生成モデルを活用した識別モデルの分布外データ検出手法に関する研究

1X17C120-4　松苗亮汰
指導教員　後藤正幸

1. 研究背景と目的

近年の深層学習技術の発展に伴い，画像識別モデルの性能は人間を超える高い性能を示すようになった．しかし，識別の対象とされていないクラスのデータが混入した場合，既知のいずれかのクラスに高い所属確率を付与してしまうという問題がある．識別モデルが対象とする分布以外から生起したこのようなデータを，分布外データと呼ぶ．それに対して，識別モデルが対象とするデータを分布内データと呼ぶ．実応用で識別モデルを活用する上で，分布外データの検出（以下，分布外検出）は重要なタスクとなっている．

分布外検出のため，生成モデルを利用した研究が進められている．これらの研究では，分布内データで学習した生成モデルで評価したデータの尤度を検出指標として用いる．しかし，分布外データに対して分布内データよりも明らかに高い尤度が算出されてしまう現象が確認され [1]，尤度による分布外検出の問題点となっている．この問題に対して，Ren らは尤度比を用いた分布外検出手法 [2] を提案した．この手法では，二つの生成モデルを用いることでデータの尤度比を算出し，これを検出指標とする．これにより，画像データの背景部分の情報（以下，背景情報）を無視し，意味のある情報（以下，有意情報）のみを評価することを試みている．その結果，単純な尤度比較よりも高い検出精度を示している．

生成モデルを用いた分布外検出において，生成モデルが学習したモデル分布は，分布内データの真の分布を精度よく近似していることが求められる．一方，分布内データは識別クラスごとに異なる分布に従うと想定され，各クラスの分布構造は，全クラスをまとめた分布構造よりもシンプルであると考えられる．したがって，従来手法のように全クラスの分布構造をまとめて一つの生成モデルで推定するのではなく，クラスごとに推定することで，推定精度の向上が期待される．

そこで本研究では，クラスごとに独立に生成モデルを学習することによる，精度の高い分布外検出手法を提案する．具体的には，分布内データを学習した生成モデルとノイズを加えた分布内データを学習した生成モデルをクラスごとに用意し，それらから算出されるクラスごとの尤度比を統合することで，新たな検出指標を作成する．さらに分布内データと分布外データとしてそれぞれ，画像のベンチマークデータセットである FashionMNIST と MNIST を用いて，提案手法の有効性を示す．

2. 準備

2.1. 問題設定

分布外検出とは，C クラス識別モデルの学習に用いられた分布内データに対して，分布外データを検出するタスクである．識別モデルの学習データを $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ とすると，本研究における分布外検出モデルは，$\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ を用いて学習される．ここで $\boldsymbol{x}_n$ は特徴量，y_n は $\boldsymbol{x}_n$ の所属クラスを表す．運用時にはこのモデルを用いて，新規の入力データ $\boldsymbol{x}$ に対して検出指標を算出し，それが閾値 ϵ を下回るならば分布外データとして検出する．

2.2. 生成モデル

生成モデルとは，学習する特徴量 $\boldsymbol{x}_n \in \mathcal{X}$ の生起する真の分布 $p^*(\boldsymbol{x})$ を，モデル分布 $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ で近似したモデルである．これにより新たなデータの生成や，入力データのモデル分布からの逸脱度が算出できる．ここでパラメータ $\boldsymbol{\theta}$ は，以下に示す $\mathcal{X}$ の対数尤度 L を最大化するように推定される．

$$L = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \tag{1}$$

2.3. 従来研究

尤度による分布外検出

尤度による分布外検出ではまず，分布内データで学習した生成モデルを用意する．そしてこのモデルに対してデータが与えられたときの尤度を，分布内データへの当てはまり度合いと捉え，分布外検出指標とする．しかし，これまでの研究において，単純な画像や背景割合の高い画像が分布外データとして混入したときに分布内データよりも明らかに高い尤度が算出されてしまい [1]，分布外として検出できないという問題点が指摘されている．

