Music Genre Classification Project

DSC 540 Shu Yang

Abstract

The project aims to classify music genres based on audio features extracted from the audio files. Support Vector Machine (SVM) and Random Forest (RF) ML methods are chosen for this task. This project also experiments with different feature selection and feature extraction techniques including low variance filter, univariate filter, recursive feature elimination, wrapper and PCA. Finally, the models with and without feature selection are evaluated and the most important features will be discussed.

Introduction

Music genre classification is a necessity in the entertainment industry. It has been well-researched in the field. Music genre classification contributes significantly to the music recommendation system, the content organization and music discovery.

Many algorithms have been adopted to solve the music genre classification problems. The most often used methods include KNN, Random Forest, SVM, and neural networks. The focus on this project is SVM model and RF model.

The SVM model is a relatively stable and flexible supervised learning model for classification tasks. It is effective in high-dimensional spaces and robust to overfitting. Random Forest is an ensemble learning method suitable for classification tasks. It reduces overfitting by randomly selecting subsets of features at each split in the decision trees. The Random Forest model is effective on larger datasets with high dimensionally. The project aims to evaluate both models on the music genre classification tasks and explore the effects of feature selection. Three types of feature selection techniques are utilized including filter, wrapper, and recursive feature elimination. Feature extraction technique Principal Component Analysis is utilized as well. Finally, since SVM model is inherently limited to binary classification, a subset of data containing only two music genre labels is tested with SVM, and its results are compared with the multiclassification model.

Literature Review

Music genre classification is considered a cornerstone of the research area Music Information Retrieval (MIR). A variety of machine learning algorithms have been applied to music genre classification tasks. Feature selection and dimensionality reduction are also topics of interest.

S. J and K.S (2022) [5] uses Logistic Regression, K-Nearest Neighbor, Random Forest, Support Vector Machine and Artificial Neural Network, along with dimensionality reduction

techniques namely, PCA, KPCA and LDA, on the GTZAN dataset. It concludes that KNN model with PCA provides the highest accuracy of 77.41% among the compared models.

D.R. Ignatius Moses Setiadi et al. (2020) use SVM as the evaluation model and adopts chi-square calculation to rank and select features. [2]

Recently, more studies have used neural networks models to address music genre classification tasks. Odle et al. (2022) [4] compares the performance of different types of neural network models, including muti-layer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN) models. It concludes that GRU (Gated Recurrent Unit) model, one type of RNN models, outperforms the others.

Other music classification studies are focused on finding the most effective audio features. Such studies emphasize audio signal processing. For example, one study [3] uses Self-Organizing Maps to map the high-dimensional musical signals into SOM features and then used PCA to further reduce the feature dimensions.

Dataset & Features

The dataset used in this project is GTZAN dataset. It was collected between 2000 – 20001 and has been widely used in machine learning research for music genre recognition (MGR). The dataset comprises 9,990 entries of music records with 60 columns, representing mostly audio features extracted from each music record. There are 10 genres in the dataset, including blues, classical, country, disco, hip hop, jazz, metal, pop, reggae and rock. Each genre contains around 1000 records, making a balanced dataset. Upon checking, there is no missing value for any feature.

The first column of the dataset contains the file name of the audio record, which is not meaningful for the study, therefore it is removed from further exploration. The remaining 58 features are listed below.

length	chroma_stft_mean	Chroma_stft_var	Rms_mean	Rms_var	Spectral_centroid_mean
Spectral_centroid_var	Spectral_bandwidth_mean	Spectral_bandwidth_var	Rolloff_mean	rolloff_var	zero_crossing_rate_mean
zero_crossing_rate_var	harmony_mean	harmony_var	perceptr_mean	perceptr_var	tempo
mfcc1_mean	mfcc1_var	mfcc2_mean	mfcc2_var	mfcc3_mean	mfcc3_var
mfcc4_mean	mfcc4_var	mfcc5_mean	mfcc5_var	mfcc6_mean	mfcc6_var
mfcc7_mean	mfcc7_var	mfcc8_mean	mfcc8_var	mfcc9_mean	mfcc9_var

mfcc10_mean	mfcc10_var	mfcc11_mean	mfcc11_var	mfcc12_mean	mfcc12_var
mfcc13_mean	mfcc13_var	mfcc14_mean	mfcc14_var	mfcc15_mean	mfcc15_var
mfcc16_mean	mfcc16_var	mfcc17_mean	mfcc17_var	mfcc18_mean	mfcc18_var
mfcc19_mean	mfcc19_var	mfcc20_mean	mfcc20_var		

