Multimedia Tools and Applications (2019) 78:11563–11584
https://doi.org/10.1007/s11042-018-6637-6

Regional classification of Chinese folk songs based on CRF model

Juan Li[1] · Jing Luo[2] · Jianhang Ding[2] · Xi Zhao[2] · Xinyu Yang[2]

Received: 15 November 2017 / Revised: 16 July 2018 / Accepted: 31 August 2018 /
Published online: 27 September 2018

Abstract
Music regional classification, which is an important branch of music automatic classification, aims at classifying folk songs according to different regional style. Chinese folk songs have developed various regional musical styles in the process of its evolution. Regional classification of Chinese folk songs can promote the development of music recommendation systems which recommending proper style of music to users and improve the efficiency of the music retrieval system. However, the accuracy of existing music regional classification systems is not high enough, because most methods do not consider temporal characteristics of music for both features extraction and classification. In this paper, we proposed an approach based on conditional random field (CRF) which can fully take advantage of the temporal characteristics of musical audio features for music regional classification. Considering the continuity, high dimensionality and large size of the audio feature data, we employed two ways to calculate the label sequence of musical audio features in CRF, which are Gaussian Mixture Model (GMM) and Restricted Boltzmann Machine (RBM). The experimental results demonstrated that the proposed method based on CRF-RBM outperforms other existing music regional classifiers with the best accuracy of 84.71% on Chinese folk songs datasets. Besides, when the proposed methods were applied to the Greek folk songs dataset, the CRF-RBM model also performs the best.

Keywords Music regional classification · Conditional random field · Restricted boltzmann machine · Temporal characteristics · Chinese folk songs

1 Introduction

With the rapid development of the Internet and music multimedia technology, the amount of music data stored and shared in the Internet grows massively. As a result, the demand for

✉ Xinyu Yang
yxyphd@mail.xjtu.edu.cn

Juan Li
lijuan@mail.xjtu.edu.cn

Extended author information available on the last page of the article.

music multimedia technology increases, and brings severe challenges and new changes. The development of music automatic classification technology [8] plays a fundamental role in the music indexing and retrieval, helping to manage the music of different categories more conveniently.

Chinese folk songs were created by local people's improvisation and passed on from one generation to the next orally. The folk tunes from the same region exhibit a particular and stable style while tunes from different areas present different regional styles [32]. For example, *MoLiHua* from southern Jiangsu tends to be lyrical, gentle, while *MoLiHua* from northeastern China shows rugged, intense and disjunct characteristics,[1] although they share almost identical lyrics and content. Various dialects, local customs and living conditions of distinct areas have profound impact on the formation of Chinese folk songs' melody style [6]. Based on these characteristics, Chinese enthnomusicologists have developed the division of Chinese folk songs according to geographic factors and named their study "Music Geography" [9, 32]. Moreover, the regional characteristics also exist in the folk songs from other countries. Greek folk songs from mainland, Islands and Asia Minor exhibit different musical styles [7], Japanese folk songs from Kanto area and Kansai area often use various melodic patterns [16]. Therefore, it is an effective methods to manage and recommend Chinese folk songs according their geographic label. Besides, research on Chinese folk songs regional classification is helpful for understanding music structure of folk songs, providing ways to automatically and quantitatively analyze folk songs, and further promoting the development of intelligent music education.

Existing approaches for music regional classification usually imitate the processing of music genre classification [5, 8, 36]. For both audio files data [1, 7, 19, 28, 29, 34, 37] and MIDI files data [4, 10, 15–17, 20], the statistical features of the whole song are usually calculated first, then the features are classified or clustered by machine learning algorithms. However, music regional classification is different from the mature music genre classification since folk songs normally have no strict creation rules. Instead, the melodic temporal structure is a key feature of folk songs, and temporal characteristic is quite important for distinguishing folk songs from different regional styles. The existing approaches for music regional classification have the problem of not considering enough temporal characteristics of music.

In this paper, we proposed a method based on Conditional Random Field (CRF) [21] to identify the regional style of Chinese folk songs. The musical audio features based on frames are regarded as the observation sequence of CRF. In addition, due to CRF's Markov hypothesis, the temporal characteristics of music are fully considered. In order to improve the probability calculation of CRF and promote the modeling and classification results, two ways of calculating the label sequence were put forward. We first used Gaussian Mixture Model (GMM) [30] to fit musical audio features and calculated the label sequence in CRF. Although GMM can fit the audio features well, it is difficult to improve the computational accuracy of label sequence when the number of Gaussian components is restricted. To solve this problem, Restricted Boltzmann Machine (RBM) [12, 31] was used instead, which has better nonlinear mapping ability and more variable space than GMM to fit the audio features. The experiment results demonstrated that the CRF-RBM model performs the best with the accuracy of 84.71%, which outperforms the CRF-GMM, CRF-DBN, LSTM and other

[1] A YouTube video of introduction to Chinese folk songs by Linna Gong (in Chinese): https://www.youtube.com/watch?v=HcBqnIHgYdg. The live singing without accompaniment of *MoLiHua* from southern Jiangsu is the part of the video from 11:53 to 12:19, while *MoLiHua* from northeastern China is from 12:28 to 13:00.

music regional classifiers. When we employed the same classifiers on Greek folk songs dataset, CRF-RBM also achieves the best performance. Moreover, we employed Wilcoxon Signed-Rank test to prove our proposed method's efficiency.

The rest of this paper is organized as follows: After introducing related work on music regional classification in Section 2, we present our method, including analysis of the music modeling based on CRF in Section 3 and the sequence labeling based on GMM and RBM(DBN) in Section 4. In Section 5, we show the experimental results and analysis. Conclusions and future research of this paper is presented in Section 6.

2 Related work

The research object of music regional classification is usually folk songs. Folk songs develop different regional styles in the process of its evolution, due to the influence of different geographic culture and languages. So researchers generally distinguish the categories of folk songs according to geographical areas. At present, there are relatively few studies on automatic music regional classification. To the best of our knowledge, current approaches for music regional classification are similar to those for music genre classification [5, 14, 18, 31, 33, 35, 42], although there are many differences between folk songs and genre songs in nature. However, the distinction of folk songs is more difficult compared to genre songs. Two main reasons are summarized as follows. Firstly, folk songs can be regarded as one kind of genre songs, i.e. the regional classification can be considered as the classification within one kind of genre songs. Secondly, the creation of folk songs lacks strict rules since folk songs are normally created by people's improvisation and songs are infected by geographical characteristics during the spread. The following is a brief review on the research of music regional classification.

Firstly, researchers usually collect folk songs audio datasets and test their classifiers on it. For instance, Bassiou et al. [1] employed Canonical Correlation Analysis (CCA) and Deep Canonical Correlation Analysis (DCCA) to calculate the correlation among lyrics, audio data and regional tags for the classification of Greek folk songs. The best result (72.9%) of folk songs regional classification based on correlation analysis of CCA was obtained. Fotiadou et al. [7] extracted auditory cortical representations from the music recordings from 8 Greek regions. Their experimental results demonstrated that the classifier of SVM obtained an average classification rate of 73.25%. Khoo et al. [19] extracted musical features from both time and frequency domains of audio files, using Regularized Extreme Learning Machine (R-ELM) classifier for the Chinese folk songs regional classification. The results showed that the classification accuracy was only 49%. Liu et al. [29] achieved the accuracy of 75.2% based on Post-Processing method and SVM classifier for the classification of Chinese folk songs. Later Liu et al. [28] proposed an active feature selection algorithm which not only reduced the dimension of musical audio features, but also improved the performance of SVM. Song et al. [37] proposed HBS and HFS feature selection algorithms, and used SVM classifier to classify the regional style of Chinese folk songs, they achieved the best performance of 78.9%.

