Research Article

Music Emotion Recognition Model Using Gated Recurrent Unit Networks and Multi-Feature Extraction

Nana Niu
Henan Finance University, Zhengzhou 450046, China

Correspondence should be addressed to Nana Niu; niunana@hafu.edu.cn

Received 18 April 2022; Revised 18 May 2022; Accepted 21 May 2022; Published 4 July 2022

Abstract

A large number of music platforms have appeared on the Internet recently. The deep learning framework for music recommendation is still very limited when it comes to accurately identifying the emotional type of music and recommending it to users. Languages, musical styles, thematic scenes, and the ages to which they belong are all common classifications. And this is far from sufficient, posing difficulties in music classification and identification. As a result, this paper uses the methods of music emotion multi-feature extraction, BiGRU model design, and music theme scene classification model design to improve the accuracy of music emotion recognition. It develops the BiGRU emotion recognition model and compares it to other models. BiGRU can correctly identify happy and sad emotion music up to 79 percent and 81.01 percent of the time, respectively. It goes far beyond Rnet-LSTM, and the greater the difference in emotion categories, the easier it is to analyze the feature sequence containing emotional features and the higher the recognition accuracy. This is especially evident in the accuracy with which happiness and sadness are recognized. It can meet users’ needs for music recognition in a variety of settings.

1. Introduction

Gated recurrent unit (GRU), as a type of recurrent neural network, is a variant of the long- and short-term memory network, which can combine the unit state and the hidden state into one state. It is also characterized by adaptively controlling dependencies on different time scales, selectively forgetting and saving past information. Recurrent neural networks, on the other hand, are usually used to process sequence data. The original GRU framework itself can not only guarantee the learning effect in terms of experimental effect, but also further simplify the framework structure. It requires less computation, which greatly improves the computing efficiency and makes it easier to train. In the network, the output state of each hidden layer is not only determined by the input at the current moment, but also related to the output state of the hidden layer at the previous moment. This memory feature can well solve problems such as sequence prediction and classification where contexts are closely related.

Music is classified at multiple levels on today’s mainstream music websites, depending on the development of the times and the interests and preferences of users. Features that reflect the characteristics of different genres, such as rhythmic features and emotion-related features, are often present in music signals for a few seconds or even longer time periods. Traditional short-term features based on millisecond-level speech frame extraction struggle to characterize the semantic information in music signals, affecting classification performance. The recurrent neural network algorithm based on the gated recurrent unit can better classify music genres and capture multi-dimensional signal features from the essential signal of music. This paper is unique in that it uses the memory neural network’s strong correlation feature to analyze the emotional expression of music in time series data. It constructs a more lightweight BiGRU model with a self-attention mechanism, reducing structural complexity and increasing recognition computational power.

2. Related Work

Music emotion recognition can organize and retrieve music effectively and efficiently, which is a challenging subject in
the field of music information retrieval. The arrival of the era of deep learning provides strong support for this identification. Hizlisoy proposed a music emotion recognition method based on the convolutional long short-term memory deep neural network (CLDNN) architecture. He found that using the proposed system with 10-fold cross-validation using features obtained by supplying the cepstral coefficients on the Mel frequency can achieve an overall accuracy of 99.19% [1]. Although his method is highly accurate, it is too complex to be widely used in practice. Nguyen proposed a music emotion recognition method combining dimension and classification. His proposed method detects 36 musical emotions and achieves the highest accuracy when using the random forest classifier. These ratios are superior to previous studies [2]. But the scope of his method is too small. It exposes flaws when it comes to testing a large number of different types of music, so it still needs to be improved. Dong proposed a novel bidirectional convolutional recurrent sparse network for music emotion recognition. His model adaptively learned emotionally salient features containing sequence information from two-dimensional time-frequency representations (i.e., spectrograms) of musical audio signals. This enables continuous emotion prediction of audio files [3]. His method is efficient and fast, but has a high error rate in the face of many different types of music, so it still needs improvement. Zhang explored a new method for personal music emotion recognition based on human physiological characteristics. He also established a feature database based on music-related emotions, including EDA, PPG, SKT, RSP, and PD variation information [4]. His research provided many valuable references for other researchers, but whether it is effective still needs to be tested for a long time. Van found a way to transform the content of the original music into music with an emotion similar to the altered emotion. To this end, he proposed an algorithm to change the emotion of playing music on a two-dimensional plane [5]. His research can not only identify musical emotion, but also provide reference for other musical emotion establishments. But his research accuracy is not particularly high, so more research support is needed. Han started with some preliminary knowledge of music emotion recognition and then proposed a three-part research framework. On the basis of this three-part research framework, he analyzed the knowledge and algorithms involved in each part in detail. It includes some commonly used datasets, emotion models, feature extraction, and emotion recognition algorithms [6]. His research is especially helpful for beginners, but his methods are too simplistic for complex situations. Chin proposed a new system for identifying emotional content in music. The system integrates three computational intelligence tools, namely, hyper-rectangular composite neural network, fuzzy system, and PSO [7]. His research has provided new ideas for music emotion recognition, but because the data used in the research are too small, it seems that it is easy to be replaced by new algorithms in the long run.

