Research on Music Emotion Classification Based on Lyrics and Audio

Wanglei Shi, Shuang Feng

School of Computer Science

Communication University of China
Beijing, China

wangleishi_cuc@163.com

***Abstract*— With the development of audio recognition and research, more and more attention has been paid to the field of music emotion classification. However, it often can not get the expected effect of classification through parameter optimization and algorithm improvement of a single classification model to improve the accuracy of emotion classification, which is often limited by the characteristics of the model itself. Based on the techniques of text processing and music content in the computer field, the paper proposes combining lyrics classification with audio classification, so as to achieve the purpose of emotion classification. In terms of lyrics classification, LSI is used to reduce the dimensionality of the lyrics matrix, which effectively removes noise and uses SVM for classification. In terms of audio classification, BP neural network is used to complete the emotion classification of music; Finally, the two classification will be fused based on the improved algorithm of LFSM. The experimental results show that the combination of multiple classification methods for music emotion recognition can achieve higher accuracy of classification.**

Keywords*— *music emotion classification; lyrics classification; audio classification; LFSM

I. INTRODUCTION

In recent years, Network technology has developed rapidly. Music, as one of the main carriers of people's emotion communication and information sharing, has been greatly improved in transmission speed and the field .It is accompanied by the problems of retrieval and management of massive online music resources, which have aroused more and more attention from academics and industry. The classification problem of music is the basis of music retrieval. Through labeling different types of music, the purpose of music classification management can be achieved. However, the artificial labeling based on musical emotion requires high cost and long time and different people have different emotion perception to the same music, which also causes the problem of low accuracy. The research of music emotion automatic classification system that combines machine learning and music emotion classification together is becoming more important.

Music emotion recognition can be roughly divided into two aspects: one is based on music content, which mainly involves the problems of the extraction of emotion-related features in music, the selection of training models, and parameter optimization. For example, Gao et al. [1] based on the music dictionary, carried out the sparse decomposition of each frame of the music spectrum, and used the music word histogram as the feature for emotion recognition. Chin et al. [2] constructed SVM emotion recognition models for different genre music respectively. Another kind music emotion recognition bases on textual statistics, such as David Torres et al. [3] who identified the emotional type of music by analyzing the lyrics information of the music. In addition, the commonly used classification models such as Gaussian Mixture Model, Bayesian Classifier, SVM, etc. [4~6] are also used as algorithm tools for musical emotion recognition. A comprehensive model-based fusion algorithm is currently widely used in the optimization of machine learning. For example, Han et al. [7] used the double-level support vector machine fusion method to complete the optimization of protein particle position prediction. Wu et al. [8] used the two-level fusion model for the optimization and promotion of authoritative advertising rankings. Hadavandia et al. [9] proposed a new neural network for the optimization of classification accuracy based on the multi-classification problem of fusion algorithms.

This article integrates the classification of lyrics classifiers and audio classifiers into the emotion recognition domain of music, and completes the emotion recognition of music based on the LFSM [10] improved algorithm, lyric emotion classification and audio emotion classification. In terms of lyric-based music emotion recognition, the pretreatment of data sets, mapping of vector feature space to semantic space, and SVM-based emotion classification are mainly introduced. In the aspect of audio-based music emotion recognition, feature selection, the construction of neural network and the evaluation of the classification model are introduced emphatically.

978-1-5386-4509-3/18/$31.00 ©2018 IEEE

II. EMOTION RECOGNITION MODEL BASED ON MULTI-TYPE CLASSIFIERS

On account of the lyrics-based and audio-based classification of music emotions, this paper introduces the LFSM improved fusion algorithm to integrate the classification results of the two methods, which greatly improves the efficiency of musical emotion recognition. The experimental process is as follows:

