International Core Journal of Engineering Volume 7 Issue 4, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

Algorithm Design of Intelligent Music Track Recommendation System based on Collaborative Filtering

Yifan Liu[1,*], Yongjun Fan[1], Mingyue Sun[1] and Junke Chen[2]

[1]Department of Electrical and Information Engineering, Shandong University of Science and Technology, Jinan, 250031, China;

[2]Department of Finance and Economics, Shandong University of Science and Technology, Jinan, 250031, China.

Abstract

For information consumers, it is very difficult to find the information they are interested in from the large amount of information; For information producers, it is also very difficult to make their own information stand out and attract the attention of the majority of users. In order to solve this contradiction, we built the de-correlation model based on principal component analysis, the prediction music track rating model, and the collaborative filtering recommendation model, so as to solve the problem that users can find high-quality music tracks from many music tracks. Through the analysis, it is concluded that the factors influencing users' evaluation of music tracks are the number of music track labels and indirect attention. Then, according to the linear regression theory, the model of predicting music track score is established, which can predict the user's score to music track. The mathematical model established in this paper has strong portability and can be extended to the fields of network, media, film and television.

Keywords

Computer Science; Algorithm Design; Speech Recognition; Collaborative Filtering; Recommended Music Tracks.

1. Introduction

With the continuous development of information technology and the Internet, a large amount of information emerges before us. Faced with this information, it is difficult for users to find the content they are really interested in, and it is difficult for information providers to accurately convey high-quality information to interested users. Therefore, studying the problem of music scoring has very important application value for music providing software to recommend high-quality music tracks for users.

The premise of solving the problem is to find out the factors that affect users' rating of music tracks. It is necessary to mine the text information and database information given by the topic, reasonably analyze and screen the data given, find out the factors that may affect the score of music tracks, and study whether the selection factors can affect the user's evaluation of music tracks through the establishment of a model.

Subsequently, the user's ratings of unwatched music tracks are collected. Based on the found factors that affect users' ratings of music tracks, as parameters, a project-based rating prediction model is established.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

2. Factor screening and data processing

2.1 A de-correlation model based on principal component analysis

According to the collected label data, relational data and music track data, it is applied to carry out data mining. Aiming at the problem of how to eliminate the interference of irrelevant information, the principal component analysis method is used to eliminate the relatively large correlation index to get the final evaluation index.

Firstly, the correlation coefficient matrix is calculated:

$$R = \begin{cases} R_{11} & R_{12} & \cdots & R_{1P} \\ R_{21} & R_{22} & \cdots & R_{2P} \\ \cdots & \cdots & \cdots & \cdots \\ R_{P1} & R_{P2} & \cdots & R_{PP} \end{cases} \tag{1}$$

In Formula (1), R is the correlation coefficient of the original variable and, and its calculation formula is as follows:

$$R_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2 \sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}} \tag{2}$$

Because R is a real symmetric matrix, you only need to calculate the upper or lower triangular elements.

Then, according to the correlation analysis in Table 1, it can be seen that the number of music track tags has a relatively large correlation with the number of times played and appreciated, and the number of times read is eliminated. Finally, this paper obtains the factors that affect users' ratings of music tracks, as shown in Figure 1.

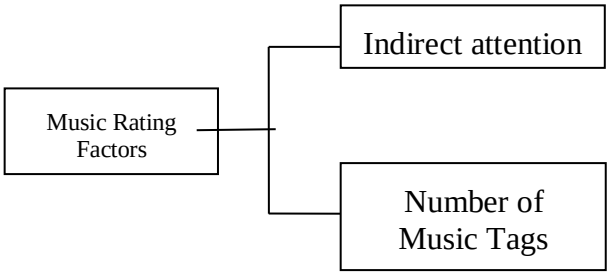

Fig. 1 Final determination of indicators

2.2 Predicting music track scoring algorithm model

2.2.1 Establishment of algorithm model

Tags can understand the reader's preference for the type of music, that is, the reader's appreciation characteristics. Appreciation features include the reader's implicit information about music tracks, and the relationship between music track scores and users can be obtained through relevant data mining. There are many methods for data mining, such as linear regression, machine learning system design, and support vector machines. All the machine learning processes involved in the literature can be regarded as the process of optimizing and solving mathematical model parameters. In a broad sense, the learning process can be transformed into an optimization problem. There are three elements in the machine learning process that affect the efficiency and effectiveness of its learning. *Hypothesis* function, cost*functions* function and *gradient−descent* function.