In addition to audio datasets, the researchers also establish folk songs MIDI datasets to study music regional classification. Conklin et al. [10] collected 3367 European folk songs in six regions, using four MIDI feature sets to extract musical features to do the regional classification of folk songs. Kawase et al. [16] divided the representative folk songs which come from 45 Japanese counties into 11 classes by extracting 24 types of interval combination patterns. These folk songs were divided into two parts corresponding to the east and

west area by hierarchical clustering method. After that, Kawase [15] further conducted a quantitative analysis of traditional folk songs from Shikoku district. He executed a classification based on tetrachords and found the geographically adjacent provinces have similar characteristics. Khoo et al. [20] selected Germany and Austria folk songs as the dataset. They proposed a method using Musical Features Density Map (MFDMap) as the music representation and the Finite Impulse Response Extreme Learning Machine (FIR-RLM) as the classifier.

Although various approaches have been proposed for music regional classification, most of them rarely take the temporal characteristics of the melody structure into consideration, it leads to their methods are not very effective. Therefore, there are still many space to be improved for music regional classification.

3 Music regional classification based on CRF

CRF was first proposed by Lafferty [21] to solve the problem of annotation and segmentation [39]. By using CRF, we take the temporal characteristics of musical audio features into account. The music of different regional types can be modeled by parametric form and parameter learning algorithm of CRF, and the regional category of the music is identified by the probability calculation of CRF.

3.1 Parametric form

The structure of music modeling based on CRF is shown in Fig. 1. Assume that j-th region of music sample in the training set is $S = \{(x^{(t)}, y^{(t)})_{t=1}^{\|T_j\|}\}$, $T_j \in \{T_1, T_2, ..., T_m\}$.

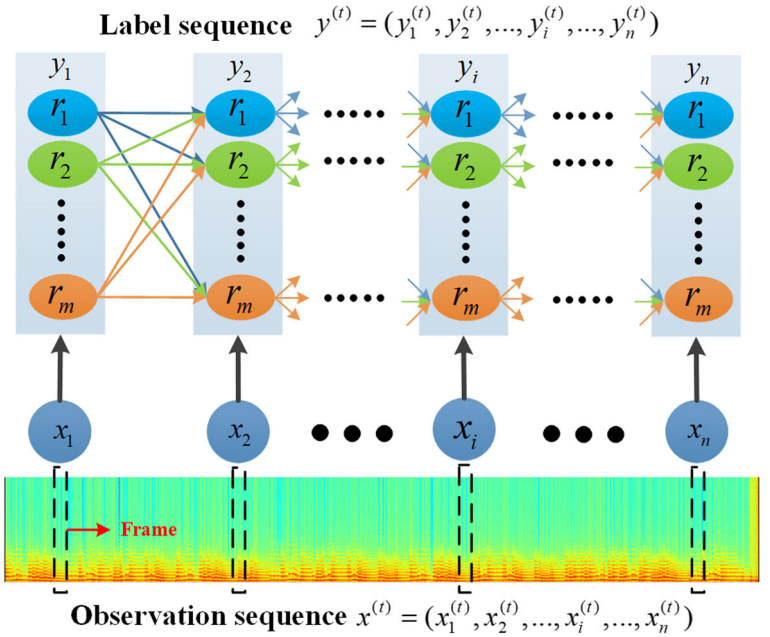


Fig. 1 Structure diagram of music modeling based on CRF

$\|T_j\|$ represents the number of songs in the j-th region. The audio sequence $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, ..., x_i^{(t)}, ..., x_n^{(t)})$ of any song is given as the observation sequence in CRF, where $x_i^{(t)}$ represents the i-th frame feature in the audio sequence, which is a d-dimensional feature vector $\Phi(x_i^t) \in R^d$. $y^{(t)} = (y_1^{(t)}, y_2^{(t)}, ..., y_i^{(t)}, ..., y_n^{(t)})$ is the label sequence in CRF, and $y_i^{(t)}$ is regional label corresponding to the i-th frame audio feature $x_i^{(t)}$. $y_i^{(t)}$ belongs to the state set $r = \{r_1, r_2, ..., r_j, ...r_m\}$ in CRF, where m is the number of regional categories in the dataset, i.e., the number of states in the CRF for each moment.

The CRF based parametric form of the sample set S is shown in following equation.

$$P(y|x) = \prod_{t=1}^{\|T_j\|} \frac{1}{Z(x^{(t)})} \exp\left(\sum_{k=1}^{K} w_k, f_k\left(x^{(t)}, y^{(t)}\right)\right) \tag{1}$$

where $Z(x^{(t)})$ is normalization factor, which is the sum of all possible label sequences $y^{(t)}$, as shown in following equation.

$$Z(x^{(t)}) = \sum_{y^{(t)}} \exp\left(\sum_{k=1}^{K} w_k, f_k\left(x^{(t)}, y^{(t)}\right)\right) \tag{2}$$

where w_k and $f_k(x^{(t)}, y^{(t)})$ represent parameters and feature function respectively. $f_k(x^{(t)}, y^{(t)})$ is the sum of feature functions of all moments in CRF.

$$f_k(x^{(t)}, y^{(t)}) = \sum_{i=1}^{n} f_k(y_{i-1}^{(t)}, y_i^{(t)}, x^{(t)}, i), \quad k = 1, 2, ..., K \tag{3}$$

$f_k(y_{i-1}^{(t)}, y_i^{(t)}, x^{(t)}, i)$ contains the state function s_l and transfer function t_k, which is defined in following equation.

$$f_k(y_{i-1}^{(t)}, y_i^{(t)}, x^{(t)}, i) = \begin{cases} t_k(y_{i-1}^{(t)}, y_i^{(t)}, x^{(t)}, i) \ , & k = 1, 2, ..., K_1 \\ s_l(y_i^{(t)}, x^{(t)}, i) \ , & k = K_1 + l; \ l = 1, 2, ..., K_2 \end{cases} \tag{4}$$

The regional label of each frame in the musical audio sequence is calculated by s_l and the transfer path between adjacent frames is calculated by t_k. K_1 denotes the number of all transfer paths, and K_2 the number of all label states.

The definitions of state function s_l and transfer function t_k in music regional classification are shown as follows.

$$s_l(y_i^{(t)}, x^{(t)}, i) = \begin{cases} 1 \ , & y_i^{(t)} = r_j \\ 0 \ , & otherwise \end{cases} \tag{5}$$

$$t_k(y_{i-1}^{(t)}, y_i^{(t)}, x^{(t)}, i) = \begin{cases} 1 \ , & y_{i-1}^{(t)} = r_{j'} \ and \ y_i^{(t)} = r_j \\ 0 \ , & otherwise \end{cases} \tag{6}$$

State function s_l indicates that the i-th frame feature $x_i^{(t)}$ is belong to the regional style represented by the state r_j. It is a vector of length $\|r\|$. When the regional label $y_i^{(t)}$ of the i-th frame feature is state r_j, s_l is recorded as 1, otherwise 0. Transfer function t_k indicates that the song is converted from regional style represented by state $r_{j'}$ into regional style represented by $r_{j'}$ between the $(i-1)$-th and i-th moments. It is initialized as a $\|r\| \times \|r\|$ matrix. When the adjacent time regional labels of frame features $y_{i-1}^{(t)}$ and $y_i^{(t)}$ are states $r_{j'}$ and r_j, respectively, t_k is recorded as 1, otherwise 0.

