

Research on Music Influence Based on PCA and Complex Network

Shuai Zhang, Yuhan Wu, Ziming Huang, Farong Wang, Aobo Li

Anhui Polytechnic University, Wuhu, Anhui 241000

Keywords: PCA Model, Complex Network Model, Euclidean distance, Music Genre

Abstract: Music is an important product of human emotional expression. In this paper, SPSS, Gephi and other software were used to establish a complex network model, and mathematical and topological methods were used to analyze the relationship between influencers and followers, and to study the changes of music over time. First, we used PCA model to reduce the dimension of music features and took the generated indicators as the measurement standard of music similarity. In order to reflect the difference and derivation of genres, we combined PCA and complex network model to analyze. Finally, we conducted Euclidean distance consistency analysis based on the PCA model, which could obtain the similarity degree between musicians.

1. Introduction

Music, with its long history and diverse types, is a part of human society and an important part of cultural heritage. Its emergence and evolution are influenced by many factors, including their innate ingenuity, current social or political events, or other personal experiences. It could also be the influence of other music artists on the creators. To understand the role music plays in the collective human experience, we should gain a better understanding of how music evolves through societies over time.

2. PCA Model

2.1 The Calculation of the Correlation Coefficient Matrix

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{bmatrix} \quad (1)$$

In formula(1), r_{ij} ($i, j=1,2,\dots,p$) is the correlation coefficient of the original variable x_i and x_j which is calculated as.

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} \quad (2)$$

Because R is a real symmetric matrix (i.e. $r_{ij}=r_{ji}$), you only need to calculate its upper or lower

triangular elements.

The *data_by_artist* is standardized *full_music_data* the original data in the data system, and the correlation coefficient matrix is calculated by formula (2).

2.2 The Calculation of Feature Values and Feature Vectors

First characteristic equation solution is $|\lambda I - R| = 0$ and the characteristic value of lambda $\lambda_i (i=1, 2, \dots, p)$, and the characteristic value of lambda, i.e., $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$; Then the eigenvectors corresponding to the eigenvalue λ_i are respectively calculated $e_i (i=1, 2, \dots, p)$. The correlation coefficient matrix is used to calculate the eigenvalues, as well as the contribution rate and cumulative contribution rate of each principal component. As can be seen from the following table, the cumulative contribution rate of 1-6 principal components has reached 81.717%, so 1-6 principal components are required. Calculate the principal component contribution rate and the accumulation.

Contribution Rate Formula:

Main ingredient z_i contribution rate: $r_i / \sum_{k=1}^p \gamma_k$ ($i=1, 2, \dots, p$), cumulative contribution rate: $\sum_{k=1}^m \gamma_k / \sum_{k=1}^p \gamma_k$.

The result of data standardization and principal component analysis in SPSS statistical analysis software is obtained.

2.3 Analysis of Index Parameters

Get 5 and 6 indicator parameters for the data de-dimensionality of *full_music_data* and *data_by_artist*. Based influence_data dataset, *data_by_artist* and *full_music_data* data set are processed to add the genre to which they belong. The music created by the default artist here is the music of the main genre created by the artist.

3. Models of Influencers and Followers

Based on a preliminary analysis of the data, the document provides a one-to-one mapping of influencers and followers across 20 music genres. According to the requirements of the topic, a complex targeted network of influencer-followers should be established using Gephi. Then, according to each different music stream, the subnetwork is established, and the degree in each subnetwork is selected as the parameter of 'music influence'. Then, we will study whether the influencer has influence on the music characteristics of his followers. Here, our team will analyze the influencer with the largest number of followers, calculate The similarity between The followers of The Beatles and The Beatles, and The similarity between The followers of The Beatles and other artists, and then make a comparison to get The results.

3.1 A Targeted Network Model for Influencers and Followers

According to the requirements of the topic, we used Gephi to build a complex directed network of influencers and followers. Because the data is too large, we filter the data of smaller nodes in the production process, so as to simplify the complex network. The influencer-follower directed network diagram after filtering the small range degree is shown in the Figure 1.



