Music Genre Classification

RE6124019
Ming-Hsuan Wu 吳明軒
Institute of Data Science
RE6124019@gs.ncku.edu.tw

Abstract

This study focuses on exploring various feature extraction techniques and their efficacy in music genre classification. We employed the GTZAN Dataset by George Tzanetakis, which includes ten distinct music genres, using the complete dataset to ensure experimental integrity. Feature extraction encompassed three primary categories: Timbral Texture Features, Rhythmic Content Features, and Pitch Content Features, as introduced in classroom discussions. We applied a 5-fold cross-validation method for model evaluation, randomly selecting 40 clips per genre for training and 10 clips for testing. The results demonstrated that the RandomForestClassifier and XGBClassifier achieved notable accuracy, with an average of approximately 0.77.

1. Introduction

The study of sound and communication forms a cornerstone of multimedia science. In the domain of music information retrieval, automated music genre classification plays a crucial role in enhancing the management and retrieval of music files, as well as augmenting the efficacy of personalised music recommendation systems. With the rapid proliferation of digital music, there is an increasing focus among researchers on developing robust methods for music genre classification.

Recent advancements in self-supervised learning approaches, such as Masked Autoencoders and other models based on masked modelling, have demonstrated promising results in both the image and audio domains [5]. These methods learn rich data representations by reconstructing masked segments of input, leveraging minimal visible inputs to infer the structure of the entire input, thus enabling the learning of useful data representations without reliance on explicit labels.

Furthermore, multi-source domain adaptation (MSDA) techniques demonstrate unique advantages when dealing with data originating from disparate probability distributions [4]. By constructing intermediate domains between sources and the target, such as through the Wasserstein barycenter transport method, these techniques effectively address the challenges posed by varying data sources. This approach aggregates the probability distributions of multiple sources and employs optimal transport strategies to transfer the aggregated source to the target domain, illustrating potential applications in visual and auditory recognition tasks.

This project does not employ the previously mentioned techniques for music genre classification. In contrast, the project examines the potential of machine learning technologies by analysing a diverse range of audio features for model training. The efficacy of various classification models is evaluated in order to assess their effectiveness in categorising music genres. The objective is to employ a range of feature extraction techniques and machine learning models to perform classification.

2. Related Work

The field of music genre classification emerged in the late 20th century, with initial studies primarily focusing on the extraction of audio features and their application in basic machine learning models. With the advent of deep learning technologies, researchers have increasingly explored the use of complex neural networks to enhance the accuracy of music genre classification.

This section introduces some of the most useful audio features and widely used classification models that have shown effective results:

2.1. Short-Time Fourier Transform (STFT)

Short-Time Fourier Transform (STFT) is a fundamental technique employed in signal processing to analyse the frequency components of a signal as they change over time, as illustrated in Figure 1. In contrast to the standard Fourier Transform, which provides the frequency spectrum of the entire signal, the STFT divides the signal into shorter segments of equal length and computes the Fourier Transform separately on each segment. This allows for the examination of temporal variations within the signal's spectral content, making it particularly useful in applications where signal properties evolve over time, such as audio and speech analysis. The technique was detailed in the influential work by Griffin and Lim, who explored signal estimation from modified STFTs, providing a robust methodological framework for enhancing applications such as speech enhancement and time-scale modification of speech [2].

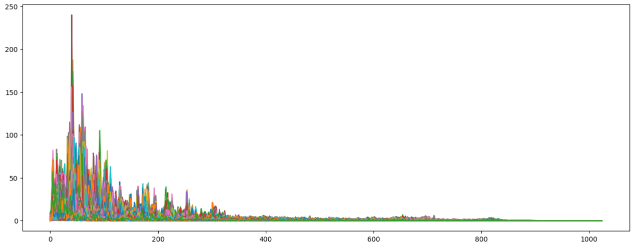


Figure 1. STFT feature of blues00.

