Art and Performance Letters (2021)
ISSN 2523-5877 Vol. 2 Num. 5

DOI: 10.23977/artpl.2021.020510
Clausius Scientific Press, Canada

Analysis of music similarity based on Pearson correlation coefficient

Zhiqun Li[1], Lusheng Wang[2], Dandan Wang[3]

[1]*School of Business Administration, Harbin University of Commerce, Harbin, Heilongjiang 150028*

[2]*School of Energy and Civil Engineering, Harbin University of Commerce, Harbin, Heilongjiang 150028*

[3]*Finance School, Harbin University of Commerce, Harbin, Heilongjiang 150028*

Keywords: K-means, Pearson correlation coefficient, Coefficient of Variation, music.

Abstract: In order to understand the role of music in our human lives, it is very important to develop a method that can quantify the evolution of music. First, we selected seven indicators of music characteristics, used **Pearson's correlation coefficient** to construct a music similarity measurement model, and performed a cluster analysis of artists within the genre to solve the Pearson correlation coefficient within and between genres. It is concluded that the music similarity within genres is higher than that between genres. Next, we selected some indicators through principal component analysis, observed the change curve of indicators over time, and analyzed how the genre changes over time. Finally, we use the **coefficient of variation** method to analyze the musical characteristics of influencers and followers, and obtain the relationship between the appeal of musical characteristics.

1. Introduction

Music as an important spiritual wealth is worth exploring the mysteries of it. In order to understand the importance of music to mankind at a deeper level, it is necessary to build a music similarity measurement model, according to which to compare whether artists within genres are more similar than artists between genres.

We used the Pearson correlation coefficient method to measure music similarity. According to the music characteristics mentioned in the question, we selected seven variables related to music characteristics, namely danceability, energy, valence, tempo, loudness, mode, key. First, we first standardize the required data. The correlation coefficient between genres can be obtained by calculation, and then the music similarity of each genres can be judged. We also divided the artists in each genre into four categories through cluster analysis, and then calculated the correlation coefficients between the four categories of artists in the genre, and compared the correlation coefficients between the artists within the genre and the correlation coefficients between different genres.

2. Musical similarity model

2.1 Data preprocessing

(1) Data consistency

Since the data in the loudness column in the data_by_artist table is negative, the data in the loudness column needs to be consistent, because the column values are all in [-60,0], we can take the opposite number of the data to turn the indicator into Positive indicators.。

(2) Data standardization

Since the dimensions of the data in each column in the data_by_artist table are different, in order to obtain numerical features that describe the correlation between variables and are independent of the dimension, we consider standardized data.

(3) Data cleaning and integration

We add each artist in the data_by_artist table to the column of the genre according to the corresponding relationship in the influence_data. Through data integration, we found that some artists do not belong to the genre, so we categorized this part of the data to avoid causing the result interference.

2.2 Build music similarity model

2.2.1 The meaning of Pearson's correlation coefficient

The Pearson correlation coefficient is used to measure the linear correlation between two variables X and Y, and its value is between -1 and 1. The definition of Pearson's correlation coefficient is shown in the figure below:

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}} \tag{1}$$

$Cov(X,Y)$ is the covariance of X, Y; $D(X)$, $D(Y)$ is the variance of X, Y; the smaller the value of $\rho(X,Y)$, the smaller the correlation of X, Y [1,2].

2.2.2 Calculate Pearson's correlation coefficient

(1) Correlation coefficients between different genres

Import pre-processed data into SPSS software, use genre as the identification code of each artist, and treat each genre as a category, and use Pearson correlation analysis to get the correlation coefficient between different genres. Thus, we can get a correlation coefficient matrix with dimension 20.

(2) Correlation coefficient within the genre

We use the k-means algorithm to divide the artists in each genre into four categories[3], and then calculate the correlation coefficient between the four categories. In the end, we calculated the correlation coefficients of 20 genres, and got 20 matrices with dimension 4.

