Analysis and Recognition of Cello Timbre Based on Deep Trust Network Model

Peng Sun*

Xi’an Conservatory of Music, China, 710061

*Corresponding author e-mail: Pengsun@163.com

Abstract. Voice color analysis and similarity calculation of music signals are the important research contents of computer music information retrieval system. In this paper, the deep trust network model is applied to the study of musical tone model. The 72-dimensional features of the cello tone are first extracted. Using the wrapper feature selection method, a 14-dimensional optimal feature subset that reflects the tone characteristics is selected, which greatly reduces the complexity of cello tone similarity calculation. In the set, SVR is used to classify and distinguish eight types of tone data, and a recognition accuracy of 62% was achieved, which is verified the feasibility of the tone model.

Keywords: Deep Trust Network Model, Feature Selection, SVR

1. Introduction
As the most important multimedia form, music has received widespread attention in the field of computer research. In recent years, with the rapid growth of digital musical tone data, audio information retrieval of musical tone signals has received widespread attention in the field of computer research. In the commercial application of music information retrieval, music software and search engines can be easily implemented. However, the retrieval of these information is essentially based on the existing text tag information of the music signal, such as song title, singer name, song style, etc.. The recommended music based on user behavior characteristics and the combination of the characteristic factors of the music itself are not enough[1]. The characteristic information of the music itself, such as tone information, melody and other music information, has yet to be tapped. In essence, it is still a traditional text retrieval. The text information corresponding to a music file can only be obtained by manual annotation. In the face of a large number of multimedia files, this method is not only laborious and time-consuming, but also almost impossible to complete. At the same time, labeling music files with text cannot represent the complete information of the music, especially
information that reflects the characteristics of the music signal itself, such as tone color, melody, pitch, pitch, etc. The loss of this information will seriously affect the accuracy of the music retrieval results, resulting in low retrieval efficiency. Audio information retrieval for musical tone signals includes multiple research directions: musical instrument recognition, singer recognition, humming retrieval, automatic beat detection, and sentiment analysis. One of the important research contents is the recognition of automatic instrument sounds, which involves the sounding principle of the musical tone and the perception mechanism of the human ear. It has important significance for the mining and application of the characteristic information contained in the musical tone signal. The problems of musical instrument recognition and speaker recognition in speech signal processing are similar. Both are based on the timbre characteristics of the audio signal to determine the sound source of the signal. However, the concept and perception of musical tone has always been vague and mysterious. In fact, its definition is not clearly defined in psychology, musicology, or computer science.\cite{2} The complexity of timbre is reflected in the following aspects: timbre is a subjective attribute of sound perception, not a pure physical attribute; timbre is a multi-dimensional attribute; no subjective scale for judging a timbre is currently found to be suitable; there is currently no unified musical tone signal Standard set for researchers to test the developed timbre calculation model.\cite{3,4,5}

2. Deep Trust SVR Algorithm

2.1. Construction and Implementation of Deep Trust Network Model
The deep trust network constructed in this paper is composed of a 1-layer Gaussian distribution function with explicit nodes of the RBM, a multi-layer hidden layer of the RBM, and a 1-layer SVR model. During the model pre-training, the calculation method for the joint distribution of data at the input layer and the conditional distribution of the hidden layer is:

\[
p (v,h;\theta) = \frac{\exp(-E(v,h;\theta))}{Z}, \tag{1}\]

Here \( Z = \int \sum_h \exp(-E(v,h;\theta)) \, dv. \)

The middle layer is the traditional RBM information conversion, that is, the (obvious layer) Bernoulli-(hidden layer) Bernoulli RBM data conversion. Its energy function is defined as:

\[
E(v,h;\theta) = - \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} v_i h_j - \sum_{i=1}^{I} b_i v_i - \sum_{j=1}^{J} a_j h_j, \tag{2}\]

Here \( \theta \) is the given model parameter \( \omega_{ij} \) represents the correlation weight between the explicit node \( v_i \) and the hidden node \( h_j \), \( b_i \) is the offset of the visible node, \( a_j \) is the offset of the hidden node, and \( I \) is the explicit layer The number of nodes in the structure, \( J \) is the number of nodes in the hidden layer structure.

2.2. SVR model
The algorithm extracts effective information by transforming the kernel function of the support vector to obtain the decision result. Figure 1 shows the model of SVR\footnote{6,7}. 

---

\[\text{doi:10.1088/1742-6596/1533/2/022015}\]
Let \( \{(x_i, y_i), i = 1, 2, ..., n\} \) be the prediction reference data sample set. There are \( n \) sample data in the sample set. Where \( x \) is the input vector and \( x_i \in \mathbb{R}^d \); \( y_i \) is the decision result and \( y_i \in \mathbb{R} \). The function expression of SVR is:

\[
\begin{align*}
f(x) &= \omega \cdot \varphi(x) + b, \\
&= \omega^T \varphi(x) + b.
\end{align*}
\]

In the formula: \( \omega \) represents the weight taken by different factors, and \( \varphi(x) \) represents the mapping function. Considering that the mapping data may still have high-dimensional spatial linear inseparability, and the high-dimensional fuzzy separability of this part of the data has little effect on the actual prediction, a relaxation variable is introduced to control the scale of fuzzy classification. The optimization of SVR can be expressed as:

\[
\begin{align*}
\begin{cases}
\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \\
\text{s.t.} & y_i - \omega \cdot \varphi(x) - b \geq \varepsilon + \xi_i, \\
& \omega \cdot \varphi(x) - b - y_i \leq \varepsilon + \xi_i^*, \\
& \xi_i \geq 0, \xi_i^* \geq 0,
\end{cases}
\end{align*}
\]

