Music Genre Classification and Recommendation by Using Machine Learning Techniques

Makine Öğrenmesi Teknikleri Kullanarak Müzik Türü Sınıflandırma ve Müzik Önerisi

Ahmet Elbir[1], Hilmi Bilal Çam[2], Mehmet Emre İyican[2], Berkay Öztürk[2], Nizamettin Aydın[1]
[1, 2]Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey
[1]{aelbir, naydin}@yildiz.edu.tr
[2]{hbilalcam, emreiyican, ozturkberkay9595}@gmail.com

Abstract—**Music genre prediction is one of the topics that digital music processing is interested in. In this study, acoustic features of music have been extracted by using digital signal processing techniques and then music genre classification and music recommendations have been made by using machine learning methods. In addition, convolutional neural networks, which are deep learning methods, were used for genre classification and music recommendation and performance comparison of the obtained results has been. In the study, GTZAN database has been used and the highest success was obtained with the SVM algorithm.**

Keywords—Music genre Classification; Acoustic features; Machine Learning; Deep Learning.

Özetçe—**Müzik türü tahmin edilmesi sayısal müzik işleme konusunun ilgilendiği konulardan bir tanesidir. Bu çalışmada sayısal işaret işleme teknikleri kullanılarak müziğin akustik özellikleri çıkarılmış ve daha sonra elde edilen bu özelliklerden makine öğrenmesi yöntemleri kullanarak müzik türü sınıflandırması ve müzik önerisi yapılmıştır. Ayrıca, sınıflandırma ve müzik önerisi için derin öğrenme yöntemlerinden olan konvolüsyonel sinir ağları kullanılmış ve elde edilen sonuçların performans karşılaştırılması yapılmıştır. Çalışma kapsamında GTZAN veri tabanı kullanılmış ve en yüksek başarı SVM algoritmasıyla elde edilmiştir.**

Anahtar Kelimeler—Müzik türü sınıflandırma; Akustik Özellikler; Makine Öğrenmesi; Derin Öğrenme

I. INTRODUCTION

The widespread usage of the Internet has brought about significant changes in the music industry as well as causing all kinds of change. Examples of these developments include the widespread use of online music listening and sales platforms, control of music copyright, classification of music genre, and music recommendations. Today, with the advancement of music broadcast platforms, people can listen to music at any time and at anywhere and can reached millions of songs through various music listening platforms such as Spotify, last fm. In this study, it is aimed to make musical recommendations and music genre classification using acoustic features obtained by digital signal processing methods from raw music without considering the user's music profile or collaborative filtering. The features to be extracted from the music has been determined as zero crossing rate, spectral centroid, spectral contrast, spectral bandwidth, spectral rolloff and Mel-frequency Cepstral Coefficients-MFCC. Furthermore, Convolutional Neural Network-CNN, which is one of the most useful methods of deep learning, has been used for music genre classification and music recommendation. The CNN is used as following; applying STFT to uploaded song, putting the spectrogram into a CNN, taking out the output of a dense layer, using this output as a feature vector to classify music genre and to calculate song similarities. To compare performance results, all classification and recommendation algorithms has been applied on the GTZAN dataset. In the second section, methods used for feature extraction, music recommendation and music genre classification has been elaborated. Experimental results has been summarized and discussed in the third section and the last section includes conclusion.

II. METHODS

A. Dataset

GTZAN which was firstly proposed by G. Tzanetakis in [1] is one of the most popular dataset used for music signal processing. It contains 1,000 music with 30-second, 22050 Hz sampling frequency and 16 bits. Genres in the GTZAN are blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock and all of these genres have 100 music. Some classification results for this dataset has been summarized in the Table-1.

978-1-5386-7786-5/18/$31.00 ©2018 IEEE

Table 1. Some Performance Results for GTZAN

Reference	Accuracy
Tzanetakis et al[1]	61.00%
Holzapfel et al. [2]	74.00%
Benetos et al. [3]	75.00%
Lidy et al. [4]	76.80%
Bergstra et al. [5]	82.50%

B. *Feature Extraction*

In this part of the paper, some important feature extraction method used in this study has been explained. In the music signal processing, feature extraction methods can be classified several aspects. Digital signal processing– on the time domain and frequency domain– is one of these. Another useful strategy used for feature extraction is statistical descriptors like mean, median, standard deviation etc. All methods which are described below divide a raw music signal into N number of windows and all of these methods are run N times.