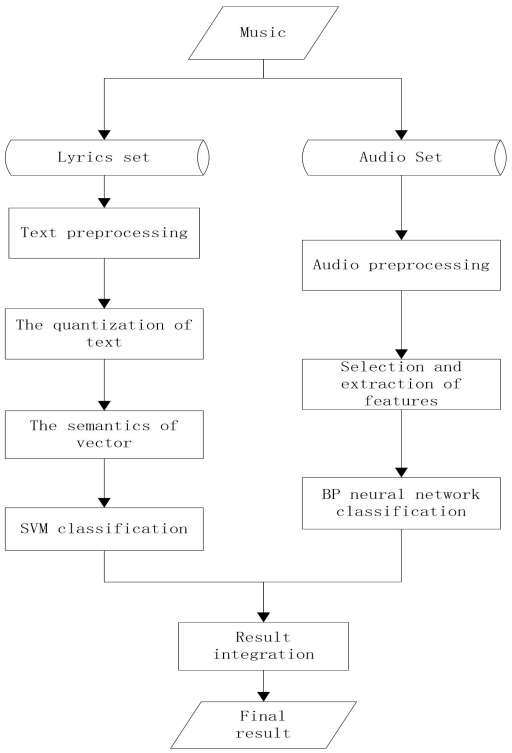

Fig. 1. The flow diagram of algorithm

The experimental process is divided into three phases:

A. *Lyrics-based music emotion recognition*

Firstly, the lyrics texts are preprocessed. The TFIDF is used to transform the processed texts into feature vectors, and the LSI model is used to reduce the dimension, mapping the feature vectors to the semantic space. The SVM is then used to classify the data sets to complete the music emotion classification.

B. *Audio-based music emotion recognition:*

The audio data is subjected to preprocessing such as frame division and windowing, and corresponding audio features are extracted according to the music emotion classification task. Finally, a BP neural network is constructed and audio-based music emotion classification tasks is completed.

C. *Improved fusion algorithm based emotion recognition*

The LFSM fusion algorithm is analyzed and the corresponding improvement method is proposed. Using the improved fusion algorithm to integrate the classification results of the two classification methods to complete the final task of emotion classification .

III. EMOTIONAL MODEL OF MUSIC

The emotional model of music is a structured and visual expression way of musical emotions. At present, the more widely used emotional models include the VA emotional model [11,12], the Hevner emotional model, and the Tellegen-Watson-Clark emotional model. The VA emotional model maps musical emotions to a point in a two-dimensional plane. Using the potency and activation as the horizontal and vertical coordinates of the emotion model respectively, the Valence is the evaluation of the emotional attributes. From left to right, it represents the negative of the emotion to positive; the activation (arousa) represents the intensity of the emotion. From bottom to top, it changes from calm to strong. As shown in fig.1, each quadrant embodies representative different emotional categories and has a good correspondence relationship with musical emotional categories. Therefore, this paper completes the selection of OVO-SVMs model and the construction of BP neural network based on the AV sentiment model.

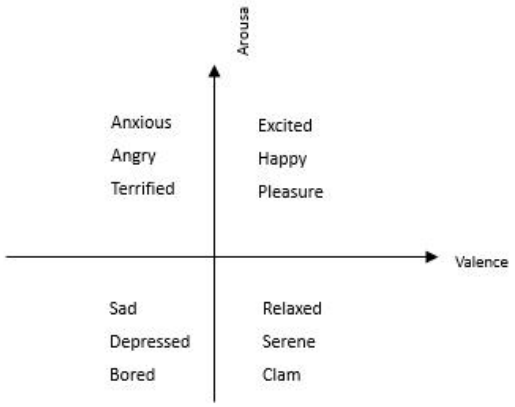

Fig. 2 *Emotional model*

IV. MUSIC EMOTION RECOGNITION BASED ON LYRICS AND AUDIO

A. *Emotional definition and construction of the data set of music*

Because emotion involves the content of psychology, it is influenced much by people's subjective factors, leading to the uncertainty of the criteria. Therefore, musical emotions are mainly divided into two main aspects: the " Expression theory" and "Evocation theory".

Emotional tags based on "Expressions" are more inclined to the intuitive expression of musical emotions, relatively more objective, lower labeling costs, and often used for emotional tagging in music retrieval. "Evocation" is more inclined to the audience's understanding of music, and the subjectivity is stronger, so it is more used in the field of music personalized recommendation.

