Research on Intelligent Recognition Method of Music Similar Segments Based on Deep Reinforcement Learning

Ni Lu1,*

1Shandong Institute of Commerce and Technology, Shandong, China

*Corresponding author e-mail: luli@sict.edu.cn

Abstract. The identification of similar music segments is of great significance for the study of online music search, content relevance, emotional expression and many other aspects. In the overall structure of music, the extraction of key frames, the identification of similar key frames for different types of music, which is to obtain better music emotion data. This paper uses in-depth reinforcement learning algorithms to analyze the music data in detail to construct music similarity. Intelligently identify the database and match the obtained music files with the music data in the database to find similar segments. Case analysis shows that this method can effectively analyze music fragments and provide a basis for subsequent music control.

Keywords: In-depth Reinforcement Learning, Similar Segments, Midi Music Files, Pattern Recognition

1. Introduction

In recent years, with the rapid development and wide application of network technology, music hardware processing technology and software processing technology, network music data has shown an explosive growth, which contains a large number of near-repetitive music clips, and quickly detects these similar music. Fragments are of great significance for online music retrieval [1-3], content association analysis, and automatic music labeling. At present, the detection of music similarity at home and abroad mainly focuses on two aspects: music segment retrieval and near-duplication music detection. Music segment retrieval refers to finding a music collection containing a certain music segment from a large number of music collections. Generally, the length of the music segment to be retrieved is small. The commonly used retrieval method is to select a moving window that is slightly larger than the music segment for a longer time. Move on the music sequence, record the similarity between the music segment in the window and the music segment to be retrieved, and select the segment with higher similarity as the final retrieval result [4-6].

In response to the above problems, according to the music data obtained by the music emotion feature extraction system, in-depth reinforcement learning is used to classify the music data, establish a standard music emotion performance model library, and use the new music file data through the method and model of closeness in fuzzy pattern recognition. The library matches to provide a basis for subsequent music control.
2. In-depth reinforcement learning algorithms

In-depth reinforcement learning is a clustering method that blurs the definition of classic division and uses the degree of membership to determine the degree of clustering. Among them, two important parameters are the number of clusters $c$; the other is the fuzzy weighted index $m$. The algorithm divides $n$ vectors $x_i \in R^d$ into $c$ groups ($k=1,2,\ldots,n$); $s$ is the vector dimension of, and find the cluster center of each group. The basic steps of in-depth reinforcement learning are as follows:

① Initialize the membership matrix $U = \{u_{ik}\}_{c \times n}$ so that it meets the constraint conditions of formula (1):

$$\sum_{i=1}^{c} u_{ik} = 1, \quad u_{ik} \in (0,1)$$  \hspace{1cm} (1)

② Calculate the cluster center $V$ according to formula (2):

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_i}{\sum_{k=1}^{n} u_{ik}^m}, \quad i = 1,2,\cdots,c$$  \hspace{1cm} (2)

③ Calculate the objective function according to formula (3):

$$d_{ik} = \sqrt{\sum_{q=1}^{s} (x_{iq} - v_{iq})^2}$$  \hspace{1cm} (3)

Is the Euclidean distance between the $k$-th fuzzy group and the $i$-th cluster center; $m$ is the fuzzy weighted index, $m \in (1, +\infty)$, $2 \leq c \leq n$, if the variable is less than the previous result For a certain threshold $\epsilon$, the algorithm stops and outputs the final membership matrix $U$ and cluster center $V$. Otherwise, the new membership matrix can be calculated according to formula (4) and return to step ②.

$$J(U, v_1, \cdots, v_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m d_{ik}^2$$  \hspace{1cm} (4)

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^m}, \quad i = 1,2,\cdots,c; \quad k = 1,2,\cdots,n$$  \hspace{1cm} (5)

From the final membership matrix $U = \{u_{ik}\}_{c \times n}$, calculated:

$$i = \arg \max_{i \in [c]} u_{ik}, k = 1,2,\cdots,n$$  \hspace{1cm} (6)

It can be seen that after learning, the samples $x_k$ are clustered into $i$ type, $i=1,2,\cdots,c$.

