International Journal of Internet, Broadcasting and Communication Vol.16 No.3 27-32 (2024)
http://dx.doi.org/10.7236/IJIBC.2024.16.3.27

IJIBC 24-3-3

Multiclass Music Classification Approach Based on Genre and Emotion

Jonghwa Kim

Professor, Department of Artificial Intelligence, Cheju Halla University, Korea
jkim@chu.ac.kr

Abstract

Reliable and fine-grained musical metadata are required for efficient search of rapidly increasing music files. In particular, since the primary motive for listening to music is its emotional effect, diversion, and the memories it awakens, emotion classification along with genre classification of music is crucial. In this paper, as an initial approach towards a "ground-truth" dataset for music emotion and genre classification, we elaborately generated a music corpus through labeling of a large number of ordinary people. In order to verify the suitability of the dataset through the classification results, we extracted features according to MPEG-7 audio standard and applied different machine learning models based on statistics and deep neural network to automatically classify the dataset. By using standard hyperparameter setting, we reached an accuracy of 93% for genre classification and 80% for emotion classification, and believe that our dataset can be used as a meaningful comparative dataset in this research field.

Keywords: *Music Dataset, Music Information Retrieval, Music Classification, Music Genre Classification, Music Emotion Classification, Machine Learning*

1. Introduction

With exponentially increasing electronic music distribution, the need for fine-grained music metadata describing the content of music catalogues becomes more important. In the last decade, the research on music information retrieval (MIR) has been growing and a number of research works have been reported to facilitate effective organization and management of the great quantity of music files available. Recently, various neural network-based machine learning algorithms such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short term memory (LSTM) have been applied to music classification and give meaningful classification results [1].

A common way to classify music is by genre. With the recent advent of various kinds of music, the boundaries between global music genres such as rock, pop, jazz, blues, and hip-hop are blurring, and music classification by genre continues to require new challenges due to the essential nature of abstract and subjective music. Most of these genres are classified in relation to the structure of musical instrument arrangements, rhythms, and harmonics, but it is not uncommon, for example, to encounter the ID3 genre tag of misclassified MP3s [2].

Manuscript Received: May. 3, 2024 / Revised: May. 10, 2024 / Accepted: May. 15, 2024
Corresponding Author: jkim@chu.ac.kr
Tel: +82-64-741-6796
Professor, Department of Artificial Intelligence, Cheju Halla University, Korea

In order to consider the rapidly changing music types depending on the times and culture and to provide high-quality music search services such as on-demand music distribution models, more comprehensive and fine-grained music metadata is inevitable. In particular, Music Emotion Recognition (MER), which recognizes the emotional type of music and uses it as an additional classification category in MIR, is also being actively studied. MER is also an essential technology to implement a personalized music recommendation system [3]. On the other hand, as in all emotion-related applications, when a classification or recommendation is given that does not match the user's emotion, the result is greater user antipathy and disappointment compared to other errors. Hence the accuracy of emotion recognition is crucial. Interest in deeper understanding of the relationship between music and emotions has motivated researchers in a variety of disciplines, including computer science, for decades [4,5,6]. Particularly MER, which involves extracting, processing, and evaluating emotion-related features from music and then associating them with specific emotions in order to automatically recognize the emotional content of music or the emotions that music triggers in listeners, is one of the most challenging problems in MIR due to the inherent subtlety and subjective nature of music [7].

In order to implement a reliable MER and genre classification system, an accurate and consistently labeled "ground-truth" music dataset is required. While large datasets labeled with various sources and multiple categories are being released [8], a standard benchmark dataset that can be cited for objective performance evaluation of emotion and genre classification systems has not yet been reported. In particular, in the case of MER dataset, the categories of music emotions such as happiness, sadness, joy, etc. perceived by the listener may vary depending on time, space, culture, age, etc. Furthermore, the expression of adjective emotional words may be felt differently depending on the language used. All these make labeling of dataset difficult.

In this paper, we present a MER dataset that was built by labeling from real listeners online over a long period of time and subjecting them to appropriate preprocessing. In order to verify the usability and quality of this dataset, we build, test, and evaluate various machine learning models for classification of each emotion and genre based on feature vector calculated.

2. Related Works

2.1 Music genre classification

Musical genre is a categorical, typological construct that allows us to categorize and distinguish different musical works into specific groups. Typically, the characterization of music in genres such as pop, jazz, classical, and blues is done by considering instrumentation, rhythmic structure, and chordal content. However, music classification by genre often causes discussion because it can be performed from various perspectives rather than a completely objective task.