Figure 1: The influencer-follower directed network

3.2 Similarity Measurement Model for Influencers and Followers

Analyze the influence with the largest number of followers, The Beatles, and then calculate the mean value of the comprehensive music features of its followers after dimensionality reduction. Finally, use the mean value to measure The Euclidean distance between the influence, Lime and The Veldt, two other artists selected at random,

$$e = \sqrt{(A_{1i} - \bar{A}_1)^2 + (A_{2i} - \bar{A}_2)^2 + \dots + (A_{ni} - \bar{A}_n)^2}$$

Where i represents the i th influencer.

A_n represents item N ; A_{ni} represents A_n indicator of the i th influencer; \bar{A}_n represents The average value of A_n about the followers of The Beatles.

The calculated values are shown in Table 1.

Table 1: Similarity measures between different artists and followers of the same artist

Influencer	A1	A2	A3	A4	A5	A6	e
The Beatles	0.07	0.5	0.45	0.76	1.47	0.31	1.785062
Lime	0.51	1.15	1.15	1.58	-0.4	0.31	2.463756
The Veldt	0.81	0.05	0.06	0.84	2.19	2.91	3.854474

It turns out that the musical characteristics of followers are closer to and more similar to the influencers they follow, meaning that "influencers" influence the music of their followers. The relationship between the three artists and the musical indicators of The Beatles' followers is reflected.

4. Models of genres

This paper introduces the similarity measurement model of the musical characteristics of artists in the same genre and the similarity measurement model of the musical characteristics of artists in

different two genres. This requires our team to integrate data of data_by_artist and in-fluence-data, then calculate the six indicators of musical characteristics of musicians in da-ta_by_artist obtained above, and finally build the model by Euclidean distance measurement.

Meanwhile, our team visualized the musical characteristics of each genre classified. Due to the limited space, we presented 8 genres here, which can reflect the characteristics of each genre and the musical differences among different genres.

4.1 Similarity Measures within and between Genres

Here, the similarity calculation of Pop/Rock, Country, and R&B is carried out with a large number of counts. First, the mean value of the six factors for all artists in the Country genre is calculated with the calculated values shown in Table 2.

Table 2: The average of 6 musical indicators for artists of the Country genre

Genre	$\overline{A1}$	$\overline{A2}$	$\overline{A3}$	$\overline{A4}$	$\overline{A5}$	$\overline{A6}$
Country	0.051152	0.391221	0.239078	-0.44523	0.124585	0.057281

Use the Euthry distance measurement to build the model

$$d = \sqrt{\sum_{i=1}^n (A_{1i} - \overline{A1})^2 + \sum_{i=1}^n (A_{2i} - \overline{A2})^2 + \dots + \sum_{i=1}^n (A_{ni} - \overline{A_n})^2}$$

Among them i represents the number of all artists in the genre; A_n Represents the NTH integrated musical feature; A_{ni} Represents a combination of musical characteristics for the i person; $\overline{A_n}$ Represents the mean value of the comprehensive music characteristics A_n of all artists of the Country genre;

Calculate the distance between each person in the Country genre and this genre, and the result is 1.543957;

The distance between Country genre and R&B genre and Pop/Rock genre is calculated according to the average of Country genre, respectively, and the resulting square values are 2.443814 and 1.831166. Obviously, artists within genres are closer together and have a higher degree of similarity, that is, the musical characteristics of artists within the same genre are more similar than those of artists across genres.

4.2 Visual Expression of Musical Characteristics of Genres

After the dimensionality reduction processing of data in data_by_artist, we will get 6 indicators, A1-A6, which comprehensively show the characteristics of each music genre. The da-ta indicators of each genre are summarized, the mean value is taken, and then the radar chart is used to show the 6 indicators of each genre's music, A1-A6, which reflects the characteristics of this genre. Due to the limited space, only eight schools are presented and analyzed.