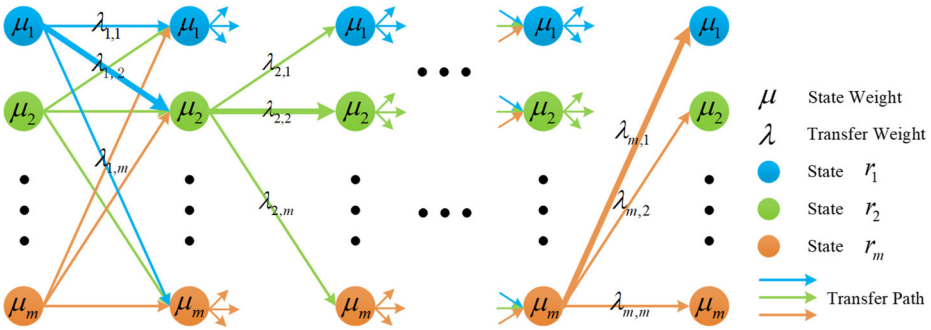


Fig. 2 States and transfer paths in CRF

The states and transfer paths of CRF are shown in Fig. 2. The frame features $y_1 = r_1$ and $y_2 = r_2$ at the first and second moments are represented by the blue nodes and green nodes respectively. Then the first dimension and the second dimension of the state function of the two moments are set to 1, and other dimensions are set to 0. At the same time, the item (1, 2) of the matrix of the transfer function is set to 1, and other items are set to 0. State r_1 transmit to state r_2 through the path represented by the bold blue line in Fig. 2.

3.2 Parameters initialization and estimation

From the above equations, w_k is the weight of the feature function f_k, which contains state weight μ_l and transfer weight λ_k, as shown in following equation.

$$w_k = \begin{cases} \lambda_k \ , \ k = 1, 2, ..., K_1 \\ \mu_l \ , \ k = K_1 + l; \ l = 1, 2, ..., K_2 \end{cases} \tag{7}$$

Parameter λ_k multiplies with transfer function t_k, representing the weights on each path during the state transfer. Such as the values on the transfer path in Fig. 2. Therefore, parameter λ is initialized as a $\|r\| \times \|r\|$ matrix. $\|r\|$ is the number of states in regional category set.

Parameter μ_l multiplies with state function s_l. The conditional probability of the label sequence in the CRF model is decided by the non-independent and mutually interacting audio features in the observation sequence, and the importance of the frames is presented by the different weights of the audio features. Therefore, parameter μ is initialized as a $\|\Phi(x_i^{(t)})\| \times \|r\|$ matrix demonstrating the weight on each state corresponding to each dimension of the audio frame feature in CRF, such as the value on the state in Fig. 2. $\|\Phi(x_i^{(t)})\|$ is the dimension of the audio frame feature. In this experiment, the parameters λ and μ is initialized by zero matrix.

Suppose that there is a set $S = \{(x^{(t)}, y^{(t)})_{t=1}^{\|T_j\|}\}$ containing $\|T_j\|$ musical audio sequences and corresponding regional label sequences, according to the maximum likelihood function estimation, the probability density function is calculated as follows.

$$P(y|x, w) = \prod_{t=1}^{\|T_j\|} P(y^{(t)}|x^{(t)}, w) \tag{8}$$

The log-likelihood function of probability density function is defined as follows.

$$L(w) = \sum_{t=1}^{\|T_j\|} \log P(y^{(t)}|x^{(t)}, w) - \frac{1}{2\sigma^2}\|w\|^2 \tag{9}$$

The first item represents the log-likelihood function for all songs. To prevent overfitting, we need to add the Gaussian priori item with mean 0 and variance σ^2. In order to obtain the optimal parameters w^* of CRF, Quasi-Newton (BFGS) method [2] is used to iteratively update the parameters of the objective function $L(w)$ in our experiment.

3.3 Music regional classification

The classification includes three steps: firstly, the testing song, considered as the observation sequences, is used as the input of m CRFs. Then, the probabilities of the observation sequence belonging to each CRF are calculated by Forward-Backward (FB) algorithm [3]. At last, according to Naive Bayesian classification criteria, the test song belongs to the CRF with the maximum probabilities, which are calculated as the following equation shown. The log-likelihood function which is defined in Section 3.2 is used here.

$$C^{(t)} = \arg \max_{1 \le j \le m} f(j) = \arg \max_{1 \le j \le m} [P(y^{(t)}|x^{(t)}, w^{(j)})] \tag{10}$$

4 The calculation of label sequence

The feature function f_k is the key to get the label sequence of the audio sequence in CRF. Since the "frame" is a very short time unit, the audio frame features extracted from music is high-dimensional, continuous and quite large. It is almost impossible to label the frame features manually. In this paper, two methods are proposed for automatically labeling. One is based on Gaussian Mixture Model (GMM), and the other Restricted Boltzmann Machine (RBM).

4.1 Sequence labeling based on GMM

GMM is a parameterized generative model, which is commonly used in the domain of speech recognition recently [30], GMM can accurately fit multi-dimensions of data with enough Gaussian components. We use GMM to fit the audio features of different regional songs. In details, we calculate the state function s_l and transfer function t_k, and finally obtain the label sequence $y^{(t)}$ of audio sequences in CRF. Detail labeling process can refer to [24], the main steps is as follows :

Step 1: Each type of regional songs $S = \{(x^{(t)})_{t=1}^{\|T_j\|}\}$ in the dataset is fitted by m GMMs which represents the state set $r = \{r_1, r_2, ..., r_j, ..., r_m\}$ in CRF. The probability distribution of j-th region songs' audio features is fitted by GMM is computed by the following equation.

$$P(x|\theta^{(j)}) = \prod_{t=1}^{\|T_j\|} \prod_{i=1}^{n} \prod_{k=1}^{k} \pi_k N(x_i^{(t)}|\mu_k, \Sigma_k) \tag{11}$$

where x is audio data, and $\theta^{(j)}$ represents the parameters of GMM, i.e., the mean μ_k, covariance Σ_k and weight of the k-th Gauss component π_k.

Step 2: When the state r_j of regional label $y_i^{(t)}$ at any time (the state function s_l) in label sequence $y^{(t)}$ is calculated. According to Naive Bayesian classification criteria, the state r_j of regional label $y_i^{(t)}$ is determined by the maximum probability, as shown in the following equation. Then the current transfer function t_k is calculated by the adjacent moment's regional labels $y_{i-1}^{(t)}$ and $y_i^{(t)}$.

$$y_i^{(t)} = \arg \max_{1 \le j \le m} f(j) = \arg \max_{1 \le j \le m} [P(x_i^{(t)} | \theta^{(j)})] \tag{12}$$

The label sequence $y^{(t)}$ of the t-th song in $S = \{(x^{(t)})_{t=1}^{\|T_j\|}\}$ can be obtained after the above steps is executed.

During the training, the classical k-means clustering is adopted to initialize the parameters of GMM, and the Expectation Maximization (EM) algorithm is used to calculate the GMM parameters. The calculation of feature function f_k and label sequence $y^{(t)}$ of song in CRF is crucial to the performance of CRF. A better modeling of music in the training phase leads to more accurate label sequence and identification. But when the number of Gaussian components in GMM is limited, there is a bottleneck in the computational accuracy of label sequence, so the performance of CRF for music region recognition is difficult to be improved.

4.2 Sequence labeling based on RBM(DBN)

In this paper, RBM is further used instead of GMM to calculate the song label sequence in CRF. RBM is an unsupervised training network. Larochelle et al. [22, 23] demonstrated that RBM can be used for supervised, semi-supervised learning and multitask learning. There are two advantages of RBM: First, it is a nonlinear mapping mechanism, which is helpful to extract high level features that increase the difference between the original inputs. So CRF can use more flexible audio features. Second, variable space is larger in RBM. It can fit the musical audio features better.

