基于决策树和改进SVM混合模型的语音情感识别

赵涓涓， 马瑞良， 张小龙
（太原理工大学 计算机科学与技术学院，山西，太原 030024）

摘 要：为有效提高语音情感识别的准确性，达到人机和谐交互的目的，本文提出了一种基于决策树和改进SVM混合模型的语音情感识别方法，有效地避免了无界泛化误差、分类器数目多、受限优化等问题，提高了悲伤、喜悦、愤怒、厌恶、惊讶、恐惧6种基本情感识别效率. 实验结果表明，该方法识别准确率为87.58%，与传统的支持向量机和人工神经网络方法相比，有更高的抗噪声能力和稳定性，能得到更高的识别准确率，而且有较强的实用性和推广能力.

关键词：人机交互；情感识别；支持向量机；决策树
中图分类号：TP 391 **文献标志码**：A **文章编号**：1001-0645(2017)04-0386-06
DOI：10.15918/j.tbit1001-0645.2017.04.011

Speech Emotion Recognition Based on Decision Tree and Improved SVM Mixed Model

ZHAO Juan-juan， MA Rui-liang， ZHANG Xiao-long
(School of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, Shanxi 030024, China)

Abstract: To effectively improve the accuracy of speech emotion recognition in intelligent man-machine harmonious interaction, a method of speech emotion recognition was proposed based on decision tree and an improved SVM mixed model. This method can avoid the tree unbounded generalization error, more the number of classifiers and other shortcomings, while taking advantage of SVM-KNN mixed model to avoid constrained optimization problems and improve the recognition efficiency. In this paper, six basic emotions were identified, including sadness, joy, anger, disgust, surprise, fear. Experimental results show that this method can effectively identify six basic emotions. Compared with the traditional support vector machine and artificial neural network method, this method can get higher recognition accuracy, better stability, strong practicability and generalization ability.

Key words: human-computer interaction; emotion recognition; support vector machine; decision tree

随着情感计算与模式识别发展，如何通过语音使人们能够与计算机和谐智能交互，已经成为智能人机交互领域的研究热点[1]. 在日常基本的语音中不仅仅包含了语音表达的信息，还隐含了说话人的情感信息，传统信息处理系统在对语音处理时主要侧重语音中词汇传达的信息是否清晰准确，忽略了其中包含的情感特征[2]，然而，情感信息的识别与处理也是信息处理系统中必不可少的一部分，因此，语音情感信息高效地识别是人机和谐交互的重要基础.

收稿日期：2015-11-12
基金项目：国家自然科学基金资助项目(61540007，61373100)；虚拟现实技术与系统国家重点实验室资助项目(BUAA-VR-15KF02，BUAA-VR-16KF-13)
作者简介：赵涓涓(1975—)，女，博士，教授，E-mail：zh_juanjuan@126.com.

目前,在国内外研究中,情感识别的主要方法有:K 最近邻方法[3]、混合高斯模型法、隐马尔科夫模型法[4]、人工神经网络方法[5]、支持向量机(SVM)以及在这些方法上的改进. 但是由于不同国家、不同语言特色以及发音的差异等,情感的表达特征也不同. 针对于不同的语言,国内外还没有建立标准、统一的语音情感数据库,而且研究者录制建立语音数据库与现实真实情感的语句在自然度上有一定程度差异,在实际应用中由于噪声等相关因素干扰,造成在不同语音数据库应用中识别率不稳定,推广能力差.

针对上述问题,本文首先通过录音与剪辑两种方式构建情感语音库,然后提取基音频率、振幅、短时能量、MFCC 等特征,采用改进的 SVM 和决策树建立混合模型对情感进行识别. 实验结果表明,该方法有更高的识别准确率、抗噪声能力,对汉语语音情感识别有较好的稳定性.