尤度比による分布外検出

Ren らの尤度比による分布外検出 [2] は，データにおける背景情報を無視して有意情報のみを検出指標に反映させることを試みている．そのために，以下に示す二つの仮定が置かれている．一つ目は，データ $\boldsymbol{x}$ は有意情報 $\boldsymbol{x}_S$ と背景情報 $\boldsymbol{x}_B$ に分割でき，それらは独立に生起するという仮定である．この仮定により，次の式 (2) が成立する．

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_S)p(\boldsymbol{x}_B) \tag{2}$$

二つ目は，データに一定割合でノイズを加えることで有意情報は破壊される一方，背景情報はほぼ破壊されないという仮定である．ここで，学習データ $\mathcal{X}$ で学習したパラメータ $\boldsymbol{\theta}$ の生成モデルと，ノイズを加えた学習データ $\tilde{\mathcal{X}}$ で学習したパラメータ $\tilde{\boldsymbol{\theta}}$ の生成モデルを考える．二つ目の仮定より，これら二つのパラメータが学習したデータの背景情報は等しいと言えることから，以下の式 (3) で表される近似が成立する．

$$p(\boldsymbol{x}_B \mid \boldsymbol{\theta}) \simeq p(\boldsymbol{x}_B \mid \tilde{\boldsymbol{\theta}}) \tag{3}$$

以上の二つの仮定のもとで，$\mathcal{X}$ で学習した生成モデルと $\tilde{\mathcal{X}}$ で学習した生成モデルで，対象データ $\boldsymbol{x}$ の尤度比 $\mathrm{LLR}(\boldsymbol{x})$ を算出する．その結果，式 (4) に示すように有意情報のみを評価することが可能となるため，これを検出指標とする．

$$\mathrm{LLR}(\boldsymbol{x}) = \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})}{p(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}})} \simeq \log \frac{p(\boldsymbol{x}_S \mid \boldsymbol{\theta})}{p(\boldsymbol{x}_S \mid \tilde{\boldsymbol{\theta}})} \tag{4}$$

3. 提案

3.1. 提案への着想

従来の分布外検出手法は，全クラスの分布内データの生成分布を一つの生成モデルで推定している．しかし識別モデルの分布内データはクラスごとに分布構造が異なると想定されるため，クラスごとに別の生成モデルで推定する方が，精度の良い推定が可能だと考えられる．そこで，クラスごとに学習した生成モデルでそれぞれ尤度比を算出し，それらを統合することで新たな検出指標を提案する．

3.2. 提案手法

クラス c に属する学習データ $\mathcal{X}_c = \{\boldsymbol{x}_{cn}\}_{n=1}^{N_c}$ で学習した生成モデルのパラメータを $\boldsymbol{\theta}_c$，ノイズを加えたクラス c の学習データを $\tilde{\mathcal{X}}_c = \{\tilde{\boldsymbol{x}}_{cn}\}_{n=1}^{N_c}$，$\tilde{\mathcal{X}}_c$ で学習した生成モデルのパラメータを $\tilde{\boldsymbol{\theta}}_c$ とする．これらを用いて式 (5) のように，入力データ $\boldsymbol{x}$ の各クラスにおける尤度比を算出する．

$$\mathrm{LLR}_c(\boldsymbol{x}) = \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_c)}{p(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}}_c)} \quad (c = 0, \ldots, C-1) \tag{5}$$

クラス c の生成モデルで算出した尤度比 $\mathrm{LLR}_c(\boldsymbol{x})$ は，入力データ $\boldsymbol{x}$ の有意情報がどれだけクラス c らしいかの指標となっている．入力 $\boldsymbol{x}$ が分布内データである場合，この指標は所属するクラスで最大値を取ることが期待される．そこで，クラスごと尤度比の最大値を検出指標とする．