2.3 Model results and analysis

(1) Correlation coefficients between different genres

According to the result, we can know that the correlation coefficient between Blues and Jazz values is 0.657, the two are the best; Blues and Religious have -.785, the two show a negative correlation. The correlation coefficient between Children's and New Age is 0.521, and the correlation coefficient between Children's and Electronic is -0.714, which appears in a significant negative correlation. The

correlation coefficient between Classical and Avant-Garde is 0.855, there is a significant positive correlation between the two, and the correlation coefficient between Classical and Latin is -.859, there is a significant negative correlation between the two, and the similarity is weak.

(2) Correlation coefficients within genres

We have obtained correlation coefficient matrices of four types of artists in 20 genres through cluster analysis and Pearson correlation coefficient analysis. Since there are many correlation coefficient matrices, we only show four as an example.

According to the correlation coefficient matrix in the table, the correlation coefficients of the artists within the Avant-Garde genre are all 1.000, indicating that the correlation within the genre is very large, while the maximum correlation coefficient between Avant-Garde and other genres is 0.855, and it has a correlation with other genres. The correlation coefficient is likewise significantly smaller than the correlation coefficient within the genre. The correlation coefficients of artists within the Blues genre are all greater than 0.500, and the overall correlation coefficients are significantly greater than those of other genres. Combining the above data and analysis results shows that artists within the genre more similar than artists between genres.

2.4 Genre changes over time

2.4.1 Selection of indicators

We take the genre as the main body, the time of each music creation as the abscissa, and some indicators about the characteristics of the music as the ordinate, draw a graph, and we can know how each genre changes with time. The following is the selection of indicators.

FAC1: Select the seven indicators of Characteristics of the music in the full_music_data table, danceability, energy, valence, tempo, loudness, mode, key. However, in order to express the simplicity of the results, we used SPSS to perform principal component analysis on the seven indicators, and finally synthesized a new indicator, namely FAC1, which can be used to measure the musical characteristics of a genre over time.

FAC2: Selected the four indicators of Type of vocals, acousticness, instrumentalness, liveness, speechiness. Similarly, we conducted a principal component analysis of these four indicators to get a new variable to measure Type of vocals, namely FAC2. This variable can be used to measure the changes of genres over time.

Number of songs: The total number of songs released by each genre in the year.

Popularity1: The average popularity of all music released by each genre in the year, and the ratio of the sum of the popularity of all music released in the year to the Number of songs for each genre. We standardise this data.

2.4.2 Result analysis

Since the final result is twenty pictures, but due to space limitations, we only take one genre for analysis. Take Comedy/Spoken as an example.