- **Zero Crossing Rate:** Zero-crossing rate is defined as the number of sign changes of a signal in a certain period of time. Sign change is defined as the transition of the signal between negative and positive values. More detailed explanations can be found in [1, 6].
- **Spectral Centroid:** Spectral centroid is a feature used on a frequency domain and indicates the point of the center of gravity of the frequencies in the frequency bin [1, 6].
- **Spectral Contrast**: Spectral contrast represents the decibel difference between the peak and the pit points on the spectrum of a signal. In audio processing, it gives information about the power changes in the sound [1, 6].
- **Spectral Bandwidth**: Spectral bandwidth gives weighted average amplitude difference between frequency magnitude and brightness. It is an indication of the frequency range in the frame.[1, 6]
- **Spectral Rolloff*:*** Spectral rolloff is the normalized frequency at which the sum of the low frequency power values of the sound reaches a certain rate in the total power spectrum. Briefly, it can be defined as the frequency value corresponding to a certain ratio of the distribution in the spectrum. This rate is generally 85%. [1,6]
- **Mel Frequency Coefficient of Cepstrum-MFCC***:* The Mel Frequency Cepstral Coefficients (MFCCs) are a small set of features which describe the overall shape of a spectral envelope. It can be considered as the feature of timbre. The purpose of the MFCC is to adapt the cepstral coefficients to the human hearing system. Cepstral coefficients are linear scale. But human can hear the frequencies below 1 KHz as linear scale and the frequencies above as logarithmic scale. Because of the fact that, MFCC is one of the commonly used features in speech and speaker recognition systems. Steps of MFCC are Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Wrapping and Spectrum, respectively. The number of coefficients- N- in the MFCC is another parameter. In this study, we assumed that N is 13. More elaborated explanations regarding MFCC can be found in [1].
- When the number of frequency bin for previously described feature extraction methods is N, returned number of centroid point by spectral centroid method is also N. However, because of variable size of raw music data, the number of returned values cannot always be N. Therefore, in this study we have used some statistical descriptors like average, median to obtain feature vectors in equal length. Various statistical functions have been applied to all the features used in the project and feature numbers given in Table-2.

Table 2. Feature Dataset Summary

Feature	Statistical Functions	# of Features
Zero Crossing Rate	Mean, Median, Standard Deviation	3
Spectral Centroid		3
Spectral Contrast		3
Spectral Bandwith		3
Spectral Rolloff		3
MFCC(13 coeff)		39
MFCC Derivation		39
TOTAL		93

C. *Music Genre Classifiation by Using Machine Learning*

In this section, to classify music according to their genre some machine learning algorithms have been discussed. The classification algorithms used in this part are K-Nearest Neighbors, Random Forest, Naive Bayes, Decision Tree and Support Vector Machine.

- **K- Nearest Neighbors-KNN**: KNN is one of the distance based supervised learning algorithms. When solving the classification problem with this method, a model is not created and the test operation is performed on the labeled samples in the data set. A new instance of the class label will be calculated from the distance from the instances in the dataset. From these calculated distances, the class tag is estimated by voting on the class labels of the nearest k. When calculating the distance, the Euclidean, Manhattan distance formulas are often used [7, 8].

- **Naïve Bayes-NB:** The Naive Bayes algorithm is a probabilistic supervised learning algorithm that generates a classification model by calculating the preliminary probabilities from the data in the data set and classifies the new data according to this model. It is an algorithm that can be used in various problems because it is compatible with every kind of data and simple statistical calculations are required [7, 8].
- **Decision Tree-DT:** Decision trees are learning algorithms that provide a supervised and model-based approach. It tries to identify the most distinctive feature in the data set as the root node of the tree. An entropy calculation is made when the most distinguishing feature is found. There are also different metrics in the literature that provide differentiating features [7, 8].
- **Support Vector Machine-SVM:** SVM is one of model based supervised learning algorithms. DVM is based on the principle of training for a decision surface that will allow the two classes to distinguish one another. This decision surface is created by optimizing the boundary regions of the two classes. SVM can be used in multi-class data sets other than two-class data sets [7, 8, and 9].
- **Random Forest-RF:** Random Forest (RF) is also utilized to the same feature set to search the success of an ensemble technique as to the music genre classification. RF can be used as a combination of multiple decision trees with bagging sample selection strategy [10, 15].