2.2. Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs) represent a critical feature, employed predominantly in the field of speech recognition and increasingly in the context of music modelling. The coefficients provide a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The mel scale is designed to reflect the way in which the human ear perceives sound, emphasising those frequencies that are more meaningful for auditory processing. The computation of MFCCs is a multi-step process that includes framing the signal, computing the power spectrum, applying the logarithmic scale, transforming to the Mel scale, and finally, using the Discrete Cosine Transform (DCT) to decorrelate the Mel-spectral vectors. This sequence of operations has been extensively validated for speech, as evidenced by Beth Logan, and is also effective for modelling music, thereby confirming the suitability of MFCCs for diverse auditory applications [3].

2.3. Random forest classifier

The Random Forest classifier is a versatile machine learning method that has been employed extensively across various domains, including remote sensing classification. This classifier operates by constructing multiple decision trees during the training process and outputting the class that is the majority vote of the individual trees for classification tasks, or the mean prediction for regression tasks. This ensemble approach enhances the overall predictive accuracy and controls overfitting, making it robust against noise within datasets. Random Forests are capable of handling large datasets with higher dimensionality effectively and can estimate which variables are important in the classification process. Furthermore, they have the capability to model the data without requiring transformation to be linearly separable, in contrast to SVMs. In the field of remote sensing, Random Forest has been successfully employed to classify land cover types derived from satellite image data. Its performance has been demonstrated to be comparable or even superior to that of other advanced classifiers, such as Support Vector Machines (SVMs), particularly in terms of classification accuracy, robustness, and computational efficiency [6].

2.4. Support Vector Machines (SVM)

Support Vector Machines (SVM) represent a robust and widely utilised classification technique in machine learning. They are known for their effectiveness in handling both linear and nonlinear data. SVMs were developed by Vapnik and his colleagues and operate by identifying the optimal hyperplane that separates data points into distinct classes with the maximum margin, thereby ensuring good generalization on unseen data. This method involves transforming the input data into a higher-dimensional space, where a linear separator is sought. One of the key strengths of SVMs lies in their use of kernel functions, which permit the adaptation of the model to various types of data distributions without the necessity for explicit transformation of space. These characteristics render SVMs particularly valuable for tasks where the decision boundary between classes is not immediately apparent in the original feature space [7].

2.5. Extreme Gradient Boosting (XGBoost)

XGBoost, or Extreme Gradient Boosting, is an effective and scalable implementation of gradient boosting that has gained popularity for its performance in various machine learning challenges. XGBoost was developed by Tianqi Chen and Carlos Guestrin and provides a robust platform for the development of advanced models and the delivery of high-performance machine learning solutions. One of the key features of XGBoost is its capacity to process sparse data, which is a common occurrence in numerous real-world datasets. Additionally, it employs a distinctive sparsity-aware algorithm that optimises performance by efficiently handling missing values. The efficacy of XGBoost is evidenced by its prevalence in the winning solutions of machine learning competitions, reflecting its capacity to effectively handle large-scale and diverse datasets [1].

3. Proposed Method

This study was developed using Python and Jupyter notebooks and employed a 5-fold cross-validation approach for music genre classification experiments. A variety of audio features were extracted and stored in comma-separated value (CSV) files for subsequent analysis. A number of classification models were tested for classification tasks. Ultimately, the evaluation metric employed was accuracy, with confusion matrices plotted to further observe the classification results.

3.1. Features Extraction

3.1.1 Spectral Centroid

The Spectral Centroid is a feature of an audio signal that describes the average centre-of-mass position of the signal spectrum, as illustrated in Figure 2. It can be defined as the geometric centre of the short-time Fourier transform (STFT) spectrum. The formula is as follows:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] \cdot n}{\sum_{n=1}^{N} M_t[n]} \tag{1}$$

where Ct represents the spectral centre of the tth frame, Mt[n] denotes the magnitude of the Fourier transform of the tth frame, corresponding to the nth frequency bin, and N is the total number of frequency bins.

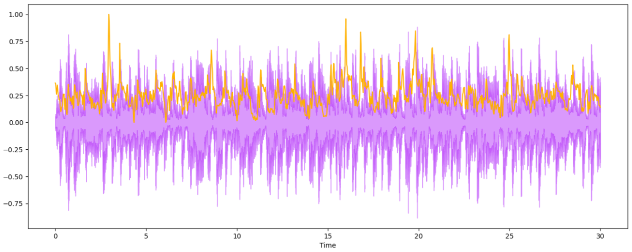


Figure 2. Spectral centroid feature of blues00.