After data processing, the user's music track rating table is obtained, which is a two-dimensional vector scale, as shown in Table 1:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

Table 1. List of predicted and known data

User ID	7245481	4156658	9977150	7625225 (To predict)	x_1	x_2
473690	4	0	0	?		
929118	4	0	0	?		
235338	4	4	5	?		
424691	4	4	5	?		
916469	4	0	4	?		
793936	4	4	0	?		

Combining the advantages and disadvantages of various methods, this paper adopts the optimized multivariate linear regression [3] to conduct the relationship between music track rating and readers' appreciation characteristics.

Parameters were introduced for each reader, a monitoring method was constructed, and linear regression was carried out on the scoring matrix column by column and the model was optimized to obtain the model parameters. Multivariable assumes that the output is determined by multidimensional, that is, the input has multidimensional characteristics. Multiple linear regression model:

$$h_\theta(x) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots \theta_n \cdot x_n$$

In this paper, two features are selected for regression prediction. In order to enhance the accuracy of the model, constant term features x_0 and $\theta^j \in R^{n+1}$ are introduced for each reader, The optimization model is as follows:

$$\min_{\theta^{(1)},\cdots\theta^{(n_u)}} \frac{1}{2}\sum_{j=1}^{n_u}\sum_{i:r(i,j)=1}\left(\left(\theta^{(j)}\right)^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}\left(\theta_k^{(j)}\right)^2 \tag{3}$$

Followed by gradient descent update:

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \tag{4}$$

$$\sum_{i:r(i,j)=1}\left(\left(\theta^{(j)}\right)^T x^{(i)} - y^{(i,j)}\right)x_k^{(i)}(fork = 0) \tag{5}$$

Gradient decreasing univariate learning method of parameters:

$$\left\{\begin{array}{l} \theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(y)}\right), \\ \theta_1 := \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(y)}\right)x^{(i)} \end{array}\right\} \tag{6}$$

2.2.2 Solution of the model

Using MATLAB linear regression processing, and based on the parameter optimization program after mechanical learning training, the solving process and results of users with ID 7625225 to 6 are shown in Table 2 and Table 3.

Table 2. Preliminary processing of data

User ID	7245481	4156658	9977150	7625225 (To predict)	x_1	x_2
473690	4	0	0	1	0.9	0
929118	4	0	0	1	1.0	0.01
235338	4	4	5	3.2	0.99	0
424691	4	4	5	3.2	0.01	1.0
916469	4	0	4	2	0.1	1.0
793936	4	4	0	2	0	0.9

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

Table 3. Rating of six pieces of music by users with predicted ID number 7625225

User ID	7245481	4156658	9977150	Estimate(7625225)	Actual value(7625225)
473690	4	0	0	4.17	4
929118	4	0	0	4.14	4
235338	4	4	5	4.25	5
424691	4	4	5	4.31	4
916469	4	0	4	4.24	5
793936	4	4	0	4.24	5

2.2.3 Model checking

The user with the ID number 7625225 was selected from it, and the predicted score value of the six music pieces was compared with the known score value, as shown in Fig. 2:

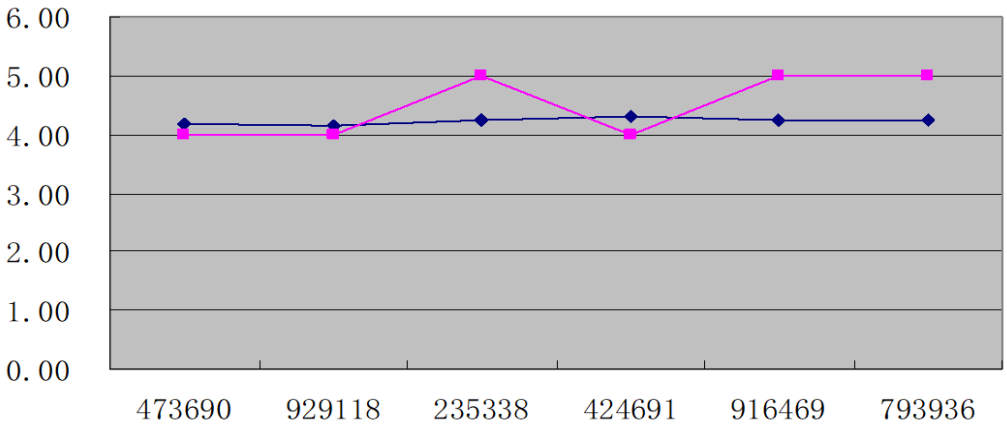

Fig. 2 Comparison of predicted and true values

It can be seen from the above figure that the predicted value in this paper fluctuates around the actual value given by the title, and the absolute error calculated by SPSS[1] is 0.015, which is relatively small. Therefore, the prediction score obtained by this model is relatively accurate.

