POLITECNICO
MILANO 1863

Project: Classify musical genre using audio files with K-Nearest Neighbors algorithm

Careddu Gianmario
Zoccheddu Sara

Prof. Edie Miglio

February 2022

1. Introduction

Music is often categorized into genres, in order to serve purposes such as song recommended applications based on genres. Therefore a classification that allows to label a song is useful.
The aim of this project is to classify songs into musical genres by using the K-Nearest Neighbors algorithm.

In order to do so, we used audio files of 30 seconds in length and extracted the most relevant features, which we used to feed the K-NN algorithm, which will be better explained later. The project explores several different possibilities of combining the features and shows the obtained classification accuracy.

2. Data set

We used a dataset with 1000 songs, divided in ten genres each with 100 songs. The songs are 30 seconds long and 22050 Hz Mono 16-bit audio files (WAV).
The WAV format contains the information regarding the amplitude of the wave composing the audio file.

3.Features

For what concerns time domain we have used Zero Crossing Rate (ZCR), Silent Ratio (SR) and the energy of the song signal (E). The analysis in frequency domain has been conducted using Mel-frequency cepstral coefficients (MFCCs) and in addition the dynamic features "delta" and "delta-delta" (first- and second-order frame-to-frame difference) coefficients.

- The **Zero Crossing Rate** indicates the frequency of the signal amplitude sign change and it is used as an indicator of how dynamic the music is. It is calculated in the following way:

$$ZCR = \frac{\sum_{n=1}^{N} |\operatorname{sgn} x(n) - \operatorname{sgn} x(n-1)|}{2N}$$

$$\operatorname{sgn}(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$

sgn $x(n)$ is the sign of current sample
sgn $x(n-1)$ is the sign of previous sample
N is the total sample

- The **Silence Ratio** is an indicator of the beat of the music and indicates the portion of the sound that is below a certain threshold (we used the 80% of the average energy of the training set) :

$$SR = \frac{\sum s}{\sum l}$$

s is the silent periods
l is the total length of the audio

- The **Average Energy** measures the loudness of the sound and it may indicate the beat and the flow of the music:

$$E = \frac{\sum_{n=0}^{N-1} X(n)^2}{N}$$

$x(n)$ is the value of sample n
N is the total sample

- **MFCCs** have been calculated for 13 frequency bands, each band has associated 1293 coefficients, the number has been selected empirically.

- "**delta**" and "**delta-delta**" are an approximation of the first and second order derivative of the signal in frequency domain, that can be used to better represent the dynamic of the signal.

4.Features processing

To reduce the dimensionality of the MFCCs space we have used two main procedures: the average energy for mel bands (allowed to reduce 1293 dimensions into 1 for each band) and the principal component analysis (to reduce the space from 13 to some n < 13 for each song). Then, we obtained the following new features:

- **PC of average energy per Mel band**
  Once we found the 13 mel frequency bands, we calculated the average energy.Then we ran the PCA to further reduce the dimensionality.

- **PC of average energy per comprehensive matrix feature**
  We built a “comprehensive” matrix by concatenating the standard mfccs with their deltas and delta-deltas. In this way we obtained 39 features. Then we calculated the average energy for each one and finally we ran the principal component analysis.

5.K-NN Algorithm

The **K-Nearest Neighbors (K-NN)** algorithm is a commonly used non-parametric classifier that, given a test point, polls the *k* nearest training points and classifies the test point according to the majority of the class label of the nearest neighbors. The distance we used in the Euclidean distance. We use 90% of the data for our training set, and 10% of the data for our testing set.

6.Results

We have run 23 versions of the algorithm combining together different features and processed features from both time and frequency domain. The most performant experiments have been reported in the tables below with the K used in the K-NN algorithm and the accuracy obtained in the test set.
Upon reading of the two provided papers and given the not brilliant results obtained classifying a large number of classes we have decided to concentrate our experiments on the more promising ones.

