Music Evaluation Based on Principal Component Analysis and Pearson Correlation Coefficient

Yike Xu, Zujun Hu

School of Software Engineering, Beijing Jiaotong University, Beijing, China

Keywords: PCA, Pearson Correlation Coefficient (PCC), k-means algorithms.

Abstract: Music and society are mutually constitutive. Initially, a preliminary data cleaning was conducted. And in order to understand the influence relations between various artists, we develop the directed influence network model. Meanwhile, ArticleRank was employed as a quantitative indicator to capture musical influence. With mounting quantities of data, we created subnetworks to explore the subsets of musical influence in more detail. Subsequently, at the request of capture the effective information accurately, we reduced the dimension of musical data by utilizing Random Forest and PCA analysis methods. To address the problem of comparing the similarity of music, Pearson Correlation Coefficient (PCC) was introduced the similarity measurement model was successfully constructed. Founded on what we have done, we accomplished the similarity analysis, in addition, we clustered genres by k-means algorithms and set up the genre development and comparison model to represent the similarities and influences between and within genres.

1. Introduction

Throughout history, music has played an essential part in mankind's daily life and has accumulated over billions of years of evolution [1]. As one of the most supreme expressions of life which builds the bridge between the spiritual and sensual world, to a certain extent, the evolutionary history of music is the mirror of human being's spiritual civilization. But how can the spiritual process be explained? For truly further analysis of music, it is necessary to view the music from a social and historical perspective [2]. A musical composition is the result of multiple factors, including subjective ideas and the objective world [3]. For instance, the tastes of the author, the artist who has a deep impact, and the song characteristics, are all constructive. However, aspects of such study are not very satisfactory. Furthermore, music is performed in a structured and continuous manner, which means it will very likely that the pieces belong to the same genre all own a specific law. This requires the use of relevant methods to verify or reveal the rules.

2. The Directed Influence Network Model

The method of knowledge graph is employed in our model construction. And we convert the influence network into a directed graph G= (V, A), where V the set of nodes are representing the influencers and followers. For further discussion, we select and build the appropriate indicators. In this model, all the persons have taken into consideration and their ID is the unique identifier so that we can avoid the same-name problem. There exists an edge $(i, j) \in A$ if and only if i influence j.

We explore a centrality algorithm to establish a parameter to capture music influence. Centrality algorithm is one of the traditional categories of graph algorithms. By implementing this, we can easily identify the important nodes, which also means the most influential artists can be found.

Rather, after strictly comparing and testing we finally pick ArticleRank algorithm to quantify the musical influence. ArticleRank, a variant of the PageRank algorithm, is original designed as an alternative of citation amount. ArticleRank is mainly used for analysing citation networks. After the contrast, we found that the influence relationship is very similar to the reference relationship. It is feasible to draw a parallel between the two networks. The algorithm can precisely reflect the following facts: the paper which has been cited by a high-impact paper is considered as an essential creation and should be given a higher weight instead of only judged by the citation amount. The citation relationship is also in line with the influence relationship built by artists.

The influence score is stated as the weight ofVi. The higher scores would indicate the greater musical influence on followers. In graphs, the influence can be represented in a more intuitive way, for the scores are visualized by node size. In this paper, ArticleRank is defined as follows:

$AR(A)=(1-d) +d(AR(T1)/C(T1) +C(AVG)) +\ldots+AR(Tn)/C(Tn)+C(AVG))$

- we assume that an artist A has artists T1 to Tn which point to it.
- d is a damping factor which can be set between 0 and 1. It is usually set to 0.85.
- C(A) is defined as the number of influences going out of artist A.
- C(A VG) is defined as the average number of influences going out of all artists.

It can be seen from the core calculation formulas that ArticleRank Algorithm does not just simply counts the number of followers every artist has. In order to make the result more reasonable, we consider not only the direct followers of each artist, but also the musical influence of the artist's followers. The AR(A) of every artist using this algorithm quantifies the music influence in the directional network we constructed. The larger the value, the greater the musical influence.

3. Similarity Measure Model

3.1 Characteristics Filtering by Calculating Feature Importance

Random forests classification algorithm is an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. In the random forest algorithm, if an attribute often becomes the best split attribute, then it is likely to be an information feature that needs to be retained. We randomly divide the song data set into 8:2 training set and test set. Then use all the characteristics of the song and use the random forest classification model to predict the genre of the song in 20 genres. The accuracy of the prediction on the test set reaches 66.29%. Therefore, we believe that these features can collectively reflect the style of a song to a large extent, and the model has a strong predictive ability on the genre of the song. The random forest model can effectively extract the style features of the song.