From the above research status, it can be seen that the research direction has become more and more inclined to use the deep learning framework [8] for classification [9, 10] and recognition [11, 12] in audio analysis. When analyzing the emotional characteristics of music, if the relevant model architecture of some neural networks can be properly used, and the neural network model and the attention mechanism can be properly improved and integrated, the emotional analysis can be carried out based on the source signal of music. It selects an appropriate amount of music for feature extraction according to the playing frequency and playing time. This can not only accurately analyze the user’s current emotional state, but also ensure the diversity and timeliness of music recommendations. The above content provides a new way and method for the research direction and content of this paper.

3. Model Building Method for Music Emotion Recognition Based on GRU Network

This paper first introduces the basic principles of recurrent neural networks before introducing how to use gated recurrent units to design the model (RNNs). The input layer, hidden layer, and output layer are the three layers of a neural network. It is equivalent to a given input, as it uses a built-in function to process the data and produce a desired output result. There will be multiple or even enormous amounts of input data in general. To process these data, a proper neural network framework is required. When the input data are arranged in sequence, RNN can be used to process the information. In this way, it is possible to better predict, analyze, and filter the previous and subsequent items of the data in the series. For a most basic RNN structure, it is shown in Figure 1.

3.1. Types of Recurrent Neural Network Memory Units

Recurrent neural networks can use past information in time series to make judgments or predictions on current inputs. But this feature also brings some intractable problems. One of the main problems is that as the time step grows, the gradient disappears or the gradient explodes easily in the process of error backpropagation. It cannot solve the problem of long-term dependence. To this end, scholars expand the nodes in the hidden layer of RNN into various memory unit structures to solve the problems in traditional recurrent neural networks.

3.1.1. Long Short-Term Memory (LSTM) Network

LSTM was first proposed in the 1990s. Its memory cell incorporates cellular states to preserve long-term memory. It also dynamically controls information retention and forgetting in the cell state through the gate structure, thus solving the long-term dependency problem. The specific internal structure of the long short-term memory network is shown in Figure 2.

From the perspective of the overall architecture, the overall composition of LSTM and RNN is similar, but the repeated parts are improved. In RNN, this repetitive part is just the most basic neural network layer, while in LSTM it is the structure of the neural network layer with multiple states.
For each state, a gate structure is used to describe and operate, including forget gate, input gate, and output gate.

3.1.2. Gated Recurrent Unit (GRU). GRU is also a kind of RNN, which can be regarded as a variant of LSTM. It should be noted that the original GRU framework itself cannot only guarantee the learning effect in terms of experimental effect, but also further simplify the framework structure. It requires less computation, which greatly improves the computational efficiency and makes it easier to train [13].

The internal specific structure block diagram of this model is shown in Figure 3.