3 Genres: Classical, Metal, Pop

Experiment number	K	Accuracy
9	4	1.0
11	4	1.0
12	1, 2, 4	1.0
13	1, 3, 4, 6	1.0
14	1, 3, 4, 5, 6	1.0
21	1, 2	1.0
3	2, 3, 4, 5, 6	0.967

4 Genres: Classical, Metal, Pop, Rock

Experiment number	K	Accuracy
12	16, 17	0.875
13	11, 12, 13, 14, 15	0.875
4	5	0.85
12	13, 14, 18	0.85

4 Genres: Classical, Metal, Pop, Country

Experiment number	K	Accuracy
7	1	0.925
14	1	0.925
7	2	0.9
7	3, 4, 5	0.875
12	7, 8	0.875

4 Genres: Classical, Metal, Pop, HipHop

Experiment number	K	Accuracy
8	6	0.925
20	11, 13, 20	0.925
7	6, 9	0.9
8	8	0.9
13	7	0.9

5 Genres: Classical, Metal, Pop, HipHop, Country

Experiment number	K	Accuracy
20	9, 11, 17	0.88
8	6	0.86
14	1	0.86

20	7, 8, 13, 15, 19	0.86
8	4, 5, 8, 9	0.84

6 Genres: Classical, Metal, Pop, Blues, HipHop, Country

Experiment number	K	Accuracy
14	2	0.817
14	7	0.8
19	9	0.8
20	13, 15, 16, 17, 19	0.8
22	9	0.8

10 Genres: Classical, Metal, Pop, Blues, Disco, Jazz, Hiphop, Country, Rock, Reggae

Experiment number	K	Accuracy
13	15	0.53
12	14	0.44
14	7	0.44
20	10	0.44
19	9	0.42

The list of the most significant experiments is reported below, with the features employed to run the K-NN algorithm:

3. PC1 and PC2 of average energy per Mel band
4. PC1, PC2, PC3 of average energy per Mel band
7. PC1, PC2 of average energy per Mel band and ZCR
8. PC1, PC2, PC3 of average energy per Mel band and ZCR
9. PC1, PC2, PC3, PC4 of average energy per Mel band and ZCR
11. PC1, PC2, PC3 of average energy per comprehensive matrix feature
12. PC1, PC2, PC3, PC4 of average energy per comprehensive matrix feature
13. PC1, PC2, PC3, PC4, PC5 of average energy per comprehensive matrix feature
14. PC1, PC2, PC3, PC4, PC5, PC6 of average energy per comprehensive matrix feature
19. PC1, PC2 of average energy per comprehensive matrix feature and ZCR
20. PC1, PC2, PC3 of average energy per comprehensive matrix feature and ZCR

21. PC1, PC2 of average energy per comprehensive matrix feature and SR
22. PC1, PC2 of average energy per comprehensive matrix feature , SR and ZCR

7.Conclusions:

The highest accuracy is obtained in the trial with 3 music genres (Classical, Metal, Pop). Several experiments, in particular the ones involving the average energy per comprehensive matrix features, reported a 100% accuracy with k up to 6.

When including a 4th genre, the accuracy dropped a little: the highest results are obtained in the experiments involving more PCs. In this case, also the ZCR plays an important role, in particular for experiments n.7 and n.8

The trials with 5 and 6 genres show again a little loss in accuracy. The highest results are obtained with experiments that make use of the average energy per comprehensive matrix features. In particular we notice how in the trial with 5 genres the highest accuracy is obtained with experiment n.20, which involves the first three PCs of the average energy per comprehensive matrix feature and ZCR, with k=9, 11, 17; for the 6 genres trial instead, the best result is obtained using 6 PCs of the average energy per comprehensive matrix feature.

Finally, we ran the trial with all the 10 musical genres. We got a top accuracy of 53% with k= 15; the most significant experiments are the ones involving five PCs of the average energy per comprehensive matrix feature.

We noticed that when the experiments using the PCs of average energy per Mel band well-performed, also the ones using PCs of average energy per comprehensive matrix features would perform with the same or eventually high accuracy, the only exception was the 4 genres classifier with country genre: this indicates that the second features are in general more reliable and stable to make predictions than the first ones.

A general trend we have noticed is that when the number of classes increases it is better to use more features from the frequency domain, in particular from the PCs of the energy per comprehensive matrix features. The features in the time domain seem to perform better in few class trials; however, the features in the frequency domain are still the most performing ones, even when considering few classes. In fact, the versions of the algorithm using only time domain features have never reached significant accuracy.

References:

1. Comparison of Music Genre Classification Using Nearest Centroid Classifier and k-Nearest Neighbours (Elizabeth Nurmiyati Tamatjita, Aditya Wikan Mahastama)
2. Music Genre Classification (John Cast, Chris Schulze, Ali Fauci)