Importance of Feature Selection Methods for Music Genre Classification Models

Anshul Hallur

Abstract

Classifying music into its respective genre has been increasingly important in today's modern music streaming algorithms. By identifying the genre of a song, listeners can better understand their likes and dislikes and can categorize the song in a meaningful way. As the music scene evolves and creates more genres and sub-genres, classifying a song into its rightful genre has become increasingly difficult due to many factors, one such factor is due to not recognizing the underlying important features that make up that song to identify the genre it should be placed in. Using the famous GTZAN dataset that is a well-known dataset for music classification models, this paper aims to contribute to the ongoing problem by introducing feature selection methods such as RFE and Boruta to further identify the most relevant features for music genre prediction. Our results showed that RFE highlighted spectral centroid and tempo as top features, while Boruta emphasized "mean_chroma", "mean_mfcc", and "spectral_centroid" as the most important and confirmed tempo as unimportant. By incorporating MFCC and Chroma features into logistic regression, random forest, and k-Nearest Neighbors (kNN) models, we observed significant performance improvements, which corroborates the findings from the Boruta method. This paper contributes to the ongoing efforts in music genre classification by introducing effective feature selection techniques that can enhance the accuracy of classification models.

Introduction

Understanding the various features that contribute to music genre identification is an ongoing discussion in the world of music information and analysis research. Efforts to automate this process have been fruitful, as popular streaming services such as Spotify and Apple Music have incorporated music genre classification models into their algorithms to help categorize music. However, over time, it has become increasingly difficult to accurately predict genres and sub-genres due to the rapid evolution of music and the increase of sub-genres. Even streaming services such as Spotify occasionally misclassify songs or struggle to incorporate the numerous sub-genres present in today's music. With over 5,071 genres in Spotify's database and counting (Barilla, 2022), classifying genres is becoming more complex, and the features that make up an audio file are increasingly important for accurate genre identification.

Previous research has explored three main approaches to music genre classification: content-based features with traditional machine learning models, spectrogram images using convolutional neural networks (CNNs), and audio embeddings using recurrent neural networks (RNNs). Popular industry-standard datasets for music genre classification include GTZAN, Extended Ballroom, FMA, and the MIREX evaluation campaign. The introduction of the GTZAN dataset in 2002 by George Tzanetakis and Perry Cook marked an important benchmark in music genre classification. Researchers have built upon this foundation, exploring various classification methods such as support vector machines (SVMs), CNNs, and RNNs (Xu, 2003). Recent studies have found that deep learning methods such as CNNs and RNNs outperform traditional machine learning approaches for music genre classification (Lau, 2021). Despite these advancements, there remains a knowledge gap in feature selection for music genre classification. Complex datasets often contain redundant and noisy features, which can lead to biased and error-prone results. For example, research conducted by Dhevan S. Lau and Ritesh Ajoodha found that their models demonstrated bias towards certain genres like rock and metal, which can be correlated to the noise and distortion found in the GTZAN dataset.

In this paper, we aim to address this knowledge gap by investigating the problem of selecting the most relevant features for music genre classification. We will use feature selection methods such as Recursive Feature Elimination (RFE) and Boruta to improve the performance and interpretability of music genre classification models by reducing the number of features and selecting the most relevant ones. Our research question is: "Can we better identify the features that help in classifying music genres?".

Methods and Materials

The dataset for this project, the GTZAN dataset, was obtained from Kaggle (Andrada, 2020). This well-established benchmark in music genre classification comprises of 1000 audio files, each 30 seconds long and evenly distributed across 10 distinct genres: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. We imported the necessary libraries, including tidyverse, dplyr, fs, class, imputeTS, tuneR, purrr and caret, then created a main data frame called “music_data”. After iterating through each genre folder and creating a genre data frame, we appended it to the “music_data” data frame. Next, we defined two functions, “calculate_tempo” and “calculate_spectral_centroid”, to extract the features (tempo and spectral centroid) from each music file and stored them in a data frame called “feature_data”. We used a 70/30 split ratio for the training and test sets to train three machine learning models: logistic regression, random forest, and k-Nearest Neighbors (kNN). For each model, we calculated the accuracy by comparing the predicted genres to the true genres in the test dataset. We also recorded confusion matrices and F1-scores for each model.

