ISSN 2616-5775 Vol. 4, Issue 2: 24-28, DOI: 10.25236/AJCIS.2021.040205

Genre Analysis based on Principal Component Analysis and Random Forest

Ziang Li

School of Economics and Management, Wuhan University, Wuhan, Hubei, 430072

Abstract: *When you listen to music, have you ever noticed that some songs are similar to the other while some are quite different? Maybe this is a universal phenomenon, however, our work try to provide fascinating insights into the characteristics and development of music. We create the adjacency matrix of 5603 musical artists to demonstrate the relationship between them, then we draw a subnetwork to show the connection between the influencers and the followers of Classical music in the 1930s. After that, we try to quantify the influences and similarities between these artists, using the adjacency matrix and principal component analysis (PCA) method, and establishing two indicators: 'influence' and 'dissimilarities'. To test the practicability of the previous indicators, we apply them to analyze the genres of music, including the a comparison of influences and similarities between and within genres, the developing process of genres and the relationship of different genres. Especially, we use the random forest method to work out what distinguishes a genre. We find that artists within genre are statitically more similar to artists between genre.*

Keywords: *Network Analysis, Adjancency Matrix, Principal Component Analysis,* Random Forest

1. Introduction

When you listen to music, you may find some melodies are similar to the other while some are quite different, and that music produced in different times may have various features. Have you ever thought that how this could happen? Music has always played an important part in human society and the development of music is an epitome of the revolution of human civilization. The study of music can give us fascinating insights into the cultural or historical factors [1]. Sometimes, profound shifts happen in music, whereas sometimes there are only a sequence of small changes. So develop a proper method to accurately quantify music revolution is a meaningful work. A key step during this process is to distinguish the impact of the influencer on his followers. In addition [2], we live in an age of rapid advance of the Internet which may exert unprecedented influences on the revolution process of music, therefore, we also need to determine these effects within the network according specific conditions in order to achieve better results.

2. A subnetwork

We use the *full_music_data* data set to create a directed network of musical influence. Because the data set is too large, visualization of the whole data set has a bad result. So we choose artists of 'Classical' genre in the 1930s to draw a subnetwork which is shown in the following figure.

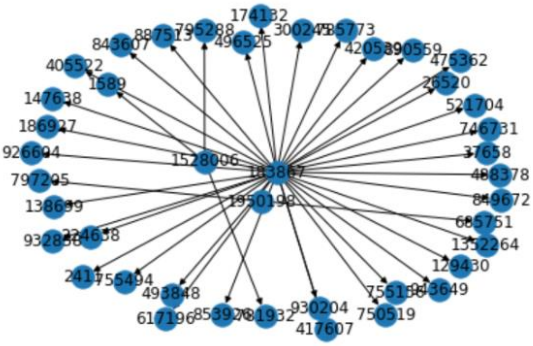

Figure 1: Musical Influence of Classical Artist in the 1930s

ISSN 2616-5775 Vol. 4, Issue 2: 24-28, DOI: 10.25236/AJCIS.2021.040205

From the subnetwork above, we can see that it is in a radial pattern. Artist 183867 has the strongest musical influence, almost the rest of artists are influenced by him. Besides, Artist 1528006 influenced another three artists, and artist 797205 and 685751 have bidirectional influences.

According to the influence we have calculated, the artist 183867 in the classical music genre of the 1930s should be in a more central position than the artists 1528006 and 1950198, which is consistent with the illustrated sub-network.

3. Similarity construction: Based on PCA analysis

To measure music similarity, we first need to analyze music characteristics [3]. From the full_music_data data set, we choose 13 variabes, including danceability, energy, valence, ect. to describe music characteristics. Because so many variables will unnecessarily increase the complexity of our work, so we adopt principal component analysis (PCA) method here to realize dimensionality reduction. The result shows that the thirteen-dimension charac- teristics is reduced to a one-dimension vector.