The general framework for automatic music genre classification involves extracting representative features from music signals, training a machine learning model, and applying it to the classification. k-NN (k-Nearest Neighbor), which classifies to the closest data point, is probably the simplest classifier [9]. Other most commonly used machine learning methods for music classification include Gaussian Mixture Model (GMM) [10], Support Vector Machine (SVM) [11], Linear Discriminant Analysis (LDA) [12], etc. There is no overwhelming classifier for music genre classification, and its performance is mainly determined by the dataset and feature values used. Therefore, researchers often use multiple classification algorithms and compare their performance. In particular, Li et al. [13] tested SVM, GMM, LDA, and k-NN on the dataset GTZAN generated in [10], and showed that SVM was superior with an accuracy of 78.5% in 10 genre classifications.

Recently, deep neural networks (DNN) have been widely applied to music genre classification [1]. In order to use the existing deep neural network algorithms optimized for image classification, spectrograms that simultaneously represent the patterns of music signals in the time and frequency domains are usually considered as images and used as inputs to convolutional neural networks (CNNs). For example, Li et al. [14] developed a CNN using raw MFCC matrices for music genre prediction, and Lidy & Schindler [15] used constant Q-transform (CQT) spectrograms as input to a CNN for the same purpose.

2.2 Music emotion classification

In emotion research and applications, there are two completely different approaches to classifying emotions. One way is to use separate adjectives to label individual categories of emotions; for example, in 1936, psychologist Hevner [16] defined eight emotional categories, as known as “adjective circle”, made up of 67 adjectives including sober, gloomy, longing, lyrical, and sprightly. Another method is to define a multi-dimensional emotion model and express all emotions as coordinates on the model. In general, Thayer's emotion plane [17] is often used to avoid ambiguity in the adjective approach. In the model, each emotion is defined by coordinates in terms of arousal (strength of emotion) and valence (negative/positive emotion).

Various traditional machine learning models were proposed for music emotion classification by extracting acoustic and psychoacoustic feature values correlated with emotion [1,18]. Recently, Xu et al. [19] reported successful music emotion recognition results through dynamic arousal and valence regression based on LSTM-RNN. Coutinho et al. [20] presented a multi-scale approach combining modified autoencoders (AEs) and deep belief networks (DBNs) at various levels, considering the high correlation between acoustic feature values.

3. Datasets

To generate a new "ground-truth" music emotion-genre dataset, we first collected songs from the internet to cover four genres, including classical (CG1), jazz (JG2), pop & rock (PG3), and hip hop (RG4), and four emotions such as joy (JE1), anger (AE2), sadness (SE3), and pleasure (PE4), that represent each of the four quadrants of Thayer’s arousal/valence emotion model. All of these music data are freely distributed by Jamendo [21] and MUSOPEN [22], so there are no copyright issues. They are converted to mono, 16kHz and 128kb/s, and the volume is normalized to 88dB. Most datasets in the literature use a uniform cut of the music signal between 3-30 seconds from a certain section of the music. However, unlike genres, the emotion of music can change over time, even within a song. Therefore, we listened to every song individually, identified the representative emotional parts, and extracted 30-second clips before and after those parts. These clips are then labeled by people of different age groups including students at the University Augsburg, Germany, and through the internet LimeSurvey (www.limesurvey.org). About 2,800 labelings were collected, and based on this, songs with high labeling reliability were selected first, and a total of 800 songs, 50 songs for each category, were created as shown in Table 1.

Table 1. Arrangement of music mood corpus (MMC800) labels

	CG1	JG2	PG3	RG4	
JE1	50	50	50	50	200
AE2	50	50	50	50	200
SE3	50	50	50	50	200
PE4	50	50	50	50	200
	200	200	200	200	800

4. Methods

4.1 Feature Extraction

We calculated various features according to the MPEG-7 [23] standard, i.e., energy, global temporal, perceptual, harmonic, and spectral shape etc. For the energy fea- tures which refer to various energy contents of a sig- nal, we calculated root mean square, spectral low energy rate, and spectral brightness with a thresh- old of 1kHz and 3kHz. As the global temporal features, log attack time, at- tack slope, and zero crossing rate are calculated from envelope waveform determined in given signal seg- ment. For the perceptual features which represent perception model of the human auditory system, we extracted the longterm loudness, specific loudness, roughness, sharpness, and MFCCs. The fundamen- tal frequency (F0) and inharmonicity are calculated from the sinusoidal harmonic modeling of the signal. From the spectrogram obtained by applying FFT to the raw signal, we extracted spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral rolloff, and spectral flux. Particularly for the spectral rolloff, we applied two thresholds of 85% and 95%. Additionally, we also calculated the octave based spectral contrast (OBSC) proposed in [24], discrete wavelet coefficient histogram (DWCH)], and common statistics of them such as the statistical means and stan- dard deviations. In total, our feature vector contains 199 features.

