価指標について，True と False の出力で 2 値分類を行う場合を想定した例を示しながら説明する．2 値分類を行う場合，表 2.1 に示したように予測の結果と正解の組み合わせができる．

予測の結果が True である場合に，正解も True のものを True Positive（真陽性，TP）といい，正解は False のものを False Positive（偽陽性，FP）という．また，予測の結果が False である場合に，正解は True のものを False Negative（偽陰性，FN）といい，正解が False のものを True Negative（真陰性，TN）という．

表 2.1　予測の結果と正解の組み合わせ

	正解が True	正解が False
予測の結果が True	True Positive（真陽性,TP）	False Positive（偽陽性,FP）
予測の結果が False	False Negative（偽陰性,FN）	True Negative（真陰性 TN）

2.5.1 Accuracy

Accuracy とは，正解率のことである．例えば，10 個の入力データがあった場合に 6 個の入力データに対する推測が正しく，4 個の入力データに対する推測が間違っていた場合，Accuracy は 0.6 となる．

表 2.1 を用いて式であらわすと以下のとおりである．

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

2.5.2 Precision

Precision は適合率といい，予測の結果が True であった時の正解が True であるものの割合のことである．

表 2.1 を用いて式であらわすと以下のとおりである.

$$Precision = \frac{TP}{TP + FP}$$

2.5.3 Recall

Recall は再現率，真陽性率といい，正解が True であるもののうち，予測の結果も True であったものの割合のことである.

表 2.1 を用いて式であらわすと以下のとおりである.

$$Recall = \frac{TP}{TP + FN}$$

2.5.4 Fall-out

Fall-out は偽陽性率といい，正解が False であるもののうち，予測の結果も False であったものの割合のことである.

表 2.1 を用いて式であらわすと以下のとおりである.

$$Fall\text{-}out = \frac{FP}{FP + TN}$$

2.5.5 AUROC

表 2.2　実際のラベルと予測の例

data number	True or False	score
1	True	0.9
2	False	0.1
3	True	0.8
4	False	0.1
5	False	0.2
6	True	0.8
7	False	0.3
8	True	0.7

AUROC とは，ROC 曲線下の面積のことである.

ROC 曲線とは，予測結果を True または False と判断する閾値を変化させたとき

表 2.3　表 2.2 から閾値ごとに求めた Recall と Fall-out

閾値 (閾値以下を False と判断)	Recall	Fall-out
0.0(All True)	1.0	1.0
0.1	1.0	0.5
0.2	1.0	0.25
0.3	1.0	0.25
0.4	1.0	0.25
0.5	1.0	0.25
0.6	1.0	0.25
0.7	0.75	0.0
0.8	0.25	0.0
0.9	0.0	0.0
1.0(All False)	0.0	0.0

の Recall（縦軸）と Fall-out（横軸）の値をプロットした曲線のことである．

すべて True と判断すると Recall，Fall-out がともに 1 となり，すべて False と判断すると Recall，Fall-out がともに 0 になる．True，False の閾値を変化させていき図 2.7 のようなグラフを描き，曲線よりも下の部分が大きいほど良いモデルであるとされている．また，完全ランダムに予測を行った場合の AUROC は 0.5（$(0, 0)$ から $(1, 1)$ への直線）となる．

例として，真偽のラベルと予測スコアが表 2.2 であった場合について述べる．閾値を 0.1 ずつ変化させていったときの Recall と Fall-out の変化を表 2.3 に示す．表 2.3 に示した値をプロットし，その値を結び，図 2.7 のようなグラフを作成する．そして，そのグラフの曲線下の面積を求めることで AUROC の値を求めることができる．