The features measure different aspects of the audio files. For example, chroma STFT measures the intensity of different pitches in an audio track. The RMS mean indicates the average power or loudness of the audio signal. The MFCC (Mel Frequency Cepstral Coefficients) features are a collection of features extracted from the audio files. They are often used in audio processing and speech recognition. All the audio features in the dataset explain an aspect of the audio file that could be meaningful for genre classification. The purpose of the feature selection is to identify the most important ones that can substitute the full features without sacrificing much the predictability of the model.

Exploratory data analysis is provided initially to understand the distribution of the features and their relationships. Visualizations are generated to help understand the features.

According to the correlation matrix, most features have weak or no correlations. Highly correlated features include spectral centroid, spectral bandwidth and rolloff. Based on the audio signal research, spectral centroid, spectral bandwidth and rolloff are typically correlated and the strength of correlation depends on the type of the audio files.

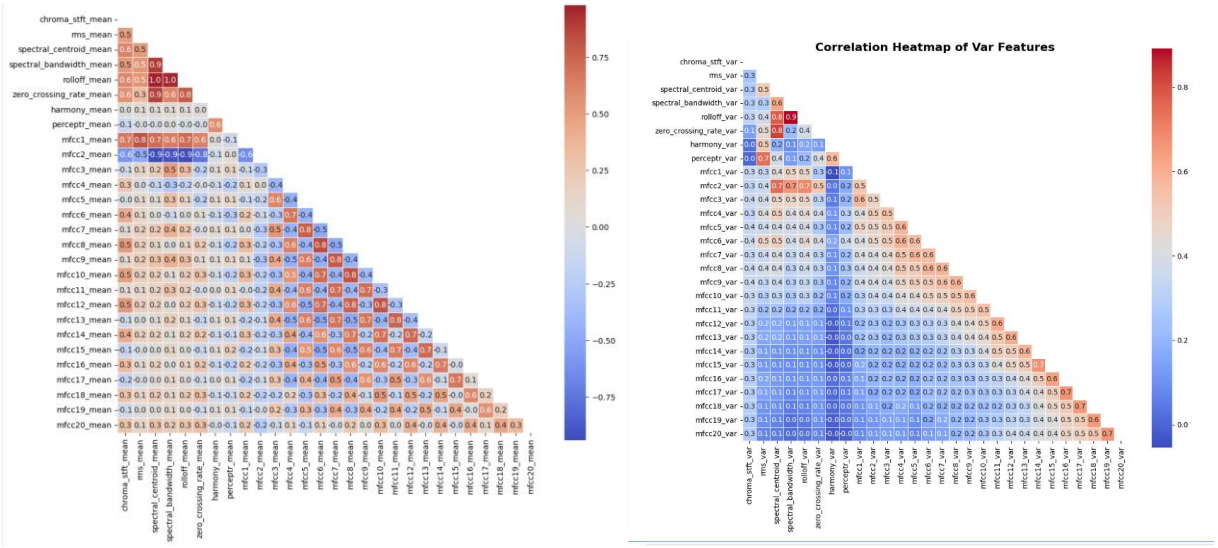


Figure 1, 2

The violin plots are generated to observe the distribution of each feature corresponding to the music genres. For example, the below violin plot demonstrated the distribution of Chroma_STFT_Mean. It can be said that classical and Jazz have lower Chroma_STFT_Mean values compared to other genres. Blues has a bimodal distribution compared to other genres.