1　相关工作

目前,语音情感识别主要的难点是如何从语音的时域和频域中选择有效的语音特征. 情感识别准确率的高低与语音特征的选取有关,同一识别模型,特征集选取不同,识别结果不同. 从众多特征中选择与各个情感显著相关的特征是提高识别准确率的根本问题. 研究表明,语音高频信息对于某些特定的情感有很好的识别度,语音低频信息与情感唤醒度有显著的相关性[6-7]. Murray 等[8]研究确定了不同情感下各个语音信号的特征状态的定性描述,Cowie 等[9]研究列出了在 14 种情感下基频、共振峰、振幅等特征的分布规律. 在国内外情感识别研究中,研究者通常采用基频、短时能量、时间等基本的韵律特征,这些特征提取方法成熟而且能很好地区分不同的基本情感. 朱菊霞等[10]在自建汉语语音情感数据集上通过支持向量机模型进行语音情感识别,徐照松等[11]在基于汉语语音情感数据库基础上利用 BP 神经网络方法进行分类取得了较好的识别率,Mao[12]和 Schuller 等[13]在语音韵律特征的基础上结合时间尺度等情感特征建立了情感特征集进行分类识别. Vlasenko 等[14]在柏林 EMO-DB 语音数据库上利用混合高斯模型对 4 种情感进行识别,取得了较高的识别率. 张石清等[15]在传统 SVM 的基础上,加入模糊隶属度,利用模糊支持向量机对 4 种情感进行识别,相比传统 SVM 取得了较好的识别性能.

2　情感语音库建立

本文实验所使用的汉语语音情感库通过录音和剪辑两种方式获取,录音时避免外界噪音干扰在安静的录音室录制,该库由 16 位专业的录音人员录制,其中男女各 8 名,录音数据采样参数为 16 kHz、16 bit 的单声道 WAV. 依据 Ekman 等[16]建立的情感分类标准,库中语句有 6 种基本情感状态:悲伤、喜悦、愤怒、厌恶、惊讶、恐惧. 基本的录音语句不包含任何情感倾向且情感自由度高. 同一个录音语句均可由上述 6 种情感表达,且与说话人无关以便于分析比较. 为保证录制的情感语句的质量,录制语句经由评判组对其情感可信度进行评价,将可信度大于 0.75 的语料作为实验数据,对易混淆情感的语句经评判无效后对语句进行重新录制. 为使数据库中语音情感更加自然地接近现实真实的情感,对电视访谈节目、影视剧、广播、演讲等语音数据的剪辑,挑选其中在安静环境中、噪声干扰小、且符合上述 6 种特征的语句进行剪辑加入情感语音库中. 最终,本文建立的情感语音库中包含情感语料共1 600句,各个情感语料组成部分如表 1 所示,其中 493 句通过剪辑获得,并且将1 100句作为训练集,500 句作为测试集.

表 1　语音数据库各情感语料组成

Tab.1　Emotional speech database of corpus

情感分类	悲伤	喜悦	愤怒	厌恶	惊讶	恐惧
语句数量	269	267	274	266	264	260

3　情感特征提取

语音情感是通过不同语音信号混合特征表现出来的,不同的情感与之相关的主要特征集也不同. 同一特征在不同情感的语音信号中数值不同、相关性不同、贡献度也不同. 研究表明,不同情感表达的区分主要表现在语音信号中的韵律特征. 那么能否从众多的语音特征中选择一个高效的、与各情感显著相关的特征集是语音情感识别准确率高低的关键问题. 本文提取以下特征为情感识别的特征集:

① 基音频率:最大值、最小值、极差、均值、标准差、方差、平均绝对斜度、上 4 分位数、中位数、下 4 分位数、内 4 分极值、基频抖动值;

② 振幅能量:最大值、最小值、极差、平均值、标准差、方差、上 4 分位数、中位数、下 4 分位数、内

4 分极值；

③ 共振峰：第1、2、3、4共振峰最大值、最小值、均值、标准差、协方差、变化率、变化率的1/3分为点和1/4分为点；

④ 短时能量：短时振幅变化率均值、最大值、最小值、极差、中值、方差；

⑤ 发音持续时间：发音持续总时间、有声发音持续时间、无声发音持续时间、有声发音持续时间与发音持续总时间之比、无声发音持续时间与发音持续总时间之比；

⑥ 语速：平均语速；

⑦ MFCC：12维MFCC均值.