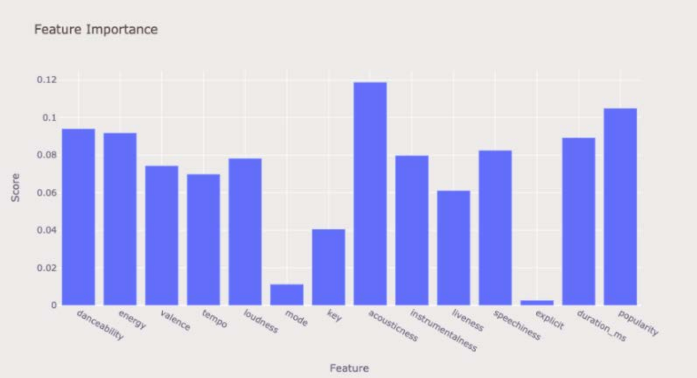


Figure. 1 feature importance

The model can calculate the importance of the features for predicting the genre of the song. As shown in Figure 1, through calculation, we can see that the three indicators of explicit, mode, and key have weak predictive capabilities for the genre of the song, meaning that these three characteristics cannot reflect song style. Impact indicators. In fact, in our common sense, explicit, mode, and key do not reflect the style of the song, so we exclude these three indicators from the consideration of song similarity analysis, and the data dimension is reduced from 14 to 11.

3.2 Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical method often used for reducing dimensionality. It transforms a group of variables with possible correlation into linearly unrelated ones through orthogonal transformation, and the converted variables are called principal components. It believes that the amount of information contained in a variable is usually measured by the variance or the sum of squared deviations, and the number of principal components is selected according to the variance contribution rate.

Generally, we select the cumulative variance contribution rate of the first few principal components as the analyzed principal components. In other words, most information can be described by these comprehensive evaluation indicators. Here, we select the principal components whose cumulative variance contribution reaches 95%.

Therefore, we merged the 11 characteristics retained after the first round of dimensionality reduction into 6 indicators, and the data dimension was reduced from 11 to 7.

3.3 Pearson Product-moment Correlation Coefficient

Among the commonly used similarity calculation methods for data mining, two types of methods, cosine similarity and Pearson correlation coefficient, are both suitable for this use scenario. After research and comparison, we chose the Pearson correlation coefficient with a higher degree of reasonableness and a wider range of applications as the standard to measure the similarity of songs.

In statistics, the Pearson correlation coefficient (PCC) is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

We use song genres as the classification standard, randomly select 50 singers from each music genre, and eliminate niche music genres with less than 100 people to reduce noise. We perform pairwise comparisons of songs within the same category to obtain the similarity between the songs and calculate the average value; also perform the same calculation between different categories to obtain the similarity and calculate the average value. The results are expressed in matrix form and visualized to generate a heat map, as shown in Figure 2.

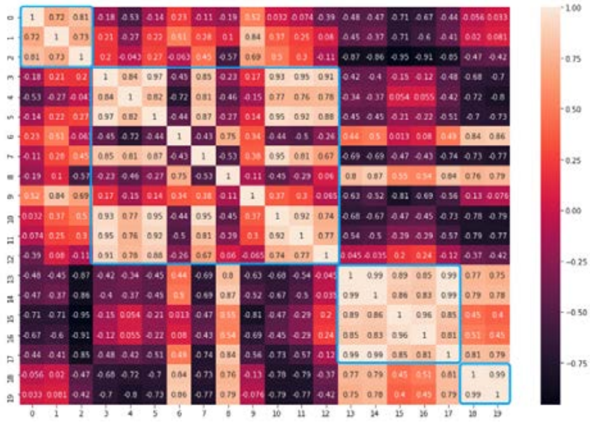


Figure. 2 feature importance

In the figure, the data on the diagonal line represents the average similarity of songs within the same category, and the data outside represents the average similarity of songs between different categories. Obviously, the value on the diagonal is always greater than the other values in the row and column. Therefore, we can know that the style similarity of singers within a category is always higher than that of singers between categories.

In the figure, we can also find that:

• The similarity between electronic and vocal is the lowest, at -0.37, indicating that electronic and vocal can be said to be two completely different music genres

• The music similarity between vocal and jazz is the highest at 0.41, indicating that these two music styles have relatively more in common.

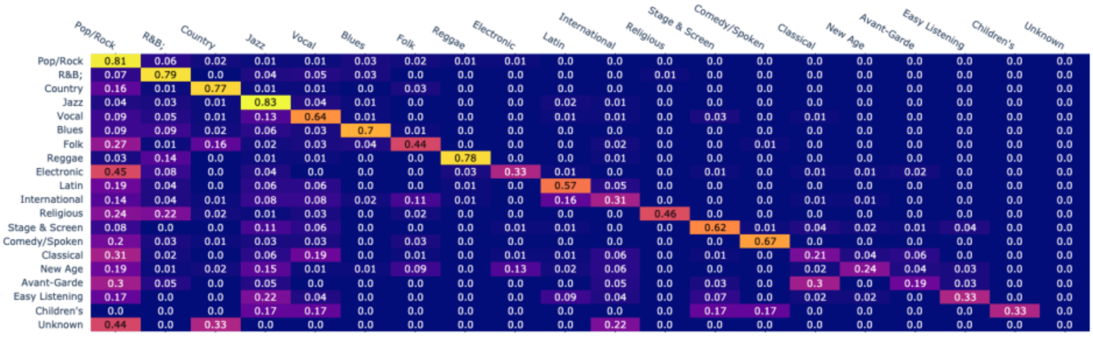
• The music similarity in the vocal music genre is the highest, at 0.8, indicating that the music style of the vocal genre is highly unified

