音響シーン分類におけるシーン抽象度によるクラスタリングを適切に行うためのクラス数の増減と分類精度の分析

Analysis of the Increase and Decrease in the Number of Classes and Classification Accuracy for Proper Clustering Based on Scene Abstraction Level in Acoustic Scene Classification

岸本 康汰
Kota Kishimoto

岡山大学 原研究室
Hara Laboratory, Okayama University

概要 本研究では，音響シーン分類において，シーン抽象度に合わせて分類クラス数を調整することを目指す. 本報告では，与えられた音データに対してクラス数を増減させ，それぞれの場合でシーン分類をおこない，その結果をもとにクラス数の変化と分類の識別精度の関係を分析した．結果から，クラス数を 2 から 6 まで増加させた場合は識別精度が徐々に低下したが, クラスを 7 以上に増やした場合, 識別精度には大きな差が見られなかった.

1 はじめに

近年では，様々な種類の音を対象とした環境音分類の検討が進められており，その中でも特に取り組み事例が多いタスクに音響シーン分類がある. 音響シーン分類は，あらかじめ決められたクラスの中から，入力信号を最もよく表す音響シーンを 1 つだけ推定するタスクである. ここで，音響シーンとは音が収録された場所や環境，周囲にいる人の行動を指す.[1]

分類タスクにおいてクラス数を増加させることはより細かな音響シーンを推定することに繋がるが，シーンの抽象度がより低くなり，分類を行う難易度が高くなってしまう．一方で，クラス数を減少させることは，より抽象度の高い分類をおこなうことに繋がる. シーンの抽象度が高くなると細かな音響シーンを比較する場合と比べ分類を行う難易度が低くなるがより詳細な識別を行うことができなくなってしまう.

本報告では，音響シーン分類を行う際にクラスの数を増減させた際の識別精度の変化をもとに，クラスの抽象度が高すぎず，分類精度が一定以上得られる際のクラス数の分析をおこなった.

2 提案方式

提案方式では，学習データを用いて学習した識別機に評価データを与え，適切にシーン分類がおこなえるクラス数の分析を行った．データの学習では，それぞれのシーンのメルスペクトログラムを音響特徴量としてニューラルネットワークに学習させ，それにより得られる出力データと正解ラベルとの差を用いて，繰り返しニューラルネットワークの各層の重みを更新するという作業をおこなった．クラス数の分析ではボトムアップのアプローチを用いた. 入力による出力と正解ラベルとの関係を混同行列 (Confusion Matrix) を用いて表し，得られた混同行列の中から誤りの分布が似ている 2 クラスの結合を行うことで 1 つクラスを減少する操作をおこなった．この操作をクラス数が 2 になるまで階層的におこない，識別精度との関係を調べた．このとき，クラス数が減少するにつれ，識別精度が上昇していくことを期待している.

3 実験条件

3.1 データセット

本実験で使用したデータセットは，TAU Urban Acoustic Scenes 2022 モバイル開発データセット [2] である. 本データセットは，10 個の音響シーン

- airport
- shopping_mall
- metro_station
- street_pedestrian
- public_square
- street_traffic
- tram
- bus
- metro
- park

からの 10 秒のオーディオセグメントが，重複しない 1 秒間のセグメントに分割され構成されている．データセットはヨーロッパの 12 都市のデータが記録されたものであるが，今回の実験ではバルセロナで録音された 10 個の音響シーンのみを用いて，学習および評価を行った.

3.2 学習条件

実験で用いる各モデルの学習は，44.1 kHz でサンプリングされたデータセットを用いておこなった. 学習に用いたメルスペクトログラムのパラメータを表 1 に示す. また，モデルのうち 14490 個を訓練データに，3580 個を評価データに割り当てた. 学習は，畳み込みニューラルネットワーク (CNN) を用いて行い，損失関数は

表 1: メルスペクトログラムのパラメータ

パラメータ	値
n_mels	80
sample_rate	44100
n_fft	4096
hop_length	441

表 2: CNN の構成 (N: バッチサイズ)

Layer type	output size	ksize	stride
input	$N \times 1 \times 80 \times 91$		
L1 convolution	$N \times 32 \times 25 \times 88$	(6,4)	(3,1)
L2 convolution	$N \times 64 \times 1 \times 87$	(3,2)	(2,1)
L3 max pooling	$N \times 64 \times 4 \times 14$	(3,6)	
L4 convolution	$N \times 128 \times 2 \times 12$	(3,3)	(1,1)
L5 average pooling	$N \times 128 \times 2 \times 1$		
L6 flatten	$N \times 256$		
L7 Linear	$N \times 10$		

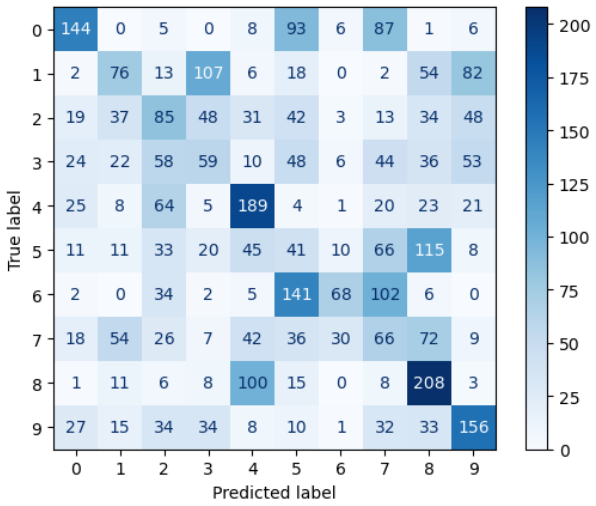

図 1: クラス数 10 の場合の混同行列

交差エントロピー誤差関数 (Cross entropy Loss) を用いた．CNN のハイパーパラメータを表 2 に示す．音響特徴量の抽出には，torchaudio (version:2.3.1) を利用し，モデルの構築と学習に pytorch (version:2.3.1) を使用した．