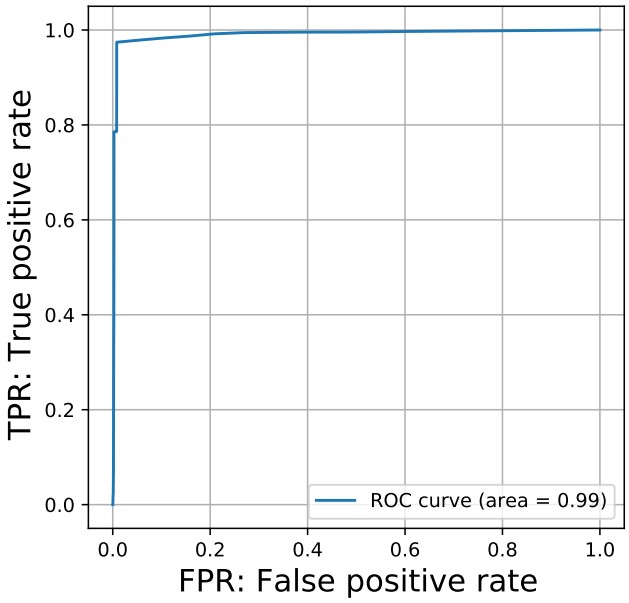
TPR: True positive rate
FPR: False positive rate
ROC curve (area = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

図 2.7　AUROC の例

第 3 章　関連研究

3.1　特徴量抽出

特徴量抽出の手法として，オートエンコーダ [4] がある．オートエンコーダは，入力と出力が同じになるようにニューラルネットワークを学習し，その中間層の出力を新たな特徴量とすることで次元削減をする手法である．入力を復元できるほどの特徴を保持したまま次元を削減できるため，特徴量抽出の手法としてよく使用される．

また，外れ値検出でも使用される．特に，多層化された畳み込みオートエンコーダは外れ値をうまく再現できないという特徴があり，正常値と同じ様な出力になる．この特徴を利用し，入力と出力の差分を取ることで外れ値である原因の箇所を特定することが可能である．

3.2　外れ値検出

2.3 節で述べたように, 外れ値検出は異常検知の一種である．外れ値検出 (Outlier Detection) に類似する問題として，分布外 (Out of Distribution) 検出がある．以下でそれぞれの違いについて述べる．

Outlier Detection	統計的におかしいデータを検出するタスク． 例：データセットのラベルを誤って振られているデータを検出する．
Out of Distribution	学習したデータセットの分布にはないデータを検出するようなタスク． 例：MNIST の学習をした機械学習モデルに FashionMNIST のデータを入力した際に検出する

というような違いがある．本研究は外れ値検出 (Outlier Detection) に位置する研究である．

外れ値検出はマハラノビス距離を用いたもの，正常データが正規分布に従うと仮

定して検出を行うもの，DBSCAN[5] を用いたものや，One Class SVM を用いたものなど，様々なものがある．

外れ値検出には 1 章でも述べたように外れ値検出には教師あり，半教師あり，教師なしの 3 種類がある．

教師ありの外れ値検出は正常値と外れ値のラベルが付いたものを学習データに使用するもののことを指す．例えば，ラベル付けされたデータを使用した SVM やランダムフォレストの外れ値検出などがそうである．

半教師ありの外れ値検出は正常値のみを学習したモデルを用いて明らかに外れた値を検出する手法のことを指す．また，少量のラベルありのデータとラベルなしのデータを使用する手法も半教師ありの外れ値検出である．例えば，SVM を用いた手法や，主成分分析を用いた手法などがある．

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もののことを指す．例えば，オートエンコーダを用いたものや，Isolation Forest[6] などがあり，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である問題点がある．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おうとすると，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．

最近では，多層ニューラルネットワークや畳み込み層を持つニューラルネットワークがよく使用されるようになり，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行うことを目的とする手法が提案されている．本研究も同様に画像のデータの外れ値検出を行うため，画像を対象とした外れ値検出の手法を以下の 3.2.1 節，3.2.2 節にて紹介する．

3.2.1 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．

$E^3Outlier$ の論文で提案されている外れ値検出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ような更新がされる (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，$E^3Outlier$ は未学習のモデルに擬似ラベルを用いて学習させるものであり，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[7] を使用した外れ値検出の手法 [8] 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．正常値は出力が同じ様になり，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．