D. Deep Learning for Music Genre Classification:

A convolutional neural networks are a tool, used to classify items that contain spatial neighborhood. Array of randomly created filters are used in the process and they are tweaked to better describe the data. Normally they are used to classify images but one dimensional filters can be utilized to classify audio. Also two dimensional filters can be used in the CNNs with spectrograms of audio. Several CNNs with general configuration given in Figure-1 are trained and fine-tuned to classify a set of songs. Numbers of convolutional layers, fully connected layers, filters in convolutional layers changes between these networks. Output of a 10 fully connected layer is saved for each song. This output is used as feature vector in similarity calculations [11].

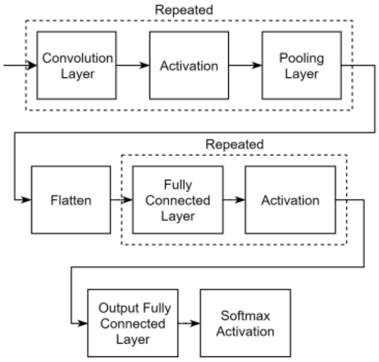


Figure 1 Convolutional Neural Network Configuration

After fine tuning, it is decided on a network with following configuration:

1- 2D Convolution Layer, 5x5 sized 32 filters, LeakyReLU activation function

2- Max Pooling Layer

3- 2D Convolution Layer, 5x5 sized 32 filters, LeakyReLU activation function

4- Max Pooling Layer

5- 2D Convolution Layer, 5x5 sized 32 filters, LeakyReLU activation function

6- Max Pooling Layer

7- 2D Convolution Layer, 5x5 sized 32 filters, LeakyReLU activation function

8- Average Pooling Layer

9- Flatten Layer

10- Dense Layer, 256 nodes, LeakyReLU activation function

11- Dense Layer, 128 nodes, LeakyReLU activation function

12- Dense Layer, 64 nodes, LeakyReLU activation function

13- Dense Layer, 10 nodes, Softmax activation function, as output layer

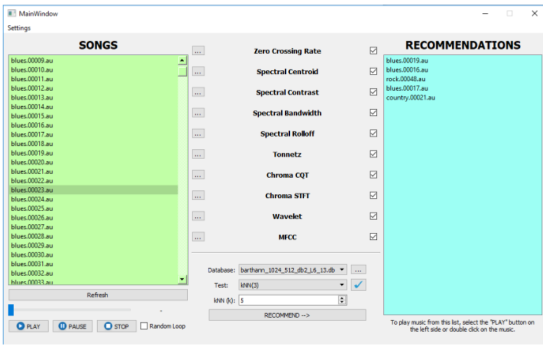
Raw audio data and acoustic features extracted from the audio is given to the network as input. Configuration of this features are given below.

- **Raw Audio:** Audio data is fed directly to the network after taking a 30 second section. 1D Convolution Layers are used instead of 2D Convolution Layers since audio data is a one dimensional vector.
- **Short Time Fourier Transform-STFT:** Short Time Fourier Transformation is applied to audio data before feeding it to the network. This transforms one dimensional time series into two dimensional frequency domain. This transformation is with hop length parameter set to 1024 and window size parameter set to 2048.
- **MFCC:** Mel Frequency Cepstrum is a representation of the short-term power spectrum of a sound, based on a mel scaled spectrogram. MFC Coefficients make up the MFC. This

transformation is with hop length parameter set to 1024, window size parameter set to 2048, both number of mel bins and mfcc coefficients set to 600 and power parameter set to 1.

