ISS-SP-018

マルチレベル特徴集約を用いた自己教師あり学習による楽曲分類

高安　雅人† 　　長尾　智晴††
† 横浜国立大学理工学部 　　†† 横浜国立大学大学院環境情報研究院

1. はじめに

近年の音楽ストリーミングサービスの普及に伴い，検索や推薦のために，楽曲の特徴を表すタグを付ける手間が急増している．そうした楽曲へのタグ付けを自動化することを目的として，教師あり学習を用いた楽曲分類モデルの研究が盛んである．しかし，教師あり学習は大量のタグ付き学習データを作成する負担が大きい．そのため，本稿では教師あり学習と比べて学習データのタグ付けが少量で済む自己教師あり学習を用いた楽曲分類モデルを提案する．

2. 提案手法

楽曲分類のための自己教師あり学習の手法は CLMR(Contrastive Learning of Musical Representations) [1]の従来手法に基づき，次の手順で行う．まず 1 つの楽曲に対して異なる Data Augmentation を行い，2 種類の学習データを作成する．この Data Augmentation には楽曲の切り取り，極性反転，ノイズ付加，音量変化，特定音域の抽出，ディレイ付加，ピッチの変更，残響付加の 8 種類の処理が含まれる．これらの内，楽曲の切り取り以外の処理は確率的に実行する．次に SampleCNN[2]をエンコーダとして学習データを特徴量化し，隠れ層が 1 層の多層パーセプトロンを用いて特徴量を射影する．この射影について，同じ楽曲を基にした学習データの射影は類似度を上げ，異なる楽曲を基にした学習データの射影は類似度を下げるように学習を行う．この学習後，エンコーダを固定し，少量の正解タグを与えて多層パーセプトロンの学習を行う．

提案手法では，エンコーダの SampleCNN に対して，[3]で提案されたマルチレベル特徴集約に用いる層を，上から順に 81×256，1×512，9×256 の層に変更して用いる．

3. 実験設定

本実験ではデータセットとして MTG-Jamendo Dataset(以下，MTG)の mood/theme のサブセットのデータを用いた．このデータセットは Jamendo という音楽配信サイトにアップロードされている楽曲と，そこにアーティスト自身が付けたタグから構成されている．タグは 59 種類からなり，1 つの楽曲データにつき複数のタグを持つ場合がある．

手法の精度評価には，ROC-AUC と PR-AUC を用いている．

4. 実験結果

表 1 に提案手法と比較手法の 3 回の学習による分類精度の平均値を示す．ただし，CLMR は自己教師あり学習を，MLFA_1 は[3]におけるマルチレベル特徴集約を，MLFA_2 は提案手法によるマルチレベル特徴集約を， SampleCNN は教師あり学習をそれぞれ表している．

表1．提案手法と比較手法による分類精度

モデル	ROC-AUC	PR-AUC
CLMR	0.85492016	0.324120675
CLMR+MLFA_1	0.847748129	0.339118243
CLMR+MLFA_2	0.856971247	0.372017489
SampleCNN	0.694846928	0.288192388
SampleCNN+MLFA_1	0.690952249	0.303063808
SampleCNN+MLFA_2	0.711054266	0.313464469

表 1 から，従来手法では ROC-AUC の値の減少と PR-AUC の値の増加が同時に起きており，精度が向上したとはいえない．それに対して，提案手法では ROC-AUC と PR-AUC の両方の値が増加しており，精度が向上しているといえる．これは，後段の層で失われた情報を補完するというマルチレベル特徴集約の特性が，マルチレベル特徴集約に用いる層を離すことでうまく働いたからではないかと考えられる．

5. まとめ

本稿ではマルチレベル特徴集約の構造を変更し，より高い精度の自己教師あり学習による楽曲分類の手法を検討した．結果として僅かだが精度の向上が見られた．この結果を受けて，これからはマルチレベル特徴集約で用いる層の選択を最適化することで精度の向上を目指す．

参考文献

[1] J. Spijkervet, and J. A. Burgoyne. ″Contrastive Learning of Musical Representations″ In Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR, 2021

[2] J. Lee, J. Park, K. L. Kim, and J. Nam. ″Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms″ SMC, 2017

[3] T. Kim, J. Lee, and J. Nam. ″Sample-level CNN Architectures for Music Auto-tagging Using Raw Waveforms″ 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)

[4] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, X. Serra. ″The MTG-Jamendo Dataset for Automatic Music Tagging″ Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)