3. Collaborative filtering recommendation model

Through the collaborative filtering recommendation algorithm [2,5], we can know the three steps that the Item-based method needs to carry out:

(1) Obtain the score data of User-item;

(2) The nearest neighbor search of the target item, that is, the similarity calculation of the item;

(3) Generate recommendations.

First, the scoring data has been obtained from Model 2. Next, this paper calculates the similarity between users by using the nearest neighbor method and Pearson, cosine and the improved cosine similarity algorithm.

(I) Pearson similarity algorithm:

$$sim(i,j) = \frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c\in I_{ij}}(R_{j,c}-\overline{R_j})^2}} \tag{7}$$

(II) Cosine similarity algorithm:

$$sim_{xy} = \frac{\sum_{i\in I_{xy}} r_{xi}\times r_{yi}}{\sqrt{\sum_{i\in I_x} r_{xi}^2}\sqrt{\sum_{i\in I_y} r_{yi}^2}} \tag{8}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

In the scoring matrix, all the scores of each item can be regarded as a column vector of the matrix, and the similarity of two items can be calculated by calculating the cosine value between the two column vectors corresponding to the two items, and the cosine value can be used to represent the similarity of the two items.

The scale of different users' ratings is not considered in the cosine similarity measurement method. Some users tend to give lower ratings and some users tend to give higher ratings. The modified cosine similarity measurement method can improve the above defects by subtracted the average ratings of users. Represents the similarity between users. Matlab is used to build and improve the similarity matrix of cosine value of user project evaluation.

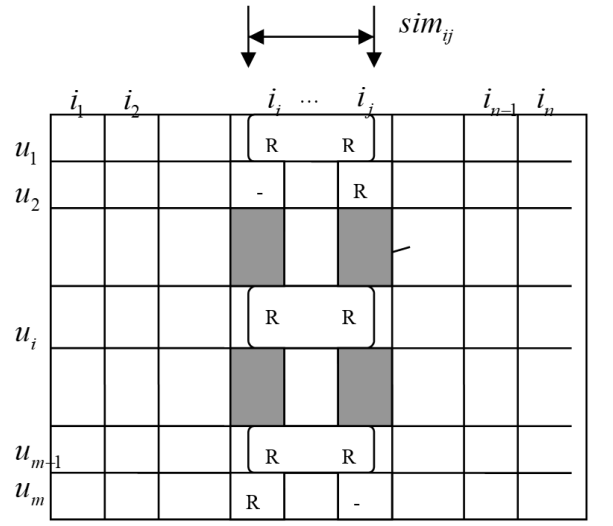

Fig. 3 Schematic diagram of similarity calculation of project-based collaborative filtering algorithm

According to the calculation method of similarity and find users - item's neighbors, the neighbors based on similarity threshold selection principle, using matlab for data selection, calculation [6], will be subject to music tracks, readers, readers historical data and the social relations between readers data segmentation, and data mining, to extract the user factor matrix and items factor matrix, the implicit information extracted from desultorily huge amounts of data, the relationship between the mining data, sorting out the music track ID and music tracks tag number corresponding relation, user -music the music score matrix, The social network relationship between users, and the similarity matrix between users, and the classification of music tracks, which provides technical support for subsequent algorithm design and data post-processing. The schematic diagram of the point set in the two-dimensional plane space as shown in Fig. 4 is obtained.