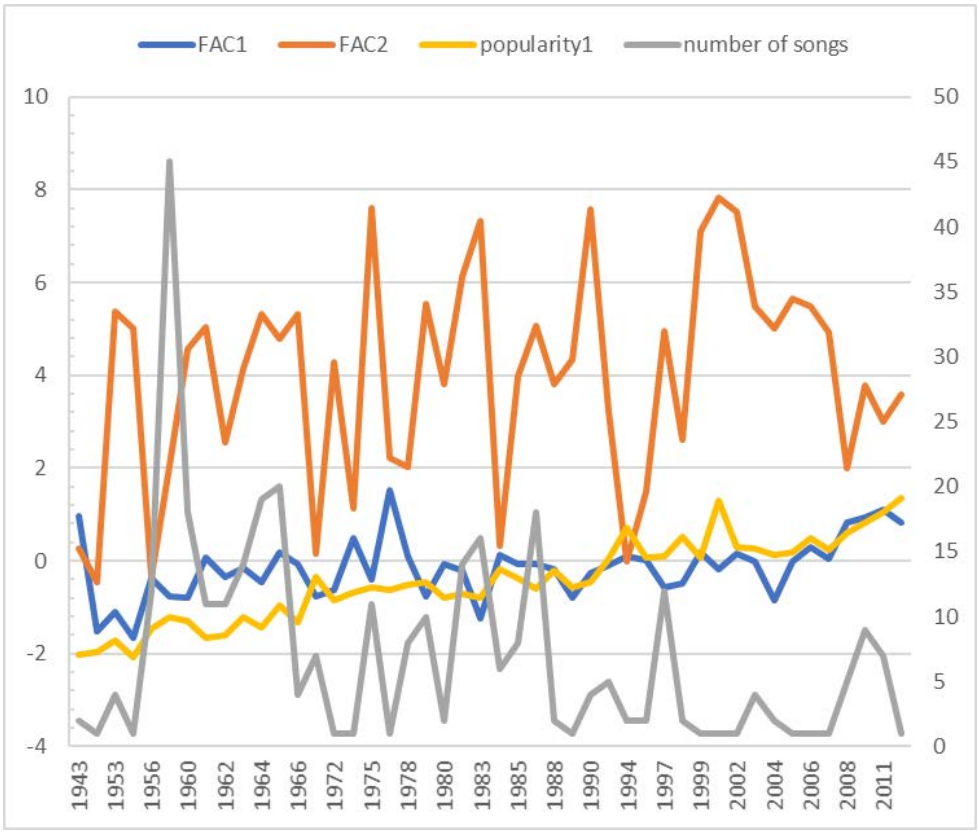

Figure 1: The trend

First, we can see that the fluctuation of FAC1 is small, which means that the indicator changes little over time and is relatively stable. However, the FAC2 indicator fluctuates greatly and frequently, which proves that the internal factors of the indicator change greatly over time and are unstable. It can be seen from the figure that the value of Popularity1 is on the rise, which proves that the popularity of its music continues to increase over time. It can be seen from the figure that Number of song released forty-five songs in this genre in 1959, which is the highest record of songs released by this genre every year.

2.5 Coefficient of variation method

In order to explore the influence of influencers on followers more clearly, we analyzed the musical characteristics of influencers and followers. According to correspondence between influencers and followers in the influence_data table, and using the music characteristics in the data_by_artist table, we can learn the music characteristics of the influencers and the corresponding followers. We have selected seven indicators of music, according to the correspondence in the influence_data table, for each indicator, we can find the deviation between the influencer and the follower, and then calculate the average and standard deviation of the deviation, and finally the coefficient of variation of each indicator can be obtained. According to the size of the coefficient of variation, we can judge which indicator is more infectious between influencers and followers. Import the data into matlab and solve it through programming. The results are shown in the following table.

Table 1: The results

f_i	danceability	energy	valence	tempo	loudness	mode	key
μ_{fi}	-0.00788	-0.06066	0.026869	-1.73969	-1.34546	0.051834	-0.51649
σ_{fi}	0.125722	0.181585	0.190491	19.23779	3.775063	0.430769	5.080928
c_v^{fi}	15.95284	2.993336	7.089681	11.05818	2.805787	8.310572	9.837409

μ_{fi} : Average of the f_i indicator.

σ_{fi} : Standard deviation of the f_i indicator.

c_v^{fi} : Coefficient of Variation of the f_i indicator.

According to statistical principles, the smaller the coefficient of variation of a certain musical feature, the stronger the stability of the musical feature between influencers and followers, that is, the more infectious the music feature. The analysis shows that the relationship between the appeal of these musical features is loudness>energy>valence>mode>key>tempo>danceability. From this we draw the conclusion that different musical characteristics have different degrees of appeal, and not all musical characteristics have similar effects.

3. Model evaluation

3.1 Strength

(1) The new music influence index makes it possible to compare and analyze the music influence of music genres and artists that cannot be directly analyzed before, and this index has wide applicability.

(2) Pearson correlation coefficient was used to accurately depict the correlation between different music genres.

(3) Use a variety of visual methods to present data, results, logic, etc.

(4) Qualitative analysis and quantitative analysis are combined to analyze and explain the problem from various aspects.

3.2 Weakness

Due to the limited data used, there will be some errors between the model and the actual situation.

References

[1] ZHANG Shiqiang, LÜ Jieneng, JIANG Zheng, et al. Study of the correlation coefficients in mathematical statistics [J]. Mathematics in Practice and Theory, 2009, 39(19): 102-107.

[2] SHAO Fan, CHEN Chen, GE Miaojia, et al. Analysis of technology innovation and application based on the power line loss Pearson algorithm [J]. Scientific and TechnologicalInnovation, 2017(14): 54-55.

[3] Jain A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.