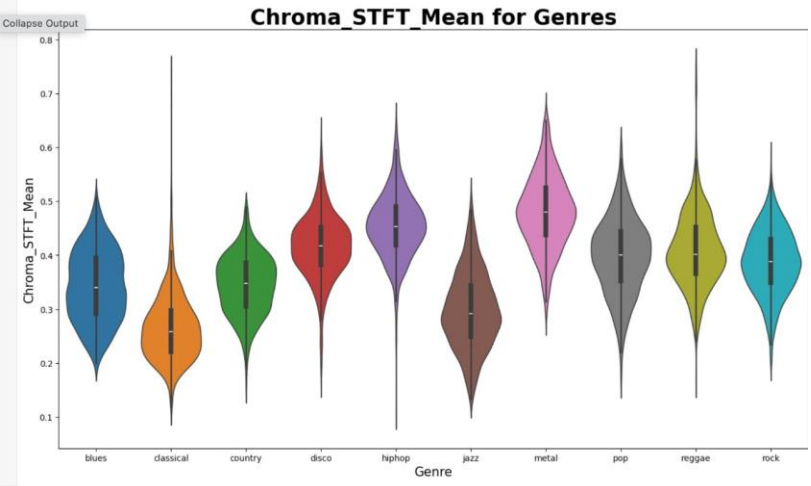


Figure 2

Machine Learning Methods

Before training the models, 58 features are standardized, and genre labels are encoded to numerical format.

For training SVM model, grid search and random search are used to find the best parameter. The random search generated the best parameter (Figure 4). The best model has C = 8.6 and kernel is RBF(default) and gamma is set as scale(default). The model achieved a high train accuracy of 0.995 and test accuracy of 0.93. The classification report and confusion matrix are as below(Figure 5,6).

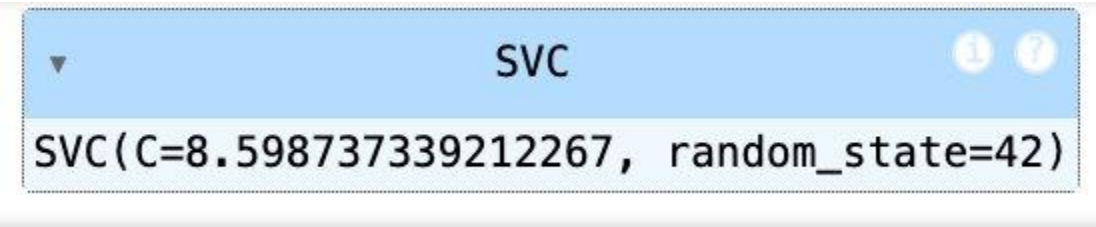


Figure 4

	precision	recall	f1-score	support
0	0.93	0.96	0.94	208
1	0.91	0.99	0.95	203
2	0.88	0.90	0.89	186
3	0.92	0.90	0.91	199
4	0.95	0.94	0.95	218
5	0.95	0.93	0.94	192
6	0.96	0.98	0.97	204
7	0.94	0.95	0.94	180
8	0.94	0.94	0.94	211
9	0.91	0.81	0.86	197
accuracy			0.93	1998
macro avg	0.93	0.93	0.93	1998
weighted avg	0.93	0.93	0.93	1998

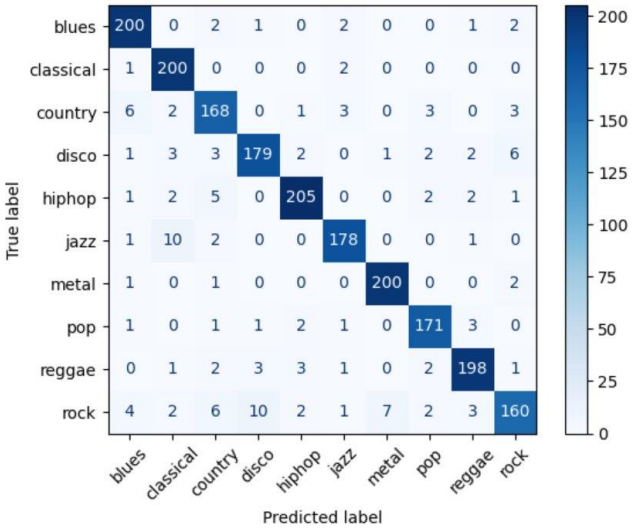


Figure 5, 6

For Random Forest model training, grid search and random search are utilized to find the best parameters. The random search is used first to narrow down the search, then the grid search is applied to find the best parameters. (Figure 7) The training accuracy for the random forest model is 0.999 and the test accuracy is 0.86. The classification report and the confusion matrix are as below. (Figure 8, 9)

