2U-05

環境音分類のためのラベルなしデータを活用したデータ拡張

庄司真都† 長尾智晴‡

横浜国立大学 大学院環境情報学府† 横浜国立大学 大学院環境情報研究院‡

1. はじめに

近年，身の回りで起こる様々な音を識別する環境音分類が研究され，ホームモニタリングや音監視システム，動画の自動タグ付けなど多くのアプリケーションでの応用が期待されている．環境音を分類する手法として深層学習を用いた手法が多く提案されている．深層学習には大量の学習データが必要であるため，環境音データはアノテーションコストが高く，ラベル付けされた環境音データのサンプル数が少ないことが課題である．一方で，ラベル付けのされていない環境音データは大量に存在する．また，多くの研究ではノイズを重畳するなどの加工を行いデータの水増しを行っている．データ加工によるデータ拡張は頑健性を高めるには有効だが，データの本質自体が変わっていないため，未知データに対する効果は低い．

そこで本稿ではラベル付けのされていない大量の環境音データを活用し，データ数の増加およびデータセットの多様性を向上させる環境音のためのデータ拡張の手法を提案する．

2. 提案手法

2.1. 概要

提案手法の流れを図 1 に示す．学習データと近い特徴をもったデータをラベルなしデータセットから探索する「近い特徴をもつデータの探索」と分類器に有効なデータだけを取り出すための「分類器を用いたデータ選別」の 2 段階に分かれている．

2.2. 近い特徴をもつデータの探索

音声データの音響特徴量を抽出し，その類似度から学習データと近い特徴をもつデータをラベルなしデータセットから探索する．音響特徴量

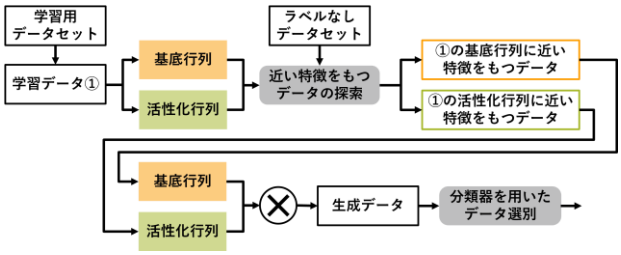

図 1. 提案手法の流れ

抽出は，まず音声データに対して短時間フーリエ変換を行いスペクトログラムに変換し，さらに非負値行列因子分解を用いて基底行列と活性化行列に分解する．次に学習済みの畳み込みオートエンコーダのエンコーダを用いて次元削減を行う．最後に UMAP[1]を用いて音響特徴量を低次元化し，ユークリッド距離から近い特徴をもつデータを探索する．

2.3. 分類器を用いたデータ選別

学習データの基底行列，活性化行列に対して，それらに近い特徴をもつ基底行列，活性化行列をラベルなしデータセットから取り出し，その行列積を求め，データを生成する．生成されたデータに対して，学習用のデータセットで学習された分類器を用いて，分類精度のしきい値を設定し分類器に有効なデータだけを選別する．

3. 実験設定

3.1. データセット

本手法で用いるラベルなしデータセットはラベル付けのされていない 250,000 個のファイルが収録されたデータセットである ESC-US[2]を用いる．また，学習および評価用のデータセットは 3 種類ある．1 つ目は 50 種類の環境音が各クラス 40 個ずつあり，計 2,000 個のファイルが収録されたデータセットである ESC-50[2]を用いる．2 つ目は 10 種類の環境音が 1 クラスあたり 1,000 個を上限とし，計 8,732 個のファイルが収録されたデータセットである UrbanSound8K[3]を用いる．3 つ目は 10 種類の音響シーンが各

Mixup Data Augmentation with Unlabeled Data for Environmental Sound Classification
†Manato Shoji ‡Tomoharu Nagao
†Graduate School of Environmental and Information Sciences, Yokohama National University
‡Faculty of Environmental and Information Sciences, Yokohama National University

