摘 要

近年来，有关新生儿疼痛的研究证实，反复的疼痛刺激会对新生儿产生一系列近期和远期的不良影响；又由于新生儿不能自述疼痛的感受，由此产生了一些针对新生儿疼痛的评估工具，其中面部表情是被广泛认同的一种最有效、可靠的评估指标，所以需要开发一种有效的新生儿疼痛面部表情自动识别评估系统，这是非常有意义的。

本文在研究了当前几种常见的多分类支持向量机（SVMs）算法的基础上，深入研究决策树方法，提出了对二叉树多分类SVM方法（BT-SVMs）的改进，并利用多分类支持向量机的方法，对新生儿的疼痛表情进行多类识别。本文的主要工作包括：（1）研究了两类支持向量机的基本理论，通过对三组数据集的实验对其核函数及其参数的选择过程进行了分析；（2）对当前常见的6种多分类SVM方法进行了系统的研究，从训练和预测两方面分析比较了这些多分类SVM方法的性能，实现了几种常用的多分类SVM分类器，并通过相同的数据库对其实验结果进行了比较；（3）在深入研究基于决策树算法的基础上，针对二叉树算法的误差累积效应问题，首先用聚类方法中的类间距离设计了二叉树结构，分析其缺点后，接着进一步提出了基于最优分类面的类间分离度，用其对二叉树多分类SVM方法的改进，同时通过实验证明了改进后算法的合理性以及可行性；（4）根据二叉树多分类SVM方法设计了新生儿面部表情识别多类分类器，并采用一对一SVMs、一对多SVMs、决策导向无环图SVMs（DDAG-SVMs）、BT-SVMs与改进后的BT-SVMs方法分别对新生儿的4类表情（安静、哭、轻度疼痛和剧烈疼痛）进行实验，根据每种多类算法选择SVMs分类器最佳核函数，交叉验证结果显示改进后的BT-SVMs方法获得了最好的识别率为86.12%，其训练时间也相对较短；折衷考虑训练时间和识别率两方面的要求，选择改进后的BT-SVMs方法较合理。通过对实验结果的分析得出如下结论：本文使用的多分类SVM表情识别方法针对新生儿疼痛面部表情的多类识别能够获得期望的性能。

关键词：新生儿疼痛，面部表情识别，支持向量机，多分类算法

Abstract

In recent years, studies on neonatal pain suggest frequent pain has a negative impact on child. Kinds of assessment tools of neonatal pain has been developed as neonates are not capable of describing their subjective experience. Above all, the change of facial expressions is regarded as the most effective, reliable assessment indicator. Therefore research on facial expression recognition is needed and very meaningful for the development of an effective auto assessment system of neonatal pain.

In this thesis, several popular multi-classification SVM(SVMs) algorithms and their characteristics are analysed. After thoroughly researching the decision-making algorithm, a improved algorithm is proposed for binary tree SVMs (BT-SVMs). Then multi-classification SVM algorithm is utilized for recognizing neonatal pain expression. The main work includes: (1)The choice of binary SVM's kernel function and parameters is analyzed by experimenting in same databases. (2) It studies exhaustively on the multiclass SVM algorithm, analyzes and compares the traditional methods' performance in the training and testing processes. Several standard multi-classification SVM classifiers are implemented, and some comparison results in two opening databases are got. (3) In the decision-making algorithm, due to the cumulating error problem of binary tree, the distance between classes of clustering is used to design binary tree SVMs. For improveing it, a measure of separability between classes is proposed based on best classifying surface of SVM. The rationality and the feasibility of the method is demonstrated by experiments. (4) BT-SVMs algorithm was used to make a multiclass classifier for recognizing neonatal pain expression. Our experiments recognizing 4 kinds neonatal facial expressions (quiet, cry, low-grade pain and acute pain) compares diversified multi-classification SVM algorithm with BT-SVMs algorithm and its improved algorithm, including one-against-one SVMs, one-against-rest SVMs and Dicision Directed Acyclic Graph SVMs (DDAG-SVMs). The best recognition rate(86.12%) was obtained from the improved BT-SVMs expending a little less runtime, better than other multi-classification SVM algorithm. To take into account the training time and the recognition rate, it is reasonable to choose the improved BT-SVMs. Analysis of experiments results get the following conclusion: the proposed SVMs algorithm has obtained

well performance for neonatal multi-expressions recognition.

Keywords: Neonatal pain, Facial expression recognition, Support vector machine, Multi-classification algorithm

南京邮电大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京邮电大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：郭昊　日期：2008.4.15

南京邮电大学学位论文使用授权声明

南京邮电大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权南京邮电大学研究生部办理。

研究生签名：郭昊　导师签名：[illegible]　日期：2008.4.15

第一章 绪论

1.1 课题背景

早期大部分麻醉学者认为新生儿不会有疼痛的感觉，所以新生儿疼痛一直未予重视和正确处理。近年来，有关新生儿疼痛的研究证实，反复的疼痛刺激会对新生儿（尤其对早产儿和危重儿）产生一系列近期和远期的不良影响，因此医疗系统正致力于开发出较好的婴儿疼痛管理协议和疼痛评估工具。

事实上病人的看护质量取决于疼痛的处理性质[1,2]。临床上，将疼痛定义为一种主观的体验，认为评估疼痛的最可靠的方法是通过自我评估报告[3]。大多数成年人都能用语言来描述疼痛的部位、时间和强度及疼痛的经历，但新生儿不能自述自己的疼痛的经历，只有通过诊断或其它方法来评估疼痛。通过研究证明，忽视疼痛可能导致新生儿中枢神经系统的缓慢变化，经常性疼痛会带来负面影响和其它显著的长期性影响[4,5]。

在评估新生儿疼痛时，专家认为仅靠生理变化来评估疼痛是远远不够的，因为其它非疼痛的刺激也可引起类似生理反应[6]，例如恐惧和兴奋。新生儿疼痛的行为反应包括躯体活动、哭泣、睡眠和饮食模式以及面部表情的改变。其中面部表情被视为是一种最好的疼痛评估指标[7]，而身体活动与哭泣对于指示疼痛不是很具体，它们在其它的情形下也会发生，如饥饿、恐惧和不安[8]。另外，新生儿疼痛的反应并不总是哭泣和活动[9]。

在研究新生儿疼痛时，目前使用的各种疼痛评估工具的测量指标各不相同，但都将面部表情作为其中最有效的一项。国际上对新生儿疼痛的评估都是由受过专门训练并熟悉各项评估技术指标的医护人员进行人工评估，这带来了一些实际问题，如需要花费大量的时间精力，而且评估结果往往受到主观因素的影响。因此，开发一种客观、快速、有效的新生儿疼痛自动评估系统是非常有意义和价值的。但目前还没有见到公开的相关报道及研究成果。

本文所选课题的研究内容是新生儿疼痛面部表情自动识别理论和方法，主要研究基于SVM的新生儿面部疼痛表情识别分类算法。其研究可为医护人员客观、可靠、有效地评估疼痛提供科学的依据，以便在临床上及时采取相应的治疗措施，以减轻新生儿的疼痛。

1.2 相关领域的研究现状及问题

1.2.1 国内外研究现状

1. 新生儿疼痛研究

在国外新生儿疼痛的相关研究已倍受关注，在国际上目前常用的评估工具有以下4种：早产儿疼痛评分（Premature Infant Pain Profile，PIPP）[10]、新生儿疼痛评分（Neonatal Infant Pain Scale，NIPS）[11]、CRIES（Crying，Required O2 for SO2 >95%，Increased vital signs，Expression，Sleeplessness）量表[12]、新生儿面部编码系统[13]（Neonatal Facial Coding System，NFCS）。上述 4 种不同新生儿疼痛评估工具的观测指标有所不同，信度、效度和可行性等的比较见表 1-1。其中"★"表示有此项观测指标；"+"表示好；"—"表示无或未验证。由下表可看出，面部表情是所有评估工具都采用的一项测评指标，而且在NFCS中还作为单一的测评指标。

表1-1 新生儿疼痛评估工具的比较

	PIPP	NIPS	CRIES	NFCS
面部表情	★	★	★	★
睡眠/觉醒状态	★	★	★	—
生理指标	★	★	★	—
哭闹	—	★	★	—
肢体	—	★	—	—
孕周	★	—	—	—
测量者间信度	+	+	+	+
内部一致性	+	+	—	+
同期/标准效度	+	+	+	+
结构效度	+	+	+	+
可行性	+	+	+	+

目前对成人的表情识别的研究工作比较多，而针对新生儿疼痛表情的研究还没有见到公开的相关报道和研究成果。

2. 表情识别技术

在计算机面部表情识别方面，国际上研究的主要有美国的麻省理工大学(MIT)，卡耐基梅隆大学[14](CMU)，马里兰大学[15](Maryland)，伊利诺大学-香槟分校[16,17](UIUC)，日本的名古屋大学[18](Nagoya)、城蟆大学[19](SEIKEI)、东京大学[20,21](Tokyo)和ATR研究所，瑞士的戴尔莫尔感知人工智能研究所[22](IDIAP)等。

MIT提出的一个新的高技术前沿研究方向——情感计算，是关于、产生于和影响于情

感方面的计算，它赋予计算机识别、理解、表达和适应人情感的能力，表情识别的研究是其中的核心内容。Mase等人[18]的表情分析思想分为从上至下和从下至上两个方向，在这两种情况下，焦点都是在计算脸部肌肉的运动，而不是特征的运动。Yacoob和Davis[15]基于FACS编码，集中分析人脸上嘴、眼睛和眉毛边缘的相关运动，在八方向上检测运动，在一张脸上预定义、手工初始化的六个矩形区域，使用简化的FACS规则识别六种表情；同时建立了一个时间模型：Beginning-Apex-Ending，规定每种表情的整个过程以中性表情作为开始和结束；并定义了变化中每个阶段的开始与结束的规则来检测各个阶段。Rosemblum和Yacoob等人[23]用RBFN（Radial Basis Function Network）结构，学习脸部特征与人类情绪之间的相关性，在最高一级识别情绪，在中间一级决定脸部特征运动，在低一级恢复运动方向；特征提取中不关注脸部的肌肉运动模型，而是关注特征部件边缘的运动。Sakaguchi等人[19]利用小波变换提取脸部图像眉眼区域和嘴部区域的频率域特征，并采用离散隐马尔可夫模型(Hidden Markov Model，简称HMM)来刻画表情出现时脸部动态时间信息，进而识别六种基本的面部表情；但是因为在建立离散隐马尔可夫模型时需对观察矢量采取矢量量化操作，因此不可避免的引入了量化误差。美国CMU的Lien.J.J[14]运用离散隐马尔可夫模型分析人脸表情的细微变化，自动区别各种基于FACS表情活动单元AU的细微面部表情；文中提出了一种构造离散隐马尔可夫模型拓扑结构的方法，对后人学习如何训练离散隐马尔可夫模型具有参考价值。Cohen等人[16]根据隐马尔可夫模型的基本理论和算法设计了一个人脸表情识别系统，结果表明HMM是一种非常有效的基于统计的识别表情的方法，缺点是离散余弦变换虽然具有压缩比高且能较好地保持原始图像信息的优点，但是在人脸表情识别系统中作为人脸表情特征向量，数据量仍然太大，影响系统速度。Hai Tao和Thomas Huang[17]基于分段Bezier体积形变模型（Piecewise Bézier Volume Deformation Model，简称PBVD），提出了一种基于解释的脸部运动跟踪算法。PBVD模型适合合成和分析脸部图像，它是线性和独立的脸部网络结构。

在国内，中科院计算所，自动化所，清华大学，中国科学技术大学，哈尔滨工业大学，华中科学技术大学等都对表情识别做了研究。哈尔滨工业大学的金辉和高文[27]等通过对若干类面部表情图像的分析，建立了基于部件分解组合的人脸图像模型。他们在对动态表情图像序列的时序分析的基础上，提出了对混合表情的识别系统；此外他们基于隐马尔可夫模型进行了面部表情图像序列的分析与识别[24]。北京自动化所研究了一种增强的基于嵌入式隐马尔可夫的实时表情识别方法[25]。清华大学的王玉波等采用了扩展的Adaboost算法构造强分类器进行实时的表情识别[26]。中国科技大学的尹星云[28]等用隐马尔可夫模型的基本理论和方法设计了人脸表情识别系统。天津大学的左坤隆、刘文耀[29]等利用活动外观模板

(AAM)提取的人脸表情特征来进行人脸表情识别(FER)。

在面部表情识别的研究中大多算法采用由Kanade和Cohn等建立的 DFAT-504 人脸表情数据库作为训练和测试样本[30]，都是针对普通成年人，采用Ekman提出的情感分类方法，定义“高兴、愤怒、厌恶、恐惧、悲伤、惊奇”六种“典型”的面部表情，但至今还未见到有关针对像“疼痛”这种特定的面部表情的识别算法。

1.2.2 面部表情识别系统

从九十年代起，人工智能与模式识别技术得到很快的发展，面部表情识别作为其中的一个领域也发挥着越来越重要的作用。面部表情识别是研究如何让机器能够辨识出人脸面部表情变化的一项工作。由于面部表情携带的是信息序列，所以面部表情是除了语言交流以外非常重要的交流方式。面部表情识别已经应用于社会生活的各个领域，如公共场合安全监控、心理学研究、精神病学、自然和谐的人机交互、远程教育、安全驾驶等方面。

一般面部表情识别是指基于视觉信息，将脸部的运动或者特征的形变进行分类。面部表情自动识别系统主要包括以下四个主要包括：图像获取、人脸检测、面部表情特征提取和面部表情的情感分类（如图1-1）。其中后三部分是核心环节，人脸检测是在脸部图像或图像序列中定位人脸，是后续部分的准备工作；面部表情特征提取是从脸部图像中提取所需的面部运动或面部特征形变信息；面部表情分类是将输入的面部图像归入某个具体类中，即给出一个类别标记。

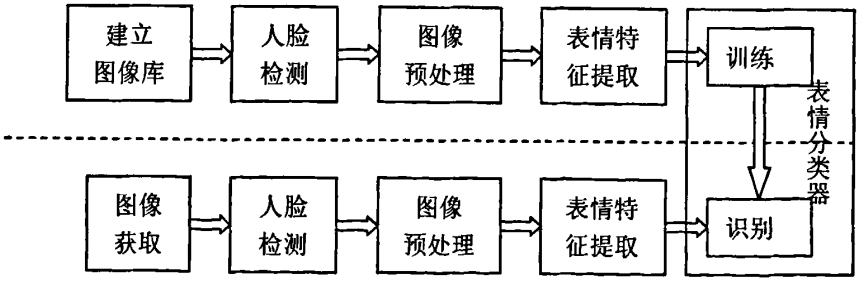


图1-1 表情识别系统框图

（1）图像的获取：该模块从外界获取图像，作为人脸识别系统的输入。该模块可以是一个摄像头或者是扫描仪等设备。

（2）人脸的检测：处理分析从图像获取模块输入的图像，检测出人脸的存在并确定其位置，并且将人脸从背景图像中分离出来。对于视频图像，不仅要求检测出人脸位置，还要求能够跟踪人脸。不同的图像输入条件，对该模块的设计提出了不同的要求。近年人脸检测和跟踪已经作为一个独立的研究课题受到很多研究者的重视。

人脸检测的众多方法都是建立在对各种面部的特征信息进行分析的基础上的。它们从分析处理各种特征信息的角度给出了不同的人脸检测思路。这些待分析的特征包括肤色特征、轮廓特征、灰度分布特征、镶嵌图划分特征、对称性结构特征等。建立在对这些特征信息进行分析基础上的解决人脸检测问题的方法包括：基于边缘的方法[31,32]、基于颜色信息(或灰度信息)的方法[33-35]、基于运动信息的方法[36,37]、基于形变模型的方法[38-40]、基于线性空间的方法[41,42]、基于机器学习的方法[43,44]等。

基于形变模型方法需要使用者为它提供较接近实际人脸区域的初始位置，这种对于初始位置的敏感性使得它的应用受到了一些限制。线性空间方法采用矩阵分析理论处理人脸检测问题，但是由于它无法体现人脸图像中各个器官之间的拓扑关系以及人脸五官的灰度信息特征使得该类方法存在一些欠缺。近些年来，机器学习在模式分类、信号估计与检测、函数拟合方面取得了较大的进展。

（3）面部表情特征提取：即采取何种表示方式表示已被检测出的表情图像或数据库中的面部表情图像，包括预处理和特征提取两部分。预处理的主要对检测到的人脸图像进行归一化、校正等处理，为后面的特征提取创造条件。人脸表情特征提取是根据人脸描绘方法不同而采用不同的特征提取方法进行的面部表情信息的提取，提取出的内容被用来进行后续的表情分类。特征提取通常会缩减输入空间的维数，缩减程序应该保留本质的信息，并拥有较高的判别能力和稳定性。基于几何的、运动的、统计的或者空间变换的特征常用来作为在分类前的面部表情的可供选择的表示方法[45]。

人脸表情特征提取的方法按其所使用图像的类型不同可分为两类：一类是静态图像中的人脸表情特征数据提取，另一类是序列图像中的人脸表情特征提取。静态图像中人脸表情特征数据提取方法包括全局人脸特征提取和人脸局部特征提取，全局人脸特征提取方法包括基于点分布模型的方法（PDM）、基于混沌调制矢量的方法（CMV）、基于小波分析的方法、PCA 分析以及使用支持向量机（Support Vector Machine）的方法。基于局部特征的人脸表情特征提取方法一般根据瞳孔和上人中的几何关系，使用图像处理的手段对突出特征如眉毛、眼睛、鼻子、嘴、轮廓和皱纹等特征进行检测。图像序列具有良好的时空特性，因此图像运动的局部参数模型（PMLP）可以用来进行特征提取。基于特征点的光流模型、基于密度流的像素跟踪模型、PCA 分析方法、势网格模型、特征脸方法、局部特征区域跟踪等方法都得到了应用。

（4）面部表情的情感分类：也可称为分类器设计。此过程结束后将生成可用于识别的参数，也就是可用于分类识别的分类器。事实上，模式识别问题可以看成是一个分类问题，即把待识别的对象归到某一类中。在表情识别问题中就是把输入的不同的表情归入某

一类。这部分的基本方法是在样本训练集基础上确定某个判决规则，使按这种判决规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小。最后根据训练所得的参数完成表情的判别工作，给出最后的识别结果，并做出相应的判断。这一过程也是表情识别的研究重点，核心是选择与所用的表情描述方式适合的分类策略。

目前用于表情分类的方法可以分为两类：时域-空域方法(Spatio-Temporal Approaches)和空域方法(Spatial Approaches)。时域-空域方法：考虑了表情随时间的变化，即时域上和空间上的表情变化信息；可用隐马尔可夫模型、自回归神经网络[25](Recurrent Neural Networks)、时空运动能量模板[46](spatio-temporal motion-energy templates)等方法对表情动态信息进行建模。空域方法：只使用特征向量不使用时间信息的分类方法，包括神经网络方法[47-50]、基于规则推理的方法[51,52]、子空间方法[54]、支持向量机方法[53,54]等。此外还有最近邻类器(NearestNeighbor，NN)[55]，Moghaddma[56]提出一种基于贝叶斯的人脸识别方法，Calder 等人[57]使用主成分分析进行表情识别，Chen 和 Huang[58]则采用聚类(Clustering)方法构造分类器。

1.2.3 基于支持向量机的表情识别

作为最有效的模式识别算法之一，支持向量机（Support Vector Machine, SVM）是近几年来发展起来的一种强有力的分类工具，在人脸识别中得到了大量应用，很多研究表明，使用支持向量机进行人脸识别能够获得很高地识别率。支持向量机是贝尔实验室研究人员V.Vpanik等人在对统计学习理论进行了三十多年研究的基础上提出的一种全新的机器学习算法，它使统计学习理论第一次对实际应用产生了重大影响，伴随SVM在模式识别领域的成功应用使核技术也获得了广泛的应用。

对于计算机来说，模式识别是一项非常困难的任务。其难点主要来源于以下两个困难：（1）维数灾难，即模式的高维度和训练机的小规模所引发的困境。模式的维数越高，所需训练样本越多。在现实应用中，我们能够拥有的训练样本数往往远远少于所需样本数；（2）训练样本缺乏代表性，即训练样本的有限性和未知测试样本的无限性所引发的困境。这导致分类器要么过拟和，要么拟和不足，难以获得令人满意的泛化性能。而SVM基于统计学习理论的结构风险最小化原则，将基于最大化边界Margin获得的分类超平面思想与基于核技术的方法结合在一起，表现出了很好的泛化能力。SVM具有一组特殊的称为支持向量的样本，可以通过寻找一个超平面使待分类的样本尽可能分开，且分类间隔最大。由于SVM有统计学习理论作为坚实的数学基础，所以它可以很好地克服维数灾难和过拟合等传

统算法所不可避免的问题。对于人脸识别来讲，SVM已经成为一种有效的分类工具。

尽管如此，目前在SVM的训练和实现技术方面仍然存在一些困难和问题：SVM中核函数的选择将影响分类器的性能，如何根据待解决问题的先验知识和实际样本数据，选择和构造合适的核函数、确定核函数的参数等问题都缺乏相应的理论指导；如何解决大规模数据集的训练速度与规模之间的矛盾，测试速度与支持向量数目之间的矛盾；虽然多类SVM的训练算法已被提出，但用于多分类问题的有效算法、多类SVM的优化设计仍是需要进一步研究的问题。

1.3 本文的研究内容

本文重点是多分类支持向量机算法及其在新生儿疼痛面部表情识别领域的应用，针对这些问题本文主要研究内容有：

（1）研究了两类支持向量机的基本理论，通过对三组数据集的实验对其核函数及其参数的选择过程进行了分析；

（2）对当前常见的6种多分类SVM方法进行了系统的研究，从训练和预测两方面分析比较了这些多分类SVM方法的性能，实现了几种常用的SVM多分类器，并通过相同的数据库对其实验结果进行了比较；

（3）在深入研究基于决策树算法的基础上，针对二叉树算法的误差累积效应问题，首先用聚类方法中的类间距离设计了二叉树结构，分析其缺点后，接着进一步提出了基于最优分类面的类间分离度，用其对二叉树SVM多分类（BT-SVMs）方法进行改进，同时通过实验证明了改进的合理性以及可行性；

（4）根据二叉树SVM多分类方法设计了新生儿面部表情识别多类分类器，并采用一对一SVMs、一对多SVMs、DDAG-SVMs、BT-SVMs与改进后的BT-SVMs方法分别对新生儿的4类表情（安静、哭、轻度疼痛和剧烈疼痛）进行实验，根据每种多类算法选择SVM多类分类器最佳核函数，交叉验证各种算法的识别率，在考虑训练时间和识别率两方面的要求后，选择其中最理想的多分类算法作为新生儿疼痛面部表情的多类识别模型。