3.2.2 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[3] 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，3.2 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*データセットで SOTA を達成している．

3.2.3 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†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，特殊な部分であると考えている．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．

3.2.2 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

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

†MNIST であれば手書き文字認識，SVHN であればカラー画像に映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

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，新しくモデルを作成すること無く，既存のものを利用することができる手法である点も利点であると考えている．

従来の外れ値検出はどれも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．学習済みモデルのクラスごとに分類できる能力を利用し，外れ値を誤りの傾向ごとに分けられる手法は他にはない手法である．どうして外れ値と判断されたのか，その外れ値は本来どうあるべきだったのかという点について議論できるようになると考えている．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．

第 4 章　提案手法

本章では，提案手法の詳細について述べる．本研究で立てた仮説である，同様の誤り方をしたデータ群であれば，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同様の境界になる．そして，決定境界を表現するための重みも同様の値になるという仮説についての詳細は 4.1.1 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．4.2 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．4.3 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．4.4 節で述べる．

4.1　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．

4.1.1　本研究の発想

例として，4 クラス分類の問題を想定した場合で説明する．学習済みモデルの決定境界が図 4.1 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 4.1 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 4.1 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

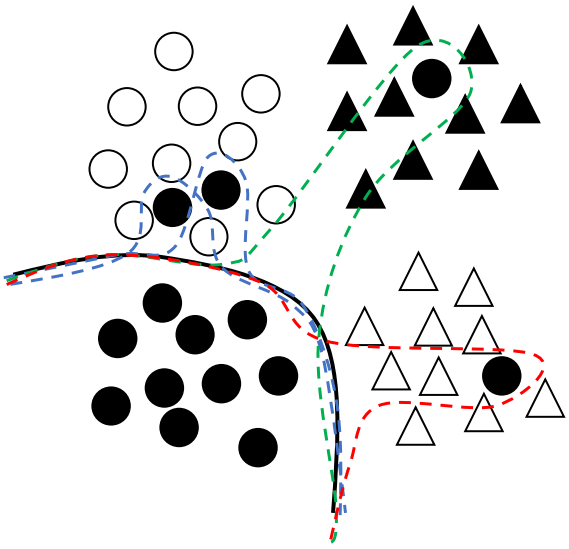

図 4.1　決定境界の変化の例

界の形が大きく変わることが予想される．図 4.1 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．

4.1.2　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することができる．

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[9] や ResNet[10] 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．

4.1.3 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，4.1.1 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．4.1.2 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 4.1 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．

4.2 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ければ高いほどモデルの表現空間はクラスごとに上手く分かれていると考えられるためである．

4.3 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振ら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 4.2 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．学習時には最終層の重み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 4.4 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．

4.4 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．

4.4.1 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方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 が振ら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，... のように誤り方毎のクラスタに分かれる．

4.4.2　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　4.4.1 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 5 章にて行う．

第 5 章　評価実験

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別れる外れ値検出ができ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 20 クラスの 5 種類を用いた．

実験の目的は二つある．

目的 1 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．

目的 2 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．

以上二つの観点から実験の考察を行う．

実験の手順についての詳細は 5.1 節，実験条件の詳細は 5.3 節で述べる．

5.1　実験手順

行った評価実験の手順について述べる．以下に手順を示す．

手順 1 誤りデータを含むデータセットを作成する．

$E^3Outlier$ の論文ではデータセットの作成は行っていない．なぜなら，一つクラスのデータに対して外れ値検出を行う手法であるためである．しかし，提案手法には事前学習を行うという過程があり，すべてのクラスのデータを使用する．そのため，事前学習と個別の学習で入力データと教師データの矛盾*を避ける必要がある．以上の理由から各実験ごとにデータセットを作成

*事前学習をデータセットをそのまま学習したものを使用して行った場合を考える．データセットの教師ラベルを変更して個別の学習を行うと，事前学習と個別の学習で同じ入力データに対して違うラベルがついているものを学習することになる．これにより，事前学習と個別の学習の間で矛盾が生じてしまう．そのため，誤ったラベルを生成した後のデータセットで事前学習を行う必要がある．