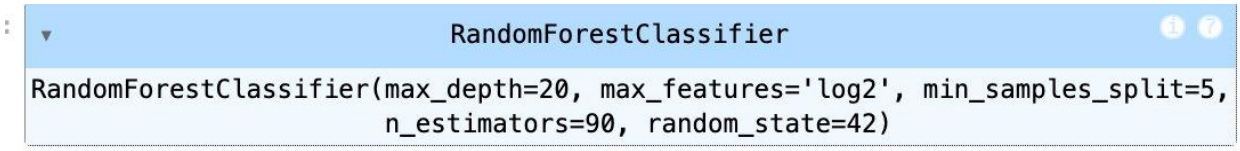


Figure 7

	precision	recall	f1-score	support
0	0.91	0.85	0.88	208
1	0.92	0.98	0.95	203
2	0.75	0.80	0.77	186
3	0.83	0.81	0.82	199
4	0.93	0.86	0.90	218
5	0.81	0.91	0.86	192
6	0.87	0.96	0.91	204
7	0.89	0.91	0.90	180
8	0.85	0.85	0.85	211
9	0.83	0.68	0.75	197
accuracy			0.86	1998
macro avg	0.86	0.86	0.86	1998
weighted avg	0.86	0.86	0.86	1998

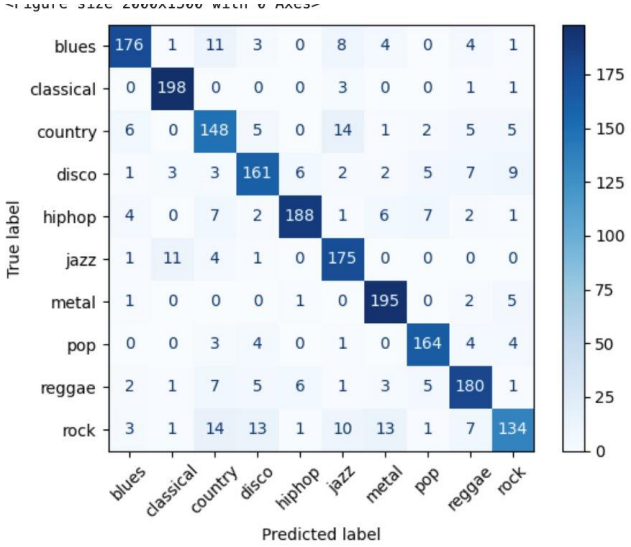


Figure 8, 9

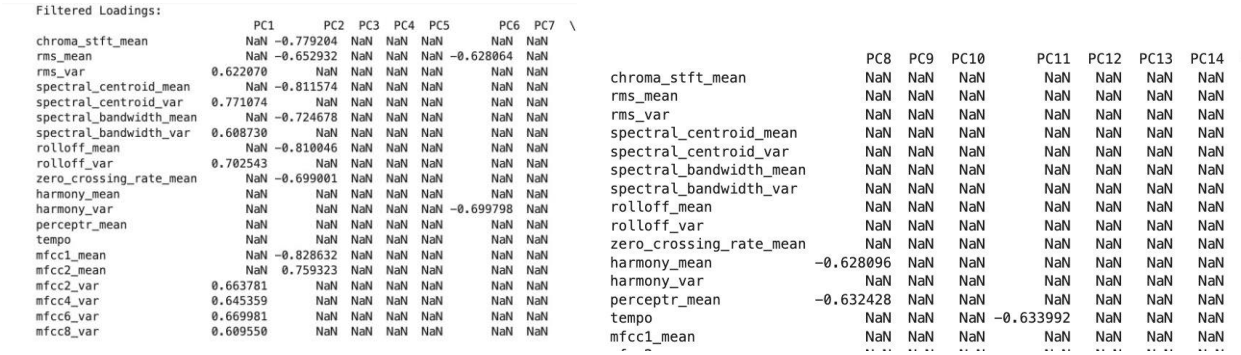
According to the accuracy scores and the classification reports, both models achieve high predictability. However, for random forest model, the difference between the train and test accuracy is 0.13, indicating that the model is slightly overfitting. The SVM has a test accuracy of 0.93, which is the better performer of the two.