We then set a binary target variable for hip hop songs and used a 70/30 split ratio for training/testing. We performed Recursive Feature Elimination (RFE) using a random forest model for the tempo and spectral centroid features and stored the results. To incorporate additional features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Chroma features into our dataset, we imported libraries, including tuneR, seewave, and audio, and used the reticulate library to import the Python library librosa. We defined two functions, “calculate_mfcc” and “calculate_chroma”, to calculate the MFCC and Chroma features for each music file. We appended the MFCC and Chroma features to the “feature_data” dataframe using the mutate function, combined and unnested the MFCC and Chroma lists, and assigned unique column names to each feature for identification. We updated the “feature_data” with the new combined dataset and performed RFE with the MFCC and Chroma features added. We then ran logistic regression, random forest, and k-Nearest Neighbors models again using MFCC and Chroma features.

Lastly, we created a new dataframe called “new_feature_data” with selected columns such as genre, tempo, spectral centroid, mean MFCC, and mean Chroma. We installed the Boruta package and performed feature selection for a binary classification of hip hop songs, storing the results. Additionally, we performed Boruta feature selection with a multinomial response using the “multinom” function from the nnet package. Finally, we visualized the rankings of the features by creating a bar chart comparing the rankings from both the RFE and Boruta models.

Results

The three machine learning models, logistic regression, random forest, and k-Nearest Neighbors (kNN) were trained and tested using the extracted features tempo and spectral centroid. Their performance was evaluated in terms of accuracy, confusion matrices, and F1-scores. For the initial models with only tempo and spectral centroid features, we can refer to Fig. 1 where we see that the logistic regression model achieved an accuracy of 29.43%, the random forest model 28.09%, and the kNN model 26.42%. Further on, the logistic regression model with only 2 features achieved F1 scores ranging from 0.23 for blues genre to 0.49 for pop. The random forest model demonstrated F1 scores from 0.04 for blues genre to 0.46 for pop. Lastly, the kNN model showed F1 scores from 0.10 for blues genre to 0.50 for pop. After incorporating the MFCC and Chroma features, we can refer to Figure 2 to see the logistic regression model's accuracy increased to 32.57%, the random forest model to 34.54%, and the kNN model to 57.73%.

Model Accuracy using 2 features (Tempo, Spectral Centroid)	
Model	**Accuracy Score (%)**
KNN	26.42
Random Forest	28.09
Logistic Regression	29.43

Figure 1. *This table outlines the model prediction accuracy score when using only tempo and spectral centroid as the features for the three models.*

Model Accuracy using 4 features (Tempo, Spectral Centroid, MFCC, Chroma)	
Model	**Accuracy Score (%)**
KNN	57.73
Random Forest	34.54
Logistic Regression	32.57

Figure 2. *This table includes the added features (MFCC and Chroma) in addition to the original two features (Tempo and Spectral Centroid) and outlines the accuracy score for each model.*

Both *Recursive Feature Elimination (RFE)* and *Boruta* feature selection methods were used to rank the importance of the features in the classification assignments. For the RFE method, the top 2 variables selected were spectral_centroid and tempo. This was concurrent with tempo and spectral centroid-only models, and for the models with all features. The Boruta feature selection method identified 3 important features: mean_chroma, mean_mfcc, and spectral_centroid, while tempo was confirmed as unimportant. Furthermore, we can look to Figure 3 and see that spectral centroid was deemed important in both methods (RFE and Boruta) as they are ranked at 1, signifying the importance of the feature. Consequently, Figure 3. also shows the average MFCC having higher importance than the average Chroma feature.