The structure of music modeling based on CRF-RBM is shown in Fig. 3. The RBM(DBN) which is shown is the red dashed box contains a visible layer v and several hidden layers H^N. Assume that j-th region of music sample in the training set is $S = \{(v^{(t)}, y^{(t)})_{t=1}^{\|T_j\|}\}$, $T_j \in \{T_1, T_2, ..., T_m\}$. $\|T_j\|$ is the number of songs of current region. musical audio sequences are $v^{(t)} = (v_1^{(t)}, v_2^{(t)}, ..., v_i^{(t)}, ..., v_n^{(t)})$, where $v_i^{(t)}$ represents the i-th frame d-dimensional feature in the audio sequence which is denoted by $\Phi(v_i^{(t)}) \in R^d$.

$x^{(t)} = (x_1^{(t)}, x_2^{(t)}, ..., x_i^{(t)}, ..., x_n^{(t)})$ is the high-level hierarchy abstract features of RBM hidden layer after learning the original audio features $v^{(t)}$. The label sequence of the CRF is $y^{(t)} = (y_1^{(t)}, y_2^{(t)}, ..., y_i^{(t)}, ...y_n^{(t)})$, and $y_i^{(t)}$ is regional label corresponding to the i-th frame abstract feature $x_i^{(t)}$. $y_i^{(t)}$ belongs to the state set $r = \{r_1, r_2, ..., r_j, ...r_m\}$ of CRF, where m still represents the number of regional categories of songs in the dataset.

The i-th musical audio feature $v_i^{(t)} = \{(v_i^{(t)})_1, (v_i^{(t)})_2, ..., (v_i^{(t)})_I, ..., (v_i^{(t)})_{n_v}\}$ is the input for the RBM visible layer. n_v is the number of RBM's input nodes. $h_i^{(t)} = \{(h^{(t)})_1, (h^{(t)})_2, ..., (h^{(t)})_J, ..., (h^{(t)})_{n_h}\}$ is RBM hidden layer. n_h is the number of RBM's hidden units.

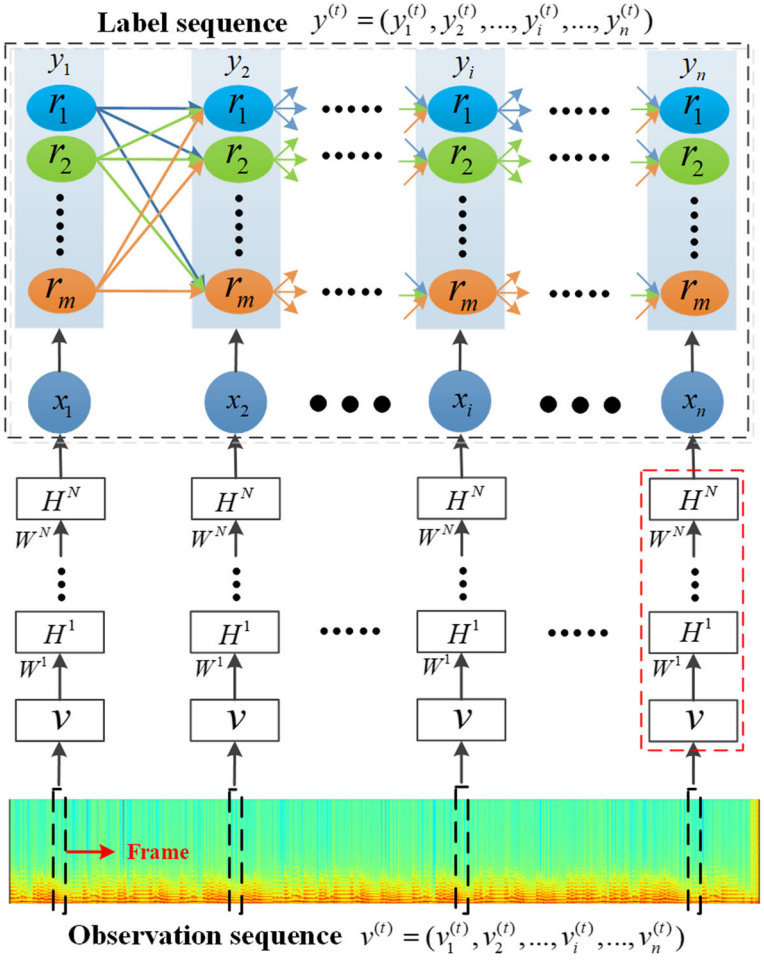


Fig. 3 Structure diagram of music modeling based on CRF-RBM

The probability distribution of the RBM fitting audio features is shown as follows.

$$P(v_i^{(t)}|\theta) = \frac{1}{Z(\theta)} \sum_h \exp(-E(v_i^{(t)}, h_i^{(t)}|\theta)) \tag{13}$$

$P(v_i^{(t)}|\theta)$ is also called likelihood function, where the partition function $Z(\theta)$ is a normalization factor which can be computed by the following equation.

$$Z(\theta) = \sum_{v,h} \exp(-E(v_i^{(t)}, h_i^{(t)}|\theta)) \tag{14}$$

$E(v_i^{(t)}, h_i^{(t)}|\theta)$ is the energy function of the RBM. The definition is shown as follows.

$$E(v_i^{(t)}, h_i^{(t)}|\theta) = -\frac{1}{2} \sum_{I=1}^{n_v} ((v_i^{(t)})_I - a_I)^2 - \sum_{J=1}^{n_h} b_J (h_i^{(t)})_J - \sum_{I=1}^{n_v} \sum_{J=1}^{n_h} (v_i^{(t)})_I w_{I,J}^1 (h_i^{(t)})_J \tag{15}$$

where $\theta = \{W^1, a, b\}$ represents the parameter set of RBM. $w^1_{I,J}$ represents the weight between visible unit $(v_i^{(t)})_I$ and hidden unit $(h_i^{(t)})_J$. a_I and b_J are their biases.

Stacking a predefined number of RBMs on top of each other build a Deep Belief Network (DBN) [12, 13], where the output from a lower-level RBM is the input to a higher-level RBM. We also employ DBN here to attain higher-level hierarchy abstract features for sequence labeling from the original audio data. RBM usually uses the Contrastive Divergence (CD-k) algorithm [11] to calculate the parameters and the backpropagation algorithm to fine-tune the parameters. For DBN, the CD-k algorithm and backpropagation algorithm are also used to compute the parameters of DBN layer by layer. The process of RBM(DBN) based label sequence calculation is as follows.

Step 1: The RBM(DBN) is created for all audio features $v_i^{(t)}$ of all regions of music sample $S = \{(v^{(t)})_{t=1}^{\|T_j\|}\}$, its likelihood function $P(v|\theta)$ is calculated by the following equation.

$$P(v|\theta) = \prod_{j=1}^{m} \prod_{t=1}^{\|T_j\|} \prod_{i=1}^{n} P(v_i^{(t)}|\theta) \tag{16}$$

The abstract feature $x_i^{(t)}$ is used as the observed value of the CRF observation sequence. Adding Softmax layer after the hidden layer makes the RBM having a discriminative mechanism. Each unit in Softmax layer represents a type of regional label.

Step 2: The state r_j of regional label $y_i^{(t)}$ at any time in label sequence $y^{(t)}$ is calculated through the following equation.

$$y_i^{(t)} = \arg \max_{1 \le j \le m} P(j) = \arg \max_{1 \le j \le m} [\frac{\exp(w_j^2 x_i^{(t)})}{\sum_{j=1}^{m} \exp(w_j^2 x_i^{(t)})}] \tag{17}$$

where $w_j^2 (w_j^2 \in W^2)$ represents the weights of all the hidden units connected to the j-th Softmax unit. Then the transfer function t_k is calculated after the adjacent moment's regional labels $y_{i-1}^{(t)}$ and $y_i^{(t)}$ are obtained through the above approach.

After the above step is executed, the label sequence $y^{(t)}$ of the t-th song in $S = \{(v^{(t)})_{t=1}^{\|T_j\|}\}$ can be obtained.