Features selection and extractions

Two types of filter method feature selection are applied in this project. The filter method evaluates the features based on their relations with target varaible. The first feature selection is the low variance filter. For this method, the data are processed through min-max scaling and the threshold for variance is 0.01. The model automatically selected 24 features out of 58. Univariate filter method uses the multual_ino_classfi function and the number of features to select is set to be 20.

For the wrapper method, the features are evaluated through training a model. Two types of wrapper methods are applied. The first is recursive feature elimination, using random forest as the estimator. The number of features to select is set to be 20. The other wrapper method uses SelectFromModel function, the estimator is random forest. The threshold for this wrapper method is “mean”, meaning that Features whose absolute importance value is greater or equal to the mean of the feature importance are kept while the others are discarded. This method automatically selected 20 features out of 58.

Principal component analysis (PCA) is applied to features, and it reduces the dimensions to 20 components. The total variance captured is 0.82. To find out the correlations between each feature and each component, factor loading is printed and filtered to show only the coefficients that have the absolute value larger than 0.6. The result is as below. (Figure 10, 11)



Filtered Loadings:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
chroma_stft_mean	NaN	−0.779204	NaN	NaN	NaN	NaN	NaN
rms_mean	NaN	−0.652932	NaN	NaN	NaN	−0.628064	NaN
rms_var	0.622070	NaN	NaN	NaN	NaN	NaN	NaN
spectral_centroid_mean	NaN	−0.811574	NaN	NaN	NaN	NaN	NaN
spectral_centroid_var	0.771074	NaN	NaN	NaN	NaN	NaN	NaN
spectral_bandwidth_mean	NaN	−0.724678	NaN	NaN	NaN	NaN	NaN
spectral_bandwidth_var	0.608730	NaN	NaN	NaN	NaN	NaN	NaN
rolloff_mean	NaN	−0.810046	NaN	NaN	NaN	NaN	NaN
rolloff_var	0.702543	NaN	NaN	NaN	NaN	NaN	NaN
zero_crossing_rate_mean	NaN	−0.699001	NaN	NaN	NaN	NaN	NaN
harmony_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
harmony_var	NaN	NaN	NaN	NaN	NaN	−0.699798	NaN
perceptr_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tempo	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mfcc1_mean	NaN	−0.828632	NaN	NaN	NaN	NaN	NaN
mfcc2_mean	NaN	0.759323	NaN	NaN	NaN	NaN	NaN
mfcc2_var	0.663781	NaN	NaN	NaN	NaN	NaN	NaN
mfcc4_var	0.645359	NaN	NaN	NaN	NaN	NaN	NaN
mfcc6_var	0.669981	NaN	NaN	NaN	NaN	NaN	NaN
mfcc8_var	0.609550	NaN	NaN	NaN	NaN	NaN	NaN

	PC8	PC9	PC10	PC11	PC12	PC13	PC14
chroma_stft_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rms_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rms_var	NaN	NaN	NaN	NaN	NaN	NaN	NaN
spectral_centroid_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
spectral_centroid_var	NaN	NaN	NaN	NaN	NaN	NaN	NaN
spectral_bandwidth_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
spectral_bandwidth_var	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rolloff_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rolloff_var	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zero_crossing_rate_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN
harmony_mean	−0.628096	NaN	NaN	NaN	NaN	NaN	NaN
harmony_var	NaN	NaN	NaN	NaN	NaN	NaN	NaN
perceptr_mean	−0.632428	NaN	NaN	NaN	NaN	NaN	NaN
tempo	NaN	NaN	NaN	−0.633992	NaN	NaN	NaN
mfcc1_mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 10, 11