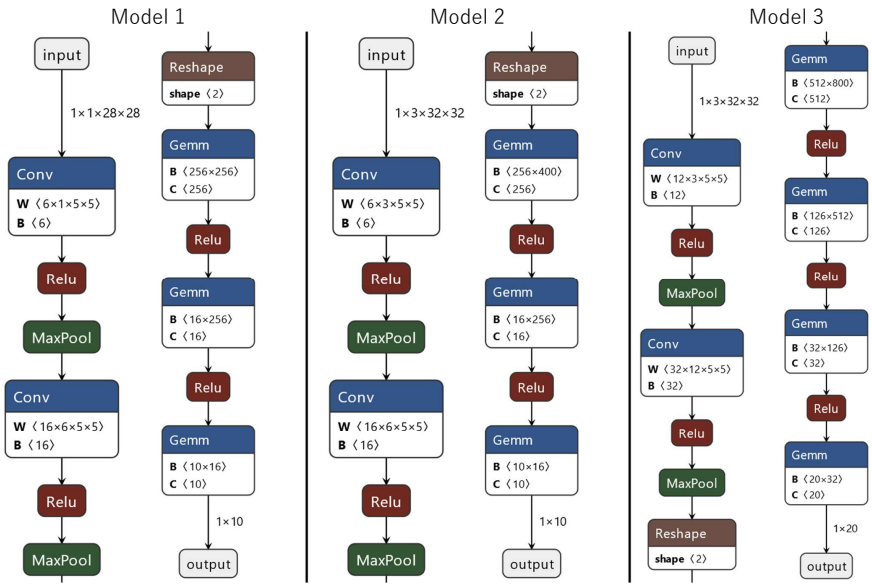

図 5.1　実験に使用したニューラルネットワークの図．各モデルを使用したデータセットは以下の通り．Model1：MNIST, FashionMNIST, Model2：SVHN, CIFAR10，Model3：CIFAR100

した．

作成したデータセットは $E^3Outlier$ の論文に倣い，inlier（正常値）となるクラス（0〜9（※ CIFAR100 は 0〜19））, 誤り率（ρ =5 %〜25 %で 5 %刻み），試行回数 5 回分の計 1500 パターンのデータセットの作成を行った．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，誤り方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．実験条件の詳細は 5.3 節にて述べる．

手順 2 事前学習のモデルを作成する．

図 5.1 に示したモデルを使用し，各データセット用に事前学習を行ったモデルを作成する．MNIST，FashionMNIST には Model1，SVHN，CIFAR10 には Model2，CIFAR100 には Model3 を使用した．

事前学習に使用するデータセットは手順 1 で作成したデータセットである．学習の内容は MNIST であれば手書き文字の数字毎にクラスを分ける，SVHN であればカラー画像にある数字毎にクラスを分けるといった，データセット本来の目的であるタスクを学習する．

手順 3 事前学習を行ったモデルを使用して，意図的な過学習を起こし，最終層の重みを保存する．

事前学習を行ったモデルを使用し，外れ値検出を行いたいクラスのラベルがついているデータを分類可能となるまで学習させる．このとき，最終層以外の勾配計算は止めて学習を行う．理由は 4 章で述べた，事前学習で学習したモデルの表現空間を変えない為である．