1.4 本文的章节安排

本课题主要是使用支持向量机（SVM）的方法来识别出新生儿的不同程度疼痛与非疼痛等多种面部表情，主要内容如下：

（1）回顾了新生儿疼痛与表情识别系统的研究状况，简要介绍了表情识别系统的三

个核心环节，在机器学习和统计学习的理论基础上详细介绍了SVM方法的理论和算法。

（2）详细的研究和分析了多分类SVM算法，尤其对基于二叉树结构的多分类SVM方法做了进一步的研究，并针对其结构上易产生误差累积的缺点，根据类间距离构造二叉树结构，并提出了基于最优分类面的类间分离度的方法对其改进，通过实验证明了改进算法的合理性以及可行性。

（3）采集新生儿图像，建立一个具有400幅新生儿图像的面部表情库。

（4）针对SVM模型选择难的问题，本文使用交叉验证与网格搜索相结合的方法来确定SVM的模型。

（5）分别对新生儿的安静、哭、疼痛时表现出的表情进行了多类识别与分析。通过对实验结果的分析得出如下结论：本文使用的SVM多分类方法针对新生儿疼痛面部表情多类识别能够获得很好的识别率。

本文的章节安排：

第一章 绪论。综述课题的选题依据与研究意义，国内外关于新生儿疼痛和表情识别方面的研究现状及发展趋势，以及基于 SVM 的表情识别，并介绍本文的主要研究内容和总体组织结构。

第二章 支持向量机的相关理论。简要介绍了机器学习的基本问题与方法、统计学习理论的基础；详细介绍支持向量机的理论：支持向量机在线性与非线性情形下获得最优分类面的理论知识、在线性不可分的情况下使用的核函数类型，以及算法的实现；并通过实验说明二分类 SVM 选择核函数与参数的过程。

第三章 多分类 SVM 算法研究。本章深入研究了现有的多种多分类 SVM 方法，分析各种算法的优缺点，对各个性能进行比较；详细介绍了基于决策树的多分类方法，并且针对基于二叉树结构的多分类 SVM 方法进行了有效的改进，最后通过仿真结果与真实数据证明了改进的合理性。

第四章 基于多分类 SVM 的新生儿疼痛表情多类识别。本章将多分类 SVM 方法运用到实际的新生儿疼痛表情识别与等级分类问题当中，应用二叉树多分类 SVM 方法及其改进算法建立了一个新的疼痛表情多类识别模型，在实验中与其它常见的多分类 SVM 方法相比，获得了更好的性能，通过实验数据证明了该结论。

第五章 总结与进一步展望。对本文工作进行归纳总结并展望了课题的进一步研究方向。

第二章 支持向量机的相关理论

2.1 统计学习理论

机器学习一直是一个非常热门的话题。而机器学习领域中的核心问题就是如何让机器有效的模仿人类的学习以及推理能力。换句话说，机器学习最主要的任务就是研究如何从已知数据（即样本集）中寻找规律，并利用这些规律对未知的或者无法观测的数据进行有效和准确的预测[59]。

传统统计学是绝大多数机器学习方法的重要理论基础。但在实际应用问题中，这些基于传统统计学的机器学习方法并没有取得令人信服的结果，主要有两个原因：

（1）由于传统统计学假定样本的分布形式是已知的，仅仅利用已知数据对给定形式的参数进行简单的预测，而在许多实际问题中，样本的分布形式是未知的，因此预测值与真实值之间可能有很大的偏差；

（2）传统统计学的另一个核心思想是渐近理论，它需要足够多的样本数目，这在实际当中往往也难以办到，所以进一步导致了预测结果的不理想。

因此产生了一种新的理论——统计学习理论，它是一种专门研究有限样本（甚至是小样本）情况下机器学习规律的理论，不再需要先假定样本的分布形式。由于这两个特点，使统计学习理论在许多实际问题中表现出了优于传统统计学的性能，从而迅速成为了研究有限样本情形下机器学习问题的最主要工具。在此基础上发展出一种新的模式识别方法——支持向量机（Support Vector Machine，简称SVM)[60,61]，该方法最早的出现是为了解决模式识别中的二值分类问题，而它在解决小样本、非线性及高维模式识别问题中表现出的许多特有的优势，使其成为了当前研究的热点。

SVM应用涵盖了三类问题：模式识别、回归分析和密度估计。本文主要研究模式识别问题，本章下面的几节将分别从机器学习、支持向量机理论及其算法实现几个方面来讨论。

2.1.1 机器学习问题的表示

机器学习的目的是利用有限数量的观测样本来估计一个系统的输入输出之间的依赖关系，使它能够对未知输出做出尽可能准确的预测。这种依赖关系称作机器学习，它能够

实现一定的函数集 $f(x,\omega),\omega\in\Psi,\Psi$ 为参数集合。

输入 x 与输出 y 之间的未知依赖关系可以看作是一个未知的联合概率分布 $F(x,y)$。根据联合分布 $F(x,y)$ 随机抽取 l 个独立同分布的观测样本

$$(x_1,y_1),(x_2,y_2),...(x_l,y_l) \tag{2.1}$$

组成训练集 T。学习问题就是在给定的函数集合 $\{f(x,\omega)|\omega\in\Psi\}$ 中选择出能够最好的逼近训练集 T 的最优函数 $f(x,\omega_0),\omega_0\in\Psi$，使期望风险泛函

$$R(\omega)=\int L(y,f(x,\omega))dF(x,y) \tag{2.2}$$

最小化。给定输入 x，$L(y,f(x,\omega))$ 是系统响应 y 与学习机器给出的预测输出 $f(x,\omega)$ 之间的损失。

对于模式识别问题，一般其输出 y 是样本类别标号，对于只有两种模式的情况 $y=\{0,1\}$ 或 $y=\{-1,1\}$，令 $\{f(x,\omega)|\omega\in\Psi\}$ 为指示函数集合。损失函数定义为：

$$L(y,f(x,\omega))=\begin{cases}0,\text{如果}y=f(x,\omega)\\1,\text{如果}y\neq f(x,\omega)\end{cases} \tag{2.3}$$

对于该损失函数，风险泛函 $R(\omega)$ 能够确定系统真实输出和指示函数 $f(x,\omega)$ 输出值不同的概率。指示函数的预测输出与系统输出不一致的情况称作分类错误。因此，关于模式识别的学习问题就是在联合概率分布 $F(x,y)$ 未知，给定训练集 T 的情况下，寻找使分类错误的概率最小的函数。

2.1.2 经验风险最小化

学习目标在于使期望风险泛函最小化，然而，实际上仅利用训练集 T 无法计算风险泛函 $R(\omega)$，期望风险由（2.2）式定义。很多传统的学习方法应用了经验风险最小化（Empirical Risk Minimization, ERM）原则来定义最小化的泛函，如最小二乘法、极大似然法等。根据概率论中大数定理的思想研究者提出了算术平均代替期望风险：

$$R_{emp}(\omega)=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(x_i,\omega)) \tag{2.4}$$

所谓ERM原则就是利用使经验风险 $R_{emp}(\omega)$ 最小的函数 $L(y,f(x,\omega_l))$ 逼近使期望风险 $R(\omega)$

最小的函数$L(y,f(x,\omega_0))$，ERM 原则的具体实现取决于特定的损失函数。

首先，$R_{emp}(\omega)$和$R(\omega)$都是ω的函数，概率论中的大数定理只说明了(在一定条件下)当样本趋于无穷多时$R_{emp}(\omega)$将在概率意义上趋近于$R(\omega)$，并没有保证使$R_{emp}(\omega)$最小的ω^*与使$R(\omega)$最小的ω^*是同一个点，更不能保证$R(\omega^*)$能够趋近于$R(\omega^*)$。

经验风险最小化原则通过最小化训练误差来实现最小化测试误差的目的。但是在实际问题求解中，样本数量是有限的。在训练样本数有限的情况下，基于经验风险最小化原则的机器学习算法普遍存在推广能力不足的问题，因此人们期望在样本有限时，经验风险最小化方法能得到较好的结果。

2.1.3 机器学习的 VC 维

为研究函数集在经验风险最小化原则下的学习一致性问题，统计学习理论定义了一系列有关函数集学习性能的指标，其中最重要的是VC维(Vapnik-Chervonenkis Dimension)。VC维是统计学习理论中的一个核心概念，也是对函数集学习性能的最好的描述指标。一般而言，VC维越大，学习机器的学习能力就越强，但学习机器也越复杂(容量越大)。函数集的VC维（而不是其自由参数个数）会影响学习机器的推广性能。为了说明VC维，本文用如下图进行说明：

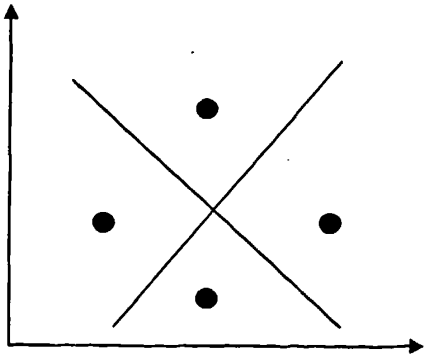


图 2-1

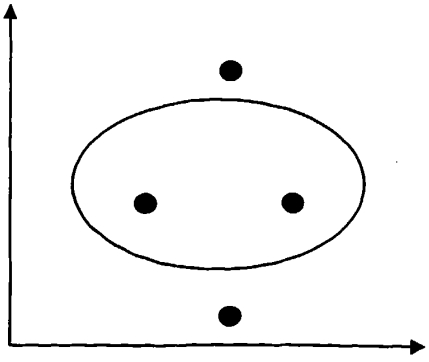


图 2-2

上图中图 2-1 中 4 个样本能完全被包含 2 个分类超平面的函数集分开，所以函数集的 VC 维等于 4。而图 2-2 中 4 个样本没有完全分开，图中函数能岔开 2 个样本，所以它的 VC 维等于 2。其实 VC 维是指对于一个预测函数集，若存在 h 个样本能够被函数集中的函数按所有可能的2^h种形式分开，则称函数集合能够把 h 个样本打散，函数集的 VC 维就是它能打散的最大样本数 h。也就是说，如果存在 h 个样本的样本集能够被函数集打散，而

不存在有 h+l 个样本的样本集能够被函数集打散，则函数集的 VC 维就是 h。如果对于任意的样本数，总能找到一个样本集能够被这个函数集打散，则函数集的 VC 维就是无穷大。应当指出，这里是存在 h 个样本的样本集能够被函数集打散，不是指任意 h 个样本的样本集能够被函数集打散。

VC维克服了“维数灾难”的问题，它以一个包含很多参数但却有较小的VC维的函数集为基础实现较好的推广性。目前尚没有通过的关于如何计算任意函数集的VC维理论，只有对一些特殊的函数集的VC维可以准确知道。对于给定的学习函数集，如何用理论或实验的方法计算它的VC维仍是当前统计学习理论中有待研究的一个问题。

2.1.4 结构风险最小化

统计学习理论从控制学习机器复杂度的思想出发，提出了结构风险(Structural Risk Minimization, SRM)最小化原则。该原则使得学习机器在可容许的经验风险范围内，总是采用具有最低复杂度的函数集。该理论针对小样本统计问题建立了一套新的理论体系，在这种体系下的统计推理规则不仅考虑了对渐近性能的要求，而且追求在有限信息的条件下得到最优结果。支持向量机方法就是在该理论的基础上发展起来的一种性能优良的机器学习方法。

根据统计学习理论中关于函数集的推广性能的讨论，对于指示函数集中所有函数，经验风险 $R_{emp}(\omega)$ 和实际风险 $R(\omega)$ 之间至少以概率 $1-\eta$ 满足如下关系：

$$R(\omega) \le R_{emp}(\omega) + \sqrt{\frac{h\left(\ln\left(2n/h\right)\right)+1-\ln\left(\eta/4\right)}{n}} \tag{2.5}$$

其中 h 是函数集的 VC 维，n 是样本总数，$R_{emp}(\omega)$ 是训练样本的经验风险，$\sqrt{\frac{h\left(\ln\left(2n/h\right)\right)+1-\ln\left(\eta/4\right)}{n}}$ 是置信范围。从（2.5）式可知：实际风险不仅与置信水平 $1-\eta$ 有关，而且与学习机器的 VC 维和训练样本数有关。在训练样本有限的情况下，学习机器的 VC 维越高，则置信范围就越大，导致实际风险与经验风险之间的误差可能就越大，这也是在一般情况下选用过于复杂的学习机器往往得不到好的效果的原因。如果要使结构化风险最小，就需要同时最小化经验风险和置信范围。在传统的机器学习中，选择学习机器模型和算法的过程就是优化置信范围的过程，如果选择了适合现有样本学习模型，就可以取得比较好的效果。因此在设计学习机器时，不但要使学习机器的经验风险最小，还要使 VC 维尽量小，以缩小置信范围，从而达到使期望风险最小的目的，也即对未来样本有较好的

推广性。

统计学习理论提出了一种新的策略来解决上述问题，也就是把函数集 $S=\{f(x,\omega),\omega\in\Psi\}$ 分解为一个函数子集序列 $S_1\subset S_2\subset\cdots S_k\subset\cdots S$，使各个子集按照置信范围的大小排列，也就是按照 VC 维的大小排列，即 $h_1\le h_2\le\cdots h_k\le\cdots h$，这样在同一个子集中置信范围就相同；在每一个子集中寻找最小化的经验风险，通常它随着子集复杂度的增加而减小。

基于以上的考虑，即把函数构造成为一个函数子集序列，使各个子集按照 VC 维的大小排列；在每个子集中寻找最小经验风险，在子集间折衷考虑经验风险和置信范围，取得实际风险最小，这种思想称为结构风险最小化，即 SRM 准则。为了使实际风险最小，需先设计函数集的某种结构使每个子集中都能取得最小的经验风险（如使训练误差为 0），然后只需选择适当的子集使置信范围最小，则这个子集中使经验风险最小的函数就是最优函数。结构风险示意图如图 2-3 所示：

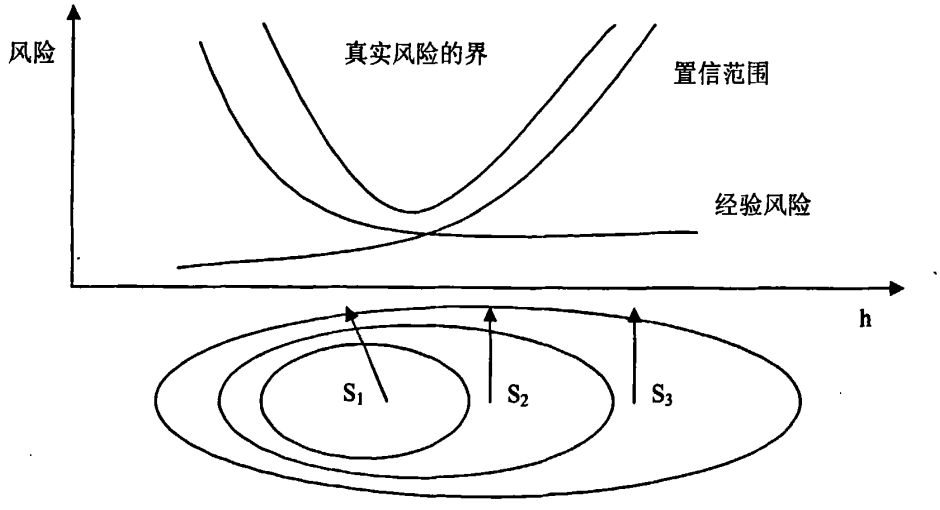


图 2-3 SRM 结构示意图

结构风险最小化原则提供了一种不同于经验风险最小化的更科学的机器学习设计原则，它最终是在式（2.5）的两个求和项之间进行折衷。实现此思想一般有两个思路：一是在每个子集中求得经验风险最小化，然后选择经验风险最小和置信范围之和最小的子集，但是此方法在子集数目很大甚至是无穷时是不可行的。第二种思路是设计函数集的某种结构使每个子集都能取得最小的经验风险，然后选择置信范围最小的子集，则这个子集中使经验风险最小的函数就是最优函数。支持向量机方法实际上就是这种思想的体现。

2.2 支持向量机概述

在上一节的统计学习理论的基础上，本节将介绍支持向量机基本原理。支持向量机(Support Vector Machine——SVM)是统计学习理论中最年轻的部分，该方法是建立在统计学习理论的VC维理论和结构风险最小化原理基础上的，根据有限的样本信息在模型的复杂性(即特定训练样本的学习精度)和学习能力(即无误识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力。其突出的优点表现在:(1)基于统计学习理论中结构风险最小化原则和VC维理论，具有良好的泛化能力，即由有限的训练样本得到的小的误差能够保证对独立的测试集仍保持小的误差。(2)支持向量机的求解问题对应的是一个凸优化问题，因此局部最优解一定是全局最优解。(3)核函数的成功应用，将非线性问题转化为线性问题求解。(4)分类间隔的最大化，使得支持向量机算法具有较好的鲁棒性。由于支持向量机自身的突出优势，因此被越来越多的研究人员作为强有力的学习工具，以解决模式识别、回归估计等领域的难题。

2.2.1 支持向量机理论

支持向量机的基本思想是：首先，在线性可分情况下，在原空间寻找两类样本的最优分类超平面。在线性不可分的情况下有两种方法，一是加入了松弛变量进行分析，二是通过使用非线性映射将低维输入空间的样本映射到高维属性空间使其变为线性情况，从而使得在高维属性空间采用线性算法对样本的非线性进行分析成为可能，并在该特征空间中寻找最优分类超平面。其次，它通过使用结构风险最小化原理在属性空间构建最优分类超平面，使得分类器得到全局最优，并在整个样本空间的期望风险以某个概率满足一定上界。

1. 线性可分的最优分类面

支持向量机是从线性可分情况下的最优分类面发展而来的。对两类问题，设训练样本集$(x_i, y_i), i=1,2,...,n$，n为训练样本个数，$x_i \in R^d$为训练样本，$y_i \in \{-1,+1\}$为输入样本x_i的类标记(期望输出)。SVM的出发点是寻找最优分类超平面，其基本思想可用图2-4的两类情况说明。

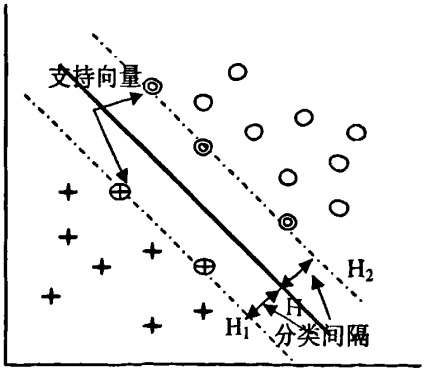


图2-4 最优分类面示意图

图2-4中，○和＋分别代表两类样本，H为分类面（线），H_1, H_2分别为过各类中离分类面（线）最近的样本且平行于分类面（线）的平面（直线），它们之间的距离叫做分类间隔（margin）。所谓最优分类面（线）就是要求分类面（线）不但能将两类正确分开（训练错误率为0，保持经验风险 $R_{emp}(\omega)$=0），而且还能使两类间的分类间隔最大（推广能力比较强）。前者是保证经验风险最小（为0），而通过后面的讨论可以看到，使分类间隔最大实际上就是使推广性的界中的置信范围最小，从而使真实风险最小。H_1, H_2上的训练样本点就称作支持向量，它们距离最优超平面（线）最近。

如样本集线性可分，d维空间中线性判别函数 $g(x)=\omega\cdot x+b$，分类面方程为 $\omega\cdot x+b=0$。我们可以对判别函数进行归一化使得对线性可分的样本集 (x_i, y_i)，$x_i \in R^d$，$i=1,2,...,n$，$y_i \in \{-1,+1\}$ 满足 $|g(x)|\geq 1$，即使离分类面最近样本的 $|g(x)|=1$。

若分类面对所有样本都能够正确分类，则满足：

$$y_i(\omega\cdot x_i+b)-1\geq 0,\quad i=1,2,...,n \tag{2.6}$$

易证，两个超平面H_1：$\omega\cdot x+b=1$和H_2：$\omega\cdot x+b=-1$间的间隔等于$2/\|\omega\|$，使边际间隔最大等价于使$\|\omega\|$(或$\|\omega\|^2$)最小。

所以，分类超平面H：$\omega\cdot x+b=0$为最优，当且仅当(ω,b)是下面优化问题的解：

$$\min:\ \frac{1}{2}\|\omega\|^2 \tag{2.7}$$

$$\text{subject to: } y_i(\omega\cdot x_i+b)-1\geq 0,\quad i=1,2,...,n \tag{2.8}$$

引入Lagrange函数：

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{n}\alpha_i\left(y_i(\omega\cdot x_i+b)-1\right) \tag{2.9}$$

其中$\alpha=(\alpha_1,\cdots,\alpha_n)^T\in R^n$为Lagrange乘子。先求Lagrange函数关于ω和b的极小值，即对ω和b求偏微分得：

$$\frac{\partial L}{\partial \omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \tag{2.10}$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.11}$$

将（2.10）与（2.11）代入（2.9）中，得到其对偶问题，即：

$$\min_{\alpha}:\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{n}\alpha_i \tag{2.12}$$

$$\text{subject to:}\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.13}$$

$$\alpha_i\geq 0, i=1,2,...,n \tag{2.14}$$

这是一个不等式约束下二次函数寻优的问题，存在唯一解$\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)^T$，解中将只有一部分（通常是少部分）$\alpha_i^*$不为零；此时原始问题(2.7)、(2.8)的解为$(\omega^*,b^*)$。

根据最优化问题的KKT(Karush-Kuhn-Tucher)条件[62]有：

$$\alpha_i^*[y_i(\omega^*\cdot x_i+b^*)-1]=0 \tag{2.15}$$

因此，$\alpha_i^*>0$所对应的样本满足式（2.8）中的等式约束，这就是支持向量。支持向量距最优超平面最近，通常只是全体样本中的很少的一部分，是对分割两类非常重要的样本点。

求得最优解$\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)^T$，则

$$\omega^*=\sum_{i=1}^{n}\alpha_i^* y_i x_i, \tag{2.16}$$

选择α^*的一个正分量α_j^*，据此计算分类阈值b^*：

$$b^*=y_j-\omega^* x_j=y_j-\sum_{i=1}^{n}\alpha_i^* y_i\left(x_i\cdot x_j\right); \tag{2.17}$$

得到的最优分类决策函数为：

$$f(x)=sgn\left((\omega^*\cdot x)+b^*\right)=sgn\left(\sum_{i=1}^{n}\alpha_i^* y_i(x\cdot x_i)+b^*\right) \tag{2.18}$$