From the table above, we find out that the first principal component can explain 99.99% of the variance, so we only need to choose PC_1as the result of our principal component analysis. Specifically, we have the eigenvector:

$$[-1.417e^{-7}, 8.022e^{-8}, -4.348e^{-7}, -5.569e^{-6}, -1.116e^{-6}, -3.433e^{-7},$$
$$-1.456e^{-7}, -3.410e^{-7}, 2.888e^{-7}, 1.049e^{-7}, -1.268e^{-8}, 1.000, 9.066e^{-6}]$$

It is the coefficient vector of the variables in the first principal component.

Next, we can develop measures of music similarity. We adopt the calculated principal component to artists in the data_by_artist dataset, and obtain the ‘comprehensive music characteristic’ of each artist. Because the distance between points in a coordinate system naturally reveals difference, or similarity, on the opposite, so we define ‘dissimilarity’ as the L1 norm of artists’ comprehensive music characteristic. Longer distance indicates that artists are less similar to each other.

4. An application to genres

4.1 Comparison between and within genres

In order to make comparison between and within genres, we partition our adjacency matrix as it is shown in the following diagram [4].

	A	B	C	D	E	F
A	1	1	2	2	2	2
B	1	1	2	2	2	2
C	2	2	1	1	2	2
D	2	2	1	1	2	2
E	2	2	2	2	1	1
F	2	2	2	2	1	1

Figure 2: The Diagram of Partitioned Adjacent Matrix

Suppose that letter A to F refers to six artists, A and B belong to one genre, C and D belong to a second one, E and F belong to a third one. The range in red indicates the similarities or influences between genres while the range in yellow indicates the similarities or influences within genres. So the similarities or influences have now been put into two categories. Notice that here we have two adjacency matrixes. The first one’s elements are used to describe ‘influence’ (the 0 and 1 binary variable, just as the adjacency matrix showed in the first part), hence it is a digraph; whereas the second one’s elements are used to describe ‘dissimilarity’(as it was defined in the last part), hence it is an undigraph.

Before our comparison work, we first join the data_by_artist data set and the full_music_data data set to match the summarized information of each artist with his genre. This process is of vital importance

ISSN 2616-5775 Vol. 4, Issue 2: 24-28, DOI: 10.25236/AJCIS.2021.040205

for our following works.We named the new data set data_by_artist_with_genre.

Then we do a t-test of the within-genre category and the between-genre category to see whether there exists significant differences between and within genres.

To finish the t-test, we should first estimate the distribution of variables, which are 'dissimilarity_within_genres','dissimilarity_between_genres','influence_within_genres' and 'influence_between_genres' . Kernel density estimation (KDE) can take full use of the data information and avoid bringing into subjective judgements, so it will approximately gain the population distribution from the sample information. Therefore, we use KDE here and specifically, adopt Gaussian kernal and Scott Bandwidth estimation. The proper gridsize here is 100. The following figures are the results.

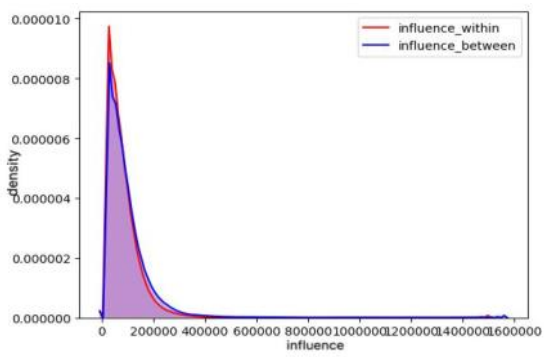

Figure 3: Distribution of Influence

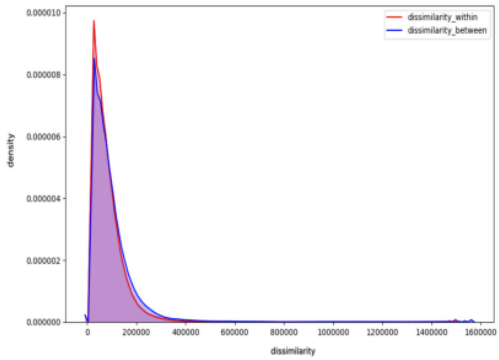

Figure 4: Distribution of Dissimilarity

The result of the t-test of dissimilarities between and within genres has a pvalue of 0.000, indicating that the similarities between and within genres are significantly different with each other. The t-statistic is 351.612, indicating that artists within genres are more similar than artists between genres.