そして，分類可能となった時点での重みを保存する．

手順 3 について具体的な例を挙げて説明する．0 とラベルが振られているクラスについて外れ値検出をしたい場合，

(1) 事前学習のモデルを読み込む．

(2) 読み込んだモデルの最終層以外の勾配計算を止める

(3) ラベルが 0 であるデータを 1 つ取り出す．

(4) その 1 つのデータを分類可能となるまで何度も学習させる．

(5) 分類可能となった時点の最終層の重みを保存する．

(1)〜(5) をラベルが 0 であるデータがなくなるまで繰り返す．

という処理を行う．

手順 4 手順 3 で保存した最終層の重みを用いて外れ値検出を行う．

k-means でクラスタリングを行い，誤りの傾向で分類する実験を行う．

手順 5 手順 4 のクラスタリング結果，ユークリッド距離，マハラノビス距離に基づいたスコアによる外れ値検出を行う．

5.2 データセット

本節では実験に使用したデータセットについて，それぞれ簡単にデータの内容を説明する．

MNIST： 0 から 9 の手書き文字認識用のテストデータを含めて 70000 件の画像からなるデータセットである．

FashionMNIST： ファッション画像のデータセットである．MNIST と同様に，テストデータを含めて 70000 件のデータがある．

SVHN： ストリートビューでの家の番地の画像のデータセット．テストデータを含めて 99289 件のデータがある．extra データとして 531131 件のデータがあるが，本論文で行う実験では使用しなかった．

CIFAR10： 乗り物や動物などの物体のカラー画像のデータセット．カラー画像認識の研究などでよく使用されている．

CIFAR100： CIFAR10 と同様に物体のカラー画像のデータセット．CIFAR10 が 10 クラスのデータセットであるのに対し，CIFAR100 は 100 クラスのデータセットである．また，20 クラスの上位クラスも設定されており，本論文で行った実験では 20 クラスの上位クラスを利用した．

5.3 実験条件

MNIST，FashionMNIST はテストデータを含む 70000 件のデータ，SVHN はテストデータを含む 99289 件のデータ，CIFAR10，CIFAR100 はテストデータを含む 60000 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 5.1 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，Accuracy 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い，外れ値検出のためのスコアは 4.4.2 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.

5.4 実験結果

5.4.1 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.

表 5.1　MNIST の実験で出来たクラスタの Accuracy(inlier:0)

ρ	ρ =0.05	ρ =0.10	ρ =0.15	ρ =0.20	ρ =0.25
True_label:0	**0.99**	**0.99**	**0.98**	**0.98**	**0.97**
True_label:1	0.99	1.00	1.00	1.00	0.99
True_label:2	0.98	0.99	0.99	1.00	0.99
True_label:3	0.99	0.99	1.00	1.00	0.99
True_label:4	0.99	0.99	1.00	1.00	1.00
True_label:5	0.99	0.97	1.00	0.99	0.99
True_label:6	0.97	0.99	0.99	1.00	0.99
True_label:7	0.99	1.00	1.00	0.99	0.99
True_label:8	0.98	0.99	1.00	1.00	0.99
True_label:9	0.99	0.99	0.99	0.99	0.99
can't_classified	0	0	0	0	0

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.

表 5.2　FashionMNIST の実験で出来たクラスタの Accuracy(inlier:0)

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

表 5.3　SVHN の実験で出来たクラスタの Accuracy(inlier:0)

ρ	ρ =0.05	ρ =0.10	ρ =0.15	ρ =0.20	ρ =0.25
True_label:0	**0.00**	**0.00**	**0.96**	**0.95**	**0.92**
True_label:1	0.00	0.00	0.70	0.77	0.78
True_label:2	0.00	0.00	0.83	0.87	0.93
True_label:3	0.00	0.00	0.76	0.87	0.81
True_label:4	0.00	0.00	0.83	0.84	0.87
True_label:5	0.00	0.00	0.80	0.72	0.84
True_label:6	0.00	0.00	0.79	0.80	0.85
True_label:7	0.00	0.00	0.90	0.86	0.85
True_label:8	0.00	0.00	0.83	0.86	0.79
True_label:9	0.00	0.00	0.75	0.79	0.76
can't_classified	5	5	4	4	4

MNIST，FashionMNIST，SVHN を使用した実験でクラスタリングを行った時のクラスタの Accuracy を求めた平均を表 5.1 から表 5.3 に示す．クラスタのラベルが 0〜9 に分かれた時の Accuracy を求めたものである．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る MNIST，SVHN でクラスタリングを行った時の一つのクラスタから 25 枚ずつランダムで表示した画像を図 5.2 と図 5.3 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