ここで，クラス c の生成モデルで算出した，クラス c に所属する分布内データの尤度比 $\mathrm{LLR}_c(\boldsymbol{x}_c)$ を考える．クラスごとにこの尤度比のスケールが異なると，比の値を単純には比較できなくなるため，所属するクラスでの尤度比が最大でなくなる可能性が生じる．そこで，クラスごとに尤度比の標準化を行う．標準化に用いる平均と標準偏差の計算には，$\mathcal{X}_c$ の尤度比を用いる．以上より，本研究で提案する検出指標 $\mathrm{mLLR}(\boldsymbol{x})$ は式 (6) のように表され，この指標が閾値 ϵ を下回るデータを分布外として検出する．

$$\begin{cases} \mathrm{mLLR}(\boldsymbol{x}) = \max\limits_c \dfrac{\mathrm{LLR}_c(\boldsymbol{x}) - \mu_c}{\sigma_c} \\ \mu_c = \dfrac{1}{N_c} \displaystyle\sum_{n=1}^{N_c} \mathrm{LLR}_c(\boldsymbol{x}_{cn}) \\ \sigma_c = \sqrt{\dfrac{1}{N_c} \displaystyle\sum_{n=1}^{N_c} \{\mathrm{LLR}_c(\boldsymbol{x}_{cn}) - \mu_c\}^2} \end{cases} \tag{6}$$

4. 評価実験

4.1. 実験条件

提案手法の有効性を示すため，5 分割交差検証法で分布外検出実験を行った．分布内データとして衣料品の画像データセットである FashionMNIST を，分布外データには手書き数字の画像データセットの MNIST を用いる．FashionMNIST は 10 クラスで構成されているため，分布外検出の前提となる識別モデルには 10 クラス識別モデルを用意する．つまり，提案手法におけるクラス数 C は 10 である．本実験では，学習データを各クラス 5,600 枚の計 56,000 枚，テストデータを各クラス 1,400 枚の計 14,000 枚とした．MNIST のテストデータは 10,000 枚とし，クラスを考慮せずまとめて分布外データとして扱う．また，深層生成モデルには，フローベース生成モデルの Real NVP[3] を使用する．フローベース生成モデルとは，データに対して可逆な写像を用いて変数変換を繰り返し，既知のベース分布へ射影することで厳密な尤度を計算する深層生成モデルである．

評価には AUC を用いる．提案手法に対する比較手法は，尤度を検出指標としたもの（以下，尤度法）と，式 (4) の尤度比を検出指標としたもの（以下，尤度比法）とした．

4.2. 結果と考察

表 1 に，5 分割交差検証による各手法の平均 AUC と標準偏差，各生成モデルの 1 エポックあたり平均学習時間を示す．

表 1: 各手法の AUC 平均（括弧内は標準偏差）と各生成モデルの 1 エポックあたり平均学習時間

手法	AUC	学習時間/epoch[秒]
尤度法	0.1053(0.0027)	678.2
尤度比法	0.7026(0.1433)	678.2
提案手法	**0.8385**(0.0902)	67.9

三つの手法を比較すると，提案手法が最も高い精度を示している．したがって，クラスごとに生成モデルを学習することで，より正確に分布内データの真の分布を捉えることに成功していると考えられる．加えて，提案手法における一つの生成モデルの学習にかかる時間計算量は，学習データをクラスごとに分割しているため，尤度比法の $1/C$ となる．クラスごとの生成モデルの学習は独立しており，並列計算が可能なため，提案手法は尤度比法と比較して時間計算量を $1/C$ に抑えることができる．

以上により，提案手法はクラスごとに生成モデルを学習することで，従来の尤度比法よりも時間計算量を抑えつつ，高い分布外検出の精度が得られることを確認した．

5. まとめと今後の課題

本研究では，従来の尤度比法を拡張したより性能の高い分布外検出手法の構築を目的として，識別クラスごとに生成モデルを学習することによる，新たな分布外検出手法を提案した．さらに，評価実験として画像データセットを用いた分布外検出を行うことで，検出精度と時間計算量の両面からその有効性を示した．今後の課題としては，クラス数の変化に対する性能の確認や，カラー画像データセットへの適用，フローベース生成モデル以外の深層生成モデルの利用等が挙げられる．

参考文献

[1] E.Nalisnick, A.Matsukawa, Y.Teh, et al, “Do Deep Generative Models Know What They Don’t Know?,” *International Conference on Learning Representations*, 2018.

[2] J.Ren, P.Liu, E.Fertig et al, “Likelihood Ratios for Out-of-Distribution Detection,” *Neural Information Processing Systems*, 2019.

[3] L.Dinh, J.Sohl-Dickstein and S.Bengio, “Density Estimation Using Real NVP,” *International Conference on Learning Representations*, 2016.