The below features are consistently selected regardless of the feature selection method used. They also have a high correlation with the principal components.

chroma_stft_mean	rms_mean	rms_var	spectral_centroid_mean	spectral_centroid_var

spectral_bandwidth_mean	rolloff_mean	rolloff_var	harmony_mean	harmony_var
perceptr_mean	perceptr_var	mfcc1_mean		

Model Evaluation

The model performances are summarized as follows.

Random Forest	Train Accuracy	Test Accuracy
58 Features	0.998	0.86
LV Filter (24 Features)	0.997	0.84
Univariate Filter (20 Features)	0.995	0.80
Wrapper (20 Features)	0.997	0.86
Recursive (20 Features)	0.997	0.86
PCA (20 Components)	0.998	0.78

SVM	Train Accuracy	Test Accuracy

58 Features	0.99	0.93
LV filter (24 Features)	0.87	0.83
Univariate Filter (20 Features)	0.88	0.81
Wrapper (20 Features)	0.93	0.88
Recursive (20 Features)	0.93	0.88
PCA (20 Components)	0.95	0.86

Discussion

Based on the model evaluation, the best model without feature selection is the SVM model with RBF kernel. The best parameters for the SVM model are generated through the random search. The best parameter for Random Forest is generated through the grid search. Although grid search is more flexible and thorough, it is computationally expensive and if the range of the parameters is not wide enough, it could end up performing worse than the random search.

With feature selection, for the random forest model, PCA performs the worst, and the wrapper method performs the best with the filter method in the middle. For the SVM model, the filter method performs the worst, and the wrapper method performs the best, with PCA in the middle. There is no obvious difference between the two wrapper methods. Since the recursive feature elimination typically takes longer to run compared to the SelectFromModel function. The winner of the model is SVM with wrapper method using random forest as the estimator.

In this project. The precision, recall, and F1 score for the country genre is generally lower than other genres with all models. To increase the prediction for country genre, the author has converted the multi-classification problem into a binary classification problem, only targeting country and other music genres. After random under sampling the dataset to achieve a balanced

dataset, the binary classification with SVM model performs better than the SVM model with multi-classification. However, it needs to be tested with all genres to draw a definitive conclusion.

Future Work

In addition to testing the predictability with binary classification models, other more comprehensive wrapper methods should be applied to compare the results. The runtime should be included to assist in model evaluation.

References

[1] Chillara, S., Kavitha, A. S., Neginhal, S. A., Haldia, S., & Vidyullatha, K. S. (2019). Music genre classification using machine learning algorithms: a comparison. Int Res J Eng Technol, 6(5), 851-858.

[2] D. R. Ignatius Moses Setiadi et al., "Effect of Feature Selection on The Accuracy of Music Genre Classification using SVM Classifier," 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2020, pp. 7-11, doi: 10.1109/iSemantic50169.2020.9234222.

[3] Jin, X., & Bie, R. (2006, June). Random Forest and PCA for Self-Organizing Maps based Automatic Music Genre Discrimination. In DMIN (pp. 414-417).

[4] Odle, Eric & Lin, Pei-Chun & Farjudian, Amin. (2022). Comparing Recurrent Neural Network Types in a Music Genre Classification Task: Gated Recurrent Unit Superiority Using the GTZAN Dataset.

[5] S. J and K. S, "Obtain Better Accuracy Using Music Genre Classification Systemon GTZAN Dataset," *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*, Vijaypur, India, 2022, pp. 1-5, doi: 10.1109/NKCon56289.2022.10126991.

[6] Raad Shariat and John Zhang. 2023. An Empirical Study on the Effectiveness of Feature Selection and Ensemble Learning Techniques for Music Genre Classification. In Proceedings of the 18th International Audio Mostly Conference (AM '23). Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/3616195.3616217