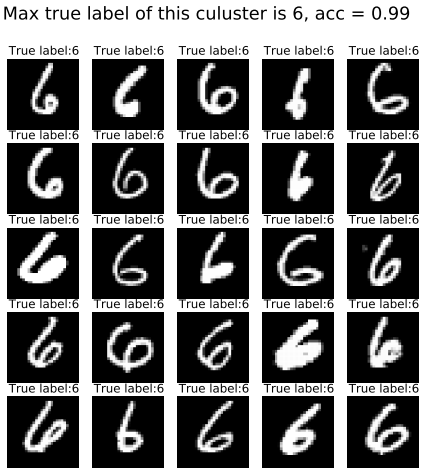

図 5.2　クラスタリングの結果
(MNIST，cluster:0)

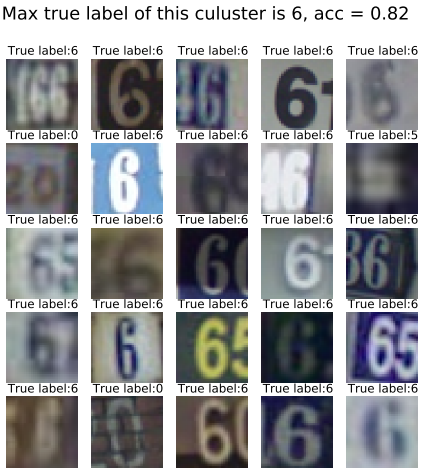

図 5.3　クラスタリングの結果
(SVHN，cluster:0)

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．

表 5.4　実験結果 (MNIST) AUROC（%）（$E^3Outlier$ の値は論文より引用）

	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
ρ =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
ρ =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
ρ =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
ρ =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
ρ =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

5.4.2　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 5.4 から表 5.8 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

表 5.5　実験結果 (FashionMNIST) AUROC（%）（$E^3Outlier$ の値は論文より引用）

	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
ρ =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
ρ =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
ρ =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
ρ =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
ρ =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

表 5.6　実験結果 (SVHN) AUROC（%）（$E^3Outlier$ の値は論文より引用）

	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
ρ =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
ρ =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
ρ =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
ρ =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
ρ =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

表 5.7　実験結果 (CIFAR10) AUROC（%）（$E^3Outlier$ の値は論文より引用）

	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
ρ =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
ρ =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
ρ =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
ρ =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
ρ =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

表 5.8　実験結果 (CIFAR100) AUROC（%）（$E^3Outlier$ の値は論文より引用）

	acc, cluster	acc, euclid	acc, mahala	loss, cluster	loss, euclid	loss, mahala	$E^3Outlier$
ρ =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
ρ =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
ρ =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
ρ =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
ρ =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

回実験を行った平均値 (μ) ±振れ幅（|| μ − μ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 5.4 から表 5.8 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.

5.5 考察

表 5.9 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

5.5.1 誤りの傾向で分ける手法についての考察

表 5.10 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 =0.05	5/5	5/5	1/5	0/5

実験の結果から，誤りの傾向で分かれた割合を表 5.10 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のではないかと考えている．そう考えた理由は，図 5.3 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．つまり 6 に対して 7 のような形が似ていないものに関しては別のクラスタに分類されているといえる．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 5.1 から表 5.3 をみてわかるように，データセットの分類難易度が上がるにつれて徐々に分類できなくなっていることがわかる．原因として考えられるのは，事前学習におけるモデルの精度の差である．それぞれのデータセットにおける事前学習の Accuracy は MNIST, FashionMNIST, SVHN, CIFAR10, CIFAR100 の順で高い（表 5.9）．つまり，モデルの表現空間をうまく学習できているものほどクラスタリングを用いて誤りの傾向で分けることができているといえる．実際に，表 5.9 の事前学習終了後の Accuracy の値が低いデータセットの実験ほど誤りの傾向で分かれにくいことがわかる．