5 Experiments and analysis

5.1 Dataset and experimental setup

This paper used 297 folk songs from Northern Shaanxi(SX), 278 from Jiangsu (JS), and 262 from Hunan (HN) as the classification datasets. Folk songs from the three region exhibit various melody characteristics. Besides, all the folk songs were collected in 1970s and 1980s, which retain the original style of Chinese folk songs. In our previous work, these datasets were also employed in the regional classification [24, 27] and music analysis of Chinese folk songs [26] .

In addition, we also collected 412 Greek folk songs recordings of 5 different geographic regions from the on-line archives[2] and research programs[3] to evaluate all the methods mentioned in this paper. There are 89 folk songs from Crete Island, 75 from Aegean Islands, 86 from Epirus, 78 from Macedonia and 84 from Asia Minor.

WAV format of the audio file is used to extract audio features. Table 1 lists the extracted musical audio features by the software Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) [38]. Seven types of audio features are extracted regarding the time domain, frequency domain and cepstrum domain of the audio signal. All the features are 86 dimensions in total. In this experiment, we extracted the musical features by frame with the length of 1024, the shift of 896 and the sampling rate of 44.1kHz.

In this paper, we systematically compared the classification performance of other methods, including traditional classifiers likes Multilayer Perception (MP), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Logistic Regression (LR), Long Short-Term Memory neural network (LSTM) and other existing approaches for music regional classification. These classifiers also used the 86 dimensions features shown in the Table 1 as inputs. We did five-fold cross-validation for all of the regional experiments of folk songs in this paper.

Table 2 shows the details of parameters used in different baseline classifiers. Compared to other baseline classifiers, LSTM can effectively handle temporal related data. In this paper, we have examined two kinds of structure with different hidden layers for LSTM. All the hidden layers per structure have the same number of nodes, and the number of hidden nodes is 64/128/256. Moreover, the initail learning rate is 0.01 and then it decreases with the ratio as 0.5 when every 10 epoches are completed. To prevent overfitting, we added a dropout layer with rate of 0.2 after each LSTM layer and the training process ceased in advance if the loss on validation data did not decrease for 5 consecutive epoches.

5.2 Parameters determination for proposed methods

The parameters of GMM include mean μ_k, covariance Σ_k and weight π_k, the number of Gaussian components and the number of parameter iterations. The last two parameters are determined by the parameter ϵ in the EM algorithm, which is set to 10^{-5} in our experiments. The detail process of GMM parameters setting can refer to [24].

For RBM(DBN), we construct three kinds of structures with different hidden layers and learning rates. We search the optimal numbers of nodes at each hidden layer with step of 50 in the range of [50:350], and all the hidden layers share the same numbers neurons. We also search the optimal unsupervised learning rate from [0.001, 0.005, 0.01] and set the supervised learning rate as 0.5. Table 3 shows all the parameters used in RBM(DBN).

In order to obtain better classification performance, we use the testing error rate as the evaluation metrics for parameters tuning of RBM(DBN) to determinate the optimal numbers of nodes at each hidden layers and its corresponding optimal learning rate. Testing error rate refers to the error recognition rate for the audio features in the testing set after RBM(DBN) has fitted the audio features of training set. Figures 4, 5 and 6 show the testing error rate for the testing set of audio features based on RBM with one hidden layer, DBN1 with two

[2]Musical Folklore Archives Melpo Merlie: http://www.mla.gr/

[3]Thrace and Macedonia: http://epth.sfm.gr/

Table 1 Extracted musical audio features

Signal domain	Audio features	Dimensions
Time	Zero Crossings Rate (ZCR)	1
Frequency	Spectral Centroid (SC)	1
	Spectral Flux (SF)	1
	Spectral Rolloff Point (SRP)	1
	Chroma (Harmonic Pitch Class Profile, HPCP)	14
Cepstrum	Linear Prediction Cepstrum Coefficient (LPCC)	12
	Mel-Frequency Cepstrum Coefficient (MFCC)	13

hidden layers and DBN2 with three hidden layers, all of them are evaluated under three kinds of learning rates.

It can bee seen from Fig. 4 that the testing error rate (22.47%) reaches the lowest when the number of hidden units is equal to 300 and learning rate is 0.005, which means the mapping features from 86-dimensional space to 300 dimensional space helps to improve the recognition of audio features and the accuracy of folk songs' label in CRF. Therefore, the number of hidden units and learning rate in RBM is set to 300 and 0.005, respectively, in the experiment. Similarly, in Figs. 5 and 6, when the number of hidden units is 250 and 200, respectively, with the learning rate is 0.01, the testing error rate of DBN1 (26.25%) and DBN2 (26.71%) reaches minimum. Therefore, the learning rate of DBN1 and DBN2 is set to 0.01, and their hidden units' number is set to 250 and 200, respectively. To sum up, the optimal structure of RBM is set to 86-300-3 with the learning rate of 0.005, and the optimal structure of DBN1 and DBN2 is set to 86-250-250-3 and 86-200-200-200-3 with the learning rate of 0.01.

We also tuned parameters of the proposed methods for Greek folk songs dataset. For the sake of brevity, we don't show the specific process of parameters determination for Greek folk songs. The optimal structure for Greek folk songs data is 86-150-5 for RBM with learning rate is 0.001, 86-150-150-5 for DBN1 with learning rate is 0.005 and 86-100-100-100-5 for DBN2 with the learning rate is 0.005.

Table 2 Parameters configuration of baseline classifiers

Baseline classifiers	Details of parameters configuration
k-NN	$k = 5$
SVM	RBF kernel, search space $2^{[-10:10]}$ with a step of 1 for C
MP	Learning rate = 0.01, momentum = 0.1
LR	$L2$-regularized, tuning the regularization in [1 : 10] with a step of 1
LSTM	Initial learning rate: 0.01
	Batch size: 50
	Epoch: 50
	Structure: **LSTM1**(one hidden layer), **LSTM2**(two hidden layers): the number of hidden nodes is 64/128/256, all the hidden layers have the same number of nodes

The model name is highlighted to distinguish different model architecture

Table 3 Detail parameters setting used in RBM(DBN) tuning

Parameters name	Detail configuration
Learning rate	Unsupervised learning rate: 0.001/0.005/0.01; supervised learning rate: 0.5
Batch size	200
Momentum	0.1
Maximum iterations	500
Structures	**RBM** (one hidden layer), **DBN1** (two hidden layers) and **DBN2** (three hidden layers): search optimal number of nodes at each hidden layer in the range of [50:350] with step of 50. Besides, all the hidden layers have the same numbers of nodes

The model name is highlighted to distinguish different model architecture

5.3 Classification performance

Table 4 shows the confusion matrix of the proposed methods for Chinese folk songs' regional classification. According to Table 4, we can calculate the accuracy of our proposed methods which is listed in Table 5.

In Table 5, we also show the average training time and testing time of these models. The training process of CRF-GMM includes GMM training, sequence labeling for training data using GMM, CRF training for each region. The time complexity of CRF-GMM is about $O(N_{tr}D^2(\frac{S(S+1)}{2}I_g + 1 + I_c))$, where N_{tr} denotes the training data scale, D denotes the dimension of features, S denotes the Gaussian components, I_g denotes the maximum iterations of EM algorithm, and I_c denotes the maximum iterations for CRF training. Besides, the testing process of CRF-GMM contains sequence labeling for testing data using GMM and the final discriminant. The corresponding time complexity is about $O(N_{te}(D^2 + m^3))$, where N_{te} denotes the testing data scale, m denotes the number of classes. In this work, the Gaussian components usually do not exceed 15, D =86, and m =3, while the sum of N_{tr} and N_{te} is nearly one million when we divided our audio dataset by frame. Therefore, the size of the input data has a direct and huge impact on running time of our method.