3.1.2 Spectral Rolloff

Spectral roll-off is an audio feature defined as a frequency in the spectrum below which a specific percentage of the total energy is contained, e.g. 85%. In other words, if we order the energy in the spectrum from low frequency to high frequency, then the spectral roll-off is that frequency below which a specific percentage of the total energy is contained, as illustrated in Figure 3. The formula is as follows:

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{m=1}^{N} M_t[n] \tag{2}$$

where the symbol Mt[n] denotes the amplitude of the Fourier transform at frame t, corresponding to the nth frequency box. N is the total number of frequency boxes. Rt is the roll-off point of the spectrum under investigation, i.e., the frequency below which the energy in the spectrum accounts for 85% of the total energy.

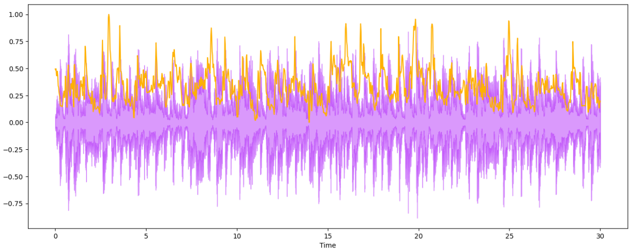


Figure 3. Spectral centroid feature of blues00.

3.1.3 Spectral Flux

Spectral flux is an audio characteristic that quantifies the rate of change in the power spectrum of an audio signal. The calculation is based on a comparison of the power spectrum of a given frame with that of the preceding frame. In more precise terms, it is typically calculated as the L2 range between two normalised spectra (also known as the Euclidean distance). The formula is as follows:

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \tag{3}$$

where Ft represents the spectral flux at frame t, while Nt[n] and Nt-1[n] are the normalised amplitude of the Fourier transform at frame t and frame t-1, respectively.

3.1.4 Zero Crossing Rate

This is a measure of the rate at which an audio signal changes from positive to negative in the time domain. In other words, it is the number of times an audio signal changes from positive to negative or vice versa. The formula is as follows:

$$Z_t = \frac{1}{2}\sum_{i=1}^{N} |sign(x[n]) - sign(x[n-1])| \qquad (4)$$

where Zt represents the zero-crossing rate at frame t, x[n] denotes the value of the audio signal at the nth sample point, and N is the total number of sample points. The sign(x[n]) function is a symbolic function that returns 1 if the value of x[n] is positive and -1 if the value of x[n] is negative.

3.1.5 Mel-Frequency Cepstral Coefficients

The Mel-Frequency Cepstrum is a linear transformation of the logarithmic energy spectrum based on a non-linear mel scale of sound frequency, as illustrated in Figure 4. The Mel-Frequency Cepstrum Coefficients (MFCCs) are the coefficients that make up the Mel-Frequency Cepstrum. The Mel-Frequency Cepstrum is derived from the cepstrum of the audio fragment. The MFCCs are the coefficients that make up the Mel-Frequency Cepstrum.

$$X_a[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N}, 0 \leqslant k < N \qquad (5)$$

$$S[m] = \ln[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]], 0 < m \leqslant M \qquad (6)$$

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\pi n(m-1/2)/M), 0 \leqslant n < M \qquad (7)$$

where the variable M represents the number of filters, while the variable N denotes the size of the FFT.

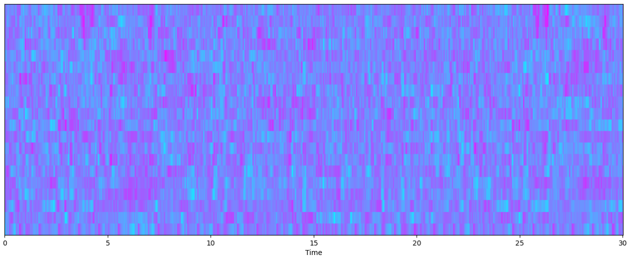


Figure 4. MFCCs feature of blues00.