以上より，外れ値を誤りの傾向で分けることは可能であるが，事前学習に使用する学習モデルの性能に依存すると考えられる．MNIST の実験においては 25 %外れ値が存在する場合でも inlier のクラスのクラスタが 97 %，他のクラスが 99 %程度の Accuracy の値が得られた．概算ではあるものの，約 90 %の外れ値を誤りの傾向で分けることができていることを確認した．

外れ値の割合はクラスタリングに影響しないと考えていたが，外れ値の割合が高いほど，クラス毎に分かれやすいという傾向があることがわかった．これは，クラスタのラベルを一番多い正解ラベルで決めているからだと考えられる．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やすいためであると考えられる．

基本的に割合が少ない外れ値の検出において，外れ値の割合が多いほど分け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したかったため，クラスタ数を指定できる k-means を用いたが，inlier となるデータ数が多いため，外れ値のクラスタに inlier であるデータが入り込む可能性がある．

そのため，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ことも考えられる．他のクラスタリング手法を用いてクラスタリングをしたり，外れ値検出をしたあとで外れ値をクラスタリングで分けたりすることで提案手法よりも精度良く誤りの傾向毎に分かれる可能性がある

と考えている．

5.5.2　外れ値検出を行うことについての考察

表 5.4 から表 5.8 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．

個別の学習は 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．そのため，類似するデータは類似する境界の形になることが考えられる．対して，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．そのため，類似するデータであってもそのデータごとに特化した境界となってしまい，類似する境界の形にはならないことが考えられる．以上の違いにより，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，クラスタを用いたスコアはクラスタのデータ群ごとにスコアが付くため角ばった AUROC を描くため，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の単純な距離を用いる．そのため，正常値の分布がどの次元においても円形でない限り外れ値のスコアの方が高くなってしまう可能性がある．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ものであり，外れ値検出において正常値は多数存在するため，外れ値の距離が大きくなる可能性が非常に高い．よって，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，外れ値検出の実験においても，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．表 5.4 から表 5.8 の結果と表 5.9 の事前学習終了後の

Accuracy の値は相関があるように見える．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．

5.5.3　全体の考察

クラスタリングを行う実験と外れ値検出のみの実験のどちら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ことが考えられる．このことは 5.5.1 節や 5.5.2 節で述べたように表 5.9 とそれぞれの実験結果を見るとモデルの性能が低いほど，AUROC やクラスタ毎の Accuracy が低いことから見て取れる．

今回使用したモデルは ResNet や EfficientNet[7] といった精度が高いことで有名なモデルを使用していない為，使用するモデルを変更することにより，どちらの実験においてもさらなる精度の向上を図ることができると考えている．

第 6 章　おわりに

本論文では，公開されているデータセットに誤ったラベルが存在していることを問題視し，問題解決のための研究を行った結果を報告した．

同様の誤り方をしたデータ群であれば，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同様の境界になる．そして，決定境界を表現するための重みも同様の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．評価実験を行った結果，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という新しい特徴量は，誤りの傾向で分けることが可能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，学習済みモデルを 1 つのデータについて過学習させ，その最終層の重みを特徴量とする手法は有効であるといえる．学習済みモデルの性能が高ければ，クラスタリングすることで誤りの傾向毎に分類可能な特徴量であることが確認できた．よって，提案手法の特徴量はこのデータにこのラベルを付けるという特徴をとらえた特徴量であると考えられる．

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．5 章で述べたように，提案した特徴量は事前学習に使用した学習モデルの性能に依存する部分があると考えられる．実際に，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．そのため，他のデータセットを用いて実験を行い，同様の結果が得られるか試したいと考えている．また，ResNet[10] や EfficientNet[7] といった分類精度が高いことで知られているニューラルネットワークのモデルを用いることで，誤りの傾向毎に分ける能力が向上することが考えられる．そのため，事前学習に使用するモデルを変更し，同じ実験を行うことも考えている．

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 E3Outlier よりも精度が高い検出器を構築することができた．外れ値検出の性能に関しても，ResNet や EfficientNet といったモデルを用いて提案手法の精度