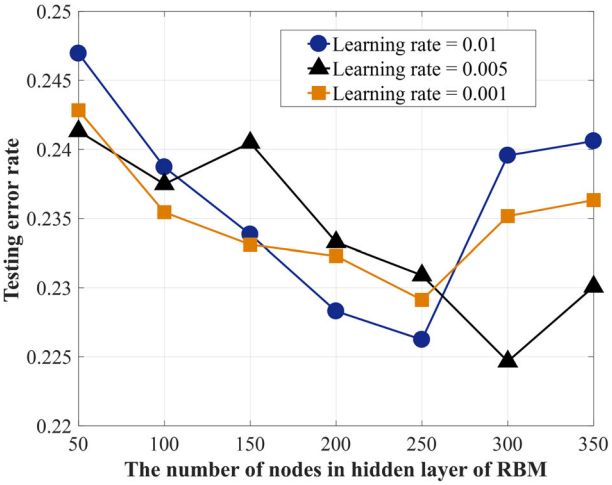


Fig. 4 Testing error rate of nodes' number in RBM under the different learning rates

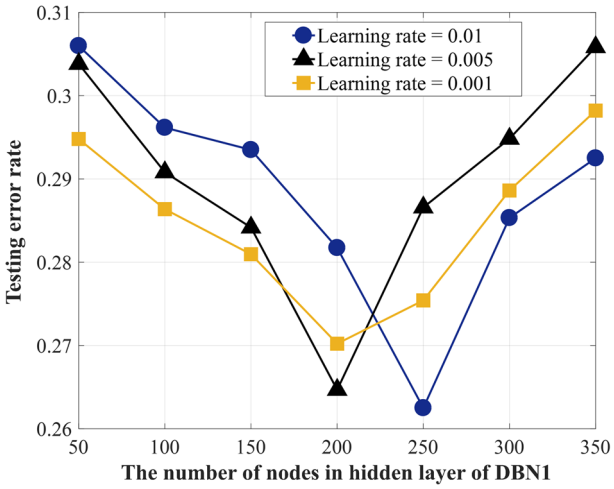


Fig. 5 Testing error rate of nodes' number in DBN1 under the different learning rates

For CRF-RBM(DBN), the training process and testing process are similar to CRF-GMM except for the RBM(DBN) training and sequence labeling. The running time of RBM(DBN) training and sequence labeling with RBM(DBN) is related to the model parameters scale. Specially, corresponding to the structure as 86-300-3, 86-250-250-3 and 86-200-200-200-3 for RBN, DBN1 and DBN2, respectively, there are totally 25800 (86×300), 84000 (86×250 + 250×250) and 97200 (86×200 + 200 × 200×2) weight parameters. The number of bias parameters of RBM, DBN1 and DBN2 are 300, 500 (250+250) and 600 (200+200+200).

In this paper, the machine with a 3.2GHz Intel(R) Core(TM) i7 and 16GB RAM was used to run our methods. Besides, nearly one million frames are processed by our proposed model, which lead to thousands of seconds to train the model and predict the label

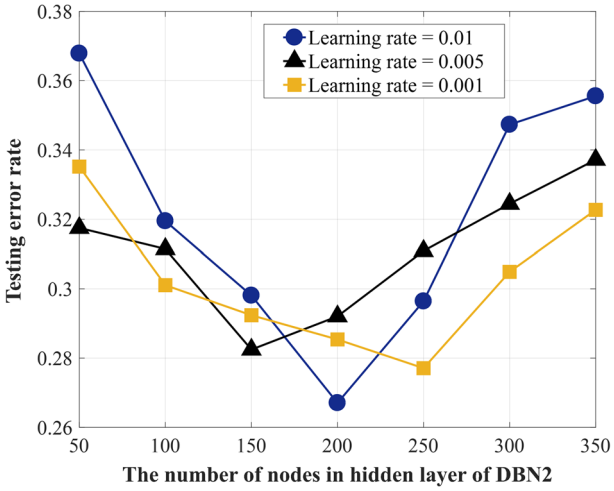


Fig. 6 Testing error rate of nodes' number in DBN2 under the different learning rates

Table 4 Confusion matrix of the average recognition results based on different label sequence calculation methods

Classifier	Regions	SX	JS	HN
CRF-GMM	SX	**52.4**	2.6	2.8
	JS	7	**39.4**	10.2
	HN	2.6	5.4	**45.0**
CRF-RBM (86-300-3)	SX	**52.2**	2.4	3.2
	JS	5.2	**43.0**	8.4
	HN	1.6	4.8	**46.6**
CRF-DBN1 (86-250-250-3)	SX	**50.6**	3.2	4.0
	JS	6.4	**41.6**	8.6
	HN	1.8	5.0	**46.2**
CRF-DBN2 (86-200-200-200-3)	SX	**51.0**	3.0	3.8
	JS	6.2	**41.2**	9.2
	HN	2.2	5.0	**45.8**

The highlighted data are used to emphasize the correct recognizing songs in each category

of new data. From Table 5, it is observed that the CRF-RBM achieves the best result with an accuracy of 84.71% and costs lowest running time. For CRF-GMM, it takes much time to determinate the number of Gaussian components and sequence labeling, and its performance is unsatisfactory. For CRF-RBM(DBN), the training time increases with the depth of model growing Besides, we can observe that the training time of CRF-DBN1 is very close to CRF-DBN2 because their parameters' scale is similar in size.

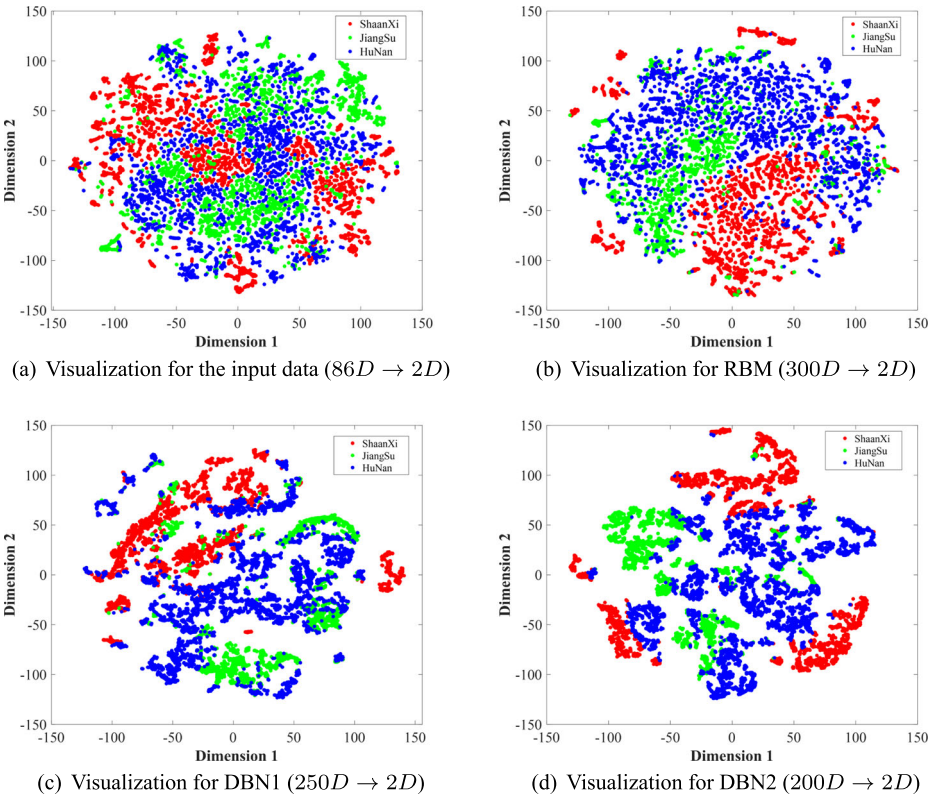
In theory, the representation abilities of deep learning methods derive mainly from the increase of the models' depth. However, CRF-DBN1 and CRF-DBN2 fail to further improve the performance although they have more hidden layers than CRF-RBM. As shown in Section 5.2, the testing error rate of RBM(22.47%) is lower than DBN1(26.25%) and DBN2(26.71%), this may directly lead to the worse classification performance for DBN1 and DBN2. To further explain this phenomenon, we visualize the original audio features and the abstract features from the last hidden layer of RBM, DBN1 and DBN2 in Fig. 7. The classical t-SNE tool [40, 41] is used to map the audio features into 2D space.