III. EXPERIMENTAL RESULTS

In this section, to compare all results including conventional acoustic features and deep learning, a graphical user interface has been implemented in Python. To extract acoustic features from raw music The Librosa library [12] has been used and Keras [13, 14] used for deep learning partition is another useful library. Figure II shows the main window of the program. User can load a raw music and extract features by means of this interface. Also, not only some parameters of feature extraction methods like window size, window type and overlap ratio but also Meta parameters of classifiers such as the kernel type of SVM, the k value of KNN can be adjusted by using interface. Furthermore, in the interface selected music from playlist can be played and user can observe recommendation results.



blues.00009.au
blues.00010.au
blues.00011.au
blues.00012.au
blues.00013.au
blues.00014.au
blues.00015.au
blues.00016.au
blues.00017.au
blues.00018.au
blues.00019.au
blues.00020.au
blues.00021.au
blues.00022.au
blues.00023.au
blues.00024.au
blues.00025.au
blues.00026.au
blues.00027.au
blues.00028.au
blues.00029.au
blues.00030.au
blues.00031.au
blues.00032.au
blues.00033.au

Database: barthann_1024_512_db2_L6_13.db
Test: kNN(3)
kNN (k): 5

blues.00019.au
blues.00016.au
rock.00048.au
blues.00017.au
country.00021.au

To play music from this list, select the "PLAY" button on the left side or double click on the music.

Figure 2 Music Classification User Interface

On the other hand, some experiments have been carried out to assess music classification and music recommendation results. Firstly, conventional acoustic features obtained from raw music has been applied on machine learning algorithms mentioned before. The results obtained using all the features are shown in Table 3 where while Hanning(hann) and Bartlett-Hann(barthann) are window type, 1024 and 4096 are used as window size. Also k for KNN is 3 and SVM uses linear kernel.

Table 3. Classification Results by Using All Features

Algorithm	hann		barthann	
	1024	4096	1024	4096
KNN	61.50%	62.99%	61.70%	62.50%
RF	61.90%	63.69%	62.80%	65.69%
NB	55.30%	56.90%	55.09%	56.20%
DT	48.50%	55.30%	50.90%	55.00%
SVM	**72.60%**	**72.70%**	**72.30%**	**72.39%**

The classification results obtained when each feature is used alone are shown in Table 4. The parameters of the features is assumed as window type hanning and windows size 4096.

Table 4. Classification Results by Using Each Feature Alone

Features	Algorithm				
	KNN	RF	NB	DT	SVM
Zero Cr.	26.20%	29.20%	29.19%	25.70%	35.60%
Spec. Ce.	25.99%	31.79%	33.09%	31.40%	35.60%
Spec. Co.	35.69%	39.40%	35.90%	37.90%	40.30%
Spec. Ba.	27.60%	29.20%	30.70%	27.70%	33.19%
Spec. Ro.	33.40%	32.10%	33.40%	30.59%	38.30%
MFCC	**57.19%**	**59.69%**	**53.20%**	**48.90%**	**69.90%**

In the Table-5, all classification results have been summarized as the success rates of individual deleting of each feature. All parameters like window type is the same as previous experiments.

Table 5. Classification Results by Deleting Each Feature

Features	Algorithm				
	KNN	RF	NB	DT	SVM
Zero Cr.	62.40%	63.20%	55.30%	52.79%	72.30%
Spec. Ce.	63.10%	62.90%	56.50%	55.70%	72.20%
Spec. Co.	59.50%	61.70%	55.30%	50.30%	69.80%
Spec. Ba.	62.70%	63.60%	55.99%	53.59%	72.40%
Spec. Ro.	62.89%	64.10%	56.30%	54.40%	73.20%
MFCC	**49.09%**	**53.60%**	**38.10%**	**45.60%**	**55.49%**

Table-6 shows music recommendation results by using some parameters obtained from the most successful classification results. For the experimental results, the recommended number of music has been selected as 5, 10 and the performance rate by type has been calculated in percentage.

Table 6. Recommendation Results by Using Conventional Features

Genre	First 5 Songs	First 10 Songs
Blues	60%	48%
Classical	88%	90%
Country	40%	50%
Disco	48%	40%
Hip-hop	72%	60%
Jazz	52%	42%
Metal	84%	76%
Pop	60%	50%
Reggae	24%	26%
Rock	60%	50%

Convolutional Neural Network – CNN also applied for music classification and music recommendation. Firstly, raw music data, STFT matrix and MFCC matrix have been trained by using CNN model mentioned before 1D, 2D and 2D, respectively. And then, feature vectors obtained from the dense layer of the trained CNN have been used for music recommendations. The following Table-7 summarize the results for classification.