For the important parameters in the learning process of GRU, the operation formulas of the training process are shown in

\[ W_r = W_{rx} + W_{rh}. \]  
\[ W_z = W_{zx} + W_{zh}. \]  
\[ W_h = W_{hx} + W_{hh}. \]  

For the output layer \( y_t \), the input and output expressions are shown in

\[ y_t^i = W_i h. \]  
\[ y_t^o = \sigma(y_t^i). \]  

In general, both LSTM and GRU can use the gate function to retain the required data, so as to ensure that no information is lost. In addition, LSTM has a more complex structure and more parameters than GRU, so the overall learning rate of GRU is higher than that of LSTM [14].

3.2. Multi-Feature Extraction of Music Emotion. In the part of music emotion classification [15], it first extracts a number of spectral features in music to form a sequence that can capture feature information from the essence of music. Afterward, a three-layer GRU combined with attention mechanism (AM) model architecture is used to classify the emotion of music.

The spectral features closely related to the music data can reveal the characteristic information of the music essence to the greatest extent. However, in the subsequent classification of music theme scenes, more feature information can improve the classification accuracy to a greater extent [16]. Therefore, the feature extraction part of this section will add a spectral feature here: the rhythm feature of music. It is also one of the basic elements of music, which is of great significance as a part of feature extraction [17]. The rhythmic feature of music is mainly composed of two parts, which are music beat and tempo. The feature extraction method uses the beat histogram, and its specific extraction method is shown in Figure 4.

Here, the beat histogram of the music happens to be based on the time domain features. It passes the time domain features through the results of several filters, thereby extracting 16-dimensional features such as the quantity, period, amplitude, and ratio of the four peaks [18]. Among them, the basic frequency domain filtering model is shown in formula:

\[ G(u, v) = H(u, v) \ast F(u, v). \]  

Among them, \( F(u, v) \) is the Fourier transform of the image, and \( H(u, v) \) is the filter used. The calculation formula of the low-pass filter is shown in

\[ H(u, v) = 1, D(u, v) \leq D_0. \]  
\[ H(u, v) = 0, D(u, v) \geq D_0. \]  

\( D_0 \) is the Euclidean distance from the pixel to the center of the image. Combining these 16-dimensional features with
The 50-dimensional features extracted in Chapter 3 can form a new 66-dimensional feature sequence. This new 66-dimensional feature sequence contains more spectral feature information than the general feature sequence.

The understanding of AM can be explained from the information transmission mechanism of human vision and brain. For example, when people are looking for an item, the human eye only needs to search according to the external characteristics of the item within the field of view; even if they see other items, they will automatically ignore it. The act of filtering data like this and focusing the brain’s vision on the item to be sought is the attention mechanism mentioned here. Generally speaking, while the neural network model solves some problems and brings some improvements, it is easy to bring other problems that cannot be ignored. When multiple models are fused, it can indeed increase the processing and analysis capabilities of the data. Whether it is the analysis of the correlation degree of the data before and after, or the improvement of learning and performance, it is well reflected, but at the same time, it will also lead to the phenomenon of information overload [19]. AM was proposed to solve this problem. Its internal structure is shown in Figure 5.

It can be seen that AM in the final analysis is a redistribution of the weight of the input data. The input data between layers are processed for the hidden state of all time steps of the encoder [20]. Its operation formula can be shown in formula (6):

\[ O = \text{softmax}(QK^T)V. \]  
(9)

As for the final output result of AM attentive\_x, the operation process is shown in Figure 6.

3.3. Design of Music Emotion Classification Model Integrating BiGRU and Self-Attention Mechanism. This paper uses a variant of the LSTM model to make the classification process faster and lighter. It combines the GRU and BiGRU models to create a BiGRU model. This neural network framework is simpler, less computationally expensive, and more efficient than the GRU model. An LDA probabilistic topic model will be introduced in the newly added topic classification section. It extracts multiple themes from each piece of music based on its spectral information. It uses the distribution probability of subject words to judge the subject scene category to which the piece of music belongs, allowing the music subject scene to be re-classified using the emotional classification of music [21].

In comparison with LSTM, the music emotion classification model combining BiGRU and self-attention proposed in this section can reduce the amount of calculation, shorten the operation time, and improve classification efficiency based on GRU. Self-attention, in contrast to AM, can calculate attention after multiple linear transformations. Finally, it stitches the results together, greatly improving the model’s fitting ability [22].