The construction of the data set directly affects the accuracy of classification and prediction in the emotional computing of music. This music emotion classification experiment is used as the basis of music retrieval. In order to ensure the accuracy of the experiment, the lyrics text of music is collected based on the principle of "Expression theory" of emotions. By setting up and collecting corresponding multi-track MIDI files, the experimental results verify the reliability of the data set.

B. Music emotion recognition based on lyrics

The lyrics contain rich emotional information. Using machine learning methods to categorize emotions of lyrics mainly involves natural language processing techniques. In order to facilitate computer understanding of lyrics information, it is necessary to map the captured lyrics information to the mathematical level to form representative features. vector. The training accuracy of the model is closely related to the choice of feature vectors.

1) Lyric data set processing and modeling method

a) Mapping of Text to Vector Space Model

The vector space model (VSM) maps documents into N-dimensional feature vectors that are key components. Because each word has a different degree of contribution to the meaning of the document, that is, the weight of the word is different, the weight of the word in the document needs to be calculated using TDIDF, and a word with a higher weight value is selected as a feature item:

Suppose a document contains n terms, denoted as (W_1, W_2, W_3, …, W_n), and the number of occurrences of each type of entry is(N_1, N_2, N_3, …, N_n). The entry appears in the document frequency is(D_1, D_2, D_3, …, D_n). Under the premise that each term is independent and regardless of the order of appearance, the keyword W_1 can be expressed as:

$$\frac{-N_i \log(D_i)}{K} = \frac{N_i}{K} \log\left(\frac{1}{D_i}\right) \tag{1}$$

Where $\frac{N_i}{K}$ is the word frequency (TF), and $\log\left(\frac{1}{D_i}\right)$ is the inverse document frequency (IDF) of the keyword. The weights of each keyword are calculated by TFIDF, and some keywords are filtered out as feature components so as to realize the mapping of the text to the VSM.

b) Invisible Semantic Index

Traditional TFIDF's calculation of text assumes that the keywords in the text are semantically independent. Simply considering keywords, such as word frequency, is influenced to a large extent by the accuracy of synonyms and polysemes.

LSI can obtain the correlation between keywords in documents and reflect the semantic relationships among documents. Through the singular value decomposition of feature matrix, the singular value vector with small semantic relevance is excluded and the expression of high-dimensional VSM is reflected to low-dimensional semantic vector space while reducing the computational complexity. The LSI calculation process is as follows:

- Decompose the SVD according to feature of matrix T calculated by TFIDF, The formula is as follows:

$$T = UXV^T \tag{2}$$

  Where T is the original characteristic matrix, U and V is the orthogonal matrix, X is the diagonal matrix, each value is singular value:

$$X = diag(a_1, a_2, a_3, \dots, a_n) \quad a_1 \geq a_2 \geq \dots \geq a_n > 0 \tag{3}$$

- Select the corresponding partial matrix according to the set singular value threshold to obtain the approximate matrix of T as Ti:

$$T_i = U_{mi} X_i V_{in}^T \tag{4}$$

- The new feature vector processed by TFIDF is applied to the mapping of the latent semantic index space by using the approximate matrix calculation formula.

After the restoration of T_i and T, there will be large differences in the data. This is a visual expression of the denoising effect by discarding the feature vectors of the lower singular values. Excluding synonyms and polysemes in the data set greatly improves the efficiency of data retrieval and classification operations.

2) Audio classification model

Based on the theory of statistical learning, support vector machine (SVM) introduces and defines the concept of "interval" and achieves the classification effect by maximizing the separation between the support vector and the separating hyperplane. It is a systematic, standardized, and complete classification theory and classification model. It is widely used in the reverse of text classification, emotion recognition, and target detection.

The SVM chosen in the experiment is based on the VA sentiment model theory, and the musical emotion is divided into different categories for emotion recognition in four quadrants. Considering that the VA sentiment model is continuously distributed, it may lead to the failure of the recognition of two types of emotions with similar potency or activation in the process of emotion recognition. Therefore, the use of OVO SVMs voting to conduct sentiment classification has improved classification accuracy to some extent.