其中 $sgn(x)=\begin{cases}1 & x>0\\ -1 & x\le 0\end{cases}$，式中的求和实际上只对支持向量求和。

2. 线性不可分的最优分类面

若训练数据集线性不可分，即某些样本不能满足（2.8）式的条件，可以在条件（2.8）中增加一个松弛项 $\xi_i\ge 0$，$i=1,2,...,n$，约束条件（2.8）变为：

$$y_i\left((\omega\cdot x_i)+b\right)\ge 1-\xi_i,\quad i=1,2,...,n, \tag{2.19}$$

显然，当 ξ_i 充分大，样本点 (x_i,y_i) 当总可以满足上述约束条件。但应该避免 ξ_i 取太大的值。所以需在目标函数里对它们进行惩罚，比如在目标函数中加入含有 $\sum_i \xi_i$ 的一项，这导致最优问题改为：

$$\min_{\omega,b,\xi}:\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i \tag{2.20}$$

$$\text{subject to: } y_i[(\omega\cdot x_i)+b]\ge 1-\xi_i,\quad i=1,2,...,n, \tag{2.21}$$

$$\xi_i\ge 0, i=1,2,...,n \tag{2.22}$$

其中，$C>0$ 是一个惩罚参数，控制对错分样本惩罚的程度。

引入Lagrange函数有：

$$L(\omega,b,\xi,\alpha,r)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(\omega\cdot x_i+b)-1+\xi_i]-\sum_{i=1}^{n}r_i\xi_i \tag{2.23}$$

其中，α_i，r_i 为Lagrange乘子，$\alpha_i\ge 0$，$r_i\ge 0$，由此得到

$$\frac{\partial L}{\partial \omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \tag{2.24}$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.25}$$

$$\frac{\partial L}{\partial r_i}=C-\alpha_i-r_i=0 \tag{2.26}$$

将（2.24），（2.25），（2.26）代入（2.23），对 α 求极大，得对偶问题：

$$\min_{\alpha}: \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{n}\alpha_i, \tag{2.27}$$

$$\text{subject to:}\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.28}$$

$$0\le\alpha_i\le C, \quad i=1,2,...,n \tag{2.29}$$

最优化求解得到的α_i^*可能的值是：①$\alpha_i^*=0$②$0<\alpha_i^*<C$③$\alpha_i^*=C$，后两者所对应的$\boldsymbol{x}_i$为支持向量。由（2.16）式可知只有支持向量对$\boldsymbol{\omega}$有贡献，也就是对最优超平面、决策函数有贡献，对应的学习方法称之为支持向量机。在支持向量中，③所对应的$\boldsymbol{x}_i$称为边界支持向量(Boundary Support Vector, BSV)，实际上是错分的训练样本点，②所对应的$\boldsymbol{x}_i$称为标准支持向量(Normal Support Vector, NSV)。

$$\text{计算}\boldsymbol{\omega}^*=\sum_{i=1}^{n}\alpha_i^* y_i\boldsymbol{x}_i \tag{2.30}$$

选择$\boldsymbol{\alpha}^*$的一个正分量$0<\alpha_j^*<C$，据此计算

$$b^*=y_j-\sum_{i=1}^{n}\alpha_i^* y_i\left(\boldsymbol{x}_i\cdot\boldsymbol{x}_j\right), \tag{2.31}$$

由此求得决策函数$f(\boldsymbol{x})=sgn\left((\boldsymbol{\omega}^*\cdot\boldsymbol{x})+b^*\right)$。

但是这种方法并不适用于所有的线性不可分情况。当不满足约束条件(2.6)的样本点超过一定数量的时候，虽然能够通过选取足够多并且足够大的松弛变量，使所有的样本点都能够满足“放宽的”约束条件(2.19)，但此时的错误率已经大大超出了允许的范围。因此需要通过其他的途径来解决该问题，但最大间隔超平面的思想仍然值得借鉴。

2.2.2 核函数

当训练样本集在原始输入空间中线性不可分时，根据核方法的思想，可以将原始空间中的样本映射到高维空间，并在高维空间中寻找能线性划分原始数据的超平面。引入从输入空间R^d到一个高维 Hilbert 空间$\boldsymbol{H}$的非线性映射变换

$$\Phi:\ \begin{matrix}\chi\subset R^d\to \mathrm{X}\subset H\\ \boldsymbol{x}\to \mathbf{x}=\Phi(\mathbf{x})\end{matrix}, \tag{2.32}$$

通过这样的映射，原来的样本集变成了$(\Phi(x_i),y_i),i=1,2,...,n$，而分类超平面变成了$\omega\cdot\Phi(x)+b=0$。这个分类面虽然在输入空间$R^d$中是非线性的，但是在高维 Hilbert 空间 H 中却可以被看作是线性的，因此可以认为它是一种广义线性超平面。

此时最优化问题的对偶问题为：

$$\min_{\alpha}:\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\Phi(x_i)\cdot\Phi(x_j)\right)-\sum_{i=1}^{n}\alpha_i, \tag{2.33}$$

$$\text{subject to:}\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.34}$$

$$0\le\alpha_i\le C,\quad i=1,2,...,n \tag{2.35}$$

已知线性支持向量机的判决函数仅仅依赖于内积(x_i,x_j)和(x_i,x)，$i,j=1,\cdots,n$，非线性不可分支持向量机的最终判定函数也仅仅依赖于空间变换后的内积$\left(\Phi(x_i),\Phi(x_j)\right)$和$\left(\Phi(x_i),\Phi(x)\right)$，$i,j=1,\cdots,n$。在许多问题当中，真正的求出从$R^d$到 H 的非线性映射$\Phi$并不是一件容易的事，通过$\Phi$来计算内积$\left(\Phi(x_i),\Phi(x_j)\right)$在很多时候都难以实现，于是就提出了核函数的方法。

核函数K是一个关于x_i和x_j的函数，它对所有的$x_i,x_j\in R^d$均满足：$K(x_i,x_j)=\Phi(x_i)\cdot\Phi(x_j)$。其中$\Phi$是从输入空间$R^d$到高维 Hilbert 空间 H 的非线性映射。根据泛函的有关理论，只要某个核函数满足 Mercer 条件[55]，那它就对应某空间中的内积。因此可以用核函数$K(x_i,x_j)$代替内积$\left(\Phi(x_i),\Phi(x_j)\right)$得到一般支持向量机的目标函数：

$$\min_{\alpha}:\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{n}\alpha_i, \tag{2.36}$$

$$\text{subject to:}\sum_{i=1}^{n}\alpha_i y_i=0 \tag{2.37}$$

$$0\le\alpha_i\le C,\quad i=1,2,...,n$$

同样求得最优解$\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)^T$，选择$\alpha^*$的一个正分量$0<\alpha_j^*<C$，据此计算

$$b^*=y_j-\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x_j), \tag{2.38}$$

由此求得决策函数

$$f(x)=sgn\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x,x_i)+b^*\right)。\tag{2.39}$$

只要满足Mercer条件的函数在理论上都可选为核函数，但不同的核函数，其分类器的性能完全不同。因此，针对某一特定问题，核函数的类型选择是至关重要的。目前常用的核函数主要有：

（1）　$K(x,x_i)=x\cdot x_i$　　(2.40)

（2）　$K(x,x_i)=[(x\cdot x_i)+1]^d$　　(2.41)

（3）　$K(x,x_i)=\exp\{-\frac{|x-x_i|^2}{\delta^2}\}$　　(2.42)

（4）　$K(x,x_i)=\tanh(\nu(x\cdot x_i)+C)$　　(2.43)

（5）　$K(x_i,x_j)=1+(x_i^T\cdot x_j)+\frac{1}{2}(x_i^T\cdot x_j)\min(x_i^T\cdot x_j)^2-\frac{1}{6}\min(x_i^T\cdot x_j)^3$　　(2.44)

其中，式（2.40）对应于线性核函数支持向量机；式（2.41）为多项式核函数，对应的支持向量机在样本空间中的分类面为多项式曲面；式（2.42）为高斯径向基核函数，对应于高斯径向基分类器，与传统的高斯径向基分类器相比，其不仅具有良好的泛化能力，而且能自动确定其各个参数；径向基形式的内积函数类似人类的视觉特性，在实际应用中经常用到；式（2.43）为Sigmoid（S形）核函数，得到的SVM是一个两层的感知器网络，其网络的权值、隐层节点数目都是由算法自动确定，而不像传统的感知器网络那样凭借个人经验确定；式（2.44）为样条核函数。

在实际分类应用中，人们发现径向基核函数表现出的性能要优于其它几种核函数。B.Baesens和S.Viaene等人曾利用LS-SVM分类器，采用UCI数据库，对线性核函数、多项式核函数和径向基核函数进行了实验比较，从实验结果来看，对不同的数据库，不同的核函数都存在优劣，而径向基核函数在多数数据库上得到了略为优良的性能。但是需要注意的是，这些比较大多数是建立在训练正确率的基础上。而判断一个支持向量机分类器性能的关键指标有两个，即学习能力和推广能力。其中推广能力的强弱更能反映分类器性能的好坏，因为设计分类器的目的就是对未知数据进行分类。所以，有必要对核函数在这两方面的能力做进一步的探究。

2.2.3 算法实现

求解支持向量机问题实际上是一个约束最优化问题，确切地说是一个带有线性约束的凸二次规划问题，可以利用非常成熟的二次规划求解方法，也可以把它转化为一个或一系

列无约束最优化问题。该问题的一个显著特征是其规模与训练集大小的平方成正比。当问题规模不大时，有很多传统的优化算法，如最速下降法、牛顿法、共轭梯度法等[63,64]。在实际处理大规模问题时，这些算法占用内存太多训练时间很长，已不再适用。但是由于最优化问题的解的稀疏性，即最优解中只有一小部分 Lagrange 参数α_i非零，于是考虑存储和计算量两方面的要求，可将大规模的原问题分解成若干小规模的子问题加以解决。此处简单介绍具有代表性的几种根据这些思想对支持向量机的求解方法。

1. 选块法（chunking）

选块算法的基本思想是，若预先知道在最优解中哪些样本对应的 Lagrange 乘子α_i的值为 0（即非支持向量），则可在支持向量训练过程中忽略这些样本，从而大大降低最优化问题的规模。由于实际上支持向量是未知的，选块算法需要通过某种迭代方式逐步排除非支持向量，选出支持向量所对应的“块”。选块算法的工作流程如下：

（1）从训练集中任意选一个子集或者“块”作为工作集；

（2）在工作集上应用某种优化算法训练支持向量机（求解最优化问题的对偶问题），得到 Lagrange 参数$\boldsymbol{\alpha}$，保留该块中与非零α_i对应的训练样本，即当前支持向量；

（3）利用得到的决策函数来检测训练集中除去该块后的其它样本，记录被错误分类的样本，即当前误差样本；

（4）以当前支持向量和当前误差样本组成新的工作集；

（5）重复步骤（2）~（4）直至算法收敛。此时工作集包含了所有的支持向量。

该方法的优点是当支持向量的数目远远小于训练样本数时，能大大提高运算速度。然而，选块方法的目标是找出所有的支持向量，因而最终需要存储相应的核函数矩阵，由此可见，当支持向量很多时，随着算法迭代次数的增多，所选的块也会越来越大，算法的执行将会变慢，选块算法会遇到困难。

2. 分解法（decomposing）

与选块法类似，分解法也是采用分解策略，但是与选块法的不同之处在于，其优化问题的规模是固定的，与支持向量的总数无关。

若将原训练集分为工作集 B 和空闲集 N，则原优化问题分解为在 B 和 N 上的两个子优化问题，Lagrange 参数$\boldsymbol{\alpha}$也分解为$\{\alpha_i \mid i \in \mathrm{B}\}$和$\{\alpha_j \mid j \in \mathrm{N}\}$两部分，且空闲集 N 固定不变。在求解定义在工作集 B 上的子优化问题后，将任一$\alpha_j, y_j f(x_j) < 0, j \in \mathrm{N}$和$\alpha_i, i \in \mathrm{B}$互换，可得以新的子优化问题，对此最优化问题的求解将使定义在原训练集上的优化问题得到进

一步优化。于是分解法步骤如下：

（1）随机选取原始训练集中的一个小子集作为工作集B，剩余样本组成空闲集N；

（2）求解定义在B上的子优化问题；

（3）检测空闲集N，若某个$\alpha_j, j \in$ N违反 KKT 条件（即$y_j f(x_j) < 0$），将其与任一$\alpha_i, i \in$ B交换，得到新的B和N。

（4）重复步骤（2）和（3）至算法收敛，N中所有的样本都满足KKT条件。

上述算法可处理任一规模的训练集，但由于每次只替换一个样本，导致收敛速度很慢。在实际应用中，通常使用复杂的启发式规则每次替换多个样本。

3. 序列最小最优化（Sequential Minimal Optimization,SMO）算法

序列最小优化算法是分解算法中选取B=2的特殊情况，即每次迭代只求解一个具有两个 Lagrange 参数α_i和α_j的最优化问题，这时工作集规模已减到最小。其优点在于，优化问题只有两个 Lagrange 参数，它用分析的方法即可解出，从而完全避免了复杂的数值解法；另外，它根本不需要巨大的矩阵存储，这样，即使是很大的 SVM 学习问题，也可在普通 PC 机上实现。SMO 方法包括两个步骤：一是求解这两个 Lagrange 参数优化问题的分解步骤；二是如何选择这两个 Lagrange 参数。

假设当前（非优）解为$\alpha_1^{old}, \alpha_2^{old}, \alpha_3, \cdots, \alpha_l$（初始化时令$\alpha = 0$），需要进一步优化的参数为$\alpha_1$和$\alpha_2$（对应训练样本$x_1$和$x_2$）。根据线性约束（2.34）可得：

$$y_1\alpha_1 + y_2\alpha_2 = y_1\alpha_1^{old} + y_2\alpha_2^{old} = constant, \tag{2.44}$$

不失一般性，首先求α_2：

$$\alpha_2^{new} = \alpha_2^{old} - \frac{y_2(E_1 - E_2)}{\eta}, \tag{2.45}$$

其中：

$$E_i = f^{old}(x_i) - y_i, \quad i = 1,2, \tag{2.46}$$

$$\eta = 2K(x_1, x_2) - K(x_1, x_1) - K(x_2, x_2), \tag{2.47}$$

α_2当前的取值范围$\{L, H\}$为：

$$\text{当 } y_1 \neq y_2 \text{ 时，} L = \max(0, \alpha_2^{old} - \alpha_1^{old}), H = \min(C, C + \alpha_2^{old} - \alpha_1^{old}), \tag{2.48}$$

$$\text{当 } y_1 = y_2 \text{ 时，} L = \max(0, \alpha_2^{old} + \alpha_1^{old} - C), H = \min(C, \alpha_2^{old} + \alpha_1^{old}), \tag{2.49}$$

因此，α_2的当前最优值为：

$$\alpha_2^{new,clipped}=\begin{cases}H, & \alpha_2^{new}\geq H\\ \alpha_2^{new}, & L<\alpha_2^{new}<H\\ L, & \alpha_2^{new}\leq L\end{cases}, \tag{2.50}$$

继而，可求得α_1的当前最优值为：

$$\alpha_1^{new}=\alpha_1^{old}+y_1y_2\left(\alpha_2^{old}-\alpha_2^{new,clipped}\right), \tag{2.51}$$

对于SMO算法中优化参数α_1和α_2的选取有如下规则：第一个参数选择任一违反KKT条件的Lagrange参数作为α_1；第二个参数选择另一个Lagrange参数作为α_2，对该参数和α_1的更新应尽可能优化原目标函数。一个简单的方法是选取α_2使$\|E_1-E_2\|$最大。

SMO算法实现简单，收敛速度快，内存需求小，是目前最好的支持向量训练算法之一。本文所有实验中的SVM分类器均采用此算法。

2.3 仿真实验

由于SVM核函数与参数选择的目标是为了得到高的训练集的检测精度即较好的预测未知数据，但识别时更关注后一种即推广性。训练集的检测精度高并不意味着对测试集的识别率也高，因为对机器学习而言，对训练集过高的精度有可能导致过学习的问题。至今都没有找到一种比较好的方法来选择核函数与参数，所以只有通过实验的方法来比较不同的参数获得的识别效果来确定参数值。

本节在Microsoft Visual C++6.0环境下，结合三组数据，对二分类支持向量机进行了分类实验，这三组数据样本分别为来自网上公用的数据库UCI（www.ics.uci.edu/~mlearn/databases）的Ionosphere数据集和Sonar数据集，Sonar数据集共有208个60维的样本，Ionosphere数据集共有351个34维的样本。

实验分为两部分：第一部分采用不同的核函数对Ionosphere数据集、Sonar数据集进行训练和测试，比较不同核函数SVM的性能；第二部分采用高斯径向基函数作为核函数，对三组数据测试不同惩罚参数C对SVM的影响。

实验步骤如下：（1）将实验中所用样本分成个数相等的10份；（2）第一次将第1份用于测试，其它的9份用于训练，获得此次的识别率，例如在50个测试样本中40个被正确分类，则识别率为40/50=80%；第二次将第2份用于测试，其它的9份用于训练，记录

此时的识别率；依此类推，将此过程重复 10 次；（3）再将 10 次实验中得到的识别率取平均，获得最终的识别率。

1. 各种核函数的比较

核函数对支持向量机的性能有很重要的影响，选择合适的核函数可以提高支持向量机的性能。在用 SVM 解决实际问题时，需要尝试多个核函数，然后根据测试结果决定采用哪一个核函数。下面我们给出一组实验结果，分别从训练和测试的时间、识别率和支持向量数三方面观察核函数的影响，实验中交叉验证的运行时间是训练时间和测试时间之和。

表 2-1 对 Ionosphere 数据集的实验

实验 1	线性核函数	多项式核函数 d=2	多项式核函数 d=3	多项式核函数 d=4	RBF 核函数	S 形核函数
运行时间(ms)	515	953	1016	1094	500	609
识别率	87.46%	93.16%	92.59%	92.59%	95.44%	88.32%
支持向量数	97	85	96	97	78	138

表 2-2 对 Sonar 数据集的实验

实验 2	线性核函数	多项式核函数 d=2	多项式核函数 d=3	多项式核函数 d=4	RBF 核函数	S 形核函数
运行时间(ms)	422	859	875	937	510	438
识别率	79.33%	88.46%	89.42%	89.42%	91.35%	77.40%
支持向量数	89	83	96	108	107	138

从上表中,综合考虑时间、识别率和支持向量数目，可以发现采用高斯径向基核函数的 SVM 其性能最佳，它的识别率最好。

2. 参数的选择实验

在选定高斯径向基函数作为 SVM 核函数后，还有两个参数需要选择：一个是惩罚参数 C，另一个是核函数 $K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\{-\frac{|\boldsymbol{x}-\boldsymbol{x}_i|^2}{\delta^2}\}$ 中的参数 δ。本文中在选择这两个参数时，使用了交叉验证与网格搜索相结合的方法，这两种方法在第四章中具体介绍。

确定参数 (C,δ) 的过程：先选定一组 (C,δ) 的范围，将它们准确率连接起来绘图，然后确定准确率出现最高的一段小步长，逐步缩小选择范围，最终确定准确率最高的那对参数。例如，对 $C=2^{-5},2^{-3},...,2^{15},\sigma=2^{-15},2^{-13},...,2^{-5}$ 分别依次进行验证，最终可以得到一组 (C,δ) 值，使得识别率达到最高，这种方法就叫做网格搜索法。根据实验，我们发现不断

验证以指数增长的(C,δ)值，是确定最佳参数的一个有效的方法。

我们对 Ionosphere 和 Sonar 数据集搜索最佳参数(C,δ)，在实验中我们用 g 代表 δ，参数选择的过程绘制了图形表示，图中不同颜色的线代表了不同识别率下的(C,δ)，其中绿色线所代表的识别率最高。

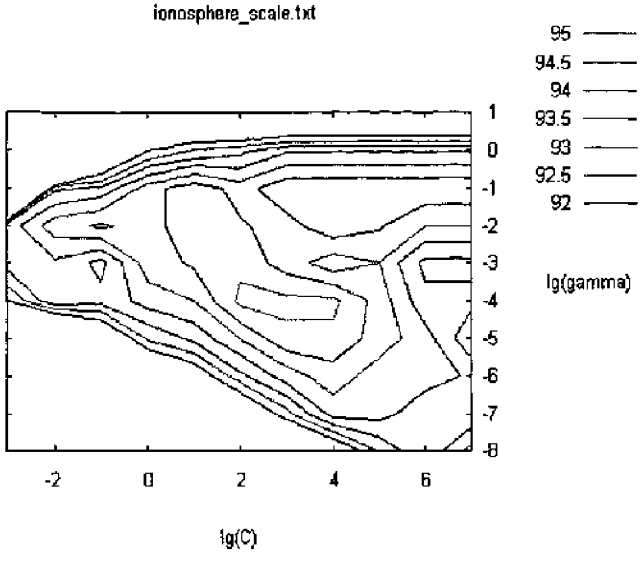


图 2-5 Ionosphere 数据集参数选择

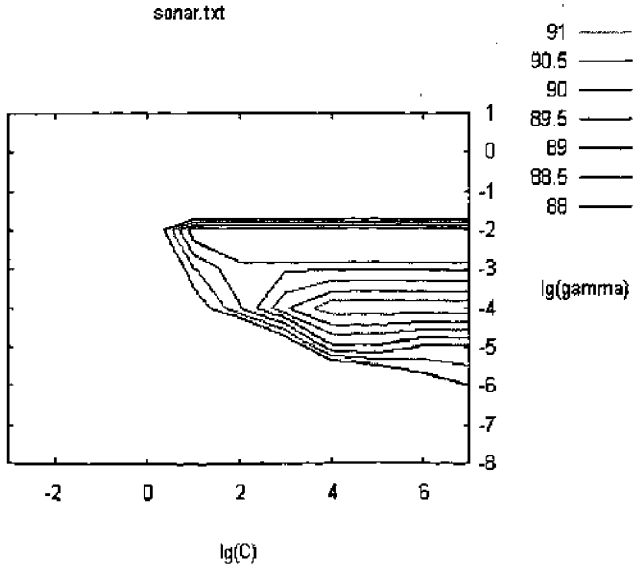


图 2-6 Sonar 数据集参数选择

当数据集 Ionosphere 参数$(C,\delta)=(8.0,0.0625)$时达到最好的识别率 95.44%，当数据集 Sonar 参数$(C,\delta)=(32,0.00625)$时达到最好的识别率 91.34%。

2.4 本章小结

本章概括介绍了统计学习理论的基本内容，并在此基础上详细分析了支持向量机方法的主要思想，包括对统计学习理论核心内容的阐述、对支持向量机算法由线性可分情况到一般的非线性情况的推导，以及核函数在 SVM 中的应用。统计学习理论和支持向量机的方法是以 VC 维理论和推广性的界为主要思想，较好的解决了小样本的学习问题。在实际应用中，SVM 将分类问题视为一个不等式约束下的二次规划问题，并用核函数代替了高维空间的非线性映射，具有全局最优的特点。

简要介绍了三种解决支持向量机最优化问题的分解算法，分解算法是针对大规模训练问题提出的，目前的 SVM 研究和应用中训练算法普遍采用的是分解算法。

最后进行了仿真实验说明 SVM 核函数及其参数的选择，在实验过程中证明采用高斯径向基函数的 SVM 分类器的性能最佳。

第三章　多分类SVM算法研究

3.1 多分类方法介绍

经典的支持向量分类机算法是针对两类的分类问题提出的，理论已经比较完善[65]，而在实际应用中，多类分类题更为普遍，例如人脸识别、语音识别，手写体数字识别问题等。因此如何将SVM方法有效的应用于多类模式识别问题中，已成为支持向量机研究的一大热点。目前存在的多分类支持向量机（SVMs）方法大致可以为两大类:

第一类方法是构造一系列的两类分类器，然后按照某种规则将它们组合起来完成多类问题的分类，类方法包括一对一（One-again-one）SVMs方法、一对多（One-again-the rest）SVMs方法、DDAG-SVMs方法、纠错码SVMs方法和SVMs决策树方法等。

第二类方法是在所有的训练样本基础上直接求解多分类决策函数，一次性将多类目标区分开。这方面的研究有Weston和Watkins(1998)提出的对两类SVM分类算法直接进行推广的方法。这类方法在思想上比第一类方法简单，但求解这样一个大的多类二次规划问题，往往会造成计算复杂度大大加，从而训练时间很长。

本章将对不同的多分类SVM算法进行具体的讨论，并对其在时间与识别率上的性能做了详细的分析。

3.1.1 一对一（one-again-one）SVMs

一对一多分类支持向量机（one-again-one SVMs）[66,67]是利用两类SVM算法在任意两类训练样本之间都构造一个最优决策面。对k类问题这种方法需要构造$k(k-1)/2$个分类平面（$k>2$）。这种方法的本质与两类SVM并没有区别，它相当于将多类问题转化为多个两类问题来求解。具体构造分类平面的方法如下：

从样本集中取出所有满足$y_i=s$与$y_i=t$的样本点（其中$1\le s,t\le k,s\ne t$），通过两类SVM算法构造最优决策函数：

$$f_{st}(x)=\omega_{st}\bullet\Phi(x)+b_{st}=\sum_{i=SV}\alpha_i^{st}y_iK(x_i,x)+b_{st}。\qquad(3.1)$$

用同样的方法对k类样本中的每一对构造一个决策函数，由于$f_{st}(x)=-f_{st}(x)$，容易知道

一个k类问题需要$k(k-1)/2$个分类平面。

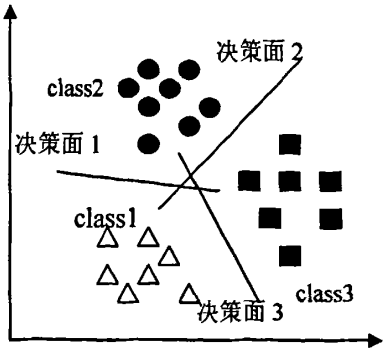


图3-1 一对一SVMs示意图

在预测时，常采取投票机制[66,68]：给定一个测试样本x，将样本x依次代入上述$k(k-1)/2$个决策函数，对每个决策函数$f_{st}(x)$，若判定x属于第s类则第s类获得一票，其计数器u_i加一，最后得票数最多的类别就是最终x所属的类别，公式如下：

$$class\ i = arg\ max\{u_1, u_2, \cdots, u_k\}。 \tag{3.2}$$

一对一SVMs方法的优点在于每次投入训练的样本相对较少，因此单个决策面的训练速度较快，同时精度也较高。但由于k类问题需要训练$k(k-1)/2$个决策面，当k较大的时候决策面的总数过多，影响预测速度。在投票机制方面，如果获得的最高票数的类别多于一类时，将产生不确定区域。

3.1.2 一对多（one-again-rest）SVMs

一对多多分类支持向量机（one-again-rest SVMs）[71]是在一类样本与剩余的多类样本之间构造决策平面，将一个k类问题转化为k个两分类问题，从而达到多类识别的目的，可以认为是两类SVM方法的推广。这种方法只需要在每一类样本和对应的剩余样本之间产生一个最优决策面，因此对一个k类问题，仅需要构造k个分类平面（$k>2$）。实际上它是将剩余的多类看成一个整体，然后进行k次两类识别。具体方法如下：

假定将第j类样本看作正类（$j=1,\cdots,k$），而将其它$k-1$类样本看作负类，通过两类SVM方法求出一个决策函数：$f_j(x)=\omega_j\cdot\Phi(x)+b_j=\sum_{i=SV}\alpha_i^j y_i K(x_i,x)+b_j$。这样的决策函数$f_j(x)$一共有$k$个。给定一个测试输入$x$，将其分别带入$k$个决策函数并求出函数值，若在$k$个$f_j(x)$中$f_s(x)$最大，则判定样本$x$属于第$s$类。

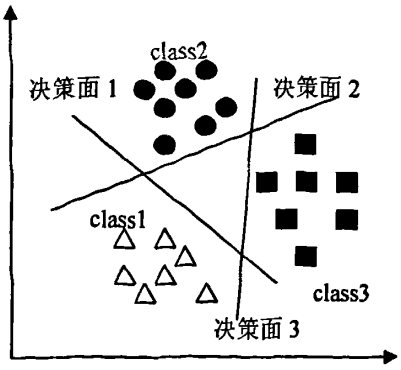


图3-2 一对多SVMs示意图

一对多SVMs方法和一对一SVMs相比，构造的决策平面数大大减少，因此在类别数目 k 较大时，其预测速度将比一对一SVMs方法快。但由于它每次构造决策平面的时候都需要用上全部的样本集，因此它的训练时间并不一定比一对一SVMs少。同时由于训练的时候总是将剩余的多类看作一类，正类和负类的训练样本数目极不平衡，这很可能影响预测时的识别率。另外，与一对一SVMs方法类似，当同时有几个 j 能取到相同的最大值 $f_j(x)$ 时，将产生不确定区域。所以这种方法的推广能力比较差；训练时间长，效率比较低，还存在决策盲区。

3.1.3 决策导向无环图（DDAG）SVMs

决策导向无环图(Dicision Directed Acyclic Graph，DDAG)算法是基于一对一 SVMs 算法提出来的一种新的学习构架。DDAG-SVMs 方法[69]与前面两种方法均不太一样，一对一 SVMs 和一对多 SVMs 通过一系列的决策函数来依次确定样本的类别，它们可以被认为是“肯定型”的算法，而 DDAG-SVMs 却是通过在每层节点处对不符合要求的类别进行排除，最后得到样本所属的类别，应该算是一种“否定型”的算法。具体算法如下：

对 k 类问题，首先利用一对一思想构造 $k(k-1)/2$ 个分类器，再结合 DDAG 将 $k(k-1)/2$ 个分类器按一定的次序构成一个有根的前向无环图 [73]，图 3-3 是使用 DDAG-SVMs 算法解决一个四类问题的完全示意图。在训练阶段 DDAG-SVMs 和一对一 SVMs 方法的步骤一样，它们的差别主要体现在预测阶段。DDAG-SVMs 首先将未知样本 x 代入根节点的决策函数进行判定，如图 3-3，此决策函数将第一类作为正样本，第四类作为负样本进行训练得到，若在此决策函数得到的值为-1，即样本 x 不属于第一类，那么将所有与第一类样本相关的决策函数全部排除，从根节点左侧继续执行判决；若在此决策函数值为 1，即样本 x 不属于第四类，那么将所有与第四类样本相关的决策函数全部排除，从根节点右侧继续执

行判决，依此类推直到到达叶节点，决出样本x的最终类别。

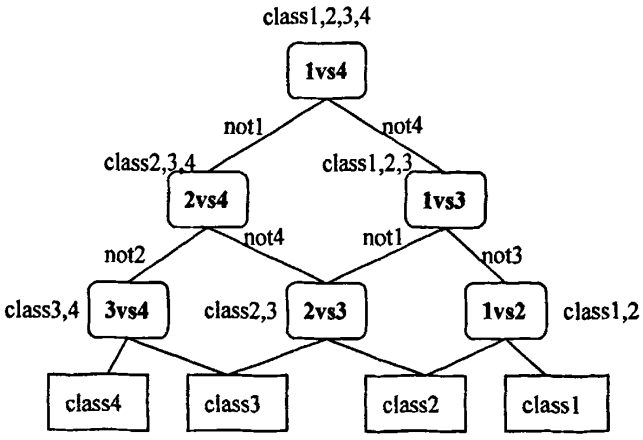


图 3-3 决策导向无环图(DDAG)算法示意图

该方法和一对一 SVMs 一样，训练的时候首先需要构造$k(k-1)/2$个分类决策面。然而和一对一方法不同的是，在每个节点预测时排除了许多类别的可能性，因此预测时用到的总分类平面只有$k-1$个，预测速度自然提高不少。但由于 DDAG-SVMs 算法采取的是排除策略，所以根节点的判别准确率对整个决策树尤为重要。若开始就决策错误，则后面的步骤就没有意义。因此如何选取判定顺序和开始阶段的判别是值得进一步研究的问题。

3.1.4 纠错编码 SVMs

纠错编码[70](Error Correcting Codes，ECC)是一种把多类分类问题进行二进制编码转化为多个两类分类问题的方法。其与SVM的结合形式主要是根据实际问题，对于k类分类问题，给每个类别赋予一个长度为L的二进制编码，形成一个k行L列的码本(下表)。对于其中第$i(i=1,\cdots,L)$列，将该位中码字为“0”的所有类别作为一类，其他码字为“1”的类别归为另一类，因此每个码位对应一个两类分类问题，k类分类问题就转化为L个两类分类问题。采用具有纠错能力的编码对类别进行编码，并将SVM作为码位分类器，这种多分类SVM方法被称为纠错编码支持向量机(ECC-SVM)[71,72]。当有未知样本x需要预测分类时，将其分别按顺序代入L个SVM决策函数，并根据L个决策函数的输出值(0或1)构成一个码字s。理论上，s一定与码本中的某一行码字完全一样，那么这一行所代表的原始类别即为未知样本的所属类别。但是考虑到实际应用中误差在所难免，因此在实际处理时，是计算编码矩阵内k个码字与s的汉明距离，距离最小者所对应的类别即为测试样本的所属类别。表3-1是一个8类分类问题的纠错编码的例子。

对于多类分类问题一个好的纠错码应该满足如下条件[73]：

（1）行$r_i(i=1,\cdots,k)$与行$r_j(j=1,\cdots,k;j\neq i)$之间不相关；

（2）列$c_i(i=1,\cdots,L)$与列$c_j(j=1,\cdots,L;j\neq i)$及其补$\overline{c}_j$之间不相关；

（3）没有全为“0”或者全为“1”的列。

根据上述条件对于一个k类分类问题，纠错码长度L的取值范围是:$\left[\log_2 k, 2^{k-1}-1\right]$。

表3-1 k=8，L=5时的分类纠错码

类别号	码字				
	1	2	3	4	5
1	0	0	0	1	1
2	0	0	1	0	1
3	0	1	0	0	1
4	0	1	1	0	0
5	1	0	0	1	0
6	1	1	0	1	1
7	1	1	1	0	1
8	1	1	1	1	0

3.1.5 多类分类的直接方法

Weston和Watkins提出了解决k类分类问题的一种直接方法[74]，实质上它是对两类SVM分类器的二次优化问题的一种自然推广形式，通过改写SVM二值分类中优化目标函数，使得其满足多值分类的需要。在构造决策函数的同时考虑所有的类，则可将原始优化问题推广为：

$$\min:\ \Phi(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\sum_{m\neq y_i}\xi_i^m \tag{3.3}$$

其约束条件为：

$$(\omega_i,x_i)+b_i\geq(\omega_m,x_i)+b_i+2-\xi_i^m, \tag{3.4}$$

$$\xi_i^m\geq 0, i=1,\cdots,n, m\in\{1,\cdots,k\} \tag{3.5}$$

由此，得到下面的k类SVM分类器的决策函数如下：

$$f(x)=\arg\max_i\left[(\omega_i\bullet x)+b_i\right], i=1,\cdots,k \tag{3.6}$$

此方法在经典支持向量分类机理论的基础上，重新构造分类模型，对新模型的目标函数进行优化，实现多类分类。它的优点十分直观，通过一步到位，即可实现多类别的分类。

但由于选择的目标函数过于复杂，要求解一个大的二次规划问题，计算量庞大，因此训练时间十分长，一般较少采用。

3.1.6 基于决策树方法的 SVMs 分类器设计

$k(k>2)$类分类问题和两类分类问题之间存在一定的对应的关系。如果一个多类分类问题k类可分，则这k类中的任何两类一定可分；另一方面，在一个k类分类问题中，如果已知其任意两类可分，则可以通过一定的组合方法最终实现k类可分。由于经典的 SVM 基于两类分类问题，可以把它和决策树思想结合起来构造多类别分类器，这种多类分类器称为 SVMs 决策树方法。

决策树方法具有以下特点:

（1）决策树分类思想简单而淳朴，建立的预测分析模型直观、易于接受。获取的知识用以规则描述，清晰、无二义性。

（2）不同于一些统计方法，决策树分类可以用相同的方法处理单模式和多模式，并且可以很容易地用于解决确定性问题和不完全确定性问题。

（3）决策树分类是从预分类实例中获取知识，避免了从领域专家得到知识所产生的瓶颈。

正是由于决策树分类方法的突出优点，该方法自从出现以后就得到了广泛的应用，如统计学、工程学(模式识别)和决策理论等。至今仍有很多学者不断地对其进行研究和改进，期望得到更好的算法以适应大规模训练样本分类的需要。

1. 决策树方法概述

决策树，或称多级分类器，是模式识别中进行分类的一种有效方法，对于多类或多峰分布问题，这种方法尤为方便。利用树分类器可以把一个复杂的多类别分类问题转化为若干个简单的分类问题来解决。它不是企图利用一种直接算法、一个决策规则去把多个类别一次分开，而是采用分级的方式，使分类问题逐步得到解决。

一个决策树由一个根节点n_1、一组非终止节点n_i和一些终止节点t_j组成，可对t_j标以各种类别标签，T 表示了决策树，那么一个决策树 T 对应于特征空间的一种划分，它把特征空间分成若干区域，在每个区域中，若某个类别的样本占优势，则可标以该类样本的类别标签。

数学上对决策树分类器作如下描述:

给定样本集Φ，其中的样本属于c个类别，用Φ_i表示Φ中属于第i类的样本集。定义一个指标集$I=\{1,2,\cdots,c\}$和一个I的非空子集的集合

$$\tau=\{I_1,I_2,\cdots,I_p\}。\tag{3.7}$$

我们可以令当$i\neq i'$时，$I_i\cap I_{i'}=\phi$。一个广义决策规则f就是Φ到τ的一个映射(记为$f:\Phi\to\tau$)。若f把第i类的某个样本映射到包含i的那个子集I_k中，则识别正确。

设$T(\Phi,I)$是由样本集Φ和指标集I所形成的所有可能的映射的集合，则$T(\Phi,I)$可表示为由对(a_i,τ_i)所组成的集合，元素(a_i,τ_i)称为一个节点，a_i是该节点上表征这种映射的参数，$\tau_i=\{I_{i1},I_{i2},\cdots,I_{ip_i}\}$是该节点上指标集$I_i$非空子集集合。令$n_i$和$n_j$是$T(\Phi,I)$的两个元素，其中

$$n_i=(a_i,\tau_i),\tau_i=\{I_{i1},I_{i2},\cdots,I_{ip_i}\},\tag{3.8}$$

$$n_j=(a_j,\tau_j),\tau_j=\{I_{j1},I_{j2},\cdots,I_{jp_j}\},\tag{3.9}$$

若

$$\bigcup_{1\le l\le p_j} I_{jl}=I_{ik},1\le k\le p_i\tag{3.10}$$

则称n_i为n_j的父节点，或称n_j为n_i的子节点。

设$B\subset T(\Phi,I)$是节点的有限集，且$n\in B$。若B中没有一个元素是n的父节点，则称n是B的根节点。当$B\subset T(\Phi,I)$满足下列条件时，它就是一个决策树分类器：

（1）B中有且只有一个根节点。

（2）设n_i和n_j是B中的两个不同元素，则

$$\bigcup_{1\le k\le p_i} I_{ik}\neq\bigcup_{1\le l\le p_j} I_{jl},\tag{3.11}$$

（3）对于每一个$i\in I$，B中存在一个节点

$$n'=(a',\tau'),\tau'=\{I'_1,I'_2,\cdots,I'_p,\},\tag{3.12}$$

且τ'中有一个元素就是i(与它对应的n'的子节点叫叶节点，又称终止节点)。

注意，在这样定义的树分类器中，每个类别标签只出现在一个叶节点上；我们当然也可以使每个类别标签出现在几个不同的叶节点上，这时前述的当$i\neq i'$时，$I_i\cap I_{i'}=\phi$的限制

条件就不再成立了。

决策树中的一种简单形式是二叉树（Binary Tree）。所谓二叉树，是指除叶节点外，树的每个节点仅分为两个分支，也就是说，每个节点n_i都有且只有两个子节点n_{il}和n_{ir}。二叉树结构分类器可以把一个复杂的多类别分类问题化为多级多个两类问题来解决，在每个节点n_i，都把样本集分为左右两个子集。分成的每一部分可能仍然包含多个类别的样本，可以把每一部分再分成两个子集……直至分成的每一部分只包含同一类别的样本，或者某一类样本占优势为止。

这种二叉树结构分类器概念简单、直观，便于解释，而且在各个节点上可以选择不同的特征和采用不同的决策规则，因此设计方法灵活多样，便于利用先验知识，从而获得一个较好的分类器。

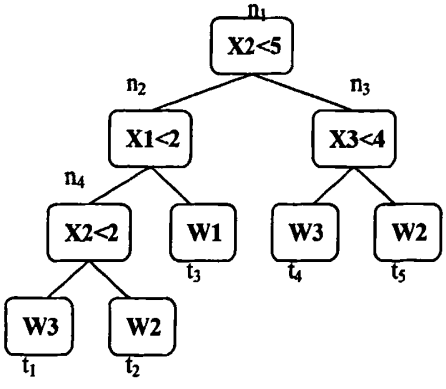


图 3-4 二叉决策树例子

图 3-4 就是一个二叉决策树的例子。该例中，在每个节点上只选择一个特征，并给出了相应的决策阈值。对于未知样本$\boldsymbol{x}$，只要从根节点到叶节点，顺序把$\boldsymbol{x}$的某个特征观察值与相应的阈值相比较，就可做出决策，把$\boldsymbol{x}$分到相应的分支，最后分到合适的类别。

设计一个决策树，主要应解决下面几个问题:

（1）选择一个合适的树结构，即合理安排树的节点和分支;

（2）确定在每个非终止节点上要使用的特征;

（3）在每个非终止节点上选择合适的决策规则。

把一个多类别分类问题转化为两类问题的形式是多样的，因此，对应的二叉树的结构也将各不相同。我们的目的是要找一个最优的决策树。

显然，一个性能优良的决策树结构应该有小的错误率和低的决策代价。但是由于很难把错误率的解析表达式和树的结构联系起来，在每个节点上所采用的决策规则也仅仅是在

该节点上所采用的特征观察值的函数，因此，即使每个节点上的性能都达到最优，也不能说整个决策树的性能达到最优。所以在实际问题中，人们往往提出其他一些优化准则，例如极小化整个树的节点数目，或极小化从根节点到叶节点的最大路程长度，或极小化从根节点到叶节点的平均路程长度等等，然后采用动态规划方法，力争设计出能最好地满足某种规则的“最优”决策树[75]。

2. 基于决策树的 SVMs 分类器设计

运用决策树方法将 SVM 推广到多类分类问题，称为基于决策树的 SVMs。SVMs 决策树方法有以下特点[76,77]：

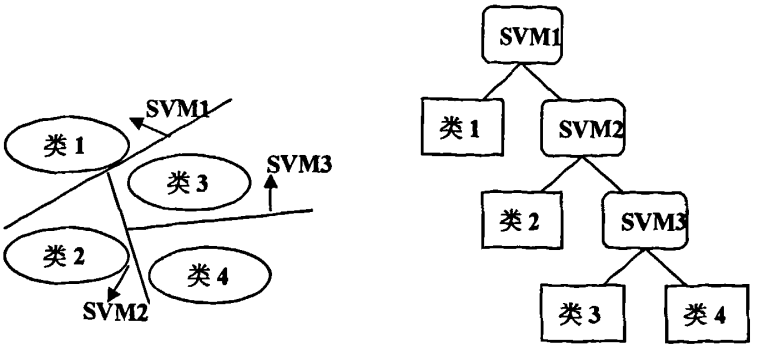
（1）决策树具有层次结构，每个层次子 SVM 的级别和重要性不同，其训练样本集的组成也是不同的。

（2）测试是按照层次完成的，对某个输入样本，可能使用的子 SVM 数目介于 1 和决策树的深度之间，测试速度快。

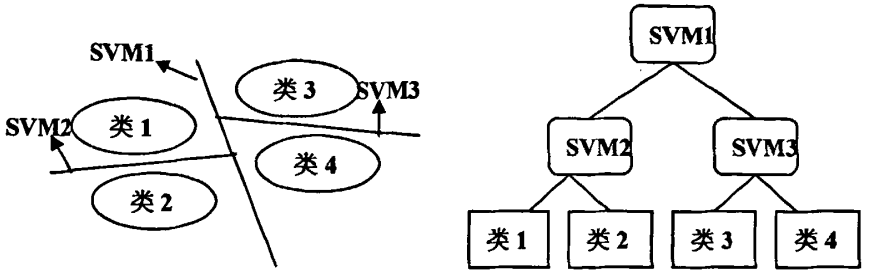
（3）决策树的各节点和叶的划分没有固定的理论指导，需要有一些先验知识。

对于分类类别不是很多的情况，我们可以提出一种简单的基于二叉树的多分类支持向量机方法（BT-SVMs），即：对于k类的训练样本，训练$k-1$个支持向量机，第 1 个支持向量机以第 1 类样本为正的训练样本，将第$2,3,\cdots,k$类训练样本作为负的训练样本训练 SVM_1，第i个支持向量机以第i类样本为正的训练样本，将第$i+1,i+2,\cdots,k$类训练样本作为负的训练样本训练 SVM_i，直到第$k-1$个支持向量机将以第$k-1$类样本作为正样本，以第k类样本为负样本训练 SVM_k-1。这种方案的优点是所需训练的支持向量机数目少，消除了在决策时存在同时属于多类或不属于任何一类的区域。在整个多类分类器训练中，总共训练样本数目和一对多方法相比减少许多。

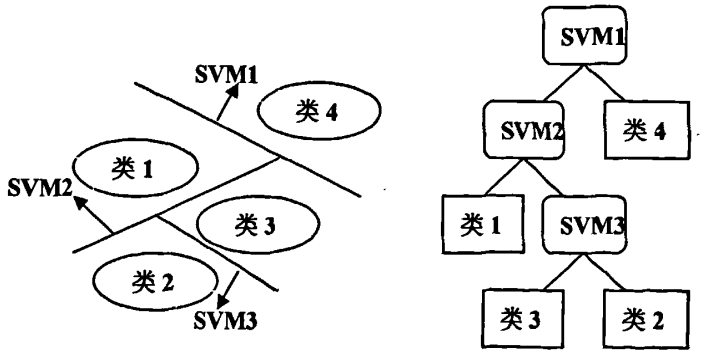
根据二叉树的定义，构造一棵有k个叶子节点的严格二叉树有多种不同方案。例如对 4 类分类问题，则有图 3-5（a）（b）（c）所示的三种二叉树结构：