In order to preview how well the extracted features conform to the genre and emotion classification, the feature dimensions were reduced to two dimensions using dimensionality reduction techniques such as PCA and plotted as shown in Figure 1.

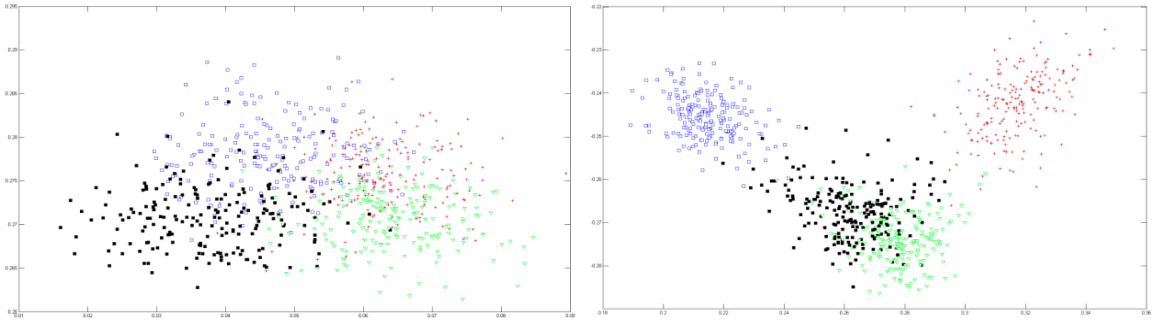

Figure 1. Fisher feature projection for genre (left) and emotion (right)

4.2 Classification

To classify music emotions and genres of MMC800, we employed k-Nearest Neighbors (k-NN), Linear Logistic Regression (LLR), Random Forests (RF), Support Vector Machines (SVM) using the Scikit-Learn Library [25], and CNN architecture of Keras [26]. Table 2 shows hyperparameter used for the classifiers.

Table 2. Hyperparameters of Classifiers

Classifier	Hyperparameter Used
k-Nearest Neighbors (k-NN)	k=3, linear search algorithm
Linear Logistic Regression (LLR)	penalty=12, multiclass=multinomial
Random Forests (RF)	#trees=1000, max depth=10
Support Vector Machine (SVM)	C(regularization)=0.17, kernel=rbf, gamma=scale, tolerance=0.001

The CNN is trained on spectrograms and has a single input layer and five convolutional blocks, each of which is composed of the following convolutional blocks;

- a convolutional layer with mirror padding, a 1x1 stride, and a 3x3 filter.
- a tuned linear activation function (ReLu).
- max pooling with 2x2 stride and window size
- dropout normalization probability of 0.2

The last layer of the CNN is a fully connected layer that implements the SoftMax activation function to output the probabilities of the 4 label classes, with the class with the highest probability being the classified label for a given input. We performed one-leave-out cross-validation for emotion and genre classification.

5. Results

The classification results for genres and emotions are summarized in Table 3. To evaluate the performance of each classifier, we used following criteria;

- Accuracy: the percentage of correctly classified test samples, i.e., accuracy = (TP/(TP+FP+FN)) × 100%.
- F-score: calculates precision and recall based on the confusion matrix and the harmonic mean value of the two.
- AUC: represents the area under the ROC (Receiver Operator Characteristics) curve and is a way to judge the performance of multiclass classification. The default model AUC for randomized prediction with equal probability is 0.5, so we generally expect a higher number.

Table 3. Classification Results

Genre Classification

Classifier	Accuracy(%)	F-score	AUC
k-NN	92.50%	0.81	0.9732
LLR	91.30%	0.78	0.9521
RF	87.25%	0.67	0.9193
SVM	93.50%	0.83	0.9874
CNN	89.45%	0.71	0.9253

Emotion Classification

Classifier	Accuracy(%)	F-score	AUC
k-NN	74.50%	0.58	0.8872
LLR	76.20%	0.60	0.8750
RF	68.50%	0.52	0.8544
SVM	77.50%	0.61	0.9088
CNN	80.50%	0.65	0.9453

6. Conclusion

In this paper, we presented a complete framework for music genre and emotion classification, including dataset generation, feature extraction, classifier modelling, and performance testing. Particularly we introduced new music genre-emotion dataset (MMC800) generated by anonymous online labeling by a large number of ordinary users instead of a few professional labelers. To evaluate the ground-truth reliability of the dataset for genre and emotion classification, we conducted classification tests using various machine learning algorithms. As a result, we achieved the highest accuracy of 93.50% using SVM for genre classification and 80.50% using CNN for emotion classification. Since the results were obtained by setting basic hyperparameters for each algorithm, there might be room to improve the accuracy.