C. Audio-based music emotion recognition

1) Selection and extraction of music features

a) MFCC

Mel Frequency Cepstrum Coefficient (MFCC) is based on the characteristics of the human auditory auditory system. Compared with other features, it has better classification stability and higher classification accuracy, and can achieve very good results in audio classification tasks. It is widely used in the current audio analysis field.

b) Base frequency

The frequency of vibration of different music in the time domain varies greatly. The lowest frequency of music in a particular period of time is called the pitch frequency. The

frequency of the gene represents the pitch of the music during that period of time. The pitch is closely related to the emotional expression of the music; The change in pitch frequency is called pitch, and music in different tones often has a greater difference in musical sentiment.

In this experiment, the fundamental frequency is calculated for different music and sub-frames, and the average frequency of the fundamental frequencies of different music is used as a model training parameter.

c) Sound length

The length of sound represents the duration of music notes, and the correlation between musical emotion and sound length is strong. In general, the shorter the sound length, the more joyful the music contains; on the contrary, the more sadness the music contains.

In the experiment, the feature of the length of sound is extracted according to the delta-time in the MIDI file, and the length of each note is quantified according to the threshold set in advance so as to complete the extraction of the length characteristics of the entire music.

d) Timbre

The timbre refers to the performance of different sound frequencies in terms of waveforms. Different sounders have different timbres due to different structures and materials. Different timbres can often be linked to musical emotions. For example, the tone of the harmonica is long and melancholy, and the expression of emotion is more sad; the tone of the Scottish bagpipes is cheerful and clear, giving people a feeling of pleasure.

The MIDI audio files used in the experiment are labeled with different timbres, and the timbre characteristics can be extracted by the numbers corresponding to different timbres.

e) Formant

The formant is the part of the frequency where the sound becomes stronger due to resonance in the process of propagation. Through resonance, it can directly reflect the physical characteristics of the channel. Channel's change in shape during transmission could reflect the different emotions contained in music, such as happy, anger, and sadness.

The peak frequency monitoring method was used to calculate the formant frequency and bandwidth, and the average values of the three formants were calculated on the basis respectively.

2) Audio classification model

BP neural network is a typical feed forward neural network, which is a multi-layer network structure with error propagation backwards. It Includes the input layer, hidden layer and output layer and all adjacent layers are fully connected to transfer information and calculate. Continuously adjusting the weight of the network by means of forward propagation of information and backward propagation of errors, the squared error sum of the network model is minimized, thereby improving the accuracy of emotion classification .

According to the selection of 5-dimensional emotion characteristics: emotional characteristics MFCC, pitch frequency, pitch length, timbre, and formant , the number of input layer nodes is set to 5; According to the VA emotion model, music emotion is divided into happiness, anger, sadness and relax 4 parts according to the coordinate quadrant, respectively marking as (0,0), (0,1), (1,0), (1,1), and set the output layer parameters to 4. On this basis completes the construction of BP neural network and music emotion recognition .

V. Sentiment Classification Mechanism Based on Fusion Algorithm

The integration of lyric-based and audio-based music emotion recognition methods can improve the accuracy of music emotion recognition to a certain extent. The implementation of ensemble learning methods is mainly to perform the operation of the integration of the classification results of the two classification methods.

Based on the VA sentiment model, the traditional subtask combined with the late fusion method (LFSM) maps sentiment to the coordinates of the two-dimensional sentiment model and considers that the sentiment classification based on lyrics has better discrimination in valence. So does the sentiment classification based on audio in arousa. Therefore the classifications of the lyrics in valence and the audio in arousa are taken separately as the results of music emotion recognition.

The LFSM expresses the musical emotions by the four quadrants in coordinates, such as happy (valence+, arousa+), angry (valence-, arousa+), sad (valence-, arousa-), relaxed (valence+, arousa-). The fusion process is shown in the following table:

TABLE I. LFSM FUSION PRINCIPLE

Lyrics	Audio	Emotion
Valence+	Arousa+	Happy
Valence-	Arousa+	Angry
Valence+	Arousa-	Sad
Valence-	Arousa-	Relaxed

The use of LFSM greatly improves the emotional classification accuracy of musical emotion. But after analysis, it is found that LFSM omits multi-modal fusion while ignoring the categorical contribution of lyric-based classification in valence and audio-based classification in arousa.