In order to realize the intelligent control of music, data processing of music elements such as pitch, timbre and speed in MIDI music files is carried out, and a music emotion feature extraction system is established to extract the key information of music emotion, obtain effective music emotion data, and then adopt in-depth enhancement Learn to classify music data, establish a standard music model library, and match new music files with the model library through the approach of fuzzy pattern recognition.
3. Application examples
According to the literature, the music-based characteristics that can embody musical emotions are pitch, length, timbre, speed and strength. This section mainly establishes a music emotion feature extraction system based on the analysis of musical elements such as pitch and length in MIDI music files.

(1) Pitch
Assuming that there are m, (m \(\geq\) 1) tracks in the current music, the pitch characteristics of each track are respectively extracted, and the pitch feature vector \(x_i\) of the music is determined according to the relevant attributes of the main track.

Assuming that the music emotion type of the intention detection is n, n \(\geq\) 1, and the characteristic value of the emotion type under the current pitch value is \(p_i(x_j), i \in [1, 2, \cdots, n]\), the part of each emotion type related to the music pitch attribute index can be expressed by the following formula:

\[
f(x_j) = \text{Max}(p_i)
\]

(7)

In the formula, \(x_i\) is the pitch feature vector of the current music.

Although pitch features are closely related to music emotion, the specific relationship between the two needs to be determined according to the selected music emotion recognition method based on basic features.

(2) Sound length
The pitch indicates the duration of the sound of the note. In a MIDI file, the pitch is mainly determined by the delta-time between the two MIDI events of note on and note off. In order to measure the sound length characteristics of the music, the function \(g\) is defined as the evaluation standard, the sound length switch threshold is defined as \(x_{\text{switch}}\), and the current note switch interval time is \(x_g\), the value of the sound length evaluation function \(g\) is as follows:

\[
g(x_g) = \begin{cases} 
\text{long}, & x_g \geq x_{\text{switch}} \\
\text{short}, & x_g < x_{\text{switch}}
\end{cases}
\]

(8)

(3) Tone
According to common sense, different instrument types have different timbre characteristics. In the MIDI standard, musical instruments with various timbre characteristics are uniformly labeled, and 128 timbres are defined in total. In this way, timbre characteristics can be obtained more conveniently. According to the corresponding relationship between timbre and emotion, 128 timbres are divided into k categories.

Different tracks of MIDI files and the same track can change the timbre at any time at different times, so the timbre characteristics need to be dynamically extracted in real time during the timbre extraction process.

(4) Speed
Tempo refers to the speed of the music beat. Even if it is the same song, if you use different beat speeds, it will give people different emotional experiences. Fast beats are emotionally happy; slow beats are depressing and dull. The beat characteristic of each song is certain. Therefore, when extracting the speed characteristic of the MIDI file, it is only necessary to read the corresponding Meta event information according to the content of the file. The extraction of the tempo feature is relatively simple, because each piece of music has a fixed number of beats, for example, 1/4 means that a quarter note is a beat, and each measure has a beat. Therefore, the speed feature can be extracted directly from the MIDI file header information and Meta events.

(5) Strength
The information that reflects the note velocity in the MIDI file is mainly contained in the data bytes in the note-off and note-on events. Strength is reflected in the numerical range of 0~127.

Obtain 40 sets of original music data, and take the first 25 sets as an example for classification
experiment, see Table 1.

| Serial number | Pitch | Pitch length | Tone | Speed | Strength |
|---------------|-------|--------------|------|-------|----------|
| 1             | 9.011 | 3.231        | 4    | 10.128| 6.237    |
| 2             | 4.239 | 4.129        | 6    | 6.091 | 5.237    |
| 3             | 3.227 | 6.113        | 4    | 7.837 | 8.999    |
| 4             | 5.871 | 8.077        | 3    | 6.129 | 4.012    |
| 5             | 2.128 | 4.126        | 7    | 9.359 | 8.129    |
| 6             | 1.821 | 3.123        | 4    | 2.120 | 1.094    |
| 7             | 6.120 | 7.228        | 2    | 9.003 | 2.239    |
| ...           | ...   | ...          | ...  | ...   | ...      |
| 25            | 2.355 | 9.298        | 8    | 1.112 | 6.390    |
| 26            | 8.476 | 3.129        | 4    | 2.659 | 7.120    |
| 27            | 3.927 | 4.773        | 4    | 5.490 | 4.992    |
| 28            | 7.012 | 8.298        | 7    | 6.121 | 4.654    |
| ...           | ...   | ...          | ...  | ...   | ...      |
| 40            | 3.226 | 7.093        | 6    | 2.578 | 1.983    |

