Music Genre Classification

Jen-Lung Hsu (徐仁瓏)
RE6121011
Institute of Data Science, National
Cheng Kung University
Tainan, Taiwan
RE6121011@gs.ncku.edu.tw

***Abstract*—This study aims to explore the problem of music genre classification and evaluate the effectiveness of different feature extraction methods and classification models. Through 5-fold cross-validation, we compared the performance of feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT), rhythmic content features, and pitch content features with classification models including logistic regression, random forest, gradient boosting, Adaptive Boosting (AdaBoost), support vector machine (SVM), and k-nearest neighbors (KNN). We ultimately found that the combination of all feature extraction methods paired with the random forest model performed best in terms of test accuracy. The code can be found in hw4.ipynb, and the execution environment can be obtained from requirements.txt.**

Keywords—feature extraction methods, classification models, k-fold cross-validation

I. INTRODUCTION

Music genre classification is an important and challenging task with significant implications for applications such as music information retrieval, recommendation systems, and music understanding. The ability to automatically categorize music into genres can facilitate various tasks in these domains, including personalized music recommendations and content organization. However, achieving accurate and robust classification poses several challenges due to the subjective nature of music perception and the complex relationships between musical features and genre labels.

In this study, we aim to investigate the impact of different feature extraction methods and classification models on music genre classification. By exploring various techniques for extracting relevant features from audio signals and employing diverse classification algorithms, we seek to enhance the accuracy and performance of genre classification systems. Through rigorous experimentation and evaluation using cross-validation techniques, we aim to identify the most effective combination of feature extraction methods and classification models for this task. Ultimately, our goal is to contribute to the advancement of music genre classification research and facilitate the development of more reliable and efficient music analysis systems.

II. METHODOLOGY

A. Feature Extractors

In this study, multiple feature extraction methods were utilized to capture different aspects of the music signal:

1) ***Mel-Frequency Cepstral Coefficients (MFCC):***

MFCC is a widely used feature extraction technique in audio signal processing. It involves computing the Mel-frequency cepstrum, which represents the short-term power spectrum of a sound. MFCCs are computed by taking the discrete cosine transform (DCT) of the logarithm of the Mel-spectrogram of the audio signal.

2) ***Fast Fourier Transform (FFT):***

FFT is a mathematical algorithm used to compute the discrete Fourier transform (DFT) of a sequence or its inverse. In the context of audio processing, FFT is applied to decompose the audio signal into its frequency components, providing information about the spectral content of the signal.

3) ***Rhythm Features:***

Rhythm features capture temporal patterns and dynamics in the music signal. These features may include tempo, beat, rhythm patterns, and other rhythmic characteristics extracted using techniques such as onset detection, autocorrelation, or rhythm histograms.

4) ***Pitch Features:***

Pitch features capture information about the pitch or tonal content of the music signal. These features may include pitch chroma, pitch histogram, pitch class distribution, or other pitch-related descriptors extracted using techniques such as autocorrelation, pitch tracking algorithms, or spectral analysis.

5) ***Combination of All Features:***

In addition to individual feature extraction methods, a combined feature set comprising all extracted features was also used. This approach aims to capture a comprehensive representation of the music signal by integrating information from multiple domains, including spectral, temporal, and tonal characteristics.

These feature extraction methods were chosen to capture various aspects of the music signal, including its spectral, temporal, and tonal properties, which are essential for effective music genre classification. Each method provides unique insights into different aspects of the audio signal, contributing to the overall discriminative power of the classification model.

B. Classification Models

In this study, the performance of the following classification models was evaluated:

1) ***Logistic Regression:***

Logistic Regression is a linear classification model that is widely used for binary and multiclass classification tasks. It estimates the probability that a given input belongs to each class using the logistic function, and then predicts the class with the highest probability.

2) ***Random Forest:***

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and

TABLE I. THE PERFORMANCE OF VARIOUS FEATURE EXTRACTION METHODS AND CLASSIFICATION MODELS

	MFCC	FFT	Rhythm features	Pitch features	Combination of all features
Logistic Regression	0.632	0.624	0.320	0.488	0.672
Random Forest	0.656	0.648	0.398	0.480	**0.676**
Gradient Boosting	0.592	0.642	0.384	0.492	0.664
AdaBoost	0.264	0.164	0.226	0.268	0.186
SVM	0.660	0.584	0.364	0.528	0.674
KNN	0.628	0.602	0.254	0.492	0.580

TABLE II. THE PERFORMANCE OF VARIOUS FEATURE EXTRACTION METHODS AND CLASSIFICATION MODELS

	n_estimators	criterion	max_depth	min_samples_leaf	min_samples_split	max_features	Test Accuracy
Grid search	200	gini	None	2	5	sqrt	0.67
Manual exploration	100	gini	None	1	2	sqrt	**0.69**

outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It combines the predictions of multiple weak learners to improve the overall accuracy and generalization performance.