In this way, it first divides the music emotion and then performs secondary classification according to the theme of the music. This can make the classification forms of music more diverse, and the recommendation effect is more accurate in the subsequent recommendation process. In the music recommendation part, this chapter also needs to collect the top \( k \) songs with the highest number of plays in a week in the user’s historical listening records. However, when dimensionality reduction is performed for audio
features, the LSTM layer passed in the dimensionality reduction process is changed to a GRU layer. AM also changed from the original attention mechanism to self-attention. The advantage of this improvement is that it can further improve the fitting ability of the model, improve the operation efficiency, and reduce the operation time. It truly combines the spectral information of music with the music recommendation model framework of emotion and topic classification to improve the recommendation accuracy.

Because both BiGRU and LSTM belong to the recurrent neural network with time series, it is also necessary to divide the music into several segments. The duration of each segment remains unchanged from Chapter 3, which is still a 3-second segment. 66-dimensional features are extracted from each segment as the input of the next BiGRU model.

BiGRU is composed of two ordinary recurrent neural networks: one is forward, and the other is reverse. This can apply both past information and future information. This just caters to the characteristics of the musical spectrum feature sequence extracted earlier. Because the spectral characteristics of music have strong time series correlation, the BiGRU model used here can have a more comprehensive analysis of the overall characteristics of a piece of music. The model frame diagram that integrates BiGRU and self-attention mechanism is shown in Figure 7.

It is the core goal of this BiGRU model to comprehensively consider the spectral feature information of the first half and the second half of a piece of music. There are only two gates in the BiGRU model structure: the update gate and the reset gate. BiGRU modifies the forget gate and input gate of LSTM to update gate $z_t$. And it can filter the information to obtain the information that needs to be remembered, and the newly remembered data will also be added during the update process. The calculation formulas of the update gate and the reset gate are shown in

$$y_t = \sigma(W_r \cdot [h_{t-1}, x_t]).$$  \hspace{1cm} (10)\\
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]).$$  \hspace{1cm} (11)

Its next step is to compute candidate hidden layer $\tilde{h}$. It holds new data at time $t$. $\gamma_t$ is used to determine the current memory that needs to be left. It can control the input of information at time $t$. If the result of $\gamma_t$ is 0, then it only contains the current information, and the output data $h_{t-1}$ of the previous hidden state will be discarded. The information needs to be remembered before the reset gate decides. The calculation formula of the candidate memory unit is

$$\tilde{h} = \tanh\left(W_r \cdot [y_t \cdot h_{t-1}, x_t]\right).$$  \hspace{1cm} (12)

The next step is to calculate the current moment memory unit $h_t$. $z_t$ can control the information $h_{t-1}$ from the hidden layer at the previous moment, and the hidden layer data $\tilde{h}$ that need to be updated finally get the current moment memory unit $h_t$. It should be noted here that if the update gate is always close to 1, the implicit state before time $t$ will be preserved and passed until time $t$. The formula for calculating $h_t$ is

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}.$$  \hspace{1cm} (13)

Finally, for the output layer $y_t$, the calculation formula is as follows:

$$y_t = \sigma(W_o \cdot h_t).$$  \hspace{1cm} (14)

3.4. Design of Music Theme Scene Classification Model

3.4.1. LDA Topic Classification Model Structure. In music, a theme is a direct association of music in an application scenario. Depending on the scene, the spectral information of the music will also contain a variety of themes. It uses spectral features extracted from the music to discover some implicit thematic information [23]. Even for the same word, it will have different probability values in different subject contexts. And LDA (latent Dirichlet allocation, text topic model) can just find the topic information implicit in the spectral features, such as audio words and topic words. The music spectrum is normalized, and the audio words contained in the speech features are obtained by clustering. In order to calculate the probability distribution of each audio word in each music, this paper adopts the LDA music theme scene classification model framework for calculation and analysis. Its frame diagram is shown in Figure 8.