（a）第一种样本特征空间划分及其对应的 SVMs 二叉决策树结构



（b）第二种样本特征空间划分及其对应的 SVMs 二叉决策树结构



（c）第三种样本特征空间划分及其对应的 SVMs 二叉决策树结构

图 3-5 四类分类问题的三种情况

无论采用何种方式生成的决策树结构，对于 k 类问题，都将构建 $k-1$ 个二分类SVM，它们与各层中的分支结点一一对应。在测试阶段，从顶层结点出发，根据各中间结点的输出值，最终推进到一个终端结点，该终端结点代表的类别既为该样本的所属类别[78,79]。

对一个具体问题，选择二叉决策树SVMs方案时，二叉树结构选择可以依据以下几点：

（1）若对 k 类问题中各类基本无先验知识，无法指导树叶节点的划分，则可采用每次决策分出一类的二叉决策树结构，如图3-5(a)所示结构。

（2）若对k类问题中各类基本有一些相关先验知识，可以指导树叶节点的划分，则可采用完全二叉决策树结构，如图3-5(b)所示结构。

（3）若对k类问题中各类基本有充分的相关先验知识，可以指导树叶节点的划分，则可采用结合二叉树理论，构造一棵测试速度最优的SVMs二叉决策树，如图3-5(c)所示结构。

我们将结合二叉树分类的SVMs方法称为BT-SVMs，它不但可以大大提高SVMs的分类速度，而且又克服了前述方法可能出现的无法分类区域的存在，从而提高了SVMs多类分类的性能。本文在后期的新生儿面部疼痛表情分类研究中，基于二叉树对SVMs进行多类分类设计，对初期研究工作进行改进，实验证明算法具有可行性。

3.2 常见多类分类方法的性能比较

多类分类器算法的性能应考虑三个方面：训练过程、测试过程和分类精度。

3.2.1 各种算法的训练过程比较

训练速度是支持向量机理论需要考虑得重要问题之一。当SVM向多分类领域推广时，由于问题规模的进一步扩大，训练速度不佳的问题进一步凸显。同时，训练算法对时间性能的影响也更明显，所以首先分析算法在训练过程中的代价（均是针对k类多分类问题）。

两类SVM分类问题中，训练时间满足以下关系式[80]：

$$T = cm^r, \tag{3.13}$$

其中m为参与训练的样本数量，c为常数，r具体大小与不同的分解算法有关，当采用SMO分解算法时$r=2$。由此可看出SVM的训练速度主要由m决定。

1. 一对一 SVMs 和 DDAG-SVMs

一对一SVMs和DDAG-SVMs训练时都需要构造$k(k-1)/2$个子分类器，为简单描述问题，假设各类含有相同数目的样本，则每次训练两个类别的样本集合大小为$2m/k$，其训练时间为：

$$T_{OAO} = T_{DDAG} = \frac{k(k-1)}{2} c\left(\frac{2m}{k}\right)^r \approx 2^{r-1} c k^{2-r} m^r, \tag{3.14}$$

当上式$r=2$时，这两种SVMs的训练速度与类别数无关。

2. 一对多 SVMs 和 ECC-SVMs

一对多SVMs和ECC-SVMs在构造每个子分类器时都是对整个训练样本集进行训练，一

对多SVMs训练k个分类器，ECC-SVMs训练L个分类器，两者的训练时间分别为：

$$T_{OAR} = kcm^r, \tag{3.15}$$

$$T_{ECC} = Lcm^r, \tag{3.16}$$

由上式可以看出，由于一对多SVMs和ECC-SVMs的训练时间取决于整个训练集的样本数目和类别数，所以其远大于一对一SVMs和DDAG-SVMs的训练时间。

3. 多类分类的直接方法

这种方法将所有样本都集中到一个优化目标函数中，其核心还是约束条件极值问题的解决，其训练时间为：

$$T = ck^r m^r, \tag{3.17}$$

可以看出其训练时间随着类别数的增加急剧增长，并且计算量庞大。

4. 基于二叉树的 SVMs

对这种算法分类器数量为k-1个，其层次数的范围为$\left(\lfloor \log_2 k \rfloor + 1, k\right)$，其训练速度受二叉树结构的影响，现考虑两种极端情况，即两种不同的二叉树结构：偏态树和正态树。偏态树（如图3-5（a）的SVMs二叉树）是从根节点开始，每个节点的子分类器指将一个类别与其它类别分开，层数取最大值；正态树（如图3-5（b）的SVMs二叉树）是从根节点开始，每个节点的子分类器将所剩类别分为两类，层数取最小值；这两种结构的BT-SVMs的训练时间分别为：

$$\text{偏态树：} T_{BT}^{'} = \sum_{i=0}^{k-2} c\left(\frac{(k-i)m}{k}\right)^r, \tag{3.18}$$

$$\text{正态树：} T_{BT}^{''} = \sum_{i=0}^{e-1} 2^i c\left(\frac{m}{2^i}\right)^r, \tag{3.19}$$

其中$e = \lceil \log_2 k \rceil$。

一般结构的BT-SVMs的训练时间介于这两种结构之间，此时训练时间取为：

$$T_{BT} = \left(T_{BT}^{'} + T_{BT}^{''}\right)/2, \tag{3.20}$$

当样本分布均匀时，BT-SVMs的层数与训练时间成正比关系。且在层数确定时，当BT-SVMs的拓扑结构由趋于正态向趋于偏态变化时，训练时间呈递增趋势。正态拓扑结构下，样本分布越均匀，则BT-SVMs的训练时间越短。在偏态树中样本分布不均匀的前提下，样本数量越大的类别在偏态树中所处的层次越高，BT-SVMs的训练时间与训练时间的极大值越接近；反之，样本数量越大的类别在偏态树中所处的层次越低，BT-SVMs的训练时间

与训练时间的极小值越接近。

图3-6为$r=2, c=1, L=2\log_2 k, k=2,4,8,16,32$时各种算法训练时间的比较。

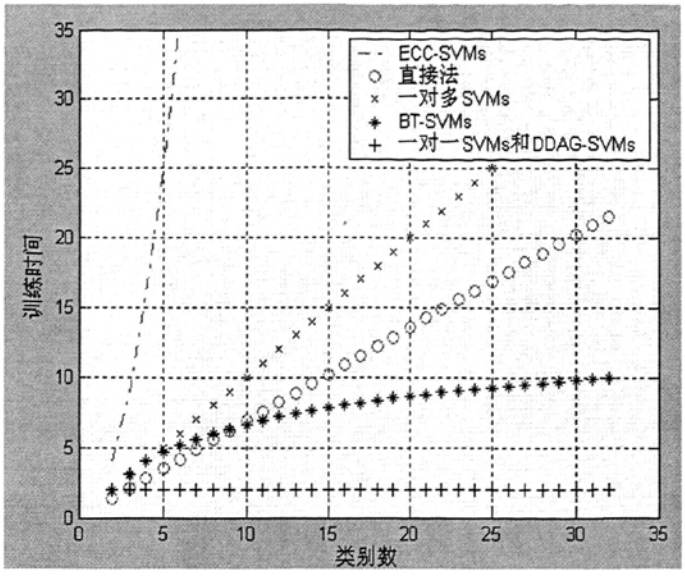


图3-6 各种算法训练时间

由图3-6可以看出，一对一SVMs和DDAG-SVMs训练速度最快，BT-SVMs次之，一对多SVMs和ECC-SVMs随着类别数的增多急剧变慢，多类分类的直接方法的训练时间随类别数增加呈直线上升。

3.2.2 各种算法的测试过程比较

当预测测试样本的类别时，将其代入已训练好的决策函数式（2.39）根据决策函数的值，确定测试样本的类别。在测试计算决策函数时，大部分时间都花费在计算$K(\boldsymbol{x}_i, \boldsymbol{x})$与求和上，因此可以判定二元支持量机的测试时间是与支持向量的个数成正比的。

推广到多分类支持向量机，设在预测一个未知样本的类别时需要b个分类器$\{f \mid f_i \in F\}(i=1,2,\cdots,b)$，$F$为SVMs算法在训练阶段得到分类器的集合。由于单个分类器所需的分类时间与其支持向量的数量成正比，所以在多分类SVM算法中对一个未知样本的测试速度取决于所需分类器中支持向量的总数$\sum_{i=1}^{b} SV(f_i)$，其中$SV(f_i)$表示第i个分类器中的支持向量个数。由此可见，SVMs算法的分类速度取决于两个因素:

（1）单个测试样本在整个分类过程中所需的分类器数量;

（2）各个分类器中支持向量的数量。

常见的SVMs算法测试时所需的分类器数量见下表：

表3-2 测试时各SVMs算法所需的分类器数量

算法	一对一SVMs	一对多SVMs	DDAG-SVMs	ECC-SVMs	BT-SVMs
分类器个数	$k(k-1)/2$	k	$k-1$	$L>\lceil\log_2 k\rceil$	$1\sim(k-1)$

现假设二元支持向量机中支持向量的个数与训练样本总数N成正比，比例系数为β。对各多类分类SVMs算法的测试过程性能分析如下：

（1）一对一SVMs和DDAG-SVMs

对k类问题，假设样本均匀分布，则一对一SVMs算法测试时所需分类器个数为$k(k-1)/2$，其支持向量的总数为：

$$SV_{OAO}=\frac{k(k-1)}{2}\beta\frac{2N}{k}=(k-1)\beta N \quad , \tag{3.21}$$

DDAG-SVMs算法测试时所需分类器个数为$k-1$，其支持向量的总数为：

$$SV_{DDAG}=(k-1)\beta\frac{2N}{k}=\frac{2(k-1)}{k}\beta N , \tag{3.22}$$

从上两式可以看出，尽管一对一SVMs和DDAG-SVMs在训练阶段所得到的子分类器个数是完全一样的，但是DDAG-SVMs在对未知样本进行分类时只需$k-1$个分类器，而一对一SVMs却需要所有的$k(k-1)/2$个分类器。在单个子分类器的支持向量数相等的情况下，DDAG-SVMs算法的测试分类速度快于一对一SVMs算法。

（2）一对多SVMs

对k类问题一对多SVMs算法测试时所需分类器个数为k，其支持向量的总数为：

$$SV_{OAR}=k\beta N , \tag{3.23}$$

由于一对多SVMs中每个分类器都是将某个类别与其他所有类别分类，需要对所有样本进行训练，而DDAG-SVMs中每个分类器都对两个类分类，所以一对多SVMs每个分类器的支持向量数要远远大于DDAG-SVMs，其分类速度较慢。与一对一SVMs相比，在样本分布均匀的情况下，两者具有相似的分类速度。

（3）ECC-SVMs

对k类问题ECC-SVMs码字长度为L时，其支持向量的总数为：

$$SV_{ECC}=L\beta N , \tag{3.24}$$

ECC-SVMs分类过程中所需分类器的个数等于纠错编码的位数L，理论上码字长度只需

$\log_2 M$，但是此时码间最小汉明距离为1，编码将丧失纠错的功能。实际往往取较长的码字以保持一定的最小码间汉明距离，通过编码的纠错能力得到较高的分类精度。当类别数较多时，如果码字长取$2\log_2 M$，分类时所需分类器的数量明显少于一对多SVMs、一对一SVMs和DDAG-SVMs，但并不一定意味着ECC-SVMs的分类速度一定快于其它三种算法。因为ECC-SVMs中每个分类器是将对应码位为0的类同其它码位为1的类分开，其中支持向量的数量与数据的可分性密切相关。因此对于某个具体问题，码本设计恰当与否将直接影响到分类速度。

（4）基于二叉树的SVMs

由于BT-SVMs的训练时间随着层次结构的不同而不同，这种特性在测试时仍然存在，即其分类速度受其层次结构的影响：

当结构为偏态树时，一个未知样本分类所需的分类器数量从1到k-1不等，平均$k/2$个，少于以一对一SVMs。当BT-SVMs最多需k-1个分类器时，分类其数量与一对多SVMs类似，但是BT-SVMs中的分类器所处理的总类别数自上而下逐渐减少，因此分类器中的支持向量数一般少于一对多SVMs，分类速度将快于一对多SVMs。

正态树，对一个未知样本分类所需的分类器数量约为$\log_2 M$，等于ECC-SVMs的理论最小值，当类别数较多时，BT-SVMs的测试时间将远远少于其它算法。

3.2.3 多分类支持向量机算法的分类精度

判断一个SVM算法优劣的标准除了训练时间和测试时间外，更重要的一个标准是分类精度，即识别率：

$$\text{识别率}=\frac{\text{测试样本中分类正确的样本数}}{\text{参与测试的样本总数}}。$$

训练时间和测试时间反映了算法的运行效率，而识别率主要衡量学习机的推广能力。

3.3 改进的二叉树 SVMs 分类器

基于决策树的SVMs方法存在的问题是：如何确定决策树的结构。例如，对四类分类问题，图3-5中3种特征空间的划分对应了三种不同结构的二叉树。此方法的缺点是，如果在某个节点上发生分类错误，则会把分类错误延续到该节点的后续节点上。因此由于误差的累积效应，分类错误在越靠近根节点的地方发生，分类性能越差，尤其是在根节点处发

生分类错误将严重影响分类性能。

为此可以考虑一开始就根据类间可分离测度将容易分开（不易产生错分）的两类作为根节点，在决策过程中的每一步不是仅仅根据决策函数的值随机决定下一个分类决策面，而是在备选决策面中选取当前最易分离的两类作为下一节点的决策面，依次类推。这样就能够使可能出现的错误尽可能的远离决策树的根，减小累积误差。

3.3.1 基于类间距离的决策树结构

根据训练数据估计类间易分性，通常的做法是用类间距离作为分离性测度[81]。在这里给出一种基于最近距离原则的决策树结构生成方法：

步骤1：分别计算各类样本之间的类间距离，并将距离按升序排序。

步骤2：取距离最近的两类分别做为正类样本与负类样本，形成新的训练集M_p，为M_p构建一个二分类SVM判别函数：

$$f^{M_p}(x)=\mathrm{sgn}\left(g^{M_p}(x)\right), \tag{3.25}$$

其中

$$g^{M_p}(x)=\sum_{i=1}^{l_p} y_i\alpha_i^* K(x,x_i)+b^*, \tag{3.26}$$

$y_i\in\{-1,1\}$，l_p为M_p包含的样本数目。此决策函数即第一个叶节点SVM分类器。

步骤3：将M_p做为一个新类放在总类中，以新的样本集作为总样本集，迭代步骤1、2，找出新的距离最近的两类，求出SVM判别函数作为上一节点的父节点，直至所有的类别聚合成一类，最终形成一个基于决策树的SVMs分类器。我们将这种算法称为BT-SVMs。

计算类与类之间的距离有多种方法：最短距离法、最长距离法、重心法和平均距离法。如重心法，求各类中所有样本的平均值作为类的重心样本，用两类的重心间的距离作为两类的距离。而两个样本间距离的计算方法有：欧氏距离法、夹角余弦距离法、二值夹角余弦法和具有二指特征的Tanimoto测度[82]，后两者都要求样本的各个特征都是以二值（0或1）表示。但这种方法的缺点是，类间距离远近并不一定能代表类间的分离性测度[83]。

3.3.2 基于最优分类面的类间分离性测度

在两类间的最优分类面训练好后，可知训练样本中能够满足最优分类面

$y_i(\omega \cdot x_i + b) - 1 \geq 0$的样本位于分类间隔之外，被正确分类，但由于分类函数并不总能正确分类所有的样本，所以存在被错分的样本。若两类样本间被错分的样本很少则可认为这两类样本较易分离，若两类样本间被错分的样本较多则可认为这两类样本较难分离，实际情况也是如此。由此，引入一种基于最优分类面的类间分离性测度，从而改进二叉树SVMs分类器。

对k类分类问题，可以定义类i与类j间的分离性测度sm_{ij}：

$$sm_{ij} = \frac{N_{ij}^s}{\left(N_i + N_j\right)}, \quad i, j \in \left(1, 2, \cdots, k\right), \tag{3.26}$$

其中，N_i为第i类的训练样本数，N_j为第j类的训练样本数，N_{ij}^s为对类i与类j进行训练后两类中本最优分类面正确分类的训练样本数。

则最难分离的两类为：

$$\left(i, j\right) = arg \min_{\substack{i=1,\cdots,k \\ j=1,\cdots,k}} sm_{ij}, \tag{3.27}$$

即分类性测度值最小的两类是最难分开的两类。

3.3.3 算法描述

1. 基于类间分离度的二叉树 SVMs 算法

对于分类类别k不是很多的情况，设训练样本集由类X_i（$i = 1, 2, \cdots, k$）组成，进行k类分类的过程如下：

（1） 训练阶段

Step1：将训练数据按类两两组合，训练$k\left(k-1\right)/2$个分类决策面，组成决策矩阵F：

$$F = \begin{bmatrix} [\,] & f_{12} & f_{13} & \cdots & f_{1,k-1} & f_{1,k} \\ f_{21} & [\,] & f_{23} & \cdots & f_{2,k-1} & f_{2,k} \\ f_{31} & f_{32} & [\,] & \cdots & f_{3,k-1} & f_{3,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{k-1,1} & f_{k-1,2} & f_{k-1,3} & \cdots & [\,] & f_{k-1,k} \\ f_{k,1} & f_{k,2} & f_{k,3} & \cdots & f_{k,k-1} & [\,] \end{bmatrix}, \tag{3.28}$$

其中，$f_{ij}\left(x\right) = \mathrm{sgn}\left(\sum_{x_p \in X_i \cup X_j} \alpha_p y_p K\left(x_p, x\right) + b\right)$，$i, j = 1, 2, \cdots, k$，　$i \neq j$，若$f_{ij}\left(x\right) = +1$，则$x$属于第$i$类，若$f_{ij}\left(x\right) = -1$，则$x$属于第$j$类，且有$f_{ij}\left(x\right) = -f_{ji}\left(x\right)$；[ ]表示空，及各

类自身无决策面。

Step2：由（3.26）式计算类与类之间的分离性测度sm_{ij}，组成分离性测度矩阵SM：

$$SM=\begin{bmatrix} [\,] & sm_{12} & sm_{13} & \cdots & sm_{1,k-1} & sm_{1,k} \\ sm_{21} & [\,] & sm_{23} & \cdots & sm_{2,k-1} & sm_{2,k} \\ sm_{31} & sm_{32} & [\,] & \cdots & sm_{3,k-1} & sm_{3,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ sm_{k-1,1} & sm_{k-1,2} & sm_{k-1,3} & \cdots & [\,] & sm_{k-1,k} \\ sm_{k,1} & sm_{k,2} & sm_{k,3} & \cdots & sm_{k,k-1} & [\,] \end{bmatrix}, \tag{3.29}$$

这是一个对称矩阵，即$sm_{ij}=sm_{ji}$，同类间的分离性测度为0。

Step3：由（3.27）式选择最难分开的两类，假设为第i_0类和第j_0类，将这两类分别做为正类样本和负类样本，形成新的训练集M_p，为M_p构建一个二分类SVM判别函数$f_{k-1}(x)$，作为叶节点第$k-1$个SVM分类器。

Step4：将i_0类和j_0类合并成新类M_p后，计算类M_p和所有其它剩余类的sm_{ip}，迭代步骤3、4，找出和类M_p分离性测度sm_{ip}最小的类i_1，将类i_1作为正类样本，类M_p作为负类样本，求出这两类的SVM判别函数作为上一节点的父节点，即第$k-2$个分类器SVM_{k-2}，依次类推，直至所有的类别聚合成一类，最终形成一个基于二叉树的SVMs分类器。

（2） 分类阶段

Step1：取测试数据x，从根节点的SVM_1开始计算$f_i(x), i=1$；

Step2：若$f_i(x)=+1$，则表示判决x属于正类所属的类别号，判决结束；若$f_i(x)=-1$，则转下一节点的SVM分类器，计算$f_{i+1}(x)$，重复Step2。

2. 算法的简化

在上述算法中，训练阶段需要不断循环计算分离性测度直到所有类别聚合成一类，这种迭代循环过程比较费时，因此我们进一步简化：在第一次计算得到分离性测度矩阵SM后，对每一类别计算其平均的$\overline{sm}$。例如，对类别k有：

$$\overline{sm_k}=\frac{1}{n-1}\sum_{i=1,i\neq k}^{n} sm_{ki},\quad k=(1,\cdots,n), \tag{3.30}$$

将$\overline{sm_k}$按降序排列，认为$\overline{sm_k}$值最小的类别于其它类别最难分开，应将其放在最底层

的分类器中进行分类。假设$\overline{sm_{i1}}$和$\overline{sm_{i2}}$值最小的两类，将$i1,i2$这两类分别做为正类样本和负类样本，形成新的训练集M_p，为M_p构建一个二分类SVM判别函数$f_{k-1}(x)$，作为叶节点第$k-1$个SVM分类器；在剩余类的$\overline{sm_k}$中再选择最小的类别，与M_p分别做为正类样本和负类样本构建新的二分类SVM，作为上一节点的父节点，依此类推，最终得到二叉树SVMs分类器。我们将这种算法称为改进的BT-SVMs。

3.4 仿真实验

本节在 Microsoft Visual C++6.0 环境下，结合两组数据，对多分类支持向量机进行了分类实验，这两组数据样本分别为来自网上公用的数据库 Statlog（http://www.liacc.up.pt/ML/old/statlog/datasets.html）的 Vehicle 数据集和 Satimage 数据集。本次实验选择了一对一 SVMs、一对多 SVMs、DDAG-SVMs、BT-SVMs 和改进的 BT-SVMs 对这两组数据进行了实验，各分类器的参数对$(C,\delta)=(15,0.5755)$。

Vehicle 数据集共有 846 个 18 维的样本，用交叉验证的方法进行测试，结果如下表所示：

表 3-3 对 Vehicle 的多分类实验

实验 3	一对一 SVMs	一对多 SVMs	DDAG-SVMs	BT-SVMs	改进的 BT-SVMs
运行时间(ms)	6031	13906	5844	5640	5579
支持向量数	432	468	440	402	411
识别率	82.51%	82.98%	82.98%	91.13%	90.31%

Satimage 数据集的样本为 36 维，它有 3104 个训练样本，2000 个测试样本。对 Satimage 样本，用 3104 个训练样本训练出 SVMs 分类器模型后，直接用 2000 个测试样本进行测试，不再进行交叉验证，结果如下表所示：

表 3-4 对 Satimage 的多分类实验

实验 4	一对一 SVMs	一对多 SVMs	DDAG-SVMs	BT-SVMs	改进的 BT-SVMs
训练时间(ms)	4278	13750	4350	4578	4375
支持向量数	1041	1112	1041	972	981
预测时间(ms)	2172	2375	2250	2046	2047
识别率	90.50%	90.35%	90.3%	92.50%	92.75%