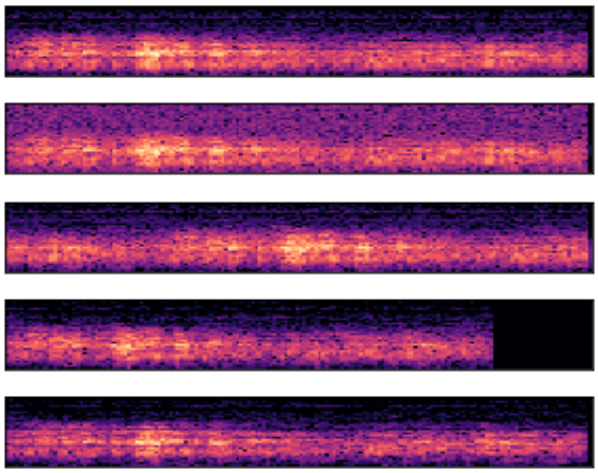

図 2. 従来のデータ拡張

864 個ずつ，計 8,640 個のファイルが収録されたデータセットである TUT Urban Acoustic Scenes 2018 development dataset[4]を用いる．

3.2. 従来手法

音声データの一般的なデータ拡張を図 2 に示す．上から順に，データ拡張なし，White noise, Time shift, Time stretch, Pitch shift を適用したスペクトログラムである．White noise は音声データにノイズを加える． Time shift は時間軸方向をシフトさせる．Time stretch は時間軸方向を伸縮させる．Pitch shift は周波数軸方向をシフトさせる．従来手法では，これら 4 つの手法を適用し，学習データを 4 倍に水増しする．

3.3. 評価方法

評価は ResNet50[5]を用いて，Accuracy，F1-Score によって精度評価を行った．

4. 結果と考察

各データセットに対する実験結果を表 1 に示す．Baseline はデータ拡張を行っていない場合，Augmentation は従来のデータ拡張を行った場合，Proposed は提案手法によるデータ拡張を行った場合である．ESC-50 と UrbanSound8K では Proposed の分類精度が Baseline を上回り，Proposed+Augmentation が一番高い分類精度となった．したがって，提案手法の有効性が示されたと言える．また，TUT では Augmentation が一番高い精度となり，Proposed は Baseline を下回り提案手法によって悪化していることがわかる．TUT は背景音が録音されているデータセットであるが，本手法で使用しているラベルなしデータセットは環境音が録音されているため，背景音を生成することが難しかったと考えられる．

5. まとめ

本稿では，ラベル付けされていないデータセットを活用した環境音のためのデータ拡張の手法を提案し，精度評価によって有効性を確認した．今後は分類精度が悪化したクラスを含め，分類精度の向上に有効な生成データの改善を行う．

表 1. 実験結果

Dataset	Method	Accuracy	F1-Score
ESC-50	Baseline	0.6716	0.667
	Augmentation	0.7866	0.7822
	Proposed	0.8091	0.8067
	Proposed+ Augmentation	**0.8525**	**0.8498**
UrbanSound8K	Baseline	0.9324	0.9382
	Augmentation	0.9621	0.9648
	Proposed	0.9484	0.9513
	Proposed+ Augmentation	**0.9623**	**0.9650**
TUT	Baseline	0.7774	0.7774
	Augmentation	**0.8550**	**0.8540**
	Proposed	0.6606	0.6647
	Proposed+ Augmentation	0..7901	0.7873

参考文献

[1] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Groberger. Umap: Uniform manifold approximation and projection. Journal of Open Source Software, Vol. 3, No. 29, p. 861, 2018.

[2] Karol J. Piczak. Esc: Dataset for environmental sound classification. Proceedings of the 23rd ACM international conference on Multimedia, 2015.

[3] Justin Salamon, Christopher Jacoby, and Juan Bello. A dataset and taxonomy for urban sound research. Proceedings - 22nd ACM International Conference on Multimedia, 11 2014.

[4] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. In Mark D. Plumbley, Christian Kroos, Juan P. Bello, Gal Richard, Daniel P. W. Ellis, and Annamaria Mesaros, editors, Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2018, Surrey, UK, November 19-20, 2018, pp. 9–13, 2018.

[5] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.