For this framework, the main goal of this paper is to find out the distribution of topics for each piece of music and the distribution of audio words in each topic. Each parameter in this frame diagram will be explained below.

3.4.2. Joint Distribution Derivation of Musical Topics and Words. In the LDA model, it sets the number of musical
topics to K. And the prior distribution is Dirichlet distribution. Therefore, all distributions will be spread out over K topics. For the spectrum document \( d \) of any piece of music, the calculation formula of its topic distribution is

\[
\theta_d \sim \text{Dirichlet}(\vec{a}), \quad (15)
\]

where \( a \) is the hyperparameter of the distribution and is a K-dimensional vector. At the same time, for any topic \( k \), the calculation formula of its word distribution is as follows:

\[
\beta_k \sim \text{Dirichlet}(\vec{\eta}). \quad (16)
\]

\( \eta \) is the hyperparameter of the distribution, which is a \( V \)-dimensional vector, where \( V \) represents the number of all audio words. Then, for the \( n \)th word in any spectrum document \( d \), the distribution of its topic numbers and the probability distribution of the audio words corresponding to each topic number can be obtained from the topic distribution. The calculation formulas of the two are as follows:

\[
Z_{d,n} \sim \text{multi}(\theta_d). \quad (17)
\]

\[
W_{d,n} \sim \text{multi}(\beta_{z_{d,n}}). \quad (18)
\]

It next calculates the conditional distribution of musical topics in the \( d \)th spectral document. The calculation process is as follows:

![Table 1: Parameter settings of recurrent neural network.](image)

![Figure 9: Experimental results and comparison of three memory units.](image)
The number of the $k$th topic in the $d$th spectrum document is denoted as $n_d^{(k)}$. Its multinomial distribution is

$$n_d = (n_d^{(1)}, n_d^{(2)}, \ldots, n_d^{(k)}).$$

(20)

According to the subject conditional distribution of a spectrum document, the subject conditional distribution probability of all spectrum documents and the conditional distribution probability of audio words corresponding to all music subjects can be calculated:

$$p(\bar{z}_d|\bar{a}) = \prod_{d=1}^{M} p(\bar{z}_d|\bar{a}).$$

(21)

### 4. Comparative Experiment and Analysis of BiGRU, an Emotion Recognition Model Based on GRU Network

In this section, three sets of comparative experiments are set up to explore the types of memory units used, the way of pooling, and the effect of classification. It then compared the efficiency of music emotion recognition. The experiments are carried out on the GTZAN and ISMIR2004 datasets. For the GTZAN dataset, it divides it into 10 equal parts in the experiment, ensuring that each equal part contains the same number of samples of different genres. The ratios are 8:1:1 as training set, validation set, and test set. For each split, tenfold cross-validation was performed for a total of 5 replicate splits. For the ISMIR2004 dataset, since it has been divided into training set and test set, there is no need to manually divide the test set. It only needs to be divided into 10 equal parts by level in the training set to ensure that the proportion
of genres contained in each equal part is the same as before the division. It uses 10% as the validation set each time and the remaining 90% as the training set. It was repeated 10 times for a total of 5 divided replicates.

4.1. Comparison between Memory Cell Types. As for the relationship between the number of cyclic network layers and performance, for convenience, Table 1 lists the structures of cyclic neural networks with different structures and their corresponding network parameter settings.

The next step is to replace the cells of the recurrent network layer in Rnet3 with three kinds of memory units: ordinary memory unit, LSTM, and GRU, and compare and analyze the recognition effect. The experimental results are shown in Figure 9.

From Figure 9, it can be found that when the network structure is the same, the recognition accuracy of LSTM and GRU is 1%~2% higher than that of ordinary RNN memory units, and the performance of GRU is about 1%~2% higher than that of LSTM. After using GRU, the performance of Rnet4 is improved slightly in GTZAN dataset, slightly better than Rnet2. This shows that the gating mechanism of GRU makes it easier for deeper networks to be optimized. In ISMIR2004, the classification performance using GRU exceeds the original performance by about 3% and also has a relatively large improvement.