The result of the t-test of influences between and within genres has a pvalue of 0.722 which shows that the influences between and within genres are indistinctive with each other. This does not necessarily mean that the artists between and within genres have the same degree of influence. Since from the original data set, the influence of artists are only represented in binary variables, which cannot reveal the exact intensity that one artist has on another. So if more accurate datas are avaliable, maybe we can judge whether there exist significant differences in influence between and within genres.

4.2 Development of genres

4.2.1 The random forest

We use the random forest method to work out what distinguishes a genre. The random decision forest is a useful sorter, so it is proper here to apply this method to select features. Using the data_by_artist_with_genre and the influence_data data sets as well as the 'com- prehensive music characteristic' of artists calculated from principal components analysis, we obtain features of music that distinguish a genre. During the machine learning process, we employ the 'Gini impurity' as the standard. There are 10000 trees in our forest, and the maximum depth is five. Here are the results.

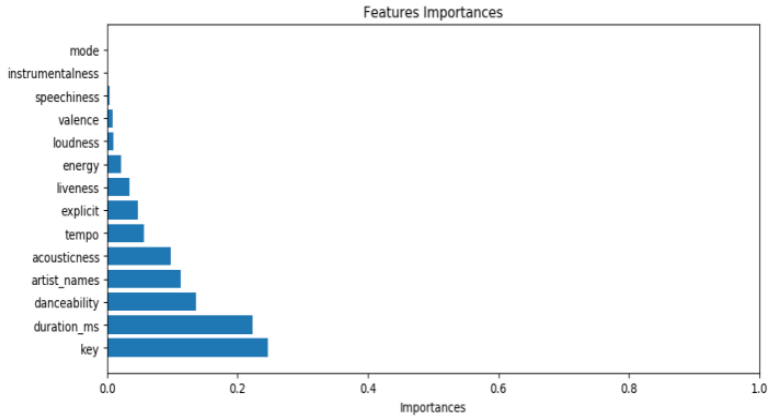

Figure 5: The Random Forest

ISSN 2616-5775 Vol. 4, Issue 2: 24-28, DOI: 10.25236/AJCIS.2021.040205

Clearly, the three main features which distinguish a genre are ‘key’, ‘duration_ms’ and ‘dancebility’

4.2.2 The change of genres over time

We draw a heat map to directly show the changes of the popularity of different genres over time. In the figure below, the color depth reflects log (∑popularity + 1).

