SVM-Based Automatic Classification of Musical Instruments

Jing Liu, Lingyun Xie
Communication University of China, Beijing, 100024, China
small__123@hotmail.com

Abstract—**Instrumental music is often classified or retrieved in terms of instruments played in it. With a large database consists of Chinese traditional music and western classical music, this paper extracted several features to automatically classify Chinese and western instruments by SVM classifier, and analyzed the classification results.**

Keywords-instrumental music; classification; SVM

I. INTRODUCTION

As early as 4th century BC, the instrument taxonomy had appeared in China, which classifies instruments according to their material [1]. From then on, musicologists around the world have brought up various schemes. Classification schemes such as the popular Sachs-Hornbostel system and grouping by playing methods have all developed for a long time. The essence of taxonomy for all science is logicality and standard, but none of them is the standard recognized all over the world or meets all application needs, which makes it hard to find a single classification scheme to satisfy all users when it comes to programs based on taxonomy [2](e.g. Chinese Musical Instrument Museum). Thus, musical instrument classification now has become not only a subject for musicologists, but also an important research field for Music Information Retrieval (MIR).

Although automatic classification of western music has received some attention in MIR, see [3]-[5], as a part of world music, study in Chinese music is very rare. This paper mixed instrument family solos of Chinese and Western music to comprehensively discuss automatic classification of musical instruments. There're more than 13 instruments of Chinese and western separately included in this paper.

Generally, automatic classification of audio is defined as two processing phase [5]: feature extraction and classification. Support Vector Machine (SVM) is one of the key techniques in machine learning, which is employed in many projects and designs in MIR, e.g. in [3] and [6]. SVM has been proved to be an efficient automatic classifier. Discussion on the application of SVM in automatic classification of musical instruments will be presented in this paper. As to identifying instrument, the vital point is discrimination of instrumental timbre. The features related to timbre, comprising STFT features, MFCC, timbral features specified in MPEG-7, and so on, will also be discussed.

In this paper, we first introduce classification schemes for Chinese and Western instruments and build a database of 2177 music clips in different instrumental families. Then the classifier of SVM and all features extracted in experiment are described, followed by analysis on results of automatic classification. The conclusion of the experiment and future directions are drawn at last.

II. INSTRUMENTAL TAXONOMY

Since ancient time, various schemes of musical instrument classification have been developed in various cultures [1]. The scheme used in modern west, groups instruments into strings, percussion and wind, which later was expanded by Martin Agricol, who divided strings into plucked strings and bowed strings [1]. Nowadays, there're two main schemes: one is by playing methods, the other is by how the sound initially produced(e.g. Sachs-Hornbostel system classifying instruments into four groups: idiophones, membranophones, chordophones, and aerophones [7]).

For appreciating the art of playing, the first scheme seems more appropriate [8], which is widely used in Chinese music world, grouping musical instruments into four families—wind, bowed string, plucked string, and percussion, similar with western classification. Therefore, we employ this scheme in this paper (Fig.1).

III. SVM

To achieve automatic classification, we use the common course in data mining and machine learning, which includes training and testing. Training uses instances in training set to build a classifier model. Testing tests new instances (called test set) on that model and completes classification. For the same data set, the key of the whole classification is the choice of the classification algorithm.

Support Vector Machine (SVM) is now popular in data mining and machine learning. The basic principle behind in [9] is finding the optimal linear hyperplane which separates data from different categories with minimum error and maximum margin.

Sequential Minimal Optimization (SMO)[10] is a fast method to train SVM, performing quite well in terms of large data sets. The experiment provided in this paper employed this method to train SVM to achieve better automatic instrument classification. The algorithm used can be described as follows [10]:

Assume that a training set X is given as:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \qquad (1)$$

i.e. $X = \{x_i, y_i\}_{i=1}^{n}$ where $x_i \in R^d$ and $y_i \in (+1, -1)$. Training SVM yields to solve a quadratic programming problem as follows

978-0-7695-4077-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICTA.2010.64

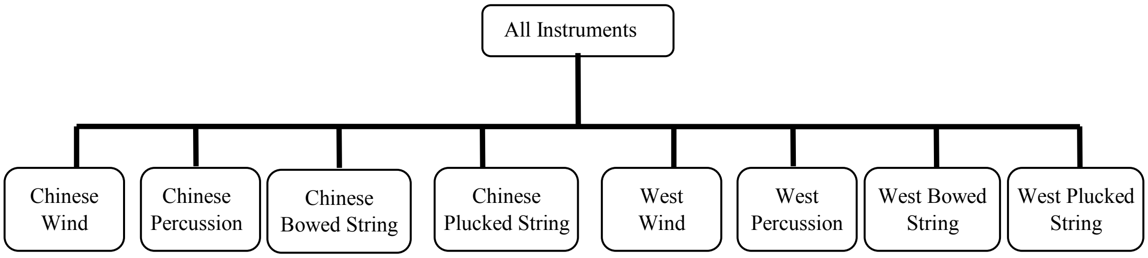

Figure 1. Structure of Musical Instrument Classification

$$\max_{\alpha_i}\left(-\tfrac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j K\left\langle x_i\cdot x_j\right\rangle+\sum_{i=1}^{l}\alpha_i\right) \tag{2}$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0, C \ge \alpha_i \ge 0, i = 1,2,\cdots l \tag{3}$$

where

$C > 0, \alpha_i = [\alpha_1, \alpha_2, \cdots, \alpha_l]^T, \alpha_i \ge 0, i = 1,2,\cdots,l$, are coeffients corresponding to x_i , x_i with nonzero α_i is called Support Vector (SV). The function K is called the Mercer Kernel which must satisfy the Mercer condition.

Let S be the index of SV, then the optimal hyperplane is

$$\sum_{i\in S}(\alpha_i y_i)K\left\langle x_i\cdot x_j\right\rangle + b = 0 \tag{4}$$

, and the optimal decision function is defined as

$$f(x) = sign\left(\sum_{i\in S}(\alpha_i y_i)K\left\langle x_i\cdot x_j\right\rangle + b\right) \tag{5}$$

where $x = [x_1, x_2, \cdots, x_l]$ is the input data, α_i and y_i are Lagrange multipliers. A new object x can be classified using (5). The vector is shown only in the way of inner product. There is a x_i Lagrangian multiplier α for each training point. When the maximum margin of the hyperplane is found, only the closed points to the hyperplane satisfy $\alpha > 0$. These points are called support vectors SV, the other points satisfy $\alpha = 0$.

IV. Feature Extraction

Features represented timbral texture include features based on the short time Fourier transform (STFT), MFCC [11], Spectral Crest Factor (SCF) and Spectral Flatness Measure (SFM) proposed in MPEG-7[12].

The following features are used in our experiment (the first 4 are based on STFT described in [11]):

A. Spectral Centroid

The spectral centroid is the center of gravity of the magnitude spectrum of the STFT

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \tag{6}$$

Where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency n . The centroid is a measure of spectral shape and higher centroid values show "brighter" sound.

B. Spectral Rolloff

The spectral rolloff is defined as the frequency R_t which covers 85% of the magnitude distribution below

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{7}$$

The rolloff also reveals an aspect of spectral shape.

C. Spectral Flux

The spectral flux is calculated as the squared difference between the normalized magnitudes of successive spectral frames

$$F_t = \sum_{n=1}^{N}\left(N_t[n] - N_{t-1}[n]\right)^2 \tag{8}$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t , and the previous frame $t-1$, respectively. The spectral flux demonstrates local spectral change.

D. Time Domain Zero Crossings

Time domain zero crossings counts the times the signal wave crossing value zero.

$$Z_t = \tfrac{1}{2}\sum_{n=1}^{N}\left|sign(x[n]) - sign(x[n-1])\right| \tag{9}$$

Where the $sign$ function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain signal for frame t .

E. Mel-Frequency Cepstral Coefficients(MFCC)

MFCC is also based on STFT, but it combines characteristics of hearing perception and the mechanism of generating speech to get cepstral coefficients, which is widely used in various occasion of audio signal processing. All signals need to transform from linear frequency domain into mel-frequency domain to get MFCC [6]

$$mel(f) = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \qquad (10)$$

where f is the linear frequency value.

The mel scale has 40 filter channels [6], which produce a measure of power of the signal, 12 linearly spaced outputs represent the spectral envelope and 27 log-spaced outputs of harmonics of the signal. Finally, a discrete cosine transform (DCT) converts these outputs to give the MFCCs. The first 13 coeffients were extracted as MFCCs in this paper.

F. Spectral Crest Factor(SCF)

SCF [12] is the ratio of the largest PSD coefficient and the mean PSD value in a frequency band. It represents the unevenness of the signal spectrum in specific frequency band.

G. Spectral Flatness Measure(SFM)

SFM describes the measure of spectral flatness in specific frequency band, representing the PSD deviation from a flat shape.

$$SFM = \frac{\left[\prod_{k=0}^{N-1} S_{xx}(k)\right]^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} S_{xx}(k)} \qquad (11)$$

Where $S_{xx}(k)$ denotes the PSD coefficients within a frequency band [12].

Extracting SCF and SFM, We divided signal into 24 frequency band by 1/4 octave from 250Hz to 16KHz, producing 24 attributes of SCF and SFM, respectively.

V. Experiment And Results

A. Dataset

We chose 170 pieces of Chinese traditional instrumental music and 160 pieces of western instrumental music from published CDs. All pieces are solo music of a certain instrument family. The sampling rate is 44.1 KHz. Since the extracted features is irrelevant to the number of channel, we resampled the music at 22.05KHz into 16-bit mono .wav file and divided them by 30 seconds for faster computation, Finally we got 2177 clips of music. Specifically, Chinese instruments cover more than 13 instruments of bamboo flute (Dizi), Chinese vertical bamboo flute (Xiao), Chinese bass drum (Dagu), gong (Luo), erheen (Erhu), Ching Hu (Jinghu), Chinese lute (Pipa), Chinese zither (Guzheng), and so on; western instruments also involve over 13 of flute, clarinet, marimba, vibraphone, violin, cello, guitar, lute and so on.

B. Classification Results and Analysis

According to taxonomy described in section II, we group Chinese and western musical instruments into 8 families, which are Chinese wind, Chinese percussion, Chinese bowed string, Chinese plucked string, western wind, western percussion, western bowed string, and western plucked string.

Feature extraction in this experiment implemented in a non-overlapped analysis window of 512 samples at 22050Hz, calculated the means and standard deviations within accumulator, and then computed the means and standard deviations of all accumulators in a clip (30s) as the final feature. Thus, the dimension of the features used in this experiment is four times of the origin. We got 16 attributes of STFT, 52 of MFCC, 96 of SCF and also 96 of SFM.

TABLE I. Average Accuracy and Standard Deviation of Accuracy in Training Set

	MFCC	STFT	SCF	SFM
AVG (%)	95.44	62.29	91.25	92.38
STD	0.0552	0.1687	0.0554	0.0560

TABLE II. Average Accuracy and Standard Deviation of Accuracy in Test Set

	MFCC	STFT	SCF	SFM
AVG (%)	87.23	60.75	74.68	78.30
STD	0.1060	0.1964	0.1549	0.1321

After feature extraction, we applied SVM to classify those extracted features upon both training set and test set. The classification results can be expressed as the mean value and standard deviation of all 8 instrumental family accuracies, showed in TABLE I and TABLE II. The results has illustrated a fact that among those four feature, MFCC performs best and STFT is the worst in our experiment, regardless the number of attributes. The average accuracy of training and test set using MFCC is as high as 91.34%, and the standard deviation of accuracy is the lowest among all used features.

Although there're 96 attributes of SCF and SFM, 44 attributes more than that of MFCC, in terms of the results, the classification performances of them are not as good as MFCC. Their average accuracies of test set is much lower than that of training set (dropping more than 10 points in percentage), which demonstrates the optimal hyperplane formed by training set can't guarantee a good performance in test set. With the least attributes, the features based on STFT didn't show the competence in average accuracy, but in classification of test set, we notice that using these features, the accuracy for western plucked string (h) is a bit higher than using MFCC (see Fig.2), which provide possibilities for combination of STFT-based features and MFCC to improve automatic classification performance.

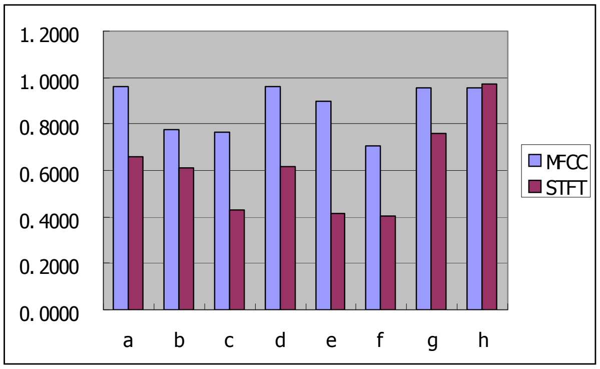

Figure 2. Classification Accuracy of Test Set on MFCC and STFT (a-h stands for the eight families in order mentioned in this section)

Confusion matrix is the raw result from automatic recognition or classification, often presented in percentage. The values in diagonal are the accuracy of classification. Through further analysis on confusion matrix of SVM-based instrument classification using the best performer MFCC, see Table III and IV, some interesting phenomena are worth of attention. The most obvious one is that most errors (red in row b and f) happen between percussion of Chinese and western, verifying the similarity of these two families, and the similar situation appeared in the bowed string family. As to the other two families, the situation is quite different. The wind instruments of Chinese and western share little common ground. Both have relatively high accuracy, and the majority of errors happened within the same cultural instrument families, i.e. Chinese instruments or western instruments. Strangely enough, western plucked string instruments are prone to be classified as Chinese percussion, and Chinese plucked string instruments are tend to be recognized as wind instruments.

VI. Conclusion

On the dataset of mixed Chinese and western instrumental music, this paper discussed automatic classification performance of four features based on SVM, a popular classifier in MIR. In accordance with the instrumental taxonomy proposed, all instrumental music was automatically classified into four families of wind, percussion, bowed string, and plucked string. The result of automatic classification proved that MFCC is the most competitive candidate, compared with the other three, in automatic instrument classification using SVM. Choosing the right feature for classifier can not only improve the accuracy, but also reduce attributes. After further analysis, we found similarity of MFCC between Chinese and western percussion instruments, not good news for automatic instrument classification.

Regarding the dimension of extracted features and the similarity between instrument families, a promising direction of future work is combining MFCC with other features or using some feature selection algorithms to reduce attributes and eliminate similarity to get maximum marginal SVM. Less attributes and higher accuracy are the basis of real-time automatic classification systems. And the future research on SVM may also explore the possibility of real-time systems to provide useful tools for application of MIR.

TABLE III. Confusion Matrix for Test on Training Set in Percentage

a	b	c	d	e	f	g	h	**Classified as**
97.47	0.00	0.00	1.90	0.00	0.00	0.00	0.63	a
0.67	**94.00**	0.00	0.00	0.00	5.33	0.00	0.00	b
2.58	1.94	**90.97**	0.00	0.00	1.29	3.23	0.00	c
0.00	0.62	0.00	**98.14**	1.24	0.00	0.00	0.00	d
0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	e
0.00	12.80	1.22	0.00	0.00	**84.15**	0.00	1.83	f
0.00	0.00	0.00	0.00	0.00	0.59	**99.41**	0.00	g
0.00	0.57	0.00	0.00	0.00	0.00	0.00	**99.43**	h

TABLE IV. Confusion Matrix for Test on Test Set in Percentage

a	b	c	d	e	f	g	h	Classified as
96.30	0.00	0.00	0.93	0.00	0.00	2.78	0.00	a
0.00	**77.39**	0.00	1.74	0.00	19.13	0.87	0.87	b
5.61	2.80	**76.64**	0.93	0.00	0.93	13.08	0.00	c
1.92	0.00	0.00	**96.15**	0.96	0.00	0.00	0.96	d
1.83	0.00	0.92	0.00	**89.91**	4.59	1.83	0.92	e
0.00	20.18	0.00	0.92	0.00	**70.64**	0.00	8.26	f
0.00	0.00	3.70	0.93	0.00	0.00	**95.37**	0.00	g
0.00	1.82	0.00	0.00	0.91	0.00	1.82	**95.45**	h

REFERENCES

[1] Wikipedia contributors, Musical Instrument Classification [OnLine]. Available:http://en.wikipedia.org/w/index.php?title=Musical_instrument_classification&oldid=325554329.

[2] Yan Wei, "Symposium Summerization of Chinese Traditional Musical Instrument Classification," Chinese Music, vol. 1, 2007, pp. 232-233.

[3] Slim Essid, Gael Richard and Bertrand David, "Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, Jan. 2006, pp. 68-80.

[4] Alicya A.Wieczorkowska, Zbiniew W.Ras, Xin Zhang and Rory Lewis, "Multi-way Hierarchic Classification of Musical Instrument Sound,"Proc. International Conference on Multimedia and Ubiquitous Engineering (MUE' 07), 2007, pp. 897-902, doi:10.1109/MUE.2007.159.

[5] Martin F. McKinney and Jeroen Breebaart, "Features for Audio and Music Classification,"Proc. 4th International Conference on Music Information Retrieval, 2003, pp. 151-158.

[6] Jeremiah D.Deng, Christian Simmermacher and Stephen Cranefield, "A Study on Feature Analysis for Musical Instrument Classification," IEEE Transactions on Systems, Man, and Cybernetics, Part B, vol. 38, Apr. 2008, pp. 429-438, doi: 10.1109/TSMCB.2007.913394.

[7] Yang Min-kang, "Comparing Research On Multi Relationship Among the Chinese-West Music Instruments and Music Instruments Classification,"Huang Zhong(Journal of Wuhan Conservatory of Music, China), (3), 2006, pp. 96-104.

[8] Yuan Jing-fang, Concise Guide to Chinese Traditional Music. Shanghai, China: Shanghai Conservatory of Music Press, 2006

[9] Perfecto Herrera-Boyer, Xavier Amatriain, Eloi Batlle and Xavier Serra, "Towards Instrument Segmentation for Music Content Description:a Critical Review of Instrument Classification Techniques,"Proc. ISMIR, 2000.

[10] Jair Cervantes, Xiaoou Li and Wen Yu, "SVM Classification for Large Data Sets by Considering Models of Classes Distribution,"Proc. Sixth Mexican International Conference on Artificial Intelligence-Special Session (MICAI 2007), 2007, pp.51-60, doi: 10.1109/MICAI.2007.27.

[11] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals," IEEE Transactions on Speech and Audio Processing, vol. 10, July 2002, pp. 293-302.

[12] Jürgen Herre, Eric Allamanche and Oliver Hellmuth, "Robust Matching of Audio Signals Using Spectral Flatness Features," Proc. IEEE Workshop on The Applications of Signal Processing to Audio and Acoustics, 2001, pp. 123-130.