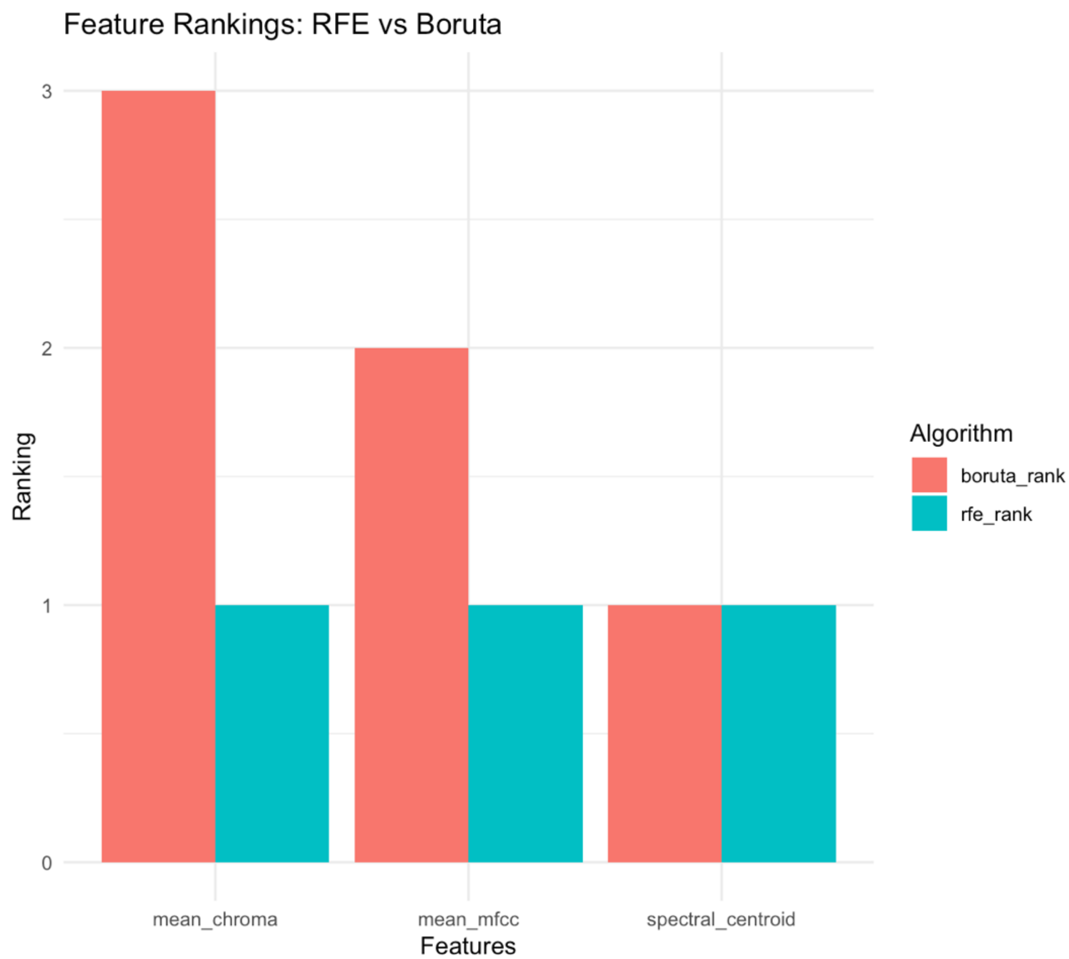

Figure 3.
The x-axis on this bar chart represents the different features used and the y-axis corresponds to the ranking of each feature based on importance (lower rank equals higher importance).

Discussion

By analyzing our results, we can confirm that incorporating MFCC and Chroma features significantly improved the performance of the machine learning models in classifying music genres compared to using only tempo and spectral centroid features. Among the three models tested, k-Nearest Neighbors (kNN) exhibited the highest accuracy of 57.73% when trained with the MFCC and Chroma features. The logistic regression and random forest models also demonstrated improved accuracy rates of 32.57% and 34.54%, respectively, with the addition of MFCC and Chroma features. Based on these results, the models with MFCC and Chroma features included perform better than the models using only tempo and spectral centroid features. This shows the importance of these features in music classification and their ability to improve predictions of genre classification models. The RFE method used a backward elimination approach, which started with all features and eliminated them one by one based on importance. The RFE method selected spectral centroid and tempo as the top 2 variables in both scenarios (with MFCC, Chroma, and without), indicating their importance for classifying music genres. However, they may not be sufficient to accurately predict the genre of a music sample. On the other hand, the Boruta model, which is designed to work with random forest, highlighted mean_chroma, mean_mfcc, and spectral_centroid as the most important features, and confirmed that tempo is not as important. This method aligns better with our initial results, which showed improvements in the three models (Logistic Regression, Random Forest, kNN) when adding the MFCC and Chroma features.

Spectral centroid is consistently selected as an important feature by both methods, highlighting its relevance in music genre classification. By using these feature classification methods, we can better identify the important features associated with classifying a music genre. Our findings are consistent with previous studies that highlighted the importance of MFCC and Chroma features in music genre classification tasks (Tzanetakis & Cook, 2002; Li et al., 2003). In comparison to these studies, our kNN model with MFCC and Chroma features achieved similar levels of accuracy, showing the effectiveness of these features for music genre classification. However, our accuracy rates are still relatively low compared to methods that use deep learning techniques for music genre classification (Choi et al., 2017; Dieleman & Schrauwen, 2014).