4.2. Comparison between Different Pooling Methods. The output state after passing through the recurrent network layer already contains contextual information. Since the network is used for whole sequence classification, it is necessary to pool and fuse the outputs at each moment into an overall representation of the sequence [24]. In this section, four pooling methods are selected: average pooling, maximum pooling, the combination of average and maximum pooling, and the output state at the last moment.

Figure 10 lists the classification accuracies corresponding to these four pooling methods. As can be seen from Figure 10, the classification performance of directly using the output state at the last moment is the best. Whether it is Rnet3-GRU or BiGRU, its output at the last moment has a performance of more than 82%, and the effect in the ISMIR2004 dataset exceeds the performance of GTZAN by about 2%. This shows that the use of max pooling and mean pooling will reduce the network classification

Table 2: Comparison of emotional discrimination effects of music.

<table>
<thead>
<tr>
<th>Experiment model</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>66-dimensional + features</td>
<td>67.2</td>
</tr>
<tr>
<td>66-dimensional + BiGRU</td>
<td>69.9</td>
</tr>
<tr>
<td>66-dimensional + BiGRU_SA</td>
<td>74.5</td>
</tr>
<tr>
<td>50-dimensional + T_LSTM + AM</td>
<td>71.6</td>
</tr>
</tbody>
</table>

Figure 11: Classification performance of output states at different times.
performance to varying degrees. The combination of the two pooling methods can improve the average accuracy slightly, but the effect is still not as good as directly selecting the output at the last moment.

Next, it replaces the pooling layer of the trained BiGRU network with the output state at each moment and then inputs it into the fully connected network. The obtained probability vector is represented by a histogram in chronological order, and Figure 11 is obtained. In Figure 11, each sub-picture is arranged in an increasing order of time from left to right and top to bottom.

As can be seen from Figure 11, from the output state earlier in time, the network cannot distinguish its specific musical emotion. As time increases, the probability vector begins to focus on certain emotions. At the last few moments, the probability distribution is concentrated in the sixth dimension. It finally successfully determines the emotional type of the input music segment. This can explain, to a certain extent, the unsatisfactory effect of max pooling or mean pooling. In the output state at the previous moment, the captured context information is insufficient, does not have globality, and cannot reflect the information of the entire piece of music. When it uses max pooling or average pooling, the previous wrong representation will be incorporated into the pooled output, thereby reducing the classification performance of the network.

When using a bidirectional recurrent neural network, the classification results of various pooling methods of BiGRU are 81.52%, 80.41%, 81.50%, and 81.85%, respectively. The bidirectional recurrent neural network can simultaneously obtain contextual information at each moment. It can reflect the global information for the output state at each moment, so the performance of maximum and average pooling has been improved.

4.3. Comparison of Music Emotion Recognition Efficiency.

In the experimental part of this section, four emotional styles of music are collected on the emotional category, namely, excitement, freshness, relaxation, and sadness. The music has one or more emotional tags and is 7 minutes or less in length.

In the user’s playback record, it sets the $k$ value to 10. It selects the top 10 most played songs in the last 7 days and performs feature extraction, respectively. When segmenting the spectrum information of music, the duration of each segment is 3 seconds, and each spectrum segment contains 66-dimensional feature information. After using these time-ordered feature sequence segments as the input of the next BiGRU model and self-attention mechanism, it also needs to normalize the obtained sequence results. It maps these data into $(0, 1)$ so that it sums to 1. It judges the emotional type of the music according to its probability value under the four emotional categories.

When discriminating emotion categories according to the user’s historical playback records, this paper conducts experiments on the following model framework. It passes 66-dimensional spectral features through BiGRU and self-attention mechanism, and has the highest accuracy in emotion discrimination. It passes 50-dimensional spectral features through Rnet-LSTM and AM, and also compares with other common framework models. It then includes the support vector machine (SVM) method and the LDA method. Under different recommendation models, the comparison of the effects of each experimental method is shown in Table 2.