4 評価実験

図 1 にクラス数 10 の場合，評価用データを入力した場合の出力と期待される出力の混同行列を例に示す．このときのシーン分類精度は 35% を示した．得られた混同行列から類似性の高いクラスをマージし，クラス数を 1 ずつ減らした際の識別精度を図 2 の青い実線に示す．結果から，クラス数が増えると識別精度が低下していることがわかる．また，結果をもとに得られたクラスタリングの階層構造を図 3 に示す．図 2，図 3 から，

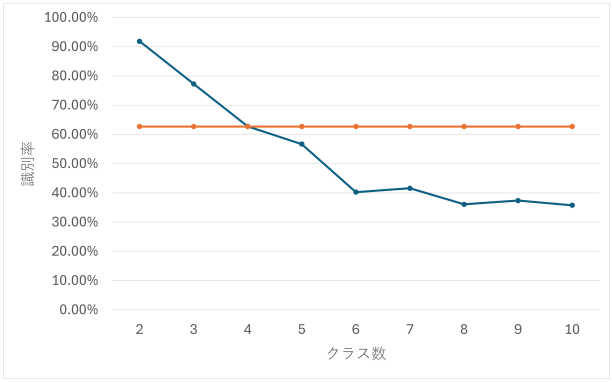

図 2: クラス数と識別精度　青は実験結果，オレンジは最高性能の Accuracy (62.7%) を示す

クラス数が 5 以下になると識別精度が上昇していることが分かる．今回のデータセットでは，クラス数 6 以上の場合において識別を行う難易度が高い，ということが分かる．

5 考察

実験結果から，音響シーン分類においてクラス数を上げることは識別精度の低下につながることが示された．そのため識別においてより具体的なシーンを分類するためには，どの程度のデータ量があればクラス数を増やした場合でも識別精度を低下させずに分類できるのかについて検討する必要があると考えられる．具体的には，学習データ数を本報告の場合より増やした場合，および減らした場合においてクラス数と識別精度の関係を調査することで，学習データ数に関係したシーン抽象度を分析することができるのでないかと考えられる．

また，今回クラスタリングを行う際に，得られた混同行列をもとに誤りの類似度の高いクラスを選別するという規則でマージをおこなったが，その他の規則に基づいてマージするクラスを選別し，本報告とは異なる順序でクラスタリングを行うことで，識別精度に差が生じるのではないかと考えられる．

6 まとめ

本報告では，音響シーン分類におけるクラス数を変化させた際の識別精度の変化について分析した．実験結果から，今回使用したデータセットではシーンがより具体的になれば識別の精度が低下する，ということが示された．今後の課題として，データ量を変化させた際の識別精度を分析することがあげられる．また，クラスタリングの方法についても今回用いた規則と異なる規則を用いた手法の検討もおこないたい．

参考文献

[1] 井本 桂右，“音響イベントと音響シーンの分析，” 日本音響学会誌, 74 巻, 4 号，pp. 198–207, 2018.
[2] Toni Heittola, Annamaria Mesaros, and Tuomas

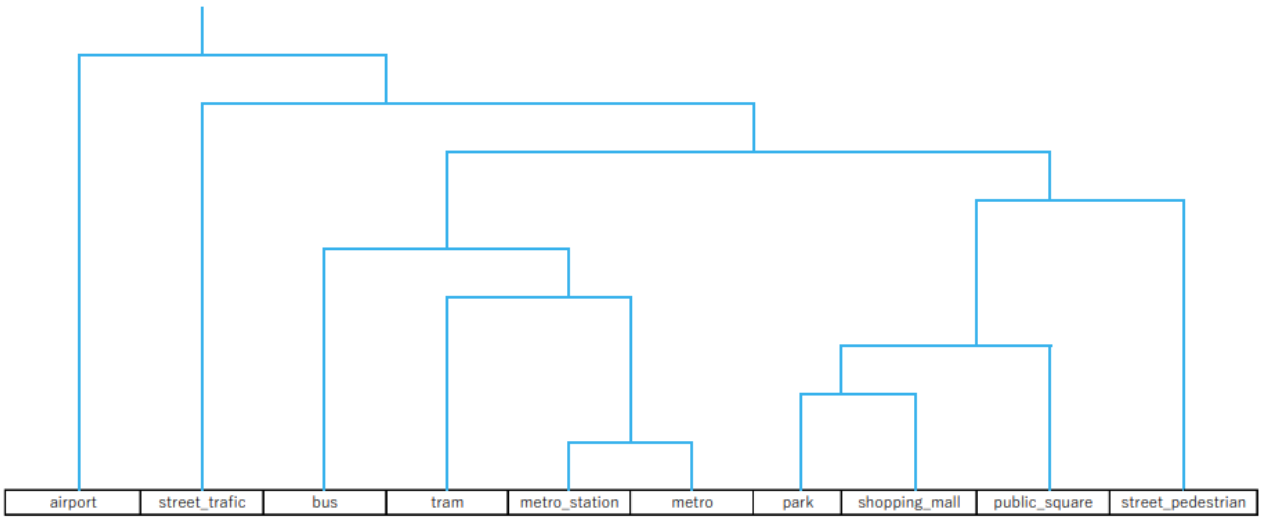

図 3: クラスタリングの結果

Virtanen. Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 56–60. 2020