Improved on the basis of LFSM, the multimodal fusion method considers that both classification methods have a certain emotional classification contribution in valence and arousa. So the coordinates of the two classification results are compared firstly. If the results are consistent, the classification is completed. If not, the inconsistent coordinate vector is judged. If it is the valence coordinate, the audio-based classification method is used for secondary classification, subject to the secondary result, and the valence coordinate vector is taken as the final classification result; if it is the arousa coordinate, then the lyric-based classification method is used for secondary classification, subject to the secondary results, and its arousa coordinate vector is taken as the final classification result.

If the lyric-based classification result is happy (valence+, arousa+), and the audio-based classification result is relaxed (valence+, arousa-), then use the audio-based emotion recognition model for secondary classification. If the classification result is relaxed (valence+, arousa-) or sad (valence-, arousa-), then the final result is relaxed (valence+, arousa-); if the classification result is happy (valence+, arousa+) or angry (valence-, arousa-), the end result is happy(valence+, arousa+).

VI. Experimental Results and Analysis

Based on the VA emotion model, this article classifies musical emotions into four categories: happy, angry, sad, and relaxed with the method of the BP neural network-based audio emotion classification and the SVM-based lyric music emotion classification .

a) Selection and division of data sets

This experiment is based on the principle of emotional expression theory to collect mature emotional tag data sets. A representative 340 MIDI audio resources and corresponding lyrics texts are selected as data sets for the four different types of emotions from a wide range of MIDI music resource websites, of which there are 70 emotionally distinct audio data with the corresponding lyrics data for training and 60 data for testing in the four emotional styles of happy、 angry、 sad、 relaxed.

b) The construction of emotional models

1） The construction of BP neural network

According to the results of the four types of music types divided by the five-dimensional audio features selected by the neural network and the VA emotion model, the input and output layers of the BP neural network are determined to be five input neural nodes and four output neural nodes, respectively.

Considering that increasing the number of hidden layers has little effect on improving the classification accuracy of the network, and it is easy to cause network complexity and over-fitting problems, the three-layer BP neural network model is selected as the emotion recognition model.

The selection of the hidden layer node reference empirical formula is as follows:

$$w = \sqrt{u + v} + l \quad (1 \le l \le 10) \tag{5}$$

w is the number of hidden layer nodes. *u* is the number of input layer nodes. *v* is the number of output layer nodes and l is a constant. After calculation, the number of hidden layer nodes are set to 6. The final structure of the network is 5-6-4.

The neural network selected activation function is sigmoid. The expected error is 0.001, and the maximum number of cycles is 1000. The training function selects the momentum gradient descending function (traingdm), and the performance function selects the mean squared error function (mse). The classification result is as follows:

TABLE II. MUSIC EMOTION CLASSIFICATION BASED ON AUDIO

Happy	Angry	Sad	Relaxed
88.34%	86.51%	89.27%	90.65%

2） Construction of SVM Model

Taking into account that the continuous distribution of emotional characteristics of the VA sentiment model will reduce the accuracy of recognition of emotion types with similar levels of valence or activation levels in the process of emotion classification , the use of separate recognition of pairs of emotions and voting to complete the final classification of emotion recognition.

Set the multi_class parameter of the SVM model to ovo, the loss function to squared_hinge, the penalty parameter to 1, and the maximum number of training iterations to 1000. The emotional classification results are as follows:

TABLE III. MUSIC EMOTION CLASSIFICATION BASED ON LYRICS

	Happy	Angry	Sad	Relaxed
Happy				
Angry	78.43%			
Sad	81.74%	74.38%		
Relaxed	78.85%	82.27%	77.21%	

Through observing the experimental results, we found that the average accuracy of audio-based music emotion recognition is 88.70%, and the average accuracy of music emotion recognition based on lyrics is 78.81%. After analysis, the emotion classification method based on music content performs better in completing the calculation of music from the perspective of hearing. It is more in line with the expression form of music, and is higher in classification accuracy. The emotion classification based on lyrics also has a good classification effect, which can be used as an auxiliary method for music emotion recognition.