向上を図りたいと考えている．

今後の展望としては，MNIST だけでなく，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．MNIST の実験と同様の性能であれば，出力されたクラスタが本来どのラベルが付けられるべきだったかを見るだけで約 90 %の外れ値を正常なラベルに修正できることになる．これにより，誤ったラベルがついている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．

1 章で述べたように，誤ったラベルがついていると報告されたデータセットが存在する．それ以外のデータセットや，今後公開されるデータセットにも誤ったラベルがついている可能性がある．データセット自体に誤りがあることは 1 章で述べたように分類器の誤った判断の原因となる可能性がある．また，機械学習モデルの性能を測るために使用されるデータセットとしても不適切であると考える．提案手法を用いることで誤ったデータをできる限り少なくし，データセットの正確性を保証できるようなものになっていけば良いと考えている．

また，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[11] 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

本研究を進めるにあたって，指導教員である，鈴木優准教授に様々なご指導，ご助言を賜りました．やりたい研究課題から，関連する部分の研究を紹介していただいたり，まず手をつけるべき課題を示していただいたりと，研究を開始するにあたっての道筋を示していただいたと感じています．全国大会への論文投稿と本論文の２テーマの研究を行いましたが，どちらの研究においても，次の実験ではどういうことをしたら良いのか，どういうことが言えたら良いのかということを一緒になって考えてくださりました．

また，本研究を始めるにあたって，相談をしたところ，「よくわからない．」とおっしゃられ，どう伝えたら良いか悩んだことを覚えています．鈴木優准教授には私の考えが伝わるまで，何度も話を聞いてもらい，「理解したけど，それうまくいくの？」と言われながら，本研究がスタートしました．実験を重ねるにつれて「面白くなってきた」とおっしゃっていただけて，うまくいくと信じて実験を続けて良かったと思ったことを覚えています．結果として，最終的には 2019 年の SOTA 手法と張り合えるほどの手法となりました．

ただ，先生にも理解してもらえなかった手法を文章にすることは難しく，何度も面談を行っていただきました．文章にしたときに伝わるようにするにはどう説明したら良いのかという話を何度も何度もしていただき，本論文を書き終えることができました．本当にありがとうございました．

事務補佐員の井尾さん，佐野さんには，外部での研究発表を行う際の様々な手続きをするにあたり，お世話になりました．井尾さんには産休に入るまで，佐野さんには井尾さんが産休に入ったあとおせわになりました．お二人には，学会の時の食事や，３年生を交えての食事会などの用意などもしていただきました．

鈴木研究室に所属する皆さんには，普段のゼミで研究内容について意見をいただいたり，実験を行うにあたって用意するデータの作成を手伝っていただいたりしました．

本論文を書き終えることができたのは，皆様が支えてくださったおかげです．心より感謝申し上げます．

最後に、本研究に限らず，大学生活の４年間にわたって経済的・心身的に支えて

下さった家族に深く感謝し、お礼を申し上げます。

参考文献

[1] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. 2021.

[2] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019.

[3] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9664–9674, June 2021.

[4] G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, Vol. 313, No. 5786, pp. 504–507, July 2006.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

[7] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[8] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. *CoRR*, Vol. abs/2005.14140, , 2020.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks

for large-scale image recognition. *CoRR*, Vol. abs/1409.1556, , 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, Vol. abs/1512.03385, , 2015.

[11] Yizhou Wang, Can Qin, Rongzhe Wei, Yi Xu, Yue Bai, and Yun Fu. Sla^{2}p: Self-supervised anomaly detection with adversarial perturbation. *CoRR*, Vol. abs/2111.12896, , 2021.

発表リスト

[1] 三島惇也，鈴木優『ツイートを用いたユーザの持つバイアスの推定』，第 20 回情報科学技術フォーラム，2021
[2] 三島惇也，鈴木優『意図しないバイアスを持った機械学習モデルの修正方法の検討』，東海関西データベースワークショップ，2021
[3] 三島惇也，鈴木優『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』，第 14 回データ工学と情報マネジメントに関するフォーラム，2021