We will now further discuss the reasons for selecting tempo, spectral centroid, MFCCs, and Chroma as features. Tempo is a feature that measures the speed or pace of an audio file, usually expressed in beats per minute (BPM). A paper on music genre classification by Snigdha Chillara emphasized the importance of this feature, as their research found that it had a high correlation with some genres in the dataset (Snigdha, 2019). Another reason for using this specific feature is its importance in Spotify's algorithm, which categorizes music into genres based on factors that include tempo (Barilla, 2022).

Spectral centroid, another feature we used, helps distinguish between different types of bright or dull sounds, as well as high-pitched or low-pitched sounds. Chillara's paper also highlighted spectral centroid as a content-based feature with high correlation to the classical and blues genres. Mel Frequency Cepstral Coefficients (MFCCs) represent the short-term spectral shape of a sound signal and were referenced in a paper by Dhevan S. Lau and Ritesh Ajoodha. By using MFCCs as one of the content-based features, the researchers found that it had a higher accuracy compared to other features (Lau, 2021). Chroma, the last feature, is defined as the pitch content of a sound signal in audio files. This feature was referenced in Tom L. H. Li and Antoni B. Chan's 2011 paper titled "Genre Classification and the Invariance of MFCC Features to Key and Tempo" and was found to have lower variance compared to other features across different keys and tempos.

Given the performance of the models, there are limitations to this project which we will address in the following paragraph. Firstly, the GTZAN dataset used for this project is relatively large, which could have impacted the performance of the machine learning models. Additionally, the feature extraction process focused on a limited set of features, which might not fully capture the nuances of different music genres. By using a smaller set of features, we also run the risk of underfitting, which decreases the quality of the results. For training and testing the models, we used "na.omit"; however, we could have considered using imputation methods to fill in missing values instead of removing data points. We can also use hyperparameter tuning for models like Random Forest and k-Nearest Neighbors to improve their performance. Due to the complexity of the dataset and the number of values, it is computationally expensive to run multiple models; in this case, we could have used sample data and improved the models instead of frequently training them on the entire dataset. By using ensemble methods such as boosting, we can also combine multiple classifiers and potentially improve overall performance. The models themselves could be improved by incorporating more advanced techniques, such as deep learning, which have shown promising results in music genre classification tasks (reference paper).

Conclusion

In conclusion, our project demonstrates the importance of selecting relevant features, such as MFCCs and Chroma, for improving the performance of machine learning models in music genre classification. By employing feature classification models like RFE and Boruta, we were able to better identify the important features that contribute to classifying music genres, thus addressing our research question. Among the three models tested—Logistic Regression, Random Forest, and k-Nearest Neighbors—the k-Nearest Neighbors classifier achieved the highest accuracy when trained with the MFCC and Chroma features. The results indicate that these features play a crucial role in predicting music genres and can significantly enhance accuracy in predicting music genres. Furthermore, the performance of the models in this project can still be improved further by incorporating advanced techniques, such as deep learning, and by addressing the limitations identified in the discussion section. In summary, the main takeaway from this project is that feature selection is critical for the success of machine learning models in music genre classification, and the use of feature classification models, such as RFE and Boruta, can provide valuable insights into identifying the most relevant features and contribute to the research done in classifying music genres.

Bibliography

1. K. S. Mounika, S. Deyaradevi, K. Swetha and V. Vanitha, "Music Genre Classification Using Deep Learning," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-7, doi: 10.1109/ICAECA52838.2021.9675685.
2. Huang, Derek A.. “Music Genre Classification.” (2018).
3. Tzanetakis, George & Cook, Perry. (2002). Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing. 10. 293 - 302. 10.1109/TSA.2002.800560.
4. Barilla, Chris. “Spotify Has a Lot of Wacky Music Genres - How Are They Determined?” *Distractify*, Distractify, 1 Dec. 2022, https://www.distractify.com/p/spotify-music-genres-explained.
5. Lau, Dhevan S. and Ritesh Ajoodha. “Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches.” Proceedings of Sixth International Congress on Information and Communication Technology (2021): n. pag.
6. Xu, Changsheng & Maddage, Namunu & Shao, Xi & Cao, Fang & Tian, Qi. (2003). Musical genre classification using support vector machines. 5. V - 429. 10.1109/ICASSP.2003.1199998.
7. Andrada. “GTZAN Dataset - Music Genre Classification.” *Kaggle*, 24 Mar. 2020, https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.
8. Snigdha Chillara, AS Kavitha, Shwetha A Neginhal, Shreya Haldia, and KS Vidyullatha. 2019. Music genre classification using machine learning algorithms: a comparison. Int Res J Eng Technol 6, 5 (2019), 851–858.