B. Comprehensive comparison of multiple classification methods

In this paper, a variety of classification methods are used to realize the musical emotion recognition, and the improvement of the LFSM fusion algorithm is proposed. On the basis of the first experiment, the second experiment uses the LFSM algorithm and the improved fusion algorithm to perform music emotion recognition research. The experimental results are shown in the following table:

TABLE IV. MUSIC CLASSIFICATION BASED ON MULTIMODAL FUSION

classification method	accuracy
SVM (lyrics)	78.81%
BP (audio)	88.70%
LFSM	90.27%
Improved LFSM	92.33%

From the experimental results, the average classification accuracy of music emotion classification based on LFSM algorithm can reach 90.27%, which greatly improves the accuracy of emotion classification. The improved fusion algorithm shows higher accuracy in emotion classification, the average classification accuracy reaching 92.33%, which reflects the feasibility and effectiveness of the improved fusion algorithm.

VII. CONCLUSION

This article mainly discusses the effect of the combination of different types of classification methods on efficiency of the emotional recognition of music. It combines the lyrics based and music content based emotion recognition method together, and an improved algorithm is proposed on the basis of the traditional fusion method. Finally, the improved fusion method is verified by analyzing the experimental data.

The combination of different types of emotion recognition methods provides a new idea for improving the accuracy of music emotion recognition, and it is worth in-depth discussion and study. In the future research, I will put more efforts in improving the accuracy of emotional recognition, and attempt more classification methods integration into the emotional recognition of music.

REFERENCES

[1] B Gao , E Dellandrea , L Chen. Music sparse decomposition onto a MIDI dictionary of musical words and its application to music mood classification[C]. International Workshop on Content-based Multimedia Indexing, 2012: 1-6

[2] YH Chin, PC Lin, TC Tai, JC Wang, et al. Genre based emotion annotation for music in noisy environment[C]. International Conference on Affective Computing and Intelligent Interaction., 2015:863-866

[3] David Torres, et al. Identifying Words that are Musically Meaningful[J]. University of California, San Diego. Austrian Computer Society, 2009: 143-152.

[4] Umapathy K, Krishnan S, Jimaa S. Multigroup Classification of Audio Signals Using Time-Frequency Parameters[J]. IEEE Trans. onMultimedia, 2005, 7 (2): 308-315.

[5] Ogihara M. Content-Based Music Similarity Search and Emotion Detection[C]. Proceedings on 2004 IEEE International Conference on Acoustics,Speech and Signal Processing,Fairmont Queen Elizabeth Hotel,Montreal,Quebec,Canada, 2004: 17-21.

[6] Wang M. User-Adaptive Music Emotion Recognition[C]. IEEE Transactions on Audio,Speech and Language Processing, 2008, 16 (2): 448-457.

[7] GS Han , ZG Yu , V Anh , AP Krishnajith , YC Tian. An ensemble method for predicting subnuclear localizations from primary protein structures[J]. Plos One, 2013, 8 (2): e57225.

[8] K Wu , C Ferng , C Ho , A Liang , C Huang, et al. A Two-Stage Ensemble of Diverse Models for Advertisement Ranking in KDD Cup 2012[Z]. ACM KDD Cup 2012 Workshop, 2012.

[9] E Hadavandi , J Shahrabi , S Shamshirband. A novel Boosted-neural network ensemble for modeling multi-target regression problems[J]. Pergamon Press, Inc., 2015, 25 (C) : 204-219.

[10] YH Yang , YC Lin , HT Cheng , IB Liao , YC Ho. Toward Multi-modal Music Emotion Classification[C]. Springer Berlin Heidelberg, 2008, 5353: 70-79.

[11] Russell, J, A. A circumplex model of affect[J]. Journal of Personality and Social Psychology, 1980, 39 (6): 1161-1178.

[12] J Posner,JA Russell,BS Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology[J]. Development and Psychopathology, 2005, 17 (3):715-734.