We take the difference between Children's and Classical genres as an example to analyze, and compare the radar charts of the two genres, so that we can get the difference between the two genres as shown in Figure 2.

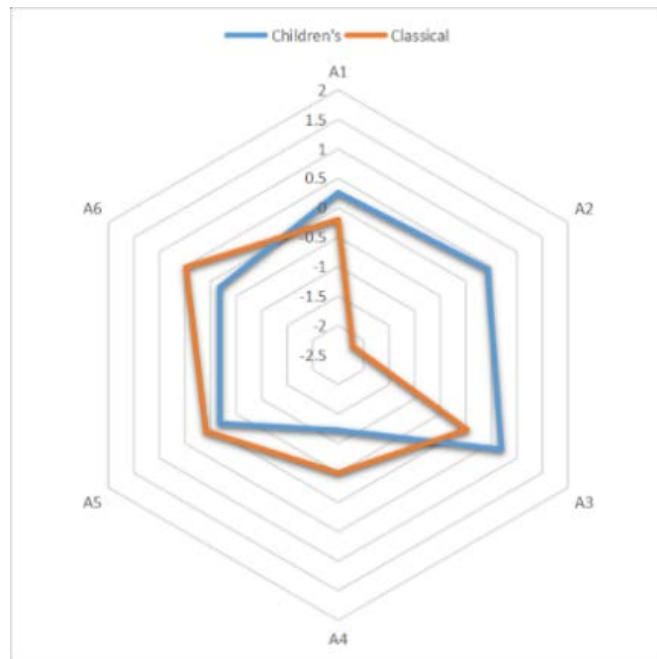


Figure 2: A comparison of the two music genres

As can be seen from the figure, the values of A1, A2 and A3 of Children's music are all higher than those of Classical music. According to the index meaning, it can be concluded that the indexes of Danceability, Valence and Energy of Children's music are higher than those of Classical music, while the indexes of Acousticness, Tempo and Duration_MS are lower than those of Classical music. This is consistent with Children's lively, short song time, suitable for dancing characteristics, and also with Classical, which pays attention to harmony and elegant style.

4.3 The Derivation of Genres

When we look at full_music_data, we find that most of the music existed in 1930 or before 1930, and some of the music was released after 1930, so we consider whether there are several existing music genres to generate a new kind of music.

We took the Electronic music from 1960, selected the genres followed by the Electronic followers from 1960 to 1980, and plotted its subnetwork.

We can see that in the early days of Electronic, Pop/Rock had the greatest influence on the genre, followed by R&B and Jazz, so it makes sense that the early Electronic style was similar to Pop/Rock.

5. Evaluation and Further

5.1 Advantages of the model

1. Uses the technique of dimensionality reduction to replace the original multiple variables with a few synthetic variables, which collect most of the information of the original variables.
2. The model evaluates the objective economic phenomenon scientifically by calculating the score of comprehensive principal component function. Thirdly, it focuses on the comprehensive evaluation of information contribution influence in application.

5.2 Disadvantages of the model

1. When the sign of factor load of principal component is positive or negative, the significance of comprehensive evaluation function is not clear. Low clarity of naming.

5.3 Further Discussion

1. Several weight indicators can be added to the influential directional network, which will make the directional network judgment more accurate.

References

- [1] Erdős, P.; Rényi, A. *On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci.* 1960, 5, 17–60.
- [2] Xuan Zhou, Fengming Zhang, Weiping Zhou, Wei Zou, Fan Yang. *Evaluation of functional robustness of complex networks based on node efficiency. Journal of physics*, 2012, 61(19): 190201. doi: 10.7498/aps.61.190201
- [3] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* 74, 47 (2002), DOI: 10.1103/RevModPhys.74.47. Crossref, ISI, ADS, Google Scholar.
- [4] Wang Xiaofan, Li Xiang, Chen Guanrong 2006 *Complex Network Theory and Its Applications* (Beijing: Tsinghua University Press)