从表 3-3、表 3-4 中可以看出，基于二叉树的 SVMs 多分类器识别率较高，训练得到的

支持向量数较少，即用较少的支持向量获得了比较好的分类效果。在实验4中，一对一SVMs训练时间最短，但 BT-SVMs 方法的训练时间和它相差不多；而在预测时，基于二叉树的SVMs方法所需时间最少。在实验3中，BT-SVMs方法的总体运行时间最少，是因为其所需预测时间较少。而改进后的BT-SVMs对Satimage样本分类时，识别率和训练时间性能有所提高，对Vehicle的分类性能也比较理想。

3.5 本章小结

多分类支持向量机理论是对传统意义上二分类支持向量机的进一步深化，在实践中具有比二分类问题更为广泛的应用前景。本章针对几种经典的多分类 SVM 方法进行了系统的讨论，首先介绍了它们的原理，接着从训练过程和测试过程两方面对各种方法进行了全面的比较，分析了各自的性能，得到如下结论：(1)BT-SVMs、一对一 SVMs 和 DDAG-SVMs的训练时间性能总体要优于一对多 SVMs 与 ECC-SVMs 算法，而直接方法的训练时间最长一般不采用；（2）当拓扑结构偏向正态时，BT-SVMs 算法拥有最理想的分类速度；（3）BT-SVMs 存在层次的累计误差，因此在拓扑结构的设计时，应尽量使分类误差远离根节点，靠近叶节点。

为了减小误差的累积效应，提出了改进的二叉树 SVMs 分类器，设计了两种不同的类间分离性测度，根据类间分离性测度来设计二叉树结构，以此来提高分类性能。

通过仿真实验将基于二叉树的SVMs方法及其改进算法与三种经典的多分类方法进行比较，证明其性能总体要优于其它三种方法。

第四章　基于多分类SVM的新生儿疼痛表情识别

自从90年代初期SVM作为一种新的机器学习方法出现在模式识别领域以来，就一直受到广泛的关注与重视，目前SVM已经被公认为是精度最高的模式分类器之一。随着SVM理论研究的进一步深入，将SVM向更广的应用领域推广已经成为一种趋势，它在各种各样的实际应用领域当中都取得了相当出色的成果。近年来，已经有学者将其引入到表情识别领域，为该领域提供了一条解决问题的新思路，同时也取得了一定的成果。本章的重点是将多分类SVM算法与表情识别模型相结合，针对新生儿疼痛表情设计了多类识别系统，可以对新生儿的不同程度的疼痛表情和其它非疼痛表情进行分类识别。

4.1 新生儿疼痛表情识别

以前人们认为新生儿传输疼痛的神经系统没有发育成熟，所以不能感觉到疼痛，其实这是一种错误的看法。事实上，新生儿感觉疼痛的灵敏性比成年人更强。又由于新生儿不能自述疼痛的感受，疼痛评估成为新生儿科学中最具挑战性的一个难题[84]。专家认为不能单靠医护人员来评估生理疼痛，因为不容易从疼痛、害怕和焦虑中区分出与疼痛相关的生理参数。

新生儿在疼痛时会表现出行为变化，包括哭声、面部表情改变、呻吟、肢体活动及行为状态（如睡眠和食欲）的改变。又因为新生儿自主神经系统并不完善，一些生理指标如心率、血压变化差异较大，病理情况反应也各异，可能导致测量结果不确定，所以不能仅用生理指标来评估新生儿疼痛，必须与行为评估方法联合应用[85]。目前广泛认同面部表情改变是评估疼痛的一种有效、可靠的指标，其中新生儿“疼痛面容”（蹙眉、挤眼、鼻唇沟加深、张口）是最可靠的疼痛指标，且持续时间最长。

在计算机领域，面部表情识别是指基于视觉信息，将脸部的运动或特征的形变进行分类。根据分类的表情数据源，可以把表情识别方法分为如下两类：基于静态与动态图像的方法。静态表情识别方法中，一般根据单一的人脸表情图像进行表情识别。本文就是使用基于支持向量机的方法对静态图像进行识别。

首先，将一幅图像输入人脸检测模块，检测是否存在人脸。本系统中人脸检测模块使用Adaboost方法来检测人脸的存在，并获得人脸位置。然后将带有人脸的图像作为特征提取模块的输入，在特征提取模块采用Gabor小波对人脸图像进行变换，得到面部表情图像

信息，所获得的特征向量的维数比较大，采用 Adaboost 方法来对特征向量进行降维。最后将提取出的特征向量输入到 SVM 表情识别模块，表情识别是将输入的面部表情归入某个具体的类中。最后，表情识别模块输出表情的识别结果和识别率，通过输出的类别来判断面部表情的类型。

4.2 基于 SVM 的表情识别模块设计

面部表情识别系统的最后一个环节就是表情识别，将根据提取的特征对面部表情进行分类，可以根据人脸活动性分类也可以根据情感进行分类，有些系统则兼容了两者。表情识别模块是以特征提取模块输出的特征向量作为输入，先对支持向量机的模型进行选择；然后用一定量的样本对该学习机进行学习和训练，最后使用测试样本进行测试，得到系统的分类效果及性能。SVM识别模块框图如图4-1所示：

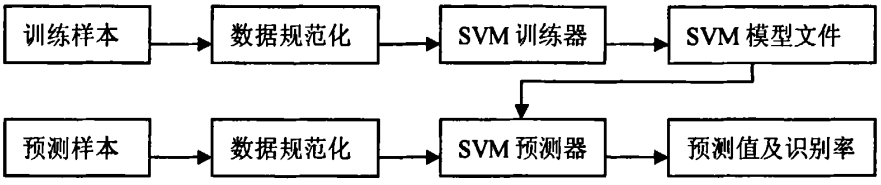


图 4-1　SVM 识别模块框图

首先，将所有样本的特征向量进行规范化，使特征向量的数值在一定的范围内，从而可以节省训练与测试的时间。然后，将已规范化处理后的训练样本数据作为 SVM 训练器的输入，得到一个模型文件。该模型文件中记录了训练时使用的核函数类型及其参数值、支持向量及其数量、计算出的拉格朗日乘子的数值、判决函数中 b 的值。再将训练器得到的模型文件与已经过规范化处理后的测试样本数据作为预测器的输入，通过 SVM 预测器得到如下结果：测试样本的预测值（类标号）及总的识别率。最后，通过人为的预先的规定及理解，将这些标号解释为不同类型的表情。

SVM的训练和预测步骤如下：

（1）将特征提取模块提取出的特征向量进行数据规范化处理；

（2）训练SVM分类器。使用交叉验证和网格搜索相结合的策略来确定SVM模型的参数，尝试使用不同核函数并比较识别结果，选择一个最佳的核函数进行训练，最后得到SVM模型文件；

（3）使用训练获得的SVM模型对测试样本进行测试，计算得到总的识别率。

4.2.1 数据的规范化

由于不同特征的物理意义不同，取值范围也常大相径庭。即使对同一特征的不同分量之间，如灰度产生的矩阵的各个分量之间，其取值范围差别也很大。这可能会使一个特征或分量的取值过小，在计算过程中似乎不起作用，若过大，使其作用过于明显。为了消除这种影响，我们需要对每一个特征分量进行规范化处理。数据规范化处理是通过线性映射将样本数据从比较大的值域放缩到指定的范围，这样既可以避免在数据域中出现数值过大的数据，又可以使支持向量机的归类计算量相对降低。

通过将属性数据按比例缩放，使之落入一个小的特定区间，将有助于加快学习阶段的速度。在 SVM 中，一些核函数要求将数据进行规范化，使数据包含于它们的定义域内。对数据的适当变换，可改善数值计算过程的问题。一般认为，对优化问题，适当地进行比例变换，可以抑制与计算误差及核计算误差界有关的错误发生。根据样本本身的特征，采用的规范化形式有所不同，一般通常采用的方法有最小－最大规范化、z-score 规范化、小数定标规范化等。本文采用最小－最大规范化方法，将所有样本的属性值都规范化到[-1, +1]之间。每个样本通过特征提取后得到了维数相同的特征向量，再将每个样本的特征值进行编号，然后再找出每个编号相同样本的特征值的最大值 Max 与最小值 Min，若此特征值为 x，使用如下式计算规范化后的特征值 y：

$$y = -1 + 2 \times (x - Min) / (Max - Min) \tag{4.1}$$

4.2.2 SVM 训练器

SVM 的训练过程如图 4-2 所示。首先，应该根据实际的训练样本选择合适的核函数和其他参数，以期获得最佳的分类其性能；接着，是解凸优化问题，算法应该简单、高效，使计算时间和内存需求量最小；最后，得到的决策函数（分类超平面）使样本特征空间风险上界最小则结束训练，否则重新训练分类器。训练器主要是提供知识给机器学习，最终得到决策函数。知识都包括在训练集合中，要想获得一个好的训练集合并不容易，它直接关系到分类器性能的好坏。对于给定的样本，支持向量机的性能主要受核参数和误差惩罚因子的影响。核函数的选择决定了特征空间的结构，惩罚因子的选择决定了经验风险的影响程度。

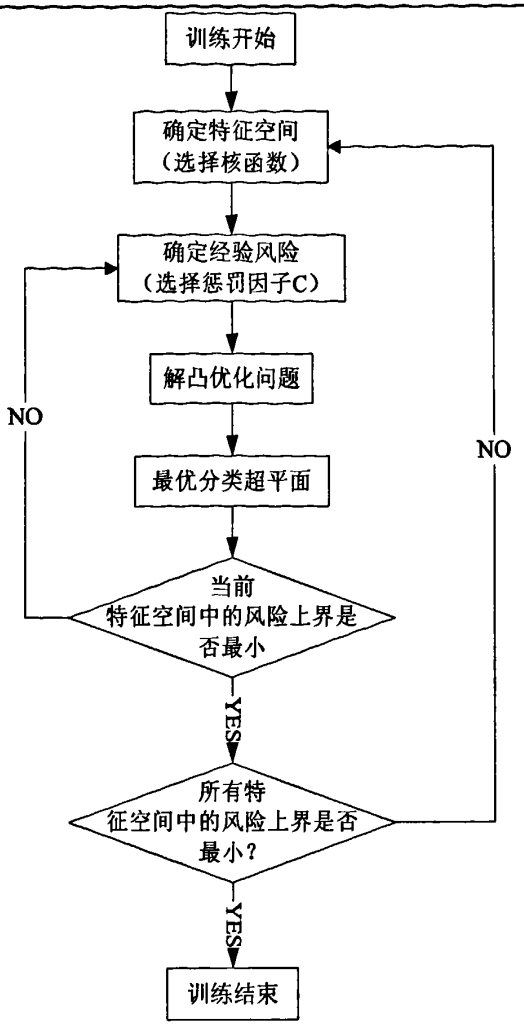


图4-2　训练的流程图

通过训练，SVM 会产生一个决策模型（训练后得到的模型文件），这个决策模型对二分类问题是一个决策函数，对多分类问题是一系列决策函数，这些决策函数叫做特征空间中的超平面。通常情况下，模型文件中的支持向量数目越多，支持向量的分类速度就越慢。可以通过测试集，来测试 SVM 决策模型的性能。分类决策模型的性能主要是通过精度、识别率及错分样本的数目等来确定。

1. 训练算法

训练 SVM 的本质是解决二次规划问题。实际应用中，如果用于训练的样本数很大，这样需要很长的训练时间与很大的存储空间，就很难应用标准的二次优化技术。最初的思想是先将整个优化问题分解为多个同样性质的子问题，再通过循环解决子问题来求得初始问题的解。由于这些方法都需要不断地循环迭代来解决每个子问题，所以需要的训练时间很长，这个问题后来逐渐得到了解决。现有的 SVM 训练算法较为著名的有选块法、分解法、序贯最小优化（SMO）算法等，这些都在第二章中介绍。

实验中我们选择了 SMO 算法，因为该算法避免了多样本情形下的数值解不稳定及耗时问题，同时也不需要大的矩阵存储空间，避开了复杂的数值求解优化问题的过程，实现简单，收敛速度快，是现在使用最广泛的算法之一。其算法流程见第二章。

2. 参数选择方法

本文中在选择核函数参数与惩罚因子时，使用了交叉验证与网格搜索相结合的方法。交叉验证又叫做交叉对比，它是一种用来评价模型推广能力的方法，主要用于模型的选择、实验设计等关于统计理论和模式识别等方面。交叉验证方法将数据集随机分成训练集、验证集和检验集三部分，利用训练集来估计给定的不同组合的参数，通过验证集来选择最优的一组参数，最后用检验集评估模型的效果。交叉验证的作用有两个：一是它可以验证分类性能，即用一组已知数据对分类器的识别率进行评测；二是还可以用它寻找最佳的 SVM 参数以及核参数，防止过学习现象的发生。此方法在计算代价和可行的参数估计之间提供了最好的折中方案，它可以较好地防止过学习现象。目前，通常使用的交叉验证方法主要有两种，一种是 Leave-One-Out/Leave-K-Out 交叉验证（Leave-One-Out Cross validation, LOOCV），另一种是 K-fold 交叉验证。

（1）Leave-One-Out交叉验证方法

该方法也称留一法，就是从N个样本中取出一个样本后，用剩下的N－1个样本做训练，来设计分类器，然后将取出的那个样本用于测试。再把原取出的样本放回去，又取出另一个样本(与前面取出的样本不同)，将剩下的N－1个样本作为训练样本集去设计分类器，再做测试。这样一共重复设计N次，测试N次，并统计被分错的样本总数K，最后用LOOCV的平均错误率 $\hat{\varepsilon}=\frac{k}{N}$ 作为在未设计好分类器的情况下，对分类错误率的估计值。并以此作为一种样本所在特征空间的分类推广能力的评价指标－LOOCV的平均误识率越小，若该特征空间中的预测分类错误率越小，则该特征空间的分类推广能力越强。

（2）Leave-K-Out交叉验证方法

将留一法推广之，每次取出K个样本作测试，剩下的（N－K）个样本作训练，共设计 C_N^k 个分类器，测试 C_N^k 次，最后得到其平均错误率，这就是Leave-K-Out交叉验证方法。

（3）K-fold交叉验证方法

K重交叉确认（K-fold Cross-Validation）的思想是：首先把 l 个样本点随机地分成互不相交的子集，即 $S_1,S_2,...,S_K$ 每个子集大小大致相等。共进行K次训练和测试，即对 $i=1,2,...,k$ 进行K次迭代，第 i 次迭代的做法是，选择 S_i 为测试集，其余子集的合集为训练集，算法根

据训练集求出决策函数后，即可对测试集S_i进行测试。记其中错误分类的样本点个数为l_i，K次迭代完成后，便得到$l_1,l_2,...,l_K$，可以推想所有K次迭代中的错误分类数$\sum_{i=1}^{k} l_i$和总样本点数之比

$$\frac{\sum_{i=1}^{k} l_i}{l} \tag{4.2}$$

作为该算法错误率的一个估计，该值称为K重交叉确认误差，它经常作为衡量算法的一个指标。根据K次迭代得到的均方差的平均值来估计期望泛化误差，最后选择一组最优的参数。当K＝N时，其中N为训练样本数，每次迭代时只留下一个样本作为测试，也就是上面所讨论的留一法。交叉验证的准确率是K次的平均值，最后可以用测试平均错误率来评价样本所在的特征空间的分类推广能力。

网格搜索是交叉确认的一种方法。网格的意思是先选定一组(C,δ)的范围，将它们准确率连接起来绘图，然后确定准确率出现最高的一段小步长，逐步缩小选择范围，最终确定准确率最高的那对参数。例如，对$C=2^{-5},2^{-3},...,2^{15},\sigma=2^{-15},2^{-13},...,2^{-5}$分别依次进行验证，最终可以得到一组$(C,\delta)$值，使得分类精度达到最高；这种方法就叫做网格搜索法。根据实验，我们发现不断验证以指数增长的(C,δ)值，是确定最佳参数的一个有效的方法。事实证明，交叉验证和网格搜索法相结合，是一种很好的提高推广能力的模型选择方法。网格搜索方法逐步缩小参数选择范围，十分直观有效。它也是一种确认最佳参数的实用算法，其简单易懂，但似乎看起来有些愚笨，但实际应用中网格搜索得到了更加广泛的应用。究其原因有以下两点：一是大家认为近似的或启发式的交叉确认算法没有经过彻底的参数搜索而从心理上感觉不太可靠；二是因每一参数对(C,δ)相互独立，网格搜索可以并行计算。

3. 模型的选择

构造出一个具有良好性能的支持向量机，模型选择是关键。这里所谓的模型选择，其实就是如何针对所给的训练样本，确定一个比较合适的核函数及其参数与惩罚因子。模型选择包括两部分工作：一是核函数类型的选择，二是确定核参数及惩罚因子。为了能够获得好的分类效果，对 SVM 中的一些参数要仔细地选择，这些参数包括：（1）核函数类型及其参数：在第三章中已介绍过，只要一个函数满足 Mercer 条件，即它是一个正积分算子的边界对称核，那么它就可以作为一个核函数。它将高维特征空间中的内积运算转化为低

维输入空间上一个简单的函数计算。（2）惩罚因子(C)：它在超平面与最近的训练点之间的距离最大与分类误差最小之间寻求最佳折衷。

（1）核函数类型及参数的选取

现有的支持向量机方法一般使用四种核函数：线性核函数、多项式核函数、径向基函数、Sigmoid核函数。径向基函数可以通过非线性变换将特征向量变换到另一高维空间中，而线性核函数的映射为线性变换，它是径向基函数的一种特例。通过选择恰当的参数，径向基函数可以获得与Sigmoid核函数同样的效果。径向基函数与其它核函数相比较，径向基函数的参数个数简单；且它不会出现象多项式核函数中经常出现的无限大或零值的情况；也不会出现象Sigmoid核函数内积不存在的现象。

然而对于核函数类型及参数的选择，现有的研究还不能从理论上给出合适的核函数和参数选择方法，国际上还没有形成统一的模式，只是凭借经验、试验对比、大范围的搜索或利用软件包提供的交叉验证功能进行寻优。很多人认为使用交叉验证和网格搜索非常消耗时间，RBF函数需要确定的参数比较少，又实际应用中RBF函数也表现出了比较好的性能，因此实际应用中比较倾向于选择RBF核函数。本论文中，采用交叉验证与网格搜索相结合的方法来选择核函数类型及参数的取值。

（2）惩罚系数 C

从第三章的分析可知，在线性不可分的情况下，求最优分类面问题等同于求解下列问题：

$$\text{minimize: } \Phi(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C(\sum_{i=1}^{n}\xi_i) \tag{4.3}$$

$$\text{subject to: } y_i[(\omega^T X_i)+b]-1+\xi_i\geq 0, \tag{4.4}$$

$$\xi_i\geq 0, i=1,2,...,n \tag{4.5}$$

其中C>0，为惩罚系数。

这是一个二次规划问题，可通过Lagrange方法求解，转化为如下问题：

$$\text{maximize: } Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{4.6}$$

$$\text{subject to: } \sum_{i=1}^{n}y_i\alpha_i=0,\quad i=1,2,...,n \tag{4.7}$$

$$0\leq\alpha_i\leq C \tag{4.8}$$

最优分类面的权系数为：

$$\omega = \sum_{i=1}^{n} y_i \alpha_i x_i \tag{4.9}$$

可见，最优分类面的权系数向量是训练样本特征向量的线性组合。$\xi_i \geq 0$为松弛项，表示错分样本的惩罚程度；C 为常数，用于控制对错分样本的惩罚的程度，实现在错分样本数与模型复杂性之间的折衷；ω和 b 为判决函数中的权向量和阈值。当无错分样本时，最小化目标函数的第一项等价于最大化两类间的间隔，可降低分类器的 VC 维，实现结构风险最小化原则。

由上述式子可知，由于C的引入，α_i的值受到了限制，C越小，α_i的值也就越小。又由式（4.9）知，$\|\omega\|$也变小，而$2/\|\omega\|$是两类之间的间隔变大。它的大小体现了SVM泛化能力的强弱，间隔越大，泛化能力越强；反之，则泛化能力越弱。所以C越小，两类之间的间隔越大，SVM泛化能力越强，但显然这时SVM的分类准确率要降低。同理，C越大，两类之间的间隔越小，SVM泛化能力就越差，但这时的SVM分类准确率可以得到提高。

从以上分析知道，惩罚系数C是控制SVM的泛化性能和错分率之间的折衷。看看对偶问题，惩罚系数C的值越大，Lagrange算子α_i的约束力就越大，这就意味着要花更长的训练时间。这两个现象是互相联系的，过长的训练时间，往往意味着过适应和低下的泛化能力。其解决办法就是减小C，但是C的值也不宜太小。在线性不可分的情况下，$0 \leq \alpha_i \leq C$，C的值过小时，所有的α_i自然也会很小，泛化能力虽然高，但准确率无法保证。从另一方面来考虑，当C较大时，获得的解应具有$\sum_i \xi_i$大的性质，即较大化分类错误，即错分的惩罚力度较大，容易造成过学习；反之，C较小时，获得的解应具有$\sum_i \xi_i$小的性质，即较小化分类错误，对错分的惩罚力度较小，容易造成欠学习现象。因此，为了实现SRM原则，选择的C应当适中。

本文中使用了网格搜索方法来确定参数对（C，σ）。当使用网格搜索法时，可以获得一组比较好的（C，σ）值，这个参数可能是局部最优的。所以还需要根据搜索出的参数对，再选择一些参数值，验证是否能得到一组更好的（C，σ）值。

4.2.3 SVM 预测器

最后根据训练器学习得到的模型对未知的预测样本进行识别，也就是检验分类器的效果。实验中，通过对测试样本进行预测，将获得的预测值与真实值进行比较，最后将正确

的预测值除以总的预测样本数就得到正确的识别率。

为了得到好的识别率，需要注意以下几个问题：

- 将数据表示成程序所能处理的格式
- 特征向量要进行规范化
- 考虑不同核函数对结果的影响
- 用网格搜索法与多次实验的方法来得到最好的参数对

4.3 数据库的建立

本文的研究目的是将 SVM 方法应用于新生儿疼痛表情的等级分类，目前国内外对此的研究较少，也没有一个公用的图像数据库可供使用，因此我们必须先建立一个新生儿表情数据库，作为以下实验的数据样本。

4.3.1 图像的获取

本文所用图像均拍摄于南京医科大学第二附属医院妇产科，得到了院方及家长的同意，拍摄过程均有医护人员在场。拍摄对象是 7 天以内的新生儿，所有的新生儿都很健康。所有照片均是用 Nikon 数码相机在室内光线充足的条件下拍摄。

由于外界刺激的不同会导致新生儿的面部表情变化不同，如何选取不同的外界刺激来促使新生儿表情变化是实验方法设计上需要考虑的关键环节之一。所以根据本课题的要求，对如下几种不同情况下新生儿的不同表情进行了拍摄：