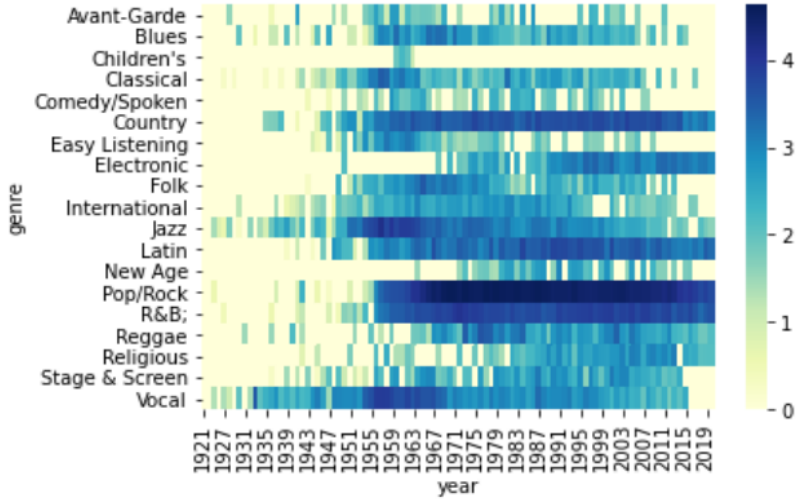

Figure 6: The Change of Genres over Time

It is clear that Pop and Rock music came into vogue from 1960s, and this trend lasts even till today. We can also discover from the figure that the changing process of R&B is almost the same as Pop and Rock music. However, Jazz and Vocal only enjoyed their heyday in the 1950s and 1960s but gradually lost popularity.Generally, during the 10 years from 1955 to 1965, there was a burst to almost all genres. All these features of the change of genres revealed from our figure are in accordance with history and reality.

4.3 Relationship between genres

Here we also use the heat map to demonstrate the relationship between genres. Assuming that the more similar one genre is with another, the closer their relationship is. So we follow our defination of ’dissimilarity’ to reveal the relationship between different genres. That is, two genres will be less related with each other if they have higher extent of dissimilarity. Our figure is shown below.

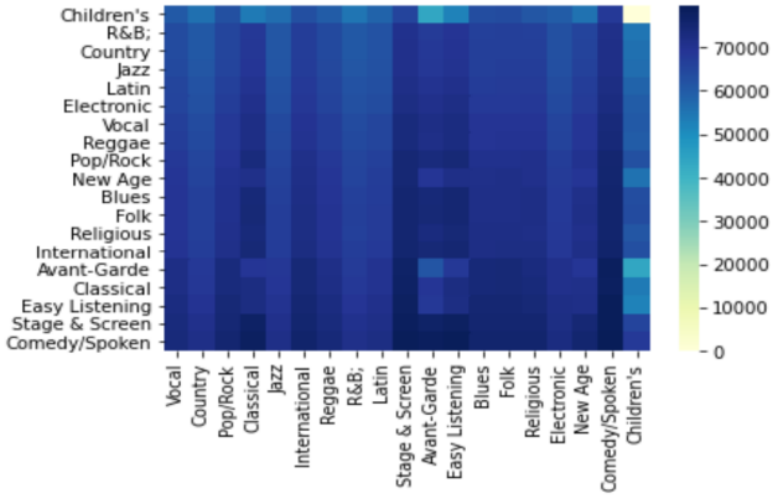

Figure 7: The relationship between genres

In the picture above, each color-filled grid indicates the mean value of the dissimilarities of all artists who belong to the same genre. Grids on the diagonal line refers to the dissimi- larities within genres while the rest of grids refers to the dissimilarities between genres. To solve this problem, we only need to focus on the later one. Childern’s, Jazz, Latin and have closer relationship with other genres, whereas

ISSN 2616-5775 Vol. 4, Issue 2: 24-28, DOI: 10.25236/AJCIS.2021.040205

Stage&Screen and Comedy/Spoken is the least related with others.

5. Conclusions

The conclusions of our model are as follows.

1) According to our definitions, artists within genre are statitically more similar to artists between genre, but when it comes to the influence, the conclusion is no longer tenable. What's more, the identified influencers do have an influence on other artists.

2) The distribution of genres change over time, and revolutions took place in 1951 and 1969.

3) We discoverd that comentropy of the distribution of genres is able to reflect revolutions.

4) Key, duration and danceability distinguish a genre. Genre, year and energy are more contagious in the casuality network.

References

[1] Erdős, P.; Rényi, A. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 1960, 5, 17–60.
[2] R.Albert and A.-L. Barabási, Rev. Mod. Phys. 74, 47 (2002), DOI: 10.1103/RevModPhys.74.47. Crossref, ISI, ADS, Google Scholar
[3] Tang, P.; Song, C.; Ding, W.; Ma, J.; Dong, J.; Huang, L. Research on the Node Importance of a Weighted Network Based on the K-Order Propagation Number Algorithm. Entropy 2020, 22, 364.
[4] Russell SJ, Norvig P (2009). Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition.