第32卷 第12期 计 算 机 工 程 2006年6月
Vol.32 № 12 Computer Engineering June 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)12—0182—02 文献标识码：A 中图分类号：TP18

基于内容的戏曲分类与分析

张一彬，周 杰，边肇祺
(清华大学自动化系，北京 100084)

摘 要：中国传统戏曲是世界艺术园地中的一支奇葩。该文使用音频分析技术和模式识别技术相结合的方法对8种典型的中国传统戏曲（京剧、昆剧、评剧、豫剧、越剧、河北梆子、黄梅戏和晋剧）进行了自动分类和相似性分析研究。实验结果表明，在一个包含了680个唱段的测试数据库上，该方法可以达到82.4%的平均分类正确率。研究结果还表明在这8种传统戏曲中，昆剧和评剧最为相似，豫剧和越剧之间差别最大。

关键词：戏曲；音频分析；模式识别

Content-based Classification and Analysis on Chinese Traditional Opera

ZHANG Yibin, ZHOU Jie, BIAN Zhaoqi
(Department of Automation, Tsinghua University, Beijing 100084)

【**Abstract**】Among all kinds of music styles, Chinese traditional opera is very familiar in China. This paper presents a study on content-based classification and analysis among eight traditional opera styles (including Beijing opera, Kunqu opera, Pingju opera, Henan opera, Shaoxing opera, Hebei opera, Huangmei opera and Jin opera), using audio analysis techniques together with pattern recognition techniques. A comparative evaluation between different classifiers is carried out on a test database of 680 audio segments, the result show that the quadratic classifier (QC) works best, and its average classification accuracy can achieve 82.4%. By analyzing the features distribution and the classification results of these traditional opera styles, some interesting conclusions are also reported.

【**Key words**】Chinese traditional opera; Audio analysis; Pattern recognition

1 概述

戏曲是中国传统文化中的瑰宝，是世界艺术园地中的一支奇葩。与其它艺术形式相比，中国传统戏曲显得丰富多彩，比较有影响力的剧种就有数十种之多。而各个剧种在其形成、发展的过程中又彼此影响、互相借鉴，因而相互之间有着千丝万缕的内在联系。比如我国的著名传统艺术京剧，它是属于皮簧声腔系统的剧种，约于公元1821年－1850年间在北京形成。它的曲调主要由二簧（徽剧）和西皮（楚调，源自西秦腔）结合而成。二簧与西皮调原本流行于安徽、湖北等地，在乾隆、嘉庆年间随着徽班进入北京，并逐渐吸取了昆曲、京腔、秦腔（山陕梆子与四川梆子）等剧种的营养，形成了同北京当地语言和风俗习惯相结合的新剧种——京剧。所以说，中国传统戏曲的剧种虽多，但彼此之间在唱腔、配器、伴奏曲调等方面却又有着或多或少的相似之处。这使得中国的传统戏曲艺术可以在整体上独立于世界上的任何一种其它艺术形式，但是这也使得我们很难对其进行基于内容的自动分类和分析。无论是与处理包含着各种相去甚远的音频类别（如语音、音乐、环境声音和静音等）的数据相比，还是同处理一般的音乐类别（如摇滚乐、钢琴曲、流行歌曲、交响乐等）的数据相比，对中国传统戏曲内部的不同剧种进行基于内容的自动分类和分析无疑都具有更大的挑战性。到目前为止，还没有任何有关这方面研究工作的文献报道。

在相关工作中，Zhang 等人提出了一种基于内容的音频分类、分割方法[1]，他们利用4个短时特征对由下面7个类别组成的音频信号进行自动分类和分割，这7个类别包括语音、纯音乐、歌曲、环境声音、带有音乐背景的语音、带有音乐背景的环境声音和静音。在文献[2]中，作者提出了一种将音频分割为语音、音乐、环境声音和静音的方法。他们首先将音频信号被分为语音信号和非语音信号两类，然后进一步将非语音信号分为音乐、环境声音和静音。在过去的工作中，我们在较大的数据库下对音频数据流的分类和分割问题也进行了较为详尽的研究[3]，所涉及的音频数据类别包括钢琴曲、交响乐、京剧、流行歌曲和语音。T. Lambrou 等人利用时域和小波变换域中的一些统计特征对爵士乐、摇滚乐和钢琴曲这3类音乐数据做了自动分类[4]，但是他们的测试数据库中仅仅包含了12首音乐。

本文选择了中国传统戏曲中的8个比较有影响力的剧种作为我们的研究对象，这8个剧种分别为：京剧，昆剧，评剧，豫剧，越剧，河北梆子，黄梅戏和晋剧。它们基本含盖了东、西、南、北、中各地的典型唱腔。我们总共收集了1 360个唱段（每个剧种170段）作为样本数据，并且采用音频分析技术和模式识别技术相结合的方法对这8个典型剧种进行

基金项目：国家自然科学基金资助项目(60205002，60332010)；北京市自然科学基金资助项目(4042020)
作者简介：张一彬(1974—)，男，博士生，主研方向：模式识别，基于内容的音频及音乐分析，信息挖掘等；周 杰，博士、教授、博导；边肇祺，资深教授、博导
收稿日期：2005-06-27 **E-mail**：zyb00@mails.tsinghua.edu.cn

了自动分类和相似性分析研究。在实验中比较了 6 种常用的分类器，发现基于正态分布的二次分类器比较适合用于戏曲分类。在这 8 个剧种包含 680 个唱段的测试集上，它所取得的平均分类正确率为 82.4%。通过对训练样本的分析，发现在这 8 个剧种当中，昆剧和评剧最为相似，豫剧和越剧之间差别最大。

2 特征提取和分类器选择

本文中每个戏曲唱段被转换成采样率为 11 025/s、量化位数为 8 位的混和单声道 WAV 文件。从每个唱段中提取一系列音频特征并组成特征向量。实验中求取音频特征时所使用的“帧”的长度为 512 个采样点，约 46ms；相邻帧之间有 112 个采样点的重叠区域，约 10ms。通过借鉴音频分析与识别领域中的前人成果以及作者的工作经验，在本文中所使用的基本音频特征为短时能量（SE）、过零率（ZCR）、Mel 频率（MF）、和谐度（HD）[3]、低短时能量值比率（LSTER）、高过零率比率（HZCRR）、谱通量（SF）[2]、谱矩（SC）、带宽（BW）和频谱滚动频率（SRF）[5]。

有了基本的音频特征后，将从每个唱段样本中提取出一个 17 维的特征向量，并用这个特征向量来代表这个音频片段。这个 17 维的特征向量中包括短时能量序列、过零率序列、Mel 频率序列、谱矩序列、带宽序列、谱通量序列以及频谱滚动频率序列的均值和标准差再加上低短时能量值比率、高过零率比率和和谐度。值得注意的是在计算 Mel 频率序列的均值和标准差时将不考虑 Mel 频率值置为 0 的那些帧。

在实验中，比较了 6 种常见的分类器。这几个分类器分别为：帕森分类器（PC），K 近邻法分类器（k-NNC），最小 Mahalanobis 距离分类器（MMDC），支持向量机（SVM），二次分类器（QC）和 BP 神经网络分类器（BPNNC）[3]。其中，K 近邻法分类器的近邻数为 3；支持向量机采用的核函数为 2 阶多项式核函数。相对其它核函数而言，它对于戏曲分类问题更加有效；BP 神经网络分类器是一个采用变步长 BP 训练算法的 3 层前馈结构神经网络，其输入层节点数等于样本特征向量的维数，输出层节点数等于类别总数，中间层包含 11 个节点，节点函数采用 Sigmoid 函数；二次分类器基于正态分布假设，其决策规则为

$$\begin{cases} g_i(x) = x^T W_i x + w_i^T x + w_{i0}, \\ if\ \ g_i(x) > 0,\ then\ x \in \omega_i \end{cases} \tag{1}$$

其中，x 表示未知样本的 d 维特征向量，W_i 表示一个 $d \times d$ 维的实对称矩阵，W_i 和 W_{i0} 为两个 d 维向量。

3 戏曲分类和相似性分析实验结果

采用基于样本训练的方法实现基于内容的戏曲自动分类。首先收集了足够多的样本数据，每个剧种 170 个唱段，每个唱段的长度均为 1min。然后将这 1 360 个属于各类戏曲的唱段平均分到训练集和测试集中。对于训练集中的每个样本，从中提取出一个 17 维的特征向量用于训练分类器。有了所需的分类器后，就可以对来自测试集的样本特征向量进行分类。基于内容的戏曲自动分类算法流程见图 1。

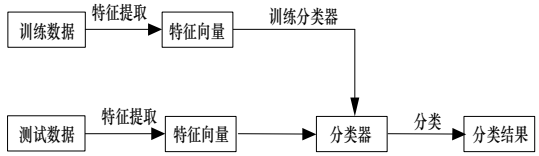

图 1 基于内容的戏曲自动分类算法流程

实验中比较了在第 2 节中所介绍的分类器，各个分类器在测试集上得到的平均分类正确率见表 1。

表 1 不同分类器在测试集上所得到的平均分类正确率

PC	K-NNC	MMDC	SVM	QC	BPNNC
54.2%	46.1%	69.1%	52.0%	82.4%	77.6%

从表 1 可以看出在区分不同戏曲剧种的分类问题中，QC 的分类效果要优于其它分类器，这一点与文献[3]中的结论有所不同。这说明分类器没有绝对的好坏之分，它的性能在很大程度上取决于所处理的研究对象的分布特性。QC 在测试集上所得到的平均分类正确率为 82.4%，其分类结果见表 2。

表 2 QC 在测试集上的详细分类结果

	1	2	3	4	5	6	7	8
1	85.9%	0	4.7%	3.5%	0	4.7%	1.2%	0
2	0	**76.7%**	**10.5%**	1.2%	7.0%	0	3.5%	1.2%
3	3.5%	**12.8%**	**69.8%**	8.1%	1.2%	2.3%	0	2.3%
4	0	4.7%	11.6%	**80.2%**	**1.2%**	0	1.2%	1.2%
5	0	1.2%	2.4%	**0**	**92.9%**	1.2%	2.4%	0
6	1.2%	0	3.5%	3.5%	0	91.9%	0	0
7	5.8%	5.8%	1.2%	1.2%	5.8%	0	77.9%	2.3%
8	0	1.2%	3.5%	1.2%	0	4.7%	5.8%	83.7%

(注：其中第 i 行第 j 列上的数字表示属于第 i 类的数据被分类器分到第 j 类中的百分比，表中 1~8 分别用来表示京剧、昆剧、评剧、豫剧、越剧、河北梆子、黄梅戏和晋剧)。

有了属于不同剧种的大量数据，就可以在此基础上对不同剧种之间的相似性做一些简单的分析。对于任何一段样本数据，都用一个 17 维的特征向量来表示它，这段样本数据就被表示为高维特征空间中的一个点。属于同一剧种的众多样本点将在这个高维特征空间中形成一个分布。这里的特征向量与戏曲分类实验中所使用的特征向量完全相同。如果两个剧种之间的相似性越高，那么由它们各自的样本点在特征空间中所形成的分布就应该拥有更多的重叠区域，或者距离更近。为了简单，可以只用各个剧种特征分布的中心点来表示这个分布。显然，如果两个剧种之间的相似性越高的话，那么它们各自特征分布的中心点之间的距离就会越近，可以将这种距离的大小作为两个剧种之间相似性的一种度量。为此，定义 i 类剧种和 j 类剧种之间的类间离散度矩阵为

$$\sum_{ij} = (m_i - m_j) * (m_i - m_j)^T \tag{2}$$

其中，m_i 为第 i 类剧种数据集的均值特征向量，m_j 为第 j 类剧种数据集的均值特征向量。则，i 类剧种和 j 类剧种之间的相似性度量可以被定义为

$$S_{ij} = \left\| \sum_{ij} \right\| \tag{3}$$

S_{ij} 越小，则表明 i 类剧种和 j 类剧种之间的相似度越高。通过对 1 360 个样本唱段进行分析，可以得到这 8 类剧种之间的相似性关系，见表 3。

表 3 不同剧种之间的相似性关系

S_{ij}	2	3	4	5	6	7	8
1	6.0	5.2	23.7	26.5	10.8	22.6	21.5
2	--	**2.3**	18.6	18.1	17.3	6.9	16.7
3	--	--	25.8	11.8	25.0	6.3	19.1
4	--	--	--	**71.9**	8.1	9.4	5.7
5	--	--	--	--	69.7	30.3	59.5
6	--	--	--	--	--	10.5	5.6
7	--	--	--	--	--	--	5.5

(注：其中第 i 行第 j 列上的数字表示第 i 类剧种与第 j 类剧种之间的相似度数值，数值越小表明两个剧种之间的相似度越大，表中 1~8 分别用来表示京剧、昆剧、评剧、豫剧、越剧、河北梆子、黄梅戏和晋剧)。

从表 3 中不难看出在这 8 个剧种当中，昆剧和评剧最为相似、而豫剧和越剧之间的差别最大。这个结论与我们前面的戏曲分类实验结果是相符合的。由于昆剧和评剧比较相似，因此这两个剧种之间的样本错分情况就比较严重。其中，有

(下转第 186 页)

表 2 不同阈值在不同训练样本集下的误识结果(%)

训练样本 \ 阈值	100	120	150	170	200
4	6.9	8.1	13.7	18.3	25.7
5	4.4	7.6	11.4	16.1	21.4
6	4.5	7.2	9.5	14.9	20.3
7	3.2	6.3	9.2	13.9	18.1

与实验 1 的结果不同，表 2 中从第 2 列起的每一列随着训练样本个数的增加是呈下降的趋势。这从反面证明了训练样本个数越多，所构成的覆盖区域与其他说话人的界限越清晰，越不容易将其他说话人的声音误认为该说话人。从第 2 行起的每一行结果却逐渐增大，说明阈值越大，把更多的有效测试样本包含进来的同时，把更多的无关测试样本包含进来了。

高阈值能带来高识别率，但同时也因为高阈值将许多不是该说话人的无关的测试样本误识为该说话人，因此阈值设置的好坏还要与误识率结合起来判断。不同环境对识别率和误识率的要求也不尽相同。在本实验中，当阈值为 100 时，是一个比较合适的选择，在保证高识别率的前提下，有较低的误识率。如果阈值的跨度设置得更精细，还可以找到更好的阈值。

综合实验 1 与实验 2 的结果，当阈值为 100 时，识别率较高，误识率较低，因此把阈值为 100 时的结果作为本文提出的新方法用于说话人识别的结果，同时与其他传统的匹配模型作比较。实验 3 的结果如表 3 所示。

从表 3 中可以看出，在这几种方法中，新方法的识别率明显高于其他方法。这证明，从“认识”的角度来学习和识别事物的方法与基于“划分”和“区别”的方法相比，在相同训练样本的情况下，能更好地对识别对象进行分析，获取信息。

表 3 不同识别方法的识别率比较(%)

训练样本 \ 识别方法	GMM	VQ	HMM	仿生模式识别
4	72.1	75.8	81.2	93.6
5	76.5	76.1	83.4	93.6
6	77.2	79.7	83.7	94.9
7	78.3	70.2	83.8	94.7

5 总结

无论是 GMM、VQ 还是 HMM 方法，在对新类别学习的时候都是封闭的，即通过打破原有类别的区域界定来形成加入新模式后的新的划分。而基于仿生模式识别的高阶神经网络方法只要通过少量样本的学习就可以在特征空间中构建出该类样本的覆盖区域范围，其优势一目了然。

同时，将神经网络应用于说话人识别中的各个细节仍需进一步地探讨。不同的神经网络性能有所差别，怎样的结构能更好地体现仿生模式识别理论也是值得继续研究的。

参考文献

1 Zhang Xinyi. Optimum Vector Quantization Codebook Design for Speaker Recognition[C]. Proc. of the ICSP, 2004: 1397-1402.

2 Abu El-Yazeed M F. On the Determination of Optimal Model Order for GMM-based Text-independent Speaker Identification[J]. Eurasip Journal on Applied Signal Processing, 2003, 2004(8): 1078-1087.

3 Astrid H, Andrew M. Recent Advances in the Multi-stream HMM/ANN Hybrid Approach to Noise Robust ASR[J]. Computer Speech & Language, 2005, 19(1): 3-30.

4 王守觉. 仿生模式识别（拓扑模式识别）—— 一种仿生识别新模型的理论与应用[J]. 电子学报, 2002, 30(10): 1417-1420.

5 Francisco A, Lopez J. A Neural-network-based Classifier for Large Pattern Recognition[J]. Neural Network World, 2003, 13(1): 3-14.

(上接第 183 页)

10.5%的昆剧样本被错分到了评剧类别中，同时有 12.8%的评剧样本被错分到了昆剧类别中，而在表 2 中错分比例超过 10%的情况却只有 3 次。同样由于豫剧和越剧之间的相似度比较低，它们之间的样本错分情况就要低得多。此外，上述结果与人们对这几个剧种的直观感受也是相同的。我们知道豫剧是典型的北方剧种，它的唱腔比较粗犷刚劲；而越剧则是典型的南方剧种，它的唱腔则更加舒缓轻柔，这两个剧种在听觉感受上是明显不同的。而昆剧虽然源自南戏系统，但也已经在北方地区流传了数百年，它与起源于北方民间说唱艺术的评剧之间具有更多的相似性也就不足为奇了。

4 结论

本文在收集了大量实验数据的基础上，对 8 种中国传统戏曲进行了自动分类和相似性分析研究。实验结果表明通过采用基于样本的方法，可以比较有效地解决这 8 类剧种之间的自动分类问题。在一个包含了 680 个唱段的测试数据库上，本文的方法可以达到 82.4％的平均分类正确率。研究结果还表明在这 8 种传统戏曲中，昆剧和评剧最为相似，豫剧和越剧之间差别最大。相似性分析的结果与戏曲分类实验的结果以及人们对戏曲欣赏的主观感受是基本一致的。将来我们会继续研究中国传统戏曲所具有的特征，希望通过学习与规则相结合的方法来进一步提高算法的效果。

参考文献

1 Zhang T, Kuo C J. Audio Content Analysis for Online Audiovisual Data Segmentation and Classification[J]. IEEE Trans. Speech and Audio Processing, 2000, 9(4): 441-457.

2 Lu L, Zhang H J, Jiang H. Content Analysis for Audio Classification and Segmentation[J]. IEEE Trans. Speech and Audio Processing, 2002, 10(7): 504-516.

3 Zhang Y B, Zhou J. Audio Segmentation Based on Multi-scale Audio Classification[C]. Proc of. IEEE ICASSP, 2004: 349-352.

4 Lambrou T, Kudumakis P, Speller R. Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains[C]. Proc of. IEEE ICASSP, Seattle, USA, 1998: 3621-3624.

5 Li D G, Sethi I K, Dimitrova N, et al. Classification of General Audio Data for Content-Based Retrieval[J]. Pattern Recognition Letters, 2001, 22(5): 533-544.