（1）安静：当新生儿处于睡眼状态或睁着眼睛观察周围的事物；

（2）更换婴儿床：更换婴儿床一分钟的过程中新生儿从睡眠到哭的表情变化；

（3）吹气：利用塑料的可挤压的相机镜头清洁器对睡眠状态的新生儿鼻子吹气；

（4）摩擦：利用蘸有 70%酒精的棉签摩擦新生儿的脚跟外侧 10 秒到 15 秒；

（5）疼痛：疼痛的表情是在两种情况下拍摄的：打针时，新生儿安静一分钟后，护士给它们打针，立即拍下此时新生儿的面部表情；采血时，当护士在新生儿足跟采血的过程中，拍下一系列表情。

图 4-3 中的 5 种不同状态下的新生儿表情图像，从左到右依次为安静、更换婴儿床、吹气、磨擦及扎针时所表现出的面部表情。表情图像如下：

图 4-3 五种基本表情

针对课题研究和应用的需要，我们利用 Nikon D100 数码相机采集了 40 个新生儿（20 个男孩，20 个女孩）进行了面部图像，一共采集了 400 多幅不同表情的彩色图像，经过严格的筛选最终选定了 350 幅图像，建立了一个规模较小的新生儿表情数据库。

4.3.2 表情分类

本课题的主要目标是识别新生儿的不同程度疼痛与非疼痛的表情，并要根据疼痛程度对疼痛表情进行等级分类。根据本课题的要求，将新生儿表情图像分为 3 种：（1）安静状态，包括新生儿安静时和吹气时所拍得表情图像；（2）哭状态，包括更换婴儿床时和摩擦时所拍表情图像；（3）疼痛状态，包括打针和采血时拍表情图像。我们发现在移动婴儿时，也会刺激他哭泣，而这并非疼痛引起的哭泣时的表情与疼痛引起的咳期的面部表情类似，所以这种刺激情况下的图像与疼痛时面部表情进行分类实验。

根据实验的要求，在最终建立的 350 幅新生儿照片数据库中，有 94 幅安静状态下面部表情照片，80 幅哭状态下的表情照片，176 幅具有疼痛的表情的照片。其中安静的表情包括安静状态下和吹气时新生儿所表现出的表情，分别有 43 幅和 51 幅；哭的表情包括更换婴儿床和摩擦时的新生儿所表现出的表情，分别有 49 幅和 31 幅。根据专家评定，将疼痛表情按疼痛程度分类，疼痛程度分为轻度和剧烈两种，分别为 96 幅和 80 幅。最后我们将新生儿面部表情图像分为四类：类 1 为安静的表情，共 94 幅图像；类 2 为哭的表情，共 80 幅图像；类 3 为轻度疼痛的表情，共 96 幅图像；类 4 为剧烈疼痛的表情，共 80 幅图像。

4.3.3 样本数据的获得

要获得分类所需的样本数据必须对图像进行预处理和特征提取，即把一组图像转化为一组向量数据表示。

面部表情图像的预处理包括表情图像的尺度归一化和灰度均衡。尺度归一化将所有人脸图像调整到统一的标准尺寸（112×92 像素的矩形区域），包括两个独立的算法：首先，需要一个算法来定义空间变换本身，即按所需图像的大小完成像素点的增删和移动；同时，采用双线性插值算法对灰度级插值，以保持图像变换前后的连续行。灰度均衡使得所有图像具有相同的灰度均值和方差，即面部图像的明暗程度及对比度得到统一，从而有效屏蔽了光照变化等因素所产生的影响。预处理后的新生儿表情图像如图 4-4 所示。

图 4-4 预处理后的四类表情图像

面部表情图像的特征提取方法：首先将新生儿表情图像进行 Gabor 小波变换，既将其与一系列 Gabor 小波函数进行卷积，使用 40 个 Gabor 滤波器（5 个尺度，8 个方向的 Gabor 小波函数簇）进行滤波，并利用快速傅立叶变换实现了其快速算法。经过采样后的特征向量维数依然很高，引入 Adaboost 算法进行特征选择，对 Gabor 特征降维，识别率根据选择的特征数目不同而有所不同，实验证明 500 维的特征向量得到的识别率最高。对图像进行预处理和特征提取的具体算法见参考文献[86]。

4.4 构造二叉树 SVMs 分类器

根据第三章对各种多分类 SVMs 算法的研究，本文采用改进的二叉树 SVMs 分类器对新生儿疼痛表情进行等级分类。在设计这个 SVMs 分类器时，选择何种结构的二叉树至关重要，它直接影响分类器的训练时间、识别率、支持向量个数等其他性能。为此我们做了以下的工作。

在构造二叉树SVMs分类器时，由于若在某个节点上发生分类错误，则会把分类错误延续到该节点的后续节点上，所以我们希望越是靠近根节点的单个分类器分类精度越高，为此在根节点处的分类器，考虑根据类间可分离测度将最容易分开（不易产生错分）的两

类作为此处分类器的训练样本，在下一个节点选取当前最易分离的两类为分类器训练决策面，依次类推，这样就能够使可能出现的错误尽可能的远离决策树的根，减小累积误差。

4.4.1 对样本数据的分析

首先我们需要对样本数据进行分析。我们将样本数据分为四类：类 1 为安静的表情；类 2 为哭的表情；类 3 为轻度疼痛的表情；类 4 为剧烈疼痛的表情。通过观察图像可以发现，新生儿安静表情，与哭的表情和疼痛表情差别很大，而哭的表情和疼痛表情非常相似，可以判定安静的表情和哭的、疼痛的表情最易分开，哭与疼痛的表情的分类精度有所下降，而两种程度的疼痛表情最难分开。

据此我们可以假设这四种类别在二维空间中有以下关系，如图 4-5 所示：

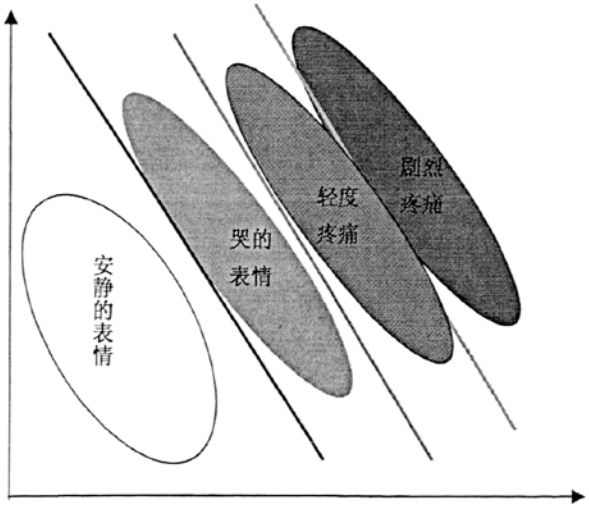


图 4-5 假设四类表情数据的二维分布

4.4.2 类间分离度的计算

我们可以用类间可分离测度对样本数据进行更直观分析，用数字描述各个类间距离即可分离程度。

1. 基于类间距离的分离性测度

设两个样本的特征值分别为 $\boldsymbol{X}_i, \boldsymbol{X}_j$，即

$$\boldsymbol{X}_i = \left(x_{i1}, x_{i2}, \cdots, x_{in}\right) \quad \boldsymbol{X}_j = \left(x_{j1}, x_{j2}, \cdots, x_{jn}\right), \tag{4.10}$$

计算类与类之间的距离采用重心法，公式如下：

$$D_{i,j}=\left|\overline{X_{w_i}}-\overline{X_{w_j}}\right|, \tag{4.11}$$

$$\overline{X_{w_i}}=\frac{1}{N_i}\sum_{X\in w_i}X,\ \overline{X_{w_j}}=\frac{1}{N_j}\sum_{X\in w_j}X, \tag{4.12}$$

w 代表某类样本的集合，N_i,N_j 分别是 w_i,w_j 类中样本的个数。

两个样本间距离的计算欧氏距离法，公式如下：

$$\text{欧氏距离}\quad D_{i,j}^2=\left(X_i-X_j\right)^{\mathrm{T}}\left(X_i-X_j\right)=\left|X_i-X_j\right|^2=\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2, \tag{4.13}$$

$D_{i,j}$ 越小则两个样本距离越近，就越相似；

按欧氏距离计算的两两类间距离为：D_{12}=1.10554，D_{13}=1.20754，D_{14}=1.26886，D_{23}=1.06673，D_{24}=1.06668，D_{34}=0.758801；将这 6 个距离按升序排列，其中 D_{14} 最大，类 1 与类 4 两类样本间距离最大，最易分开；D_{34} 最小的即两类样本间距离最小，最难分开。

2. 基于最优分类面的分离性测度

对于这四类表情，对每两类分别单独进行二类 SVM 分类，共需训练 6 个二类 SVM 分类器，得到 6 个分类决策函数为 $f_{12},f_{13},f_{14},f_{23},f_{24},f_{34}$，决策矩阵 F 如下：

$$F=\begin{bmatrix}[\ ] & f_{12} & f_{13} & f_{14}\\ f_{12} & [\ ] & f_{23} & f_{24}\\ f_{13} & f_{23} & [\ ] & f_{34}\\ f_{14} & f_{24} & f_{34} & [\ ]\end{bmatrix}, \tag{4.15}$$

分别记录每个决策函数正确分类的样本数，计算分类性测度，例如对类别 1 和 2 有如下计算：

$$sm_{12}=\frac{N_{12}^s}{\left(N_1+N_2\right)}, \tag{4.16}$$

其中，N_1 为类 1 的训练样本数，N_2 为类 2 的训练样本数，N_{12}^s 为对类 1 与类 2 进行训练后两类中本最优分类面正确分类的训练样本数。

实验计算得到 sm_{12}=0.8571，sm_{13}=0.8796，sm_{14}=0.7486，sm_{23}=0.6404，sm_{24}=0.6728，sm_{34}=0.8652；分离性测度矩阵

$$SM=\begin{bmatrix} [\,] & 0.8571 & 0.8796 & 0.7486 \\ 0.8571 & [\,] & 0.6404 & 0.6728 \\ 0.8796 & 0.6404 & [\,] & 0.8652 \\ 0.7486 & 0.6728 & 0.6728 & [\,] \end{bmatrix}$$

按升序排列，sm_{13}最大，表明类 1 与类 3 是最易分开的两类，sm_{23}最小，表明类 2 与类 3 是最难分开的两类。最后计算各类的平均分离度$\overline{sm_1}=\frac{1}{6-1}\sum_{i=2}^{6}sm_{1i}=0.8284$，$\overline{sm_2}=0.7234$，$\overline{sm_3}=0.7951$，$\overline{sm_4}=0.7622$。

4.4.3 疼痛表情等级分类器

根据基于类间欧氏距离的$D_{i,j}$的计算结果来设计分类器：在根节点 SVM 分类器选择与其它类的类间距离最大的类 1 为正类，将类 2、类 3 和类 4 合并为一类作为负类，训练分类器得到第一个分类器 SVM1；其子节点选类 2 为正类，类 3 和类 4 合并为一类作为负类，训练分类器得到第二个分类器 SVM2；最后类间距离最近的类 3 和类 4 作为正和负类，训练分类器得到叶节点分类器 SVM3，结构如图 4-6。这种结构符合图 4-5 中的假设。

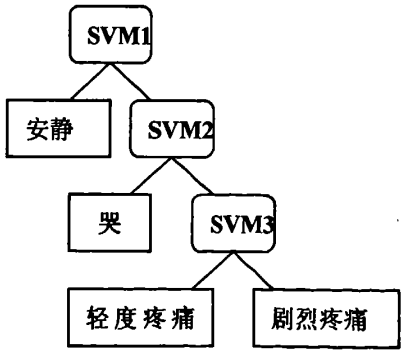


图 4-6 基于类间距离的分类器结构

由于类 1 的样本最易与其它类分离，将它安排在第一层的分类器，两种不同程度的疼痛最不易分类，尽可能将放在底层的结点分类器，以提高整个决策树分类器的分类精度。

根据基于类间可分离测度sm的计算结果来设计分类器：在根节点 SVM 分类器同样选择平均分离度最大的类 1 为正类，将类 2、类 3 和类 4 合并为一类作为负类，得到第一个分类器 SVM1；其子节点选平均分离度次之的类 3 为正类，类 2 和类 4 合并为一类作为负类，训练得到第二个分类器 SVM2；最后将类 4 作为正类，类 2 作为负类，训练分类器得到叶节点分类器 SVM3，结构如图 4-7。

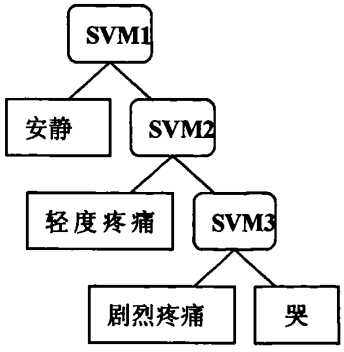


图 4-7 基于类间分离度的分类器结构

在后面的实验中，我们将根据基于类间距离得到的二叉树结构的多类 SVM 分类器称为 BT-SVMs，将根据基于类间可分离测度得到的二叉树结构的多类 SVM 分类器称为改进的 BT-SVMs。

4.5 实验过程与结果分析

实验的硬件配置为：512M 的内存、Intel(R) Pentium(R)4 CPU 2.66GHZ。

编程环境和语言为：WindowsXP 系统，编译器为 Microsoft VC6.0，编程语言为 C/C++。

实验的过程分为：预处理、特征提取和分类，如图 4-8 所示。

在预处理阶段，先对原始图像进行裁减、翻转，尺度归一化和灰度均衡。原有 2592×1944 像素的图像处理后变为 112×92 像素。

在特征提取阶段，先将新生儿表情图像进行 Gabor 小波变换，经过采样后的特征向量维数依然很高，再利用 Adaboost 算法对采样后的 Gabor 特征进行特征选择，最后得到 500 维的特征数据[86]。

最后是分类，本文采用 SVMs 多类分类器对特征向量进行多类分类。首先，将所有样本的特征向量进行规范化，使特征向量的数值在一定的范围内。然后，将规范化处理后的训练样本数据作为 SVMs 训练器的输入，得到一个模型文件。该模型文件中记录了训练时使用的核函数类型及其参数值、支持向量及其数量、拉格朗日乘子和判决函数的值。再将训练器得到的模型文件与规范化处理后的测试样本数据作为预测器的输入，通过 SVMs 预测器得到如下结果：测试样本的预测值（类标号）及总的识别率。最后，通过人为的预先的规定及理解，将这些标号解释为不同类型的表情。

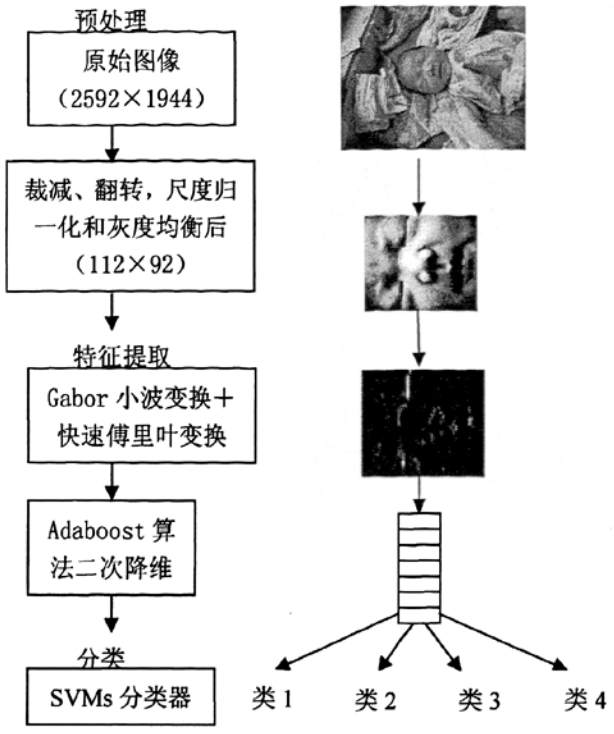


图 4-8 实验过程

4.5.1 样本数规范据化

一幅图像经特征提取后得到一个 500 维的样本特征向量，现仅列出其中一个数据的一部分前 50 维数据，如下所示：

1:1.6873836e+017	2:1.7967833e+063	3:2.4037668e+010	4:8.0668794e+092
5:6.2677926e+000	6:9.4173329e+010	7:6.8084181e+004	8:1.6554176e+017
9:2.8630678e+134	10:2.4037668e+010	11:1.1136555e+011	12:1.6039936e+011
13:1.1564749e+017	14:1.6873836e+017	15:1.7967833e+063	16:3.1767476e+004
17:4.6891137e-001	18:2.5222480e+010	19:5.9153184e-007	20:8.3122012e+092
21:6.8251064e+004	22:6.0431711e+009	23:2.4037668e+010	24:7.9579851e+044
25:8.7180572e+010	26:2.4293815e-007	27:5.3599337e+010	28:6.8084181e+004
29:1.6554176e+017	30:5.9177222e+134	31:8.2329325e+009	32:5.7847545e-001
33:2.8630678e+134	34:8.2454530e-007	35:1.6039936e+011	36:1.6873836e+017
37:1.7967833e+063	38:2.4037668e+010	39:9.8816826e-001	40:5.5195830e-001
41:5.9153184e-007	42:8.3122012e+092	43:6.8251064e+004	44:1.7967833e+063

45:3.6321887e+000 46:2.4037668e+010 47:8.7180572e+010 48:5.1849920e-008 49:2.4293815e-007 50:8.3122012e+092

上述特征向量中的第一个数字代表此特征向量所属类别为 1，而 1:、2:等表示特征值的序号（索引号），后面的数值代表提取出的特征值大小。本文中所有实验中使用的特征向量都是将特征值规范化在[-1,1]范围内，规范化后的特征向量如下：

1 1:0.343108 2:0.0921285 3:-0.766022 4:0.350559 5:1 6:0.261016 7:0.435578 8:0.318284 9:-0.320639 10:-0.766022 11:-0.551643 12:-0.321495 13:-0.342292 14:0.343108 15:0.0921285 16:-0.433613 17:-0.858052 18:-0.752814 19:0.0851365 20:0.393026 21:0.393985 22:-1 23:-0.766022 24:0.378526 25:0.240813 26:-0.623033 27:-0.501819 28:0.435578 29:0.318284 30:0.403533 31:-0.984919 32:-0.778922 33:-0.320639 34:0.943568 35:-0.321495 36:0.343108 37:0.0921285 38:-0.766022 39:-0.88407 40:-0.801214 41:0.0851365 42:0.393026 43:0.393985 44:0.0921285 45:0.160432 46:-0.766022 47:0.240813 48:-0.471781 49:-0.623033 50:0.393026

从上面规范后的特征数据中可以看出，每一列中总有一个数据为+1，同时也有一个数据为-1。+1 与-1 分别表示在原始数据中，它们是这一列中的最大及最小值，其它的数值都是通过公式（4-1）所计算出的结果。

数据规范化的目的是：减少训练与预测过程中核函数的计算时间。因为在进行训练与预测时，都需要使用特征向量来计算核函数的值，原始特征向量的数值范围广，且它远大于规范化后的数据值，因此增加了计算量和程序的运行时间。

4.5.2 参数与核函数选择实验

参数的寻找过程：本文中使用了网格搜索法来搜索最佳参数对(C,g)，其中C表示惩罚因子，g表示核函数前面的系数(前面所提到的σ)。在搜索的过程中，并绘制了图形用来理解整个搜索过程。图 4-9 为使用 Gabor 小波+Adaboost 方法提取出的特征向量进行实验所绘制图像：

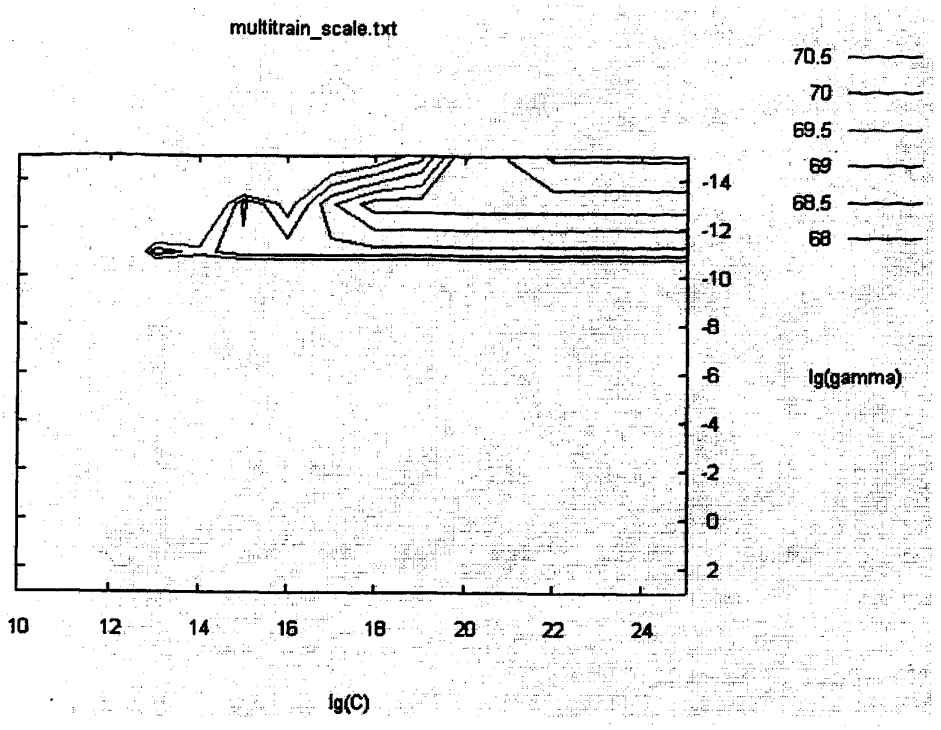


图 4-9 多项式核函数的参数选择

上图中的横、纵坐标分别为取惩罚因子C与核函数的系数g以 10 为底的对数。图中六条不同颜色的实线代表了不同识别率下的(C,g)等高线，其中绿色的线索代表的识别率最高，(C,g)就是在最高识别率下获得的。搜索到最佳参数对为：C=1048576，g=3.0517578125e-005；本文以后的实验中参数对(C,g)均采用此值。

针对一对一 SVMs、一对多 SVMs、DDAG-SVMs、BT-SVMs 和改进的 BT-SVMs 分类器，本文使用 K 重交叉认证方法确定核函数。实验方法如下：先将 350 个样本的特征向量进行规范化；再分别使用线性、多项式、RBF、Sigmoid 四种核函数进行 10 重交叉验证。交叉验证的实验步骤如下：

（1）先将所有的样本分成数量相同的 10 份，每份 35 个样本；

（2）第一次将第 1 份用于测试，其它的 9 份用于训练，获得此次的识别率；第二次将第 2 份用于测试，剩下的 9 份用于训练，记录此时的识别率，例如在 35 个测试样本中 28 个被正确分类，则识别率为 28/35=80%；依此下去将此过程重复 10 次。最后计算这 10 次的平均值就得到了此次实验的平均识别率；

（3）将步骤（1）与步骤（2）重复 10 次；

（4）再将 10 次实验中得到的识别率取平均，获得最终的识别率。

实验中，取惩罚系数 C＝1048576，系数 g =3.0517578125e-005，使用 Gabor 小波+Adaboost 方法提取出的特征向量进行实验，结果如表 4-1 至表 4-5 所示：