Figure 7a shows the visualization for the original audio features, the features from three regions are mixed together randomly with little distinction. After the processing of RBM, it can be observed that most of the abstract audio feature correspond to the same region cluster together in in Fig. 7b. Figure 7c and d shows the visualization for the abstract features after the mapping of DBN1 and DBN2. Compared to Fig. 7b, the points with the same color (from the same region) in Fig. 7c and d are split into several smaller clusters, although

Table 5 Performance comparison of music regional classification for Chinese folk songs

Classifier	Accuracy	Training time	Testing time
CRF-GMM	81.72%	11488.2s	522.8s
CRF-DBN2	82.44%	7961.2s	3.8s
CRF-DBN1	82.68%	7719.6s	3.6s
CRF-RBM	**84.71%**	4602.8s	3.2s

The highest recognition accuracy is highlighted in bold



(a) Visualization for the input data ($86D \rightarrow 2D$)

(b) Visualization for RBM ($300D \rightarrow 2D$)

(c) Visualization for DBN1 ($250D \rightarrow 2D$)

(d) Visualization for DBN2 ($200D \rightarrow 2D$)

Fig. 7 Visualization for input data, the output of last hidden layer data for RBM, DBN1 and DBN2

all of the points looks more focused. It means the spatial distinction of the three regional abstract features is reduced with the hidden layers increasing, which contributes to higher testing error rate of DBN1 and DBN2 in sequence labeling and worse final classification performance.

On the other hand, it should be noticed from Table 4 that folk songs from northern Shaanxi have the highest recogniton rate while Jiangsu folk songs the lowest recognition rate. Besides, Jiangsu folk songs tend to be misclassified into the other two classes, especially into Hunan folk songs no matter which classification method is used. These phenomena can be explained from some aspects of Chinese music theory as follows.

Firstly, in terms of scales, northern Shaanxi folk songs tend to use 7-tone scales (*gong*, *shang*, *jue*, *qingjue*, *zhi*, *yu*, *biangong*) which contains the 5-tone scales (*gong*, *shang*, *jue*, *zhi*, *yu*) commonly used in Hunan folk songs and Jiangsu folk songs [9].

Secondly, for melodic progression, the melodies of northern Shannxi folk songs tend to be more angular in shape. The tunes from northern Shaanxi often use the pitch intervals which are more than perfect fourth, and one of the famous intervallic emphasis is the "double perfect fourth" [6, 25]. The characteristic melodic progression of Hunan folk songs is the combination of major third and minor third [6]. Jiangsu folk songs tend to be more smooth and curved, compared to the folk songs from the former two regions. The frequent melodic progressions are the consecutive use of major second, minor third and perfect fourth

[6]. However, these melodic progressions are also very common both in folk songs from northern Shaanxi and Hunan.

Therefore, it can be concluded from the above analysis that northern Shaanxi folk songs are the most distinguishable among folk song from the three regions. In addition, scales similar to Hunan folk songs and the common melodic progressions result in high misclassified rate for Jiangsu folk songs.

5.4 Comparisons with other methods

To thoroughly evaluate the performance of our proposed methods, we compared them with the baseline classifiers listed in Table 2 and other approaches for music regional classification. All of the experiments were evaluated both on Chinese folk songs dataset and Greek folk songs dataset.

Table 6 shows the results of all classifiers with the optimal parameters. It can be seen from the results that CRF-RBM outperforms the baseline classifiers and other approaches for music regional classification. All the classifiers employing temporal structure obtain better recognition accuracy than those non-temporal structure. Both CRF-DBN1 and CRF-DBN2 have the similar accuracy on Chinese folk songs and Greek folk songs, and CRF-GMM model shows poor classification performance among all temporal structures, especially on the Greek folk songs dataset. For LSTM neural networks, LSTM1 with one hidden layer obatains higher recognition accuracy than LSTM2 with two hidden layers both on the two folk songs dataset. LSTM1 obtains the second highest recognition accuracy both on the two datasets. However, LSTM2 gets the third highest recognition accuracy on Greek folk songs and obtains a relatively low accuracy on Chinese folk songs. The two structures of LSTM behaves variously on different folk songs datasets.

In order to further validate whether CRF-RBM presents a significant improvement over approaches, we employed the non-parametric two-tailed Wilcoxon Signed-Rank test at a significance level of 5% (α=0.05). A pairwise comparison is made between CRF-RBM and

Table 6 Recognition accuracies of the proposed methods and other existing approaches

	Methods	Accuracy	
		Chinese folk songs	Greek folk songs
Non-temporal structure	MP	75.31%	53.34%
	SVM	77.44%	57.67%
	k-NN	76.13%	54.28%
	LR	78.54%	56.81%
	Liu Y et al. [29]	74.95%	53.96%
	Liu Y et al. [28]	79.28%	58.72%
Temporal structure	Li J et al. [27]	80.13%	63.38%
	CRF-GMM [24]	81.72%	61.56%
	LSTM2	82.25%	64.13%
	CRF-DBN2	82.44%	63.52%
	CRF-DBN1	82.68%	63.84%
	LSTM1	83.18%	66.03%
	CRF-RBM	**84.71%**	**67.38%**

The highest recognition accuracy is highlighted in bold

Table 7 Results of Wilcoxon Signed-Rank test at α=0.05, compared CRF-RBM with other temporal structures

	p-value	Null hypothesis
Li J et al. [27]	0.00988	Rejected
CRF-GMM [24]	0.00694	Rejected
LSTM2	0.02202	Rejected
CRF-DBN2	0.01242	Rejected
CRF-DBN1	0.01640	Rejected
LSTM1	0.03662	Rejected

other temporal structure for the two datasets based on recognition accuracy. The null hypothesis was set as H0: There is no significance difference occurred in performance between two approaches. The results shown in Table 7 provide strong evidence for CRF-RBM's efficiency.

6 Conclusions and future work

In this paper, we proposed the music regional classification methods based on CRF with three kinds of sequence labeling methods, which fully considering the temporal characteristics of musical audio features. In our experiments, GMM is used to fit music audio features and calculate the label sequence of songs in CRF. This method has achieved a good result on the music regional classification. But the accuracy of the label sequence is difficult to be improved when the number of Gaussian components in GMM is limited. To solve the problem, RBM(DBN) is used instead of GMM to calculate the label sequence of songs in CRF. All the proposed methods were evaluated on Chinese folk songs and Greek folk songs. The experimental results demonstrated that CRF-RBM outperforms the baseline classifiers and other existing approaches for music regional classification. Moreover, the results of the Wilcoxon Signed-Rank test show that CRF-RBM produced statistically significant better results in terms of predictive accuracy at the conventional 0.05 threshold.

In the future, we will try other well-designed temporal structures to further improve the classification performance of music region recognition for Chinese folk songs. In addition, we will develop new approaches to make a new attempt for the characteristic pattern mining of different regions.