3.1.6 Energy

The energy signature of a sound is a reliable indicator for detecting silence, helping to segment audio sequences and determine segment boundaries. It can be approximated by the root-mean-square of the signal in each frame. The volume of the audio signal depends on the gain value of the recording and digitisation equipment. In order to standardise the volume of a frame, it is necessary to compare it with the maximum volume of some previous frames.

$$v(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} s_n^2(i)} \qquad (8)$$

where the energy of the tth frame is denoted by Et, the value of the audio signal of the tth frame at the nth sample point is designated by st[n], and N is the total number of sample points.

3.1.7 Tempo&Beats

The librosa.beat.beat_track tool is employed to utilise the Tempo and Beats music features. The librosa.beat.beat_track tool is a dynamically planned beat tracker. The detection of beats is conducted in three stages. Initially, the intensity of the audio is measured, after which the estimated tempo is calculated based on the correlation between the intensity and the estimated tempo. Finally, the peaks in the intensity are selected that align with the estimated tempo. The function returns the estimated global tempo (expressed in beats per minute) and the position of the beat event.

3.1.8 Pitches&Magnitudes

In the field of audio processing, the librosa.piptrack function is capable of providing two crucial features: pitches and magnitudes. The term ”pitch” refers to the fundamental frequency of an audio signal, which is an essential aspect that reflects the audio’s frequency. Conversely, the magnitudes feature describes the amplitude of the audio, which is a pivotal aspect that reflects the strength of the audio. The librosa.piptrack function employs a parabolic interpolation method based on thresholds to track the pitches. The function initially determines the strength of the audio onset and subsequently estimates the tempo based on the onset correlation. Finally, the function selects peaks in the onset intensity that are approximately in accordance with the estimated tempo. The function returns an estimate of the global tempo (expressed in beats per minute) and the position of the beat event.

3.1.9 Chroma

The librosa.feature.chroma_stft function is employed to compute the spectral feature representation of an audio signal. Its principal objective is to compute a

Table 1. Accuracy of Validation.

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average Accuracy	Total Accuracy
Naive Bayes	0.58	0.57	0.575	0.65	0.6	0.595	0.54667
SGDClassifier	0.535	0.615	0.575	0.685	0.605	0.603	0.62667
KNN	0.615	0.6	0.645	0.65	0.59	0.62	0.54333
Decission trees	0.545	0.58	0.62	0.61	0.56	0.583	0.59
Random Forest	0.755	0.78	0.81	0.815	0.78	**0.788**	**0.72**
SVM	0.675	0.645	0.695	0.73	0.735	0.696	0.62667
Logistic Regression	0.685	0.695	0.64	0.735	0.71	0.693	0.63667
MLPClassifier	0.71	0.63	0.63	0.695	0.66	0.665	0.65667
XGBClassifier	0.795	0.78	0.785	0.775	0.755	**0.778**	**0.70333**
XGBRFClassifier	0.685	0.695	0.715	0.715	0.69	0.7	0.66667

chromaticity diagram from a waveform or power spectrogram. The function is utilized to compute the spectral feature representation of an audio signal, as illustrated in Figure 5.

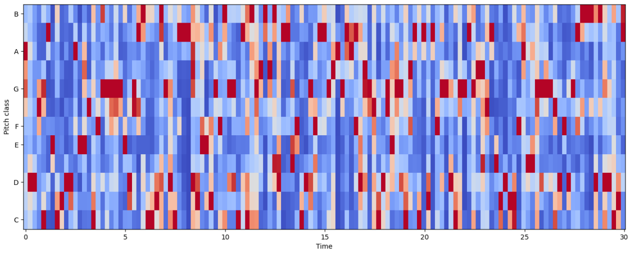


Figure 5. Chroma Frequencies feature of blues00.