There are formulas (7)–(8) to get 40 sets of music data. Take 25 groups of data, namely 25 5-dimensional vectors, as the input matrix of FCM clustering, and divide the music into \( c = 8 \) groups. Suppose the fuzzy weighting index \( m = 2 \), the maximum number of iterations is 100 steps, and the threshold \( \varepsilon = 0.00001 \), the classification result can be obtained from the formula 2.1, and the in-depth reinforcement learning result is shown in Figure 1.

![Figure 1](image_url)

**Figure 1.** Results of in-depth reinforcement learning.

Put the music characteristic curve \( G \) to be recognized into the 8 classification results that the fuzzy C mean algorithm has divided into, as shown in Figure 2.
It can be seen from Figure 2 that the shape of the musical characteristic curve of "Spring Festival Overture" is relatively close to the first category, and indeed it belongs to the first category. According to the various closeness values above, it can be seen that the lattice closeness δ1(A,B)=0.6382, the maximum and minimum closeness δ5(A,B)=0.7349, it is more appropriate to use these two closeness degrees, see Table 2.

Table 2. Comparison of accuracy of different recognition methods.

| Recognition methods                  | Sample size | Correct amount | Accuracy  |
|--------------------------------------|-------------|----------------|-----------|
| In-depth reinforcement learning      | 15          | 13             | 86.67     |
| BP neural network                    | 15          | 11             | 73.33     |

The experimental results show that for music with close musical emotions or with multiple musical emotions, the accuracy of the BP neural network recognition method is significantly lower, and the use of in-depth reinforcement learning methods still has a good recognition effect, highlighting the advantages of this method. The recognition rate of the method in this paper reached 86.67%, which basically met the requirements of emotion recognition and achieved the expected effect.

4. Conclusion

Aiming at many factors such as the length, pitch and timbre of different types of music, this paper constructs a database of music emotions and a database of music performances for feature matching. Through the use of in-depth enhancement algorithms, the music membership matrix and clustering center are calculated, the membership matrix is used to classify the obtained music data to solve the problems of slow manual classification, strong subjective awareness and low work efficiency. The experimental results show that the intelligent recognition method for music similar segments proposed in this paper can compare different music similarities, it can provide a basis for the intelligent control of music.

References

[1] Chen, H., He, Z., Shi, B., & Zhong, T. (2019). Research on recognition method of electrical components based on yolo v3. IEEE Access, 7, 157818-157829.
[2] Jiang, M., Xu, L., Tao, C., & Lu, X. (2014). Research on interaction and evaluation method of learning intelligent system based on knowledge memory. LECTURE NOTES IN ELECTRICAL ENGINEERING, 271(4), 173-181.
[3] Xing, L., Zhang, J., Liang, Hui’e, & Li, Z. (2018). Intelligent recognition of dominant colors for chinese traditional costumes based on a mean shift clustering method. Journal of the Textile...
[4] Lei, X., Zhang, Z., & Dong, P. (2018). Dynamic path planning of unknown environment based on deep reinforcement learning. Journal of Robotics, 2018, 1-10.

[5] Wang, H., Yang, Z., Zhou, W., & Dalin, L. I. (2019). Online scheduling of image satellites based on neural networks and deep reinforcement learning. Chinese Journal of Aeronautics, 32(04), 245-253.

[6] Wu, Y., Wei, J., Pan, J., & Chen, P. (2019). Research on microseismic source locations based on deep reinforcement learning. IEEE Access, 39962-39973.