Here \( \xi \) and \( \xi^* \) are both relaxation variables. The optimization problem of this function can be solved by Lagrange function:

\[3\]
\[
\begin{align*}
\max & \sum_{i=1}^{n} y_i(a_i - a_i^*) - \frac{1}{2} \sum_{i,j=1}^{n} y_i(a_i - a_i^*)(a_j - a_j^*) - \varepsilon \sum_{i=1}^{n} (a_i + a_i^*), \\
\text{s.t.} & \sum_{i=1}^{n} (a_i - a_i^*) = 0, \\
& a_i \geq 0, a_i^* \leq C,
\end{align*}
\]

From equation (5) to solve equation (3), the SVR prediction model is:

\[
f(x) = \sum_{i=1}^{n} (a_i - a_i^*)K(x_i, x) + b, \quad (6)
\]

In the formula, \(K(x_i, x)\) is the kernel function of SVR. The kernel function with appropriate accuracy can be selected according to actual requirements.

### 2.3. Deep trust SVR model building

The deep trust SVR model built in this paper is different from the traditional SVR shallow model. The model consists of a deep learning model consisting of a 1-layer RBM with a Gaussian distribution node, a multi-layer hidden layer RBM, and a 1-layer SVR machine (the schematic diagram of the model is shown in Figure 2).

![Figure 2. Schematic diagram of deep belief support vector regression](image)

### 3. Experiment

#### 3.1. database

The tone samples required for this article are all from the Instrument Sampling Library of McGill University in Canada (McGill University Master Samples, MUMS).

#### 3.2. Voice characteristic analysis

Using Wrapper feature selection, 30-dimensional and 18-dimensional feature vectors are obtained. In
the third experiment, the improved GMM idea is used to obtain the average results of 10 experiments. The final optimal feature subset was 21 dimensions. In these three experiments, feature vectors with a probability of 100% are formed to form a 14-dimensional feature subset. We call this feature subset the smallest core feature subset vector that reflects the timbre. The specific weights are: the spectral attenuation cutoff frequency (4), MFCC 3rd (19), 4th (20), LPC 3rd (46), mean square error (8), LPC 10 Dimension (52), fifth dimension of MFCC (21), first dimension of Method of Moment (63), degree of spectral variation (9), first dimension of LPC (43), mean square error of low energy frame ratio (14), MFCC The second dimension (18), the third dimension of the Mean of variance of the Method of Moment (70), and the RMS (11).

Note: 4, 8, 9 and so on are the core feature subset numbers

3.3. Tone Similarity Experiment Results

After feature selection, the selected features in test set 2 are selected as the test set for testing. Note that in our data set, we split the test set into two. Test set 1 is used as the test set for training the model, and test set 2 is used as the test for the overall model. Figure 3 shows the tone recognition results of test set 2 after feature selection and multiple GMM models under different dimensional feature vectors. We can see from the figure that after feature selection, the timbre characteristics expressed by the Gaussian model in the 30-, 18-, and 21-dimensional feature subsets have achieved good timbre discrimination. Under the 14th dimension of the smallest feature subset, a result close to the best recognition rate is also achieved.

![Figure 3. Voice recognition rate of different feature vectors](image)

4. Conclusion

Starting from the deep learning model, this chapter takes the temporal integration of the timbre feature sequence as the input of the deep learning model to realize the instrument recognition. Deep learning models have greatly improved the recognition effect of wind instruments, while improving the overall
performance of instrument recognition and suppressing confusion between instruments and instrument families.

References

[1] Liangliang Liu, Shaowu Chen, Xiaofeng Zhu, Xing-Ming Zhao, Fang-Xiang Wu, Jianxin Wang. Deep convolutional neural network for accurate segmentation and quantification of white matter hyperintensities[J]. Neurocomputing, 2019.

[2] Yoonchang Han, Jaehun Kim, Kyogu Lee. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2017, Vol. 25 (1), pp. 208-221 ACM

[3] Yuan-Chao Wang, Ming-Tao Li, Zhi-Chen Pan, Jian-Hua Zheng. Pulsar candidate classification with deep convolutional neural networks[J]. Research in Astronomy and Astrophysics, 2019, 19(09): 119-128.

[4] Kin Wah Edward Lin, B. T. Balamurali, Enyan Koh, Simon Lui, Dorien Herremans. Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy[J]. Neural Computing and Applications, 2020, 32(4).

[5] Rajib Sarkar, Sombuddha Choudhury, Saikat Dutta, Aneek Roy, Sanjoy Kumar Saha. Recognition of emotion in music based on deep convolutional neural network[J]. Multimedia Tools and Applications, 2020, 79(1).

[6] Chang-Hao Zhu, Jie Zhang. Developing Soft Sensors for Polymer Melt Index in an Industrial Polymerization Process Using Deep Belief Networks[J]. International Journal of Automation and Computing, 2020, 17(1).

[7] Zhang Xin, Guo Shunsheng, Li Yibing, Jiang Li. Semi-supervised fault recognition based on Laplacian feature map and deep belief network [J / OL]. Journal of Mechanical Engineering: 1-13[2020-02-17]. http://kns.cnki.net/kcms/detail/11.2187.TH.20191224.1509.168.html.