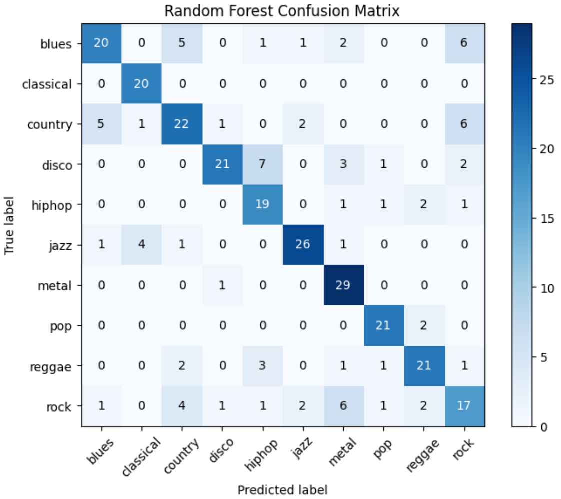
All the aforementioned sound features are extracted utilising the librosa suite, and the mean and variance are subsequently calculated as the final training parameters.

4. Experimental Results

The following classifiers were employed for the study: GaussianNB, SGDClassifier, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, SVC, LogisticRegression, MLPClassifier, XGBClassifier and XGBRFClassifier. A 5-Fold cross-validation was performed, with accuracy serving as the evaluation criterion, as shown in 1 . The predicted results were then used to plot the confusion matrix.

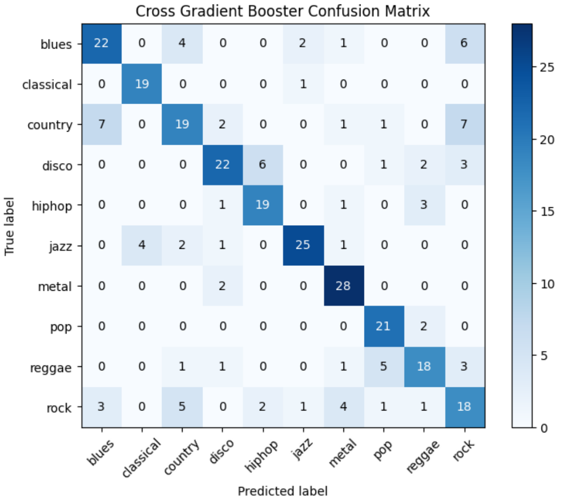
The table presents the accuracies of all classification methods for each fold, the average accuracies for the five-fold cross-validation, and finally the accuracies of all the data randomly divided into 70% training sets and 30% validation sets.

The two most effective confusion matrices are the Random Forest in Fig 6 and the XGBClassifier in Fig 7.



	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	20	0	5	0	1	1	2	0	0	6
classical	0	20	0	0	0	0	0	0	0	0
country	5	1	22	1	0	2	0	0	0	6
disco	0	0	0	21	7	0	3	1	0	2
hiphop	0	0	0	0	19	0	1	1	2	1
jazz	1	4	1	0	0	26	1	0	0	0
metal	0	0	0	1	0	0	29	0	0	0
pop	0	0	0	0	0	0	0	21	2	0
reggae	0	0	2	0	3	0	1	1	21	1
rock	1	0	4	1	1	2	6	1	2	17

Figure 6. Confusion Matrices of Random Forest.



	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	22	0	4	0	0	2	1	0	0	6
classical	0	19	0	0	0	1	0	0	0	0
country	7	0	19	2	0	0	1	1	0	7
disco	0	0	0	22	6	0	0	1	2	3
hiphop	0	0	0	1	19	0	1	0	3	0
jazz	0	4	2	1	0	25	1	0	0	0
metal	0	0	0	2	0	0	28	0	0	0
pop	0	0	0	0	0	0	0	21	2	0
reggae	0	0	1	1	0	0	1	5	18	3
rock	3	0	5	0	2	1	4	1	1	18

Figure 7. Confusion Matrices of XGBClassifier.

5. Conclusion

The present study sought to investigate the efficacy of various machine learning models in the context of music genre classification. To this end, audio features from the GTZAN dataset were subjected to analysis, with the Random Forest and XGB classifiers emerging as particularly promising due to their impressive accuracy performance. In the future, it would be beneficial to consider the application of deep learning techniques and more detailed feature analyses in order to enhance the accuracy and applicability of the music genre classification system.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 3

[2] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 2

[3] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, page 11. Plymouth, MA, 2000. 2

[4] Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multisource domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793, 2021. 1

[5] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked modeling duo: Learning representations by encouraging both networks to model the input. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[6] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005. 2

[7] Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020. 2