4 语音情感识别模型

对于多分类语音情感问题，传统SVM有两种解决方案，即一对一和一对多. 通常一对一方法中易出现无界泛化误差，一对多方法易造成分类器数目多、分类效率低的缺点. 针对上述问题，本文将6种情感分类问题进行分解，建立基于决策树的多级SVM分类器，对于样本集，每一级的SVM识别出一种情感，剩余的样本集进入下一级的SVM进行识别，如图1所示，逐级递减，最后决策树的叶子节点是所得到的情感分类.

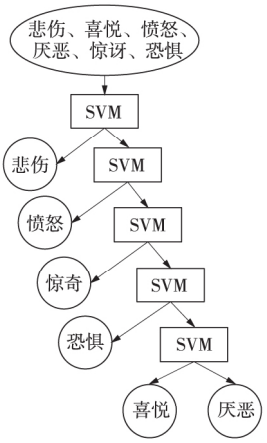

图1 决策树SVM多分类示意图

Fig.1 The decision tree and the SVM classification schemes

然而，在面对不同的应用时，通常无法选择合适的核函数，而且传统SVM对于复杂问题的分类准确率低，对大规模分类问题训练时间复杂度高等[17]，对SVM分类时错分样本的分布进行分析发现，错误样本多集中在分界面的附近[18-19]，而远离分界面的样本基本能够得到正确地分类. 提高超平面附近样本的分类正确率是提高SVM精度的关键. 因此，对于SVM超平面附近产生的错分样本，利用KNN算法进行结合，构建了SVM-KNN组合分类模型，如图2所示，由于初始的SVM分类对样本的分类准确度和可信度低，于是本文通过计算分界面两边样本的相似度，选择超平面两边分类不明确、模糊的n个样本，由于分界面附近的样本基本上都是支持向量，所以结合SVM和KNN，可对样本在空间的不同分布使用不同的分类方法，提高SVM分类的准确性.

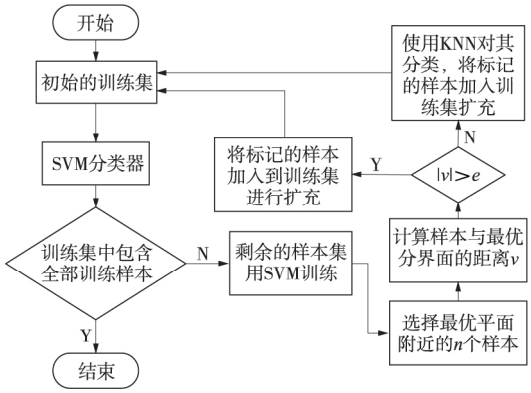

图2 SVM-KNN混合模型

Fig.2 SVM and KNN hybrid model

构建SVM-KNN分类器步骤：

① 初始认为训练集中的样本都为被标记，从训练集中随机选择选择少量样本，构造一个小样本训练集，确保初始训练样本集中每种情感都至少包含一个样本；

② 根据初始训练样本得到一个情感A的弱SVM分类器，然后确定其最优分类超平面，支持向量集T、分类决策函数的系数W和常数b；

③ 用弱SVM标记语音情感样本集中所有的未标记样本，选择超平面附近分类模糊、准确率低的样本：从A类情感中任选一个样本，计算它与非A类情感所有样本的相似度，挑选出n个最可能是非A类情感的样本，记为样本集A. 从非A类样本中任选一个样本，计算其与A类情感所有样本的相似度，挑选出n个最可能是A类情感的样本，记为样本集B；

④ A和B中的样本为超平面附近的点，将A、B中的样本x代入决策函数

$$g(\boldsymbol{x})=\sum\nolimits_i y_i a_i K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b,$$

计算得到样本点与最优分类面之间的距离v；

⑤ 若$|v|>e$，则通过SVM对样本点的分类准确度、可信度高，因此，可以通过决策函数$f(\boldsymbol{x})=$

$\mathrm{sgn}g(x)$确定样本点所属类别;

⑥ 若$|v|<e$,则样本点在超平面附近,分类可信度低、易错分,因此,通过 KNN[4]方法确定样本 x 所属类别. 将 A 类与非 A 类的支持向量集 T 作为训练样本,计算样本 x 与 T 中每一个向量之间的距离 $\mathrm{d}(x,x_i)$,将距离最近的向量所属的类别作为样本 x 的类别,

$$d(x,x_i)=\|\Phi(x)-\Phi(x_i)\|=\sqrt{k(x,x)-2k(x,x_i)+k(x_i,x_i)},$$

式中:x_i 为支持向量;$k(\ )$为一阶多项式核函数;阈值 e 的范围为[0,1],具体值可以根据实验结果进行动态调整,初始值一般设置为 1,如果调整为 0,则算法为传统的 SVM 算法;

⑦ 将 SVM 分类得到的样本和 KNN 分类的样本放入初始训练集对其进行扩充,从而在扩充后的训练集基础上训练一个新的 SVM2;

⑧ 迭代下去,直到训练集中所有样本都加入初始训练集中时,停止迭代. 利用最终的训练集得到一个对 A 类情感分类精度高的 SVM 分类器;

⑨ 此时训练出的决策树中一级的 SVM 分类器,然后利用非 A 类的样本集作为下一级 SVM 的训练器. 这样逐级训练得出各个情感类别相对应的 SVM 分类器.

5　实验结果

首先本文在自制的语音情感数据库上提取基音频率、振幅能量、短时能量、共振峰、语速、发音持续时间、MFCC 等情感特征,在 SVM 的基础上,设计了基于决策树和 SVM-KNN 混合模型进行情感识别方法,在自建情感语音数据库上对 6 种基本情感识别,结果如表 2 所示,该方法对于悲伤、恐惧、愤怒情感的识别正确率高、稳定性好,在某些语料中厌恶、惊讶情感的识别率错误率较高,这可能是由于厌恶、惊讶两种情感的音频特征相似以及数据库中两种情感语音表现的情感特征不显著. 与此同时,本文通过传统的支持向量机方法和人工神经网络方法对自建语音情感数据库中 6 种情感进行识别,其结果比较如表 3 所示,本文方法的情感平均识别率为 87.58%,比传统 SVM 方法提高了 6.96%,比人工神经网络方法提高了 9.22%,优于传统 SVM 和 ANN 方法,而且对于愤怒、恐惧、悲伤 3 种情感,本文识别准确率比传统 SVM 和 ANN 方法有了较高的识别率.

表 2　本文 6 种情感识别率结果

Tab.2　Six kinds of emotion recognition results

情感	情感识别率/%					
	悲伤	喜悦	愤怒	厌恶	惊讶	恐惧
悲伤	88.35	1.57	1.67	4.74	4.67	5.36
喜悦	1.74	86.25	7.36	2.38	6.35	1.45
愤怒	2.33	3.23	91.44	3.96	5.23	5.86
厌恶	8.41	3.41	3.26	83.54	4.74	6.74
惊讶	3.20	7.17	3.85	1.26	82.76	8.32
恐惧	6.74	4.66	3.66	3.17	5.48	93.17

表 3　情感识别率比较

Tab.3　Emotion recognition rate

识别方法	情感识别率/%						平均识别率/%
	悲伤	喜悦	愤怒	厌恶	惊讶	恐惧	
SVM	79.54	83.34	79.37	78.63	80.51	82.35	80.62
ANN	73.62	81.52	78.65	77.39	82.37	78.25	78.63
本文	88.35	86.25	91.44	83.54	82.76	93.17	87.58

为检验本文方法对其他语音情感数据库分类的准确性,通过本文方法及 SVM、ANN 方法对北京航空航天大学情感语音数据库中 6 种基本情感进行识别,结果表明,本文将 3 种方法的平均识别率进行比较,结果如表 4 所示,ANN 方法其识别率变化小,但识别率低,传统 SVM 方法识别率变化幅度较大,整体而言,本文方法有较好的稳定性和识别准确率.

为验证不同的信噪比对本文方法识别率的影响. 本文选取无噪声的 800 句测试样本添加高斯白噪声,分别在不同的信噪比下进行实验,结果如表 5 所示.

表 4　北航情感语音数据库实验比较

Tab.4　Emotional speech database experimental comparison

数据库	平均识别率/%		
	SVM	ANN	本文
本文自建情感语音数据库	80.62	78.63	87.58
北京航空航天大学情感语音数据库	75.65	77.91	85.43

表 5　不同信噪比下情感识别率

Tab. 5　Emotion recognition rate under different signal-to-noise ratio

方法	信噪比/dB							
	100	80	60	40	20	10	0	−10
本文	86.43	83.27	79.62	74.85	69.21	61.53	47.53	14.85
SVM	80.15	76.62	71.54	67.34	61.35	46.75	35.62	4.62
ANN	77.68	70.94	62.86	56.72	51.82	40.51	24.97	0

实验结果表明,随着信噪比的增加,3 种方法的识别准确率都在不同程度逐渐降低,但是本文方法的识别率下降速率、下降幅度都小于传统 SVM 和 ANN 方法,而且在相同的信噪比下平均识别率都

高于后两者. 因此,本文方法有更好的稳定性和抗噪声能力.

6 结 论

在传统SVM的基础上,设计了基于决策树和改进SVM混合模型的情感语音识别方法,通过SVM与KNN混合分类模型避免了大规模训练样本下受限优化的问题,同时提高SVM分类精度以及识别速度. 通过在两个不同的汉语语音数据库上分别对悲伤、喜悦、愤怒、厌恶、惊讶、恐惧6种基本情感进行识别,平均识别率达到85%以上,优于传统的SVM和ANN方法,具有良好的稳定性和抗信噪比的能力. 但对于厌恶、惊讶两种情感的识别率较低,这可能是由于在自建的情感数据库中两种情感的语料有一部分区分度小、某些音频特征相似,或者本文在特征选择时有缺陷. 而且在决策树上SVM分类的准确率会受到上一级SVM分类准确率的影响. 因此,最佳特征组的选择、情感分类的增加、识别率的提高仍为今后研究的重点.

参考文献:

[1] 赵腊生,张强,魏小鹏.语音情感识别研究进展[J].计算机应用研究,2009,26(2):34-38.
Zhao Lasheng, Zhang Qiang, Wei Xiaopeng. Research and development of speech emotion recognition[J]. Application Research of Computers, 2009, 26(2): 34-38. (in Chinese)

[2] 张石清,李乐民,赵知劲.人机交互中的语音情感识别研究进展[J].电路与系统学报,2013,18(2):440-451.
Zhang Shiqing, Li Lemin, Zhao Zhijing. Research and development of speech emotion recognition in human-machine interaction[J]. Journal of Circuits and Systems, 2013, 18(2): 440-451. (in Chinese)

[3] Lee C M. Classifying emotions in human-machine spoken dialogs [C]// Proceedings of 2002 IEEE International Conference on Multimedia and Expro-Proceeding. [S.l.]: IEEE, 2002: 737-740.

[4] 林奕琳,韦岗,杨康才.语音情感识别的研究进展[J].电路与系统学报,2007,12(1):90-98.
Lin Yilin, Wei Gang, Yang Kangcai. Research and development of speech emotion recognition[J]. Journal of circuits and systems, 2007, 12(1): 90-98. (in Chinese)

[5] Khanchandani K B, Hussain M A. Emotion recognition using multilayer perceptron and generalized feed forward neural network[J]. Journal of Scientific and Industrial Research, 2009, 68(5): 367.

[6] Huang C W, Jin Y, Wang Q Y, et al. Speech emotion recognition based on decomposition of feature space and information fusion[J]. Signal Processing, 2010, 26(6): 835-842.

[7] Lee C M, Narayanan S S, Pieraccini R. Classifying emotions in human-machine spoken dialogs [C] // Proceedings of IEEE International Conference on Multimedia and Expo. [S.l.]: IEEE, 2002: 737-740.

[8] Murray I R, Arnott J L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion[J]. The Journal of the Acoustical Society of America, 1993, 93(2): 1097-1108.

[9] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal Processing Magazine, 2001, 18(1): 32-80.

[10] 朱菊霞,吴小培,吕钊.基于SVM的语音情感识别算法[J].计算机系统应用,2011,20(5):87-91.
Zhu Juxia, Wu Xiaopei, Lü Zhao. Speech emotion recognition algorithm based on SVM[J]. Computer Systems & Applications, 2011, 20(5): 87-91. (in Chinese)

[11] 徐照松,元建.基于BP神经网络的语音情感识别研究[J].软件导刊,2014,13(4):11-13.
Xu Zhaosong, Yuan Jian. Research on speech emotion recognition based on BP neural network[J]. Software Guide, 2014, 13(4): 11-13. (in Chinese)

[12] Mao X, Chen L, Fu L. Multi-level speech emotion recognition based on HMM and ANN[C]// Proceedings of Wri World Congress on Computer Science and Information Engineering. [S.l.]: IEEE, 2009: 225-229.

[13] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE, 2004: 577-580.

[14] Vlasenko B, Schuller B, Wendemuth A, et al. Combining frame and turn-level information for robust recognition of emotions within speech[C]// Proceedings of INTERSPEECH 2007, Conference of the International Speech Communication Association. Antwerp, Belgium: [s.n.], 2007: 2249-2252.

(下转第395页)

6 结　论

本文在研究非采样 Contourlet 变换和归一化理论的基础上,提出一种新的数字图像水印算法,充分利用归一化的抵抗能力,可以很好地抵抗几何攻击,在水印领域应用非采样轮廓变换,这也是一个大胆尝试,实验数据表明,本文算法对平移变换的抵抗力最强,对旋转和缩放也有很好的抵抗力,而且经过 JPEG 压缩之后,数字水印也不会失真. 并且在抽取水印图像的过程中,不需要借助于原始图像,可以做到真正的盲检测.

参考文献:

[1] Run R S, Horng S J, Lai J L, et al. An improved SVD-based watermarking technique for copyright protection [J]. Expert Systems with Applications, 2012, 39(1): 673-689.

[2] Hajjara S. Digital image watermarking using localized biorthogonal wavelets[J]. European Journal of Scientific Research, 2009,26(4):594-608.

[3] Cvejic N, Seppanen T. Spread spectrum audio watermarking using frequency hopping and atack characterization[J]. Signal Processing, 2004,84(1):207-213.

[4] Kurosaki M, Kiya H. Error concealment using a data hiding technique for mpeg video[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2002,85(4):790-796.

[5] Li Leida, Guo Baolong. Localized image watermarking in spatial domain resistant to geometric attacks [J]. AEU-International Journal of Electronics and Communications, 2009,63(2):123-131.

[6] Hajjara S. Digital image watermarking using localized biorthogonal wavelets[J]. European Journal of Scientific Research, 2009,26(4):594-608.

[8] Ping Dong, Brankov J G, Galatsanos N P, et al. Digital watermarking robust to geometric distortions[J]. IEEE Transactons on Image Processing, 2005, 14 (12): 2140-2150.

[9] Lee H Y, Kim H. Robust image watermarking using local invariant features [J]. Optical Engineering, 2006, 45(3):037002,1-11.

[10] Parameswaran L A, Anbumani K.A robust image watermarking scheme using image moment normalization [J]. Transactions on Engineering, Computing and Technology, 2006,13:1305-5313.

[11] Da Cunha A L, Zhou J, Do M N. The nonsubsampled contourlet transform: theory, design, and applications [J]. IEEE Transactions on Image Processing, 2006, 15(10):3089-3101.

[12] Cunha A L, Zhou J, Do M N. Nonsubsampled contourlet transform: filter design and applications in denoising[C]// Proceedings of IEEE International Conference on Image Processing. [S.l.]: IEEE, 2005, 9 (1):49-52.

(责任编辑:李兵)

(上接第 390 页)

[15] **张石清. 基于模糊支持向量机的语音情感识别[J]. 台州学院学报**, 2007, 28(6): 52-55.
Zhang Shiqing. Speech emotion recognition based on fuzzy support vector machine[J]. Journal of Taizhou University,2007,28(6):52-55. (in Chinese)

[16] Ekman P. An argument for basic emotions [J]. Cognition & Emotion, 1992, 6(3-4): 169-200.

[17] Vlachosa A. Active learning with support vector machines[D]. Scotland: University of Edinburgh, 2004: 12-14.

[18] Rong L, Shiwei Y, Zhongzhi S. A effective classified algorithm of support vector machine with multi-representative points based on nearest neighbor principle[C]// Proceedings of International Conferences on Info-Tech and Info-Net. Beijing: IEEE,2001:113-119.

[19] Chin K K. Support vector machines applied to speech pattern classification [D]. Cambridge: Cambridge University, 1998.

(责任编辑:李兵)