As can be seen from Table 2, after adding the feature sequence of music rhythm, compared with the Rnet-LSTM model, the BiGRU model fully considers the temporal correlation of the feature sequence. It strengthens the context of feature sequences in temporal order. Later, self-attention strengthened the computing power of the model. The accuracy rate of emotion discrimination for music is higher, and it is better than the traditional neural network model. It has improved accuracy compared to using only SVM model or only LDA for emotion discrimination of music. At the same time, because it incorporates 66-dimensional features, the BiGRU model and the neural...
network framework of self-attention have higher accuracy and better effect on music emotion discrimination. When classifying emotion for music, it is not only necessary to compare the accuracy of emotion discrimination, but also the diversity of emotion categories is a problem that needs to be considered. The music emotion category is divided into four types: happy, refreshing, relaxing, and sad. The accuracy of emotion recognition is shown in Figure 12.

From Figure 12, it can be found that the BiGRU model’s recognition accuracy for emotion has increased. This is particularly evident in the recognition accuracy of happiness and sadness. The Rnet-LSTM model recognizes music with happy and sad emotions up to 71.34% and 75.45%, which are higher than other types of recognition. And BiGRU achieves 79% and 81.01%, which are far more than Rnet-LSTM. And the greater the difference between emotion categories, the easier it is to analyze the feature sequence containing emotional features, and the higher the recognition accuracy. Even for the two emotions of freshness and relaxation with high similarity, the emotion recognition accuracy after incorporating self-attention is improved. This shows that BiGRU can easily identify musical emotions in everyday situations, and it can perform better in complex situations.

It is followed by accuracy detection when classifying the user’s emotion. Here, the comparison of various collection methods will still be performed based on the user’s historical playback records. However, compared with the experimental content in Chapter 3, one more collection method is added here. It takes the audio features of the top 1, 3, 5, 7, and 10 music with the highest number of plays in the historical playback records in the past seven days as the input of the sequence in the subsequent model. It is then combined with BiGRU and self-attention to compare the accuracy of each acquisition method, and the experimental results obtained are shown in Table 3.

From the comparison in Table 3, it can be found that compared with Rnet-LSTM, the BiGRU model has higher accuracy in discriminating user emotions even if the same collection method is used. This further illustrates the high accuracy of the newly proposed model in emotion discrimination. Of course, the input sequence here also follows the weighted average method. The input sequence in each collection method is the weighted average of the feature sequences of the currently collected music. At the same time, it can be found that when the feature sequence obtained by the weighted average of the audio features of the first 10 pieces of music is used, the accuracy of this collection method is the highest compared to other collection methods. From this, it can be explained that when the music recommendation is performed based on the user’s historical playing records, the more the amount of music used, the richer the types of music in the music recommendation. It is closer to the user’s preference, and the higher the real-time performance, the better able to grasp the user’s emotional changes, which makes the music recommendation more accurate.

5. Conclusions

This paper explains the research needs of music classification and recommendation methods from the current rhythm of people’s lives and analyzes the current situation to draw out the problems existing in other music recommendation methods. Therefore, it proposes a multi-feature extraction music emotion recognition model based on GRU network. This model learns features from both the forward and reverse directions of the time series. Its classification efficiency has been improved, but there are still some work contents worthy of further improvement and in-depth exploration, such as the user’s emotional state and life scene problem. At present, people’s living conditions and scenes are more diverse. As for how to ensure the accuracy of music emotion and theme classification when adding emotion and scene classification items, the author looks forward to using more data and experiments in the future to further explore more accurate recognition efficiency.

Data Availability

The data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author does not have any possible conflicts of interest.

References


---

**Table 3: User emotion discrimination accuracy.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Average accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st song + BiGRU + self-attention</td>
<td>61.8</td>
</tr>
<tr>
<td>Top 3 songs + BiGRU + self-attention</td>
<td>65.7</td>
</tr>
<tr>
<td>Top 5 songs + BiGRU + self-attention</td>
<td>67.2</td>
</tr>
<tr>
<td>Top 7 songs + BiGRU + self-attention</td>
<td>71.6</td>
</tr>
<tr>
<td>Top 10 songs + BiGRU + self-attention</td>
<td>73.4</td>
</tr>
</tbody>
</table>