Table 7. CNN Classification Results

Data Type	Accuracy	Recall	Precision	F-Measure
Raw Music	15.00 %	19.00%	15.00%	13.00%
STFT	**66.00 %**	**65.00 %**	**69.00 %**	**65.00 %**
MFCC	63.00 %	63.00 %	64.00 %	62.00 %

According to the genre, the music recommendation for 5 music and 10 music performance also has been evaluated in percentage. In the Table-8, the recommendation performances obtained from CNN trained with STFT have been showed.

Table 8. CNN Recommendation Results

Genre	First 5 Songs	First 10 Songs
Blues	63.00%	49.70%
Classical	90,80%	87.70%
Country	56.80%	49.40%
Disco	53.60%	45.90%
Hip-hop	64.20%	57.30%
Jazz	65.00%	52.70%
Metal	79.40%	75.10%
Pop	78.40%	73.80%
Reggae	57.20%	48.90%
Rock	49.80%	42.20%

IV. CONCLUSION

This study aims to classify and recommend songs using acoustic features, extracted by digital signal processing methods and convolutional neural networks. Study has been conducted over two steps; determining how features that will be used in recommendation are obtained and developing a service that recommends songs to user requests. Firstly, feature extraction has been carried out by means of digital signal processing methods and then CNN has been trained as an alternative feature extraction. Then acoustic features of songs are used in classification to determine the best classification algorithm and the best recommendation results. According to the results summarized in previous tables, SVM achieved better classification results than other methods. In addition, changing the window size and window type caused very small performance changes. When it comes to the classification performance of the effects of the features used, according to the Table-4 and Table-5 MFCC has better effect than the other methods. Using deep learning method showed that there is no considerable performance chance on music genre classification. Furthermore, SVM has achieved higher success than the CNN algorithm. As for music recommendations, because there is no objective metric about music recommendation, music recommendation by genre can be used a solution this challenge. According to the recommendation results, for some genres of music like Classical, the music recommendation is highly successful, while in some species the performance falls. In the future works, we will study on high level feature extraction methods and several deep learning architectures. Especially, since deep learning methods need high performance computing architectures, we will concentrate on this issue.

REFERENCES

[1] A. Tzanetakis, G. and Cook, P. “Musical genre classification of audio signal”, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 3, pp. 293-302, July 2002.

[2] Holzapfel, A. and Stylianou Y. “Musical genre classification using nonnegative matrix factorization-based features”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, pp. 424-434, 2008.

[3] Benetos, E. and Kotropoulos C. “A tensor-based approach forautomatic music genre classification”, Proceedings of the European Signal Processing Conference, Lausanne, Switzerland, 2008

[4] Lidy, T., Rauber, A., Pertusa, A. and Inesta, J. “Combining audio and symbolic descriptors for music classification from audio”, Music Information Retrieval Information Exchange (MIREX), 2007.

[5] Bergstra, J., Casagrande, N., Erhan, D., Eck, D. and Kegl B. “Aggregate features and AdaBoost for music classification”, Machine Learning, Vol. 65, No. 2-3, pp. 473-484, 2006.

[6] A. Karatana and O. Yildiz, “Music genre classification with machine learning techniques,” in Signal Processing and Communications Applications Conference (SIU), 2017 25th, IEEE, 2017, pp. 1–4.

[7] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

[8] Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[9] Cortez P., In press. RMiner: Data Mining with Neural Networks and Support Vector Machines using R. In R. Rajesh (Ed.), Introduction to Advanced Scientific Softwares and Toolboxes.

[10] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5– 32, Oct. 2001.

[11] G. Gwardys and D. Grzywczak, “Deep image features in music information retrieval,” Intl Journal of Electronics and Telecomunications, vol. 60, no. 4, pp. 321–361, 2014.

[12] https://librosa.github.io/librosa/ (visited on 15/06/2018)

[13] https://keras.io (visited on 15/06/2018)

[14] S. Dieleman. (2014). Recommending music on spotify with deep learning,[Online].Available: http://benanne.github.io/2014/08/05/spotifycnns.html (visited on 03/31/2018

[15] Ahmet Elbir, Hamza Osman Ilhan, Gorkem Serbes, Nizamettin Aydin. "Short Time Fourier Transform based music Genre classification" , 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018