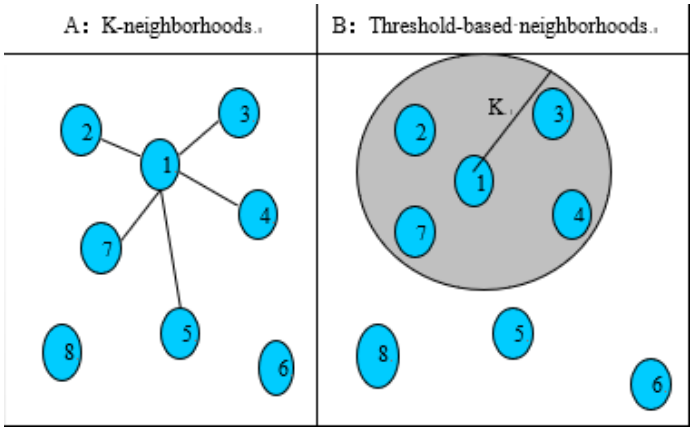

Fig. 4 Schematic diagram of similar neighbor calculation

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

The neighbor calculation based on the similarity threshold is to limit the maximum distance of the neighbors. All the points in the region with the distance of K centered on the current point are the neighbors of the current point. The number of neighbors calculated by this method is uncertain, but the similarity will not appear a large error. As shown in Fig.6, B starts from point 1 and calculates the neighbors whose similarity is within K to obtain points 2, 3, 4 and 7. The degree of similarity obtained by this method is better than that obtained by a fixed number of neighbors, especially for the treatment of outlier points.

The nearest neighbor of the target user is obtained by the proposed similarity measurement method, and the next step is to generate the corresponding recommendation. The calculation method is as follows:

$$P_{u,i} = \overline{R_u} + \frac{\sum_{n \in NBS_u} sim(u,n) \times (R_{n,i} - \overline{R_n})}{\sum_{n \in NBS_u} (|sim(u,n)|)} \tag{9}$$

Where $sim(u,n)$ represents the similarity between user u and user n; $R_{n,i}$ represents the user's rating of the project; $\overline{R_n}$ respectively represents the average rating of user u and user n to the project.

4. Conclusion

The measurement standards for evaluating the recommendation quality of recommendation systems mainly include statistical accuracy measurement methods and decision support accuracy measurement methods. The average absolute deviation in the statistical accuracy measurement method is easy to understand and can intuitively measure the quality of the recommendation. It is the most commonly used recommendation quality measurement method. This article uses the average absolute deviation as the metric. The average absolute deviation is calculated by calculating the predicted user score The deviation from the actual user rating measures the accuracy of the prediction. The smaller the deviation, the higher the recommendation quality.

The mean absolute deviation is defined as:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{10}$$

Calculate the recommended accuracy value and analyze the experimental error. First find the category attribute similarity of all new items and other items, get the nearest neighbor of the new item through the category similarity, calculate the value through the nearest neighbor prediction, and draw a line graph.

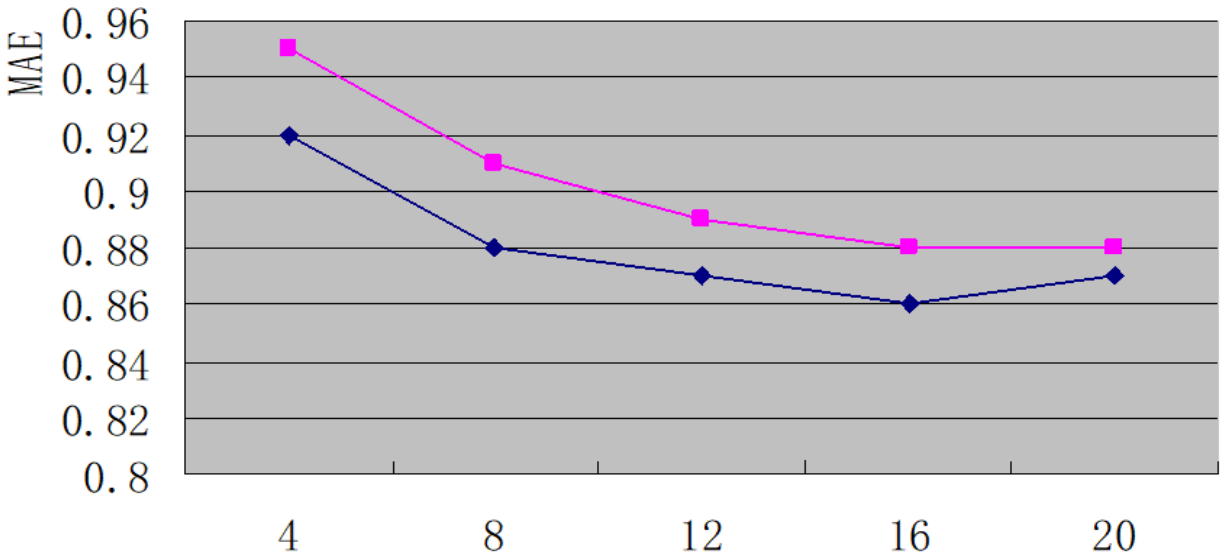

Fig. 5 Comparison of the modified cosine and Pearson's prediction formula with the line chart (the modified curve is shown in pink).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0006