References

1. Bassiou N, Kotropoulos C, Papazoglou-Chalikias A (2015) Greek folk music classification into two genres using lyrics and audio via canonical correlation analysis. In: 2015 9th international symposium on image and signal processing and analysis(ISPA), pp 238–243
2. Byrd RH, Hansen SL, Nocedal J, Singer Y (2014) A stochastic quasi-newton method for large-scale optimization. Siam Journal on Optimization 26(2):1008–1031
3. Chouzenoux E, Pesquet JC, Repetti A (2014) Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. ACM Trans Multimed Comput Commun Appl 162(1):107–132

4. Conklin D (2013) Multiple viewpoint systems for music classification. Journal of New Music Research 42(1):19–26
5. Corrêa DC, Rodrigues FA (2016) A survey on symbolic data-based music genre classification. Expert Syst Appl 60:190–210
6. Du YX (1993) The music dialect area and its divisions of Han folk songs (in Chinese). Chin Music 1:14–16
7. Fotiadou E, Bassiou N, Kotropoulos C (2016) Greek folk music classification using auditory cortical representations. In: 2016 24th European signal processing conference (EUSIPCO), pp 1133–1137
8. Fu ZY, Lu GJ, Ting KM, Zhang DS (2011) A survey of audio-based music classification and annotation. IEEE Trans Multimedia 13(1):303–319
9. Han KH (1989) Folk songs of the Han Chinese: characteristics and classifications. Asian Music 20(2):107–128
10. Hillewaere R, Manderick B, Conklin D (2009) Global feature versus event models for folk song classification. In: 2009 10th international society for music information retrieval conference, pp 729–734
11. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800
12. Hinton GE (2012) A practical guide to training restricted boltzmann machines. Momentum 9(1):599–619
13. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
14. Huang YF, Lin SM, Wu HY, Li YS (2014) Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. Data Knowl Eng 92:60–76
15. Kawase A (2017) Quantitative analysis of traditional folk songs from Shikoku district. In: 2017 international conference on culture and computing, pp 170–177
16. Kawase A, Tokosumi A (2010) Regional classification of traditional Japanese folk songs. Kansei Engineering International Journal 10(1):19–27
17. Kedyte V, Panteli M, Weyde T, Dixon S (2017) Geographical origin prediction of folk music recordings from the United Kingdom. In: 2017 18th international society for music information retrieval conference, pp 23–27
18. Kereliuk C, Sturm BL, Larsen J (2015) Deep learning and music adversaries. IEEE Trans Multimedia 17(11):2059–2071
19. Khoo S, Man Z, Cao Z (2012) Automatic Han Chinese folk song classification using the musical feature density map. In: 2012 6th international conference on signal processing and communication systems(ICSPCS), pp 1–9
20. Khoo S, Man Z, Cao Z, Zheng J (2013) German vs. Austrian folk song classification. In: 2013 8th IEEE conference on industrial electronics and applications(ICIEA), pp 131–136
21. Lafferty J, Mccallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: International conference on machine learning, pp 282–289
22. Larochelle H, Bengio Y (2008) Classification using discriminative restricted boltzmann machines. In: International conference on machine learning, pp 536–543
23. Larochelle H, Mandel M, Pascanu R, Bengio Y (2012) Learning algorithms for the classification restricted boltzmann machine. J Mach Learn Res 13(1):643–669
24. Li J, Ding J, Yang X (2017) The regional style classification of Chinese folk songs based on GMM-CRF model. In: 2017 9th international conference on computer and automation engineering, pp 66–72
25. Li J, Dong L, Ding J, Yang X (2015) Exploring the general melodic characteristics of XinTianYou folk songs. In: 2015 12th sound and music computing conference, pp 393–399
26. Li J, Wang Y, Yang X (2016) General characteristics analysis of Chinese folk songs based on layered stabilities detection(LSD) audio segmentation algorithm. In: 2016 42nd international computer music conference(ICMC), pp 16–20
27. Li J, Wang Y, Yang X (2017) Regional recognition of Chinese folk songs based on LSD audio segmentation algorithm. In: 2017 9th international conference on computer and automation engineering, pp 60–65
28. Liu Y, Wei L, Liu ZL, Wang P (2008) The feature selection of regional style classification of Chinese folk songs. Acta Electronica Sinica 36(S1):152–156
29. Liu Y, Xu JP, Wei L, Tian Y (2007) The study of the classification of Chinese folk songs by regional style. In: International conference on semantic computing(ICSC), pp 657–662
30. Mannepalli K, Sastry PN, Suman M (2015) MFCC-GMM Based accent recognition system for Telugu speech signals. Int J Speech Technol 19(1):87–93
31. Martel J, Nakashika T, Garcia C, Idrissi K (2013) A combination of hand-crafted and hierarchical high-level learnt feature extraction for music genre classification. In: International conference on artificial neural networks, pp 397–404

32. Miao J, Qiao JZ (1985) A study of similar color area divisions in Han folk songs(in Chinese). Journal of Central Conservatory of Music 1(1):26–33
33. Nanni L, Costa YMG, Lucio DR, Silla CN Jr, Brahnam S (2017) Combining visual and acoustic features for audio classification tasks. Pattern Recogn Lett 88:49–56
34. Panteli M, Benetos E, Dixon S (2016) Learning a feature space for similarity in world music. In: 2016 17th international society for music information retrieval conference, pp 538–544
35. Rajan R, Murthy HA (2017) Music genre classification by fusion of modified group delay and melodic features. In: 2017 Twenty-third national conference on communications, pp 1–6
36. Scaringella N, Zoia G, Mlynek D (2006) Automatic genre classification of music content: a survey. IEEE Signal Proc Mag 23(2):133–141
37. Song H, Sun K, Li B, Liu X (2011) HBS And HFS feature selection methods for Chinese folk music classification. In: IEEE international conference on transportation, mechanical, and electrical engineering, pp 2441–2444
38. Tzanetakis G, Cook P (2000) Marsyas: a framework for audio analysis. Organised Sound 4(3):169–175
39. Uzunbas MG, Chen C, Metaxas D (2016) An efficient conditional random field approach for automatic and interactive neuron segmentation. Med Image Anal 27:31–44
40. Van Der Maaten L, Hinton GE (2012) Visualizing non-metric similarities in multiple maps. Mach Learn 87(1):33–55
41. Van Der Maaten L (2014) Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 15(1):3221–3245
42. Wu MJ, Jang JSR (2015) Combining acoustic and multilevel visual features for music genre classification. ACM Trans Multimed Comput Commun Appl 12(1):1–17

Juan Li received her master's degree from Xi'an Jiaotong University in 2008, and now she is a Ph.D candidate. She is also the associate professor in the center of Music Education of Xi'an Jiaotong University. Her research interests include folk music analysis, multimedia data processing.

Jing Luo received his B.S. in computer science from Xi'an Jiaotong University in 2015, and now he is a Ph.D candidate in the department of computer science in Xi'an Jiaotong University. His research fields include music data processing and analysis, music composition.

Jianhang Ding receive his B.S. in computer science from Southwest Jiaotong University in 2014. He is currently a post-graduate in the department of computer science in Xi'an Jiaotong University. His research interests include music information retrieval and machine learning.

Xi Zhao is a lecturer in the department of computer science, Xi'an Jiaotong University. She received her Ph.D from Edinburgh University in 2014. Her research interests include multimedia data processing and computer graphics.

Xinyu Yang received his Bachelor, Master and Ph.D degrees from Xi'an Jiaotong University in 1995, 1997 and 2001. He is currently a professor in the department of computer science in Xi'an Jiaotong University. His research interests include multimedia modeling and mining, big data privacy protection.

Affiliations

Juan Li[1] · Jing Luo[2] · Jianhang Ding[2] · Xi Zhao[2] · Xinyu Yang[2]

Jing Luo
luojingl@stu.xjtu.edu.cn

Jianhang Ding
jh.ding@stu.xjtu.edu.cn

Xi Zhao
xi.zhao@mail.xjtu.edu.cn

1 Center of Music Education, Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, People's Republic of China

2 Department of Computer Science and Technology, Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, People's Republic of China