As a future work, we will reanalyze the features introduced in this paper in depth to identify the features

that are dominantly correlated to genre and emotion classification respectively, and explore ways to improve the learning model for emotion classification, especially those with relatively low accuracy. In addition, the original dataset and feature vectors introduced in this paper will be shared with researchers in related fields through appropriate channels.

References

[1] N. Ndou, R. Ajoodha and A. Jadhav, "Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches," IEEE Intern. IOT, Electronics and Mechatronics Conf., Toronto, pp. 1-6, 2021

[2] K. Kosina, “Music genre recognition,” Fach-hochschule Hagenberg, Tech. Rep., 2002.

[3] JS. Gómez-Cañón et al. “TROMPA-MER: an open dataset for personalized music emotion recognition.” Journal of Intelligent Information System, 60, pp. 549–570, 2023.

[4] P. N. Justin, Musical Emotions Explained, Oxford University Press, 2019.

[5] J. Kim and E. Andre, “Emotion recognition based on physiological changes in listening music,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30 (12), pp. 2067-2083, December, 2008

[6] K. R. Scherer, “Which emotions can be induced by music? what are the underlying: Mechanisms? and how can we measure them?” Journal of New Music Research, vol. 33, no. 3, pp. 239–251, 2004.

[7] Y. Chen, Y. Yang, J. Wang and H. Chen, “The AMG1608 dataset for music emotion recognition.” In: ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 693–697, 2015

[8] B. L. Sturm, “A Survey of Evaluation in Music Genre Recognition. Adaptive Multimedia Retrieval, 2012

[9] M. Genussov and I. Cohen, “Musical genre classification of audio signals using geometric methods. In: EUSIPCO, IEEE, S. 497–501, 2010

[10] J. Burred and A. Lerch, “A Hierarchical Approach To Automatic Musical Genre Classification.” In: in Proc. Of the 6 th Int. Conf. on Digital Audio Effects, S. 8–11, 2003

[11] S. Brecheisen, H. Kriegel, P. Kunath and A. Pryakhin, “Hierarchical Genre Classification for Large Music Collections.” In: Proceedings of the ICME 2006, July 9-12, Toronto, Canada, 1385–1388, 2006

[12] C. Lee, J. Shin, K. Yu and J. Su, “Automatic Music Genre Classification using Modulation Spectral Contrast Feature.” In: Multimedia and Expo, 2007 July, 2007

[13] T. Li, M. Ogihara and Q. LI, “A Comparative Study on Content-based Music Genre Classification.” In: Proceedings of the 26th ACM SIGIR, NY, USA, 2003

[14] T. Li, A. Chan and A. Chun, “Automatic musical pattern feature extraction using convolutional neural network.” In Proc. Int. Conf. Data Mining and Applications. 2010

[15] T. Lidy and A. Schindler. “Parallel convolutional neural networks for music genre and mood classification”, MIREX2016, 2016.

[16] K. Hevner, “Experimental Studies of the Elements of Expression in Music.” XLVII (1936), S. 246–268, 1936

[17] R. R. Thayer, The Biopsychology of Mood and Arousal. Oxford University Press, USA, 1990

[18] D. Han, Y. Kong, J. Han et al. A survey of music emotion recognition. Front. Comput. Sci. 16, 166335, 2022

[19] M. Xu, X. Li, H. Xianyu, J. Tian, F. Meng and W. Chen. Multi-scale Approaches to the MediaEval 2015 "Emotion in Music" Task. MediaEval Workshop, Sept.14-15, 2015

[20] E. Coutinho, G Trigeorgis, S. Zafeirious and B. Schuller. “Automatically estimation emotion in music with deep long-short term memory recurrent neural networks.” In:Proceeding of the MediaEval Workshop, Sept.14-15, 2015

[21] JAMENDO, www.jamendo.com.

[22] MUSOPEN, www.musopen.org

[23] E. Allamanche, J. Herre and O. Hellmuth, “MPEG-7 audio low level descriptors for audio identification,” Proposal 6832, ISO/IECJTC1/SC29/ WG11(MPEG), 2001

[24] D.N. Jiang, L. Lu, H.J. Zhang, J.H. Tao, and L.H. Cai, “Music Type Classification by Spectral Contrast Features.” In: Multimedia and Expo, 2002. ICME ’02. Proceedings. 2002

[25] F. Pedregosa et al. “Scikit-learn:Machine learning in python,” 2018.

[26] F. Chollet et al., “Keras,” https://github.com/fchollet/keras, 2015.