3) ***Gradient Boosting:***

Gradient Boosting is an ensemble learning technique that builds a strong learner by iteratively adding weak learners to the ensemble. It minimizes a loss function by optimizing the predictions of the model with respect to the residuals of the previous iteration, resulting in a highly accurate and robust classifier.

4) ***Adaptive Boosting (AdaBoost):***

AdaBoost is a boosting algorithm that combines multiple weak classifiers to form a strong classifier. It assigns higher weights to misclassified instances in each iteration, allowing subsequent weak learners to focus more on the difficult instances. By iteratively adjusting the weights, AdaBoost creates a strong classifier that performs well on the entire dataset.

5) ***Support Vector Machine (SVM):***

Support Vector Machine is a powerful supervised learning algorithm used for classification and regression tasks. It constructs a hyperplane or set of hyperplanes in a high-dimensional space that separates the classes with the maximum margin, thus maximizing the margin between classes and improving generalization.

6) ***K-Nearest Neighbors (KNN):***

K-Nearest Neighbors is a non-parametric classification algorithm that assigns a class label to an input based on the majority class among its k nearest neighbors in the feature space. It does not require training and makes predictions based on the similarity between instances in the feature space.

Each classification model employs different strategies to classify input instances into predefined categories. They utilize various mathematical techniques, such as logistic regression, decision trees, boosting, and distance metrics, to learn the underlying patterns in the data and make accurate predictions.

C. K-Fold Cross-Validation

To assess the performance of the models, we employed the 5-fold cross-validation method. The dataset was divided into 5 equal-sized folds, and in each iteration, 4 folds were used for training while the remaining fold was used for testing. This process was repeated 5 times, ensuring that each fold served as both training and testing data. The performance metrics were then averaged across the 5 folds to obtain a robust evaluation of the model's performance.

K-fold cross-validation is a widely used technique in machine learning for model evaluation and selection. It helps to mitigate the risk of overfitting by using multiple train-test splits of the dataset, allowing for a more reliable estimation of the model's generalization performance. By repeating the process across different partitions of the data, k-fold cross-validation provides a more accurate assessment of the model's performance on unseen data.

III. EXPERIMENT

We utilized the GTZAN dataset, which comprises 10 distinct music genres, with each genre containing 50 music clips, each lasting 30 seconds. Through our experimentation, we compared the performance of various feature extraction methods and classification models, ultimately identifying the optimal combination.

The experimental results, as shown in Table 1, revealed that the rhythm and pitch feature extraction methods performed poorer compared to MFCC and FFT. The combination of all features yielded the highest accuracy across all models. Among the classification models, AdaBoost exhibited the weakest performance, while other models performed similarly. Particularly, Random Forest showed the best performance when combined with all feature extraction methods.

Therefore, we adopted this combination and utilized 5-fold cross-validation to search for its optimal hyperparameters,

yielding our results. However, during the hyperparameter tuning phase, the test results achieved only 0.67 accuracy, which was lower than expected. Consequently, further manual exploration was conducted to improve the performance, resulting in a final test accuracy of 0.69. Details of the hyperparameter settings are presented in Table 2.

Furthermore, we presented a visual representation of the predicted classification outcomes of our ultimate model on the test dataset, as illustrated in Figure 1.

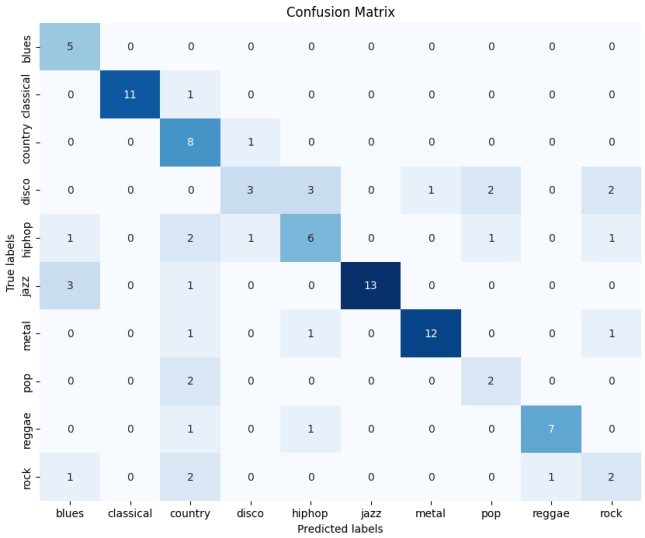

Fig. 1. The confusion matrix of the final model on the test set.

IV. Conclusion

Through experimental comparison, we found that combining all feature extraction methods with the Random Forest model achieved the best performance in music genre classification tasks. Furthermore, we fine-tuned the hyperparameters of the optimal model, further enhancing classification accuracy. These results hold significant reference value for applications such as music information retrieval and music recommendation systems. Ultimately, the performance reached 0.69 on the test set.