• The music similarity within the R&B genre is the lowest, only 0.16, indicating that R&B’s music style is very extensive and contains a very high degree of inclusion

4. Conclusions

The trends of the music genres in the first cluster (jazz, vocal, blues, folk, international) are similar, and they tend to decline, indicating that people are gradually losing interest in music with higher acousticness.

We can also explore the degree of influence between each genre, by counting how many artists in each music genre are affected by the total number of artists from each genre, to calculate the proportion of influencers of each genre, as shown in Figure 3.



	Pop/Rock	R&B;	Country	Jazz	Vocal	Blues	Folk	Reggae	Electronic	Latin	International	Religious	Stage & Screen	Comedy/Spoken	Classical	New Age	Avant-Garde	Easy Listening	Children's	Unknown
Pop/Rock	0.81	0.06	0.02	0.01	0.01	0.03	0.02	0.01	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R&B;	0.07	0.79	0.0	0.04	0.05	0.03	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Country	0.16	0.01	0.77	0.01	0.01	0.0	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jazz	0.04	0.03	0.01	0.83	0.04	0.01	0.0	0.0	0.0	0.02	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vocal	0.09	0.05	0.01	0.13	0.64	0.01	0.0	0.0	0.0	0.01	0.01	0.0	0.03	0.0	0.01	0.0	0.0	0.0	0.0	0.0
Blues	0.09	0.09	0.02	0.06	0.03	0.7	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Folk	0.27	0.01	0.16	0.02	0.03	0.04	0.44	0.0	0.0	0.0	0.02	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0
Reggae	0.03	0.14	0.0	0.01	0.01	0.0	0.0	0.78	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Electronic	0.45	0.08	0.0	0.04	0.0	0.0	0.0	0.03	0.33	0.01	0.0	0.0	0.01	0.0	0.01	0.01	0.02	0.0	0.0	0.0
Latin	0.19	0.04	0.0	0.06	0.06	0.0	0.0	0.01	0.0	0.57	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
International	0.14	0.04	0.01	0.08	0.08	0.02	0.11	0.01	0.0	0.16	0.31	0.0	0.0	0.0	0.01	0.01	0.0	0.0	0.0	0.0
Religious	0.24	0.22	0.02	0.01	0.03	0.0	0.02	0.0	0.0	0.0	0.0	0.46	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stage & Screen	0.08	0.0	0.0	0.11	0.06	0.0	0.0	0.0	0.01	0.01	0.0	0.0	0.62	0.01	0.04	0.02	0.01	0.04	0.0	0.0
Comedy/Spoken	0.2	0.03	0.01	0.03	0.03	0.0	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.67	0.0	0.0	0.0	0.0	0.0	0.0
Classical	0.31	0.02	0.0	0.06	0.19	0.0	0.01	0.0	0.01	0.01	0.06	0.0	0.01	0.0	0.21	0.04	0.06	0.0	0.0	0.0
New Age	0.19	0.01	0.02	0.15	0.01	0.01	0.09	0.0	0.13	0.02	0.06	0.0	0.0	0.0	0.02	0.24	0.04	0.03	0.0	0.0
Avant-Garde	0.3	0.05	0.0	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.03	0.0	0.3	0.0	0.19	0.03	0.0	0.0
Easy Listening	0.17	0.0	0.0	0.22	0.04	0.0	0.0	0.0	0.0	0.09	0.04	0.0	0.07	0.0	0.02	0.02	0.0	0.33	0.0	0.0
Children's	0.0	0.0	0.0	0.17	0.17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.17	0.17	0.0	0.0	0.0	0.0	0.33	0.0
Unknown	0.44	0.0	0.33	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure. 3 The Influence between Genres

Through this figure, we can figure out the influence and evolution process between the genres. For example, Country is greatly influenced by Pop/Rock, and Pop/Rock has an influence on almost every music genre due to its wide spread and wide audience.

References

[1] Lena, Jennifer C. Extreme Metal: Music and Culture on the Edge. [J]. American Journal of Sociology, 2007.

[2] Y u Runyang. Musicological Analysis of the Prelude and Finale of Tristan and Isolde (Part II) [J]. Music Studies,1993.

[3] Garcia. A History of Western Music[M]. High School Journal, 2004.