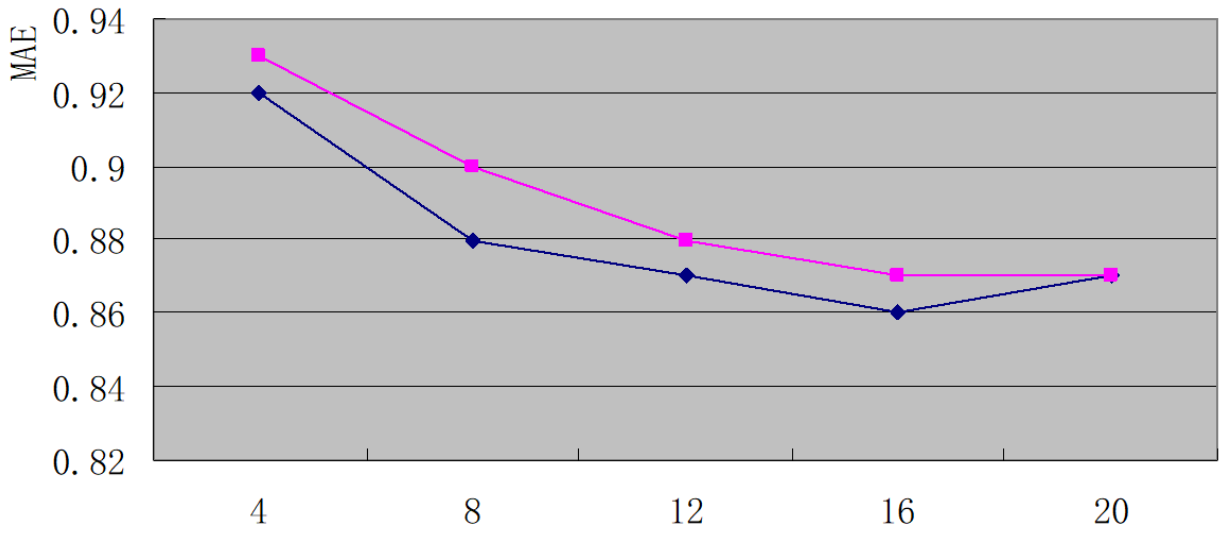

Fig. 6 Comparison of cosine prediction formulas before and after improvement (pink is improved).

Analyzing Figures 5 and 6, we can see that according to the three similarity calculation formulas: cosine, adjusted cosine and Pearson's formula, the improved prediction formula calculated is lower than that obtained by Pearson's and the improved prediction formula, which proves The improved prediction formula is effective in improving the recommendation accuracy of the system, which proves that the improved prediction formula is better than the Pearson prediction formula before the improvement, which improves the system to a certain extent when the scoring data is sparse Recommended accuracy.

Acknowledgements

I would like to thank Mr. Zhu Qigang from Shandong University of Science and Technology for his guidance in innovation.

References

[1] Gao Xiangbao, Dong Hanqing. Data Analysis and SPSS Application. Beijing: Tsinghua University Press, 2007.

[2] Deng Ailin, Zhu Yangyong, Shi Bo. Collaborative Filtering Recommendation Algorithm Based on Project Score Prediction. Journal of Software, 14(09) 1621:1624-1626, 2003.

[3] Information on: http://blog.csdn.net/lifeitengup/article/details/9174419#comments

[4] Ji Yun. Construction of Movie Website Based on Collaborative Filtering Recommendation Algorithm. Harbin University of Technology, 2009.

[5] Yao Zhong, Wei Jia, Wu Yue. Collaborative Filtering Recommendation Algorithm Based on High-dimensional Sparse Data Clustering.Journal of Information Systems, Vol. 2 (3): 78-96, 2008.

[6] DONG Lin. Practical Explanation of MATLAB --Fundamentals, Development and Engineering Applications. Beijing: Publishing House of Electronics Industry, 2009.