表 4-1 各算法多项式核函数阶数的选择

各种算法	多项式核函数						
	d=1	d=2	d=3	d=4	d=5	d=6	d=7
一对一 SVMs	65.44%	65.16%	66.29%	66.01%	64.87%	64.02%	63.17%
一对多 SVMs	80.17%	80.74%	80.45%	77.90%	75.92%	73.37%	69.12%
DDAG-SVMs	66.86%	67.14%	67.42%	68.27%	69.12	67.14%	64.02%
BT-SVMs	82.15%	81.02%	82.44%	82.44%	82.72%	82.15%	78.75%
改进的 BT-SVMs	85.55%	86.12%	85.27%	84.99%	81.59%	79.60%	76.77%

表 4-1 是各种多分类算法在选择选多项式为核函数时对不同阶数 d 进行的实验，对其作图表显示如下：

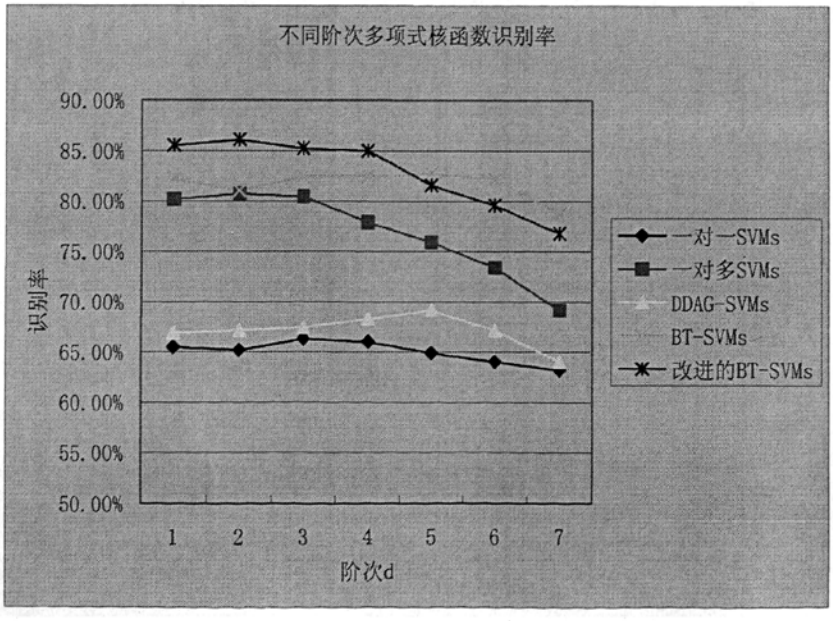


图 4-10 各算法多项式核函数不同阶数的识别率

从上图中可以发现识别率随着多项式的阶数 d 的变化而变化，每种算法达到最好识别率时的阶数都各不相同，需分别选择：对一对一 SVMs 阶数 d=3 时识别率最好，对一对多 SVMs 阶数 d=2 时的识别率最好，对 DDAG-SVMs 阶数 d=5 时的识别率最好，对 BT-SVMs 阶数 d=5 时获得最高识别率，对改进的 BT-SVMs 阶数 d=2 时的识别率最好，而当阶数分别增大或减小时识别率均减小。

表 4-2 一对一 SVMs 核函数选择

核函数类型	线性核函数	多项式核函数 d=3	径向基函数	Sigmoid 核函数
识别率	66.01%	66.29%	66.01%	66.57%
总时间(ms)	150266	21188	34109	29031
支持向量数	156	192	171	176

表 4-2 是对一对一 SVMs 选择核函数的实验中，表中总时间是指交叉验证的总运行时间。可以看出在各种函数中，Sigmoid 核函数所获得识别率最高，多项式核函数次之。在采用一对一 SVMs 方法实现分类器时选择 Sigmoid 核函数。

表 4-3 一对多 SVMs 核函数选择

核函数类型	多项式核函数 d=2	径向基函数	Sigmoid 核函数
识别率	80.74%	81.30%	79.89%
总时间(ms)	69875	104429	87828
支持向量数	226	212	226

表 4-3 是对一对多 SVMs 选择核函数的实验中，结果显示，径向基函数作为核函数时所获得识别率最高为 81.3%，在采用一对多 SVMs 方法实现分类器时选择径向基函数作为核函数。

表 4-4 DDAG-SVMs 核函数选择

核函数类型	线性核函数	多项式核函数 d=5	径向基函数	Sigmoid 核函数
识别率	65.44%	69.12	69.12%	66.57%
总时间(ms)	123781	12078	32516	35906
支持向量数	151	196	172	176

表 4-4 是对 DDAG-SVMs 选择核函数的实验，从实验结果可以看出，当多项式核函数阶数 d=5 时获得与径向基函数相同的识别率，最高为 69.12%，与径向基函数相比，多项式核函数的速度更快，因此在采用 DDAG-SVMs 方法实现分类器时选择阶数 d=5 的多项式核函数。

表 4-5 BT-SVMs 核函数选择

核函数类型	多项式核函数 d=5	径向基函数	Sigmoid 核函数
识别率	82.72%	81.59%	80.17%
总时间(ms)	12859	38516	33828
支持向量数	180	147	155

表 4-5 是对 BT-SVMs 选择核函数的实验，从表中可以看出，当多项式核函数阶数 d=5 时获得最高，为 82.72%，且速度最快，因此在采用 BT-SVMs 方法实现分类器时选择阶数 d=5 的多项式核函数。

表 4-6 改进的 BT-SVMs 核函数选择

核函数类型	多项式核函数 d=2	径向基函数	Sigmoid 核函数
识别率	86.12%	84.70%	86.12%
总时间(ms)	33750	46500	40750
支持向量数	171	153	167

表 4-6 是对改进的 BT-SVMs 选择核函数的实验，当多项式核函数阶数 d=2 时获得与 Sigmoid 核函数相同的识别率，最高为 86.12%，但多项式核函数的运行时间少于 Sigmoid 核函数，其训练和预测速度更快，因此在采用改进的 BT-SVMs 方法实现分类器时选择阶数 d=2 的多项式核函数。

在对这 5 种 SVMs 方法进行核函数类型的选择实验过程中，发现使用线性核函数进行实验时，迭代的次数非常高，所花费的时间也比其它方法长的多，以至于在一对多 SVMs、BT-SVMs 和改进的 BT-SVMs 的实验中无法正常训练样本数据，而在一对一 SVMs 和 DDAG-SVMs 的实验中其运行时间几乎是多项式核函数的 10 倍以上，因此不再考虑线性核函数作为 SVM 的核函数。而当使用多项式核函数时，阶数越高分类器运行速度越快，所以当折衷考虑分类精度和分类速度两方面的因素时，在保证识别率的前提下，应尽量选用阶数较高的多项式核函数。

4.5.3 各种多类分类 SVMs 算法的比较

在对各 SVMs 分类器选择最佳核函数后，再重新对新生儿的表情进行多类分类实验，从训练时间，识别率和支持向量个数方面比较这几种方法的实验结果。实验中不再采用交叉验证的方法，在 350 个样本中，从 90 幅类 1 安静的表情中随机选 9 幅，80 幅类 2 哭的表情中随机选 8 幅，100 幅类 3 轻度疼痛的表情中随机选 10 幅，80 幅类 4 剧烈疼痛的表情中随机选 8 幅，共 35 幅表情图像样本组成测试样本集，其余样本作为训练样本；参数使用了参数对 C=1048576，g=3.0517578125e-005，获得的实验结果如下：

表 4-7 各种多分类 SVMs 算法的比较结果

多分类 SVMs 算法	一对一 SVMs	一对多 SVMs	DDAG-SVMs	BT-SVMs	改进的 BT-SVMs
支持向量数	182	223	201	163	199
训练时间(ms)	3016	10734	1156	1187	1234
识别率(%)	68.57%	80.00%	71.43%	82.86%	88.57%

在此实验中对改进的 BT-SVMs 分类器选择了更高阶的多项式分类器（d=5），获得了很好的识别率。从实验中我们得到结论，对新生儿疼痛表情进行多类分类时，在这几种多类分类 SVMs 方法中，基于二叉树的 SVMs 分类器的识别率最高，基于类间距离构造二叉树结构的 SVMs 分类器的识别率可达 82.86%，而按本文提出的类间分离度进行改进后的二叉树 SVMs 分类器的识别率可达到 88.57%；在训练时间方面，DDAG-SVMs、BT-SVMs 和改进的 BT-SVMs 三种分类器所消耗时间基本相同，它们的训练速度较快，一对多 SVMs 分类其所需时间最长，几乎是最快速度的 10 倍。在进行大规模样本训练时，训练时间是评判分类器性能的一个重要标准，在对训练时间和识别率两方面进行考虑后，改进后的二叉树 SVMs 分类器与其它几种方法相比，即可得到较好的识别分类效果，又花费较少的训练时间，其总体性能较其它几种方法理想。

4.6 本章小结

本章主要讨论了基于多类 SVM 的新生儿疼痛表情等级识别模块的建立。首先分三步设计了 SVM 识别系统，接着根据所建新生儿面部表情数据库的特点，构造了基于改进的二叉树结构的 SVMs 多类分类器，实现了对新生儿疼痛表情的多类别（安静、哭、轻度疼痛和剧烈疼痛）识别，最后通过实验验证该分类器的性能，并且和其他多类分类 SVMs 分类器进行比较，证明其优越性，它不仅训练时间短，而且能获得较好的识别效果。

对于新生儿表情的多类分类采用改进的二叉树结构的 SVMs 多类分类方法获得较好的识别率，但还有很多问题待改善，例如哭和疼痛表情很接近，怎样选取其特征使两者更易分类等。

第五章 总结与展望

随着表情识别技术的迅猛发展，作为表情识别系统核心部件的表情识别分类器得到了广泛的应用。本课题根据新生儿疼痛表情识别目前存在的主要问题，结合近年来发展迅猛的机器学习工具—支持向量机(SVM)技术，进行了将SVM技术应用于新生儿疼痛表情识别多类分类领域的研究。支持向量机拥有一套坚实的理论基础，被广泛的用于求解模式识别、回归分析等问题。但是，虽然SVM在理论上有很突出的优势，但其还存在很多问题尚未解决，尤其是在多分类方面，其应用研究也相对滞后；另外在新生儿疼痛表情识别领域应用SVM技术目前公开的实验、应用报道还不多见，因此本文所开展的研究具有理论和现实两方面的重要意义。

支持向量机作为一种基于统计学习理论的机器学习方法，较好地解决了小样本学习问题。表情识别属于样本有限的模式识别问题，所以本文将支持向量机的方法用于本课题的研究中。本文的工作主要包括以下几个部分：

（1）本课题主要探索新生儿疼痛面部表情多类识别的难点和关键技术问题，为开发基于面部表情识别的新生儿疼痛自动评估系统打下基础。本文阐述了新生儿的表情识别系统的研究意义及现状，简要介绍了表情识别系统的四个部分：图像获取、人脸检测、面部表情特征提取、面部表情的情感分类。基于课题的要求及支持向量机方法的优点，本文将支持向量机的方法应用于新生儿面部表情识别的研究中。

（2）在机器学习、统计学习的理论基础上详细介绍了SVM方法的概念和特点，全面系统地阐述了如何将SVM方法识别分类研究中。因为模型选择是是否能获得好的SVM分类器的关键因素，所以本文使用交叉验证与网格搜索相结合的方法来确定SVM的模型。

（3）针对多类分类情况，对目前常用的SVMs多类分类算法进行了深入研究，从训练和测试两方面比较了各种分类方法的性能。在基于决策树的SVMs分类方法的基础上，详细介绍了基于二叉树的SVMs分类器，并根据类间距离来设计二叉树的结构。

（4）为了减小BT-SVMs分类器误差的累积效应，提出了改进的二叉树SVMs分类器，根据SVM最有分类面的性质设计类间分离性测度，根据此类间分离性测度设计二叉树结构，这样结合了样本数据的特点可提高分类器的准确率和速度。实验结果证实了此模型的正确性和有效性。

（5）建立了新生儿面部表情数据库，将新生儿表情分为安静、哭、轻度疼痛和剧烈

疼痛四种类别。根据此样本数据的特点，构造了基于改进的二叉树结构的SVMs多类分类器，实现了对新生儿疼痛表情的多类识别，在与其他多类分类SVMs分类器进行比较中，实验结果证明其节约了训练和预测时间，用较少的支持向量数也能获得较好的识别效果。

综上所述，本论文在支持向量机多分类的应用方面取得了一定的进展，并获得了一些成果。但仍存在很多问题，在以下方面还需要做进一步的研究：

（1）建立一个大规模的新生儿表情库。本文中选用的200幅新生儿的面部图像进行实验，图像的数量是远远不够的。为了进一步研究新生儿的面部表情，应该建立一个表情丰富，且图像比较多的新生儿面部表情库。

（2）本课题研究对新生儿疼痛程度的分类。但目前只对疼痛作了最初步的分类，因为引起新生儿疼痛的因素很多，疼痛的程度也有所不同，因此需要进一步对新生儿的疼痛表情做细致的分类，这样可以为医护人员客观、可靠、有效地评估疼痛提供科学的依据。

（3）本课题的目标是能快速识别出新生儿的疼痛表情并进行分类。因此需要研究支持向量机的快速算法，在对大规模的图像库进行训练时，能够提高训练速度。

（4）目前SVM方法中最重要的问题之一就是核函数的选取与构造。如何选择最优核函数及其参数，大多以经验或反复试算，虽然已经有许多研究人员提出了各种各样的选取与构造的方法，但总的来说这方面的研究还处于探索的阶段，还有许多值得深入研究的问题。例如在本课题中，如果能根据新生儿面部表情的特性，构造出一个适合该问题的核函数，则分类效果会得到改进。

（5）对于多类分类问题，在模型参数的选择上，本文采取的方法是所有子分类器采用统一的参数。为每个子分类器优化自身参数能得到更好的分类精度，但是这样会进一步加大训练时间。如何在子分类器的优化与训练时间之间寻求一个平衡还有待探讨。

（6）多类支持向量机的构建缺乏一个最优方案，大多数研究表明，目前的几种多类支持向量机的构建方法，还没有一种能在所有方面稳定的胜过其它方法，这在一定程度上阻碍其应用于实际。因此需要对构建多类支持向量机的最优方法做进一步研究。

致 谢

特别感谢卢官明老师，他在这三年时间中给予我悉心指导与帮助，使我终于能够完成毕业论文的撰写，并且从中获得了许多知识。卢老师的渊博学问及认真的工作态度，为我树立了良好的榜样。而他平和的态度接近了师生之间的距离，给我留下了深刻的印象。

感谢教研室的其他的老师和同学，我从他们那里学到了许多东西。同时感谢邹婵洁、罗智兰、陈大伟、焦方飞、唐红及其他的师弟与师妹们平时的帮助与支持。

最后，感谢答辩委员会的各位专家老师，对他们的评审表示诚挚的谢意。

作者

2008.03

参考文献

[1] C. R. Chapman, K. L. Syrjala. Measurement of pain, in: J.J.Bonica, J.D.Loeser, C.R. Chapman, W.F.Fordyce (Eds.), The Management of Pain, Vol. 1. Lea and Febiger, Philadelphia, 1990: 580-594.

[2] D. Harrison, C. Evans, L. Johnston, P. Loughnan, Bedside assessment of heel lance pain in the hospitalized infant, Journal of Obstetric, Gynecologic, and Neonatal Nursing , 2002, 31(5): 551-557.

[3] H.Merskey, D. G. Albe_fessard, J. J. Bonica, A. Carmon, R. Dubner, F. W. L. Kerr, U. Lindblom, J. M. Mumford, P. W. Nathan, W. Noordenbos, C. A. Pagni, R. A. Sternbach, S. S. Sunderland. Pain terms: a list with definitions and notes on usage recommended by the IASP subcommittee on taxonomy, Pain, 1979, 6: 249-252.

[4] K. J. S. Anand, R. E. Grunau, T. Oberlander. Developmental character and long-tern consequences of pain in infants and children, Child and Adolescent Psychiatric Clinics of North America, 1997, 6: 703-724.

[5] R. E. Grunau. Long-term consequences of pain in human neonates, in: K. J. S. Anand, B. J. Stevens, P. J. Mcgrath (Eds), Pain in Neonates: 2nd Revised and Enlarged Edition, Elsevier, New York, 2000: 55-76.

[6] B. Stevens, C. Johnston, S. Gibbins. Pain assessment in neonates, in: K. J. S. Anand, B. J. Stevens, P. J. McGrath (Eds.), Pain in Neonates, 2nd Revised and Enlarged Edition, Elsevier, New York, 2000: 101-134.

[7] K. D. Craig. The facial display of pain in infants and children, in: G. A. Finley, P. J. McGrath (Eds.), Measurement of Pain in Infants and children, Pain Research and Management, vol. 10 IASP Press, Seattle, 1998: 103-121.

[8] B. M. Lester. A biosocial model of infant crying, in: L. Lipsill (Ed.), Advances in Infancy Research, Ablex, New York, 1984: 167-212.

[9] C. C. Johnston, M. E. Strada. Acute pain response in infants: a multidimensional description, Pain, 1986, 24: 373-382.

[10] B. Stevens, C. Johnston, P. Petryshen, et al. Premature infant pain profile: Development and

initial validation. The Clinical Journal of Pain, 1996, 12(1): 13-22.

[11] J. Lawrence, D. S. Alcock, P. J. McGrath, et al. The development of a tool to assess neonatal pain. Neonatal Network, 1993, 12(6): 59-66.

[12] C. Pasero. Pain assessment in infants and young children: neonates. The American Journal of Nursing, 2002, 102(8): 61-64.

[13] R. E. Grunau, T. Oberlander, L. Holsti, et al. Bedside application of the neonatal facial coding system in pain assessment of premature neonates. Pain, 1998, 76(3): 277-286.

[14] J. J. Lien. Automatic recognition of facial expression using Hidden Markov Models and estimation of expression intensity: [Ph. D. Dissertation]. Pittsburgh: University of Pittsburgh, 1998.

[15] Y. Yacoob, L. S. Davis. Computing spatio-temporal representations of human faces. In: Proc Of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Seattle, 1994: 70-75.

[16] I. Cohen, Garg Ashutosh, S. Thomas. Huang. Emotion Recognition using Multilevel-HMM. NIPS Workshop on Affective Computing, Colorado, 2000.

[17] Hai Tao, Thomas Huang. Explanation-based facial motion tracking using a piecewise Bezier volume deformation model. In Proc Of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99). Ft. Collins, CO, USA, 1999: 611-617

[18] K. Mase, A. Pentland. Recognition of facial expressions from optical flow. IEICE Transactions, Special Issue on Computer Vision and its Applications, 1991, E74(10): 3474-3483.

[19] T. Sakaguchi, S. Morishima. Face feature extraction from spatial frequency for dynamic expression recognition. In: Proc Of 13th International Conference on Pattern Recognition. Vienna, 1996: 451-455.

[20] T. Otsuka, J. Ohya. Recognition of facial expressions using HMM with continuous output probabilities. In: Proc Of 5th IEEE International Workshop on Robot and Human Communication. Roman, 1996: 323-328.

[21] T. Otsuka, J. Ohya. Recognizing multiple persons'facial expressions using HMM based on automatic extraction of significant frames from image sequences. In: Proc Of IEEE International Conference on Image Processing. Santa Barbara, CA, USA, 1997: 546-549.

[22] B. Fasel, J. Luettin. Automatic facial expression analysis:a survey. Pattern Recognition,

2003, 36(1): 259-275.

[23] M. Rosenblum, Y. Yacoob, L. S.Davis. Human emotion recognition from motion using a radial basis function network architecture. In: Proc Of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, TX, 1994: 43-49.

[24] 金辉，高文. 基于HMM的面部表情图像序列的分析与识别. 自动化学报，2002，28(4): 646-650.

[25] Xiaoxu Zhou, Xiangsheng Huang, Bin Xu, Yangsheng Wang. Real-time Facial Expression Recognition Based on Boosted Embedded Hidden Markov Model. IEEE Trans. On Proeeedings of the Third International Conference on Image and Graphics, 2004.

[26] Y B Wang, H Z Ai, B Wu, et al. Real time facial expression recognition with adaboost. In: Proc Of the 17th International Conference on Pattern Recognition (ICPR'04). Cambrige, UK, 2004, 3: 23-26.

[27] 高文，金辉. 面部表情图像的分析与识别. 计算机学报，1997，20(9)：782-789.

[28] 尹星云，王洵. 用隐马尔可夫模型设计人脸表情识别系统. 电子科技大学学报，2003，32(6)：725-728.

[29] 左坤隆，刘文耀. 基于活动外观模型的人脸表情分析与识别. 光电子激光，2004，15(7): 53-857.

[30] T. Kanade, J. F. Cohn, Y. L. Tian. Comprehensive database for facial expression analysis, in: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00). Grenoble, France, March, 2000: 46-53.

[31] C. L. Huang, Y. M. Huang. Facial Expression Recognition using Model based Feature Extraction and Action Parameters Classification[J]. Visual Communication and Image Representation, 1997, 8(3): 278-290.

[32] W. J. Guo, T. T. Niu. A new face detection method based on shape information[J]. Pattern Recognition Letters, 2000, 21(6): 463-471.

[33] R. Hoogenboom, M. Lew. Face detection using local maxima[C]. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996: 334-339.

[34] R. L. Hsu, M. Abdel-Mottaleb, A. K. Jain. Face detection in color images[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(5): 696-706.

[35] W. Haiyuan, C. Qian, M. Yachida. Face Detection From Color Images Using a Fuzzy

Pattern Matching Method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1999, 21(6): 557-564.

[36] J. Steffens, E. Elagin, H. Neven.Person Spotter-fast and robust system for human detection, tracking and recognition[C]. Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998: 516-521.

[37] C. H. Lee, J. S. Kim, K. H. Park. Automatic Human Face Location In A Complex Background[J]. Pattern Recognition, 1996, 29(11): 1877-1889.

[38] T. F. Cootes, G. J. Edwards, C. J. Taylor. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 681-685.

[39] X. W. Hou, S. Z. Li, H. J. Zhang, Q. S. Cheng. Direct Appearance Models[C]. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2001.

[40] S. Kimura, M. Yachida .Facial expression recognition and its degree estimation[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997: 295-300.

[41] A. Pentland, B. Moghaddam, T. Starner. View-based And Modular Eigenspaces For Face Recognition[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994: 21-23.

[42] M. H. Yang, N. Abuja, D. Kriegman. Face detection using mixtures of linear subspaces[C]. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000: 70-76.

[43] Y. L. Cun, P. Haffner, L. Bottou, et al. Gradient-Based Learning for Object Detection, Segmentation and Recognition [T]. AT&T Shannon Lab, 1999: 1-26.

[44] P. Viola, M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[45] G. Donato, M. S. Barlett, J. C. Hager, P. Ekman, T. J. Sejnowski. Classifying Facial Actions. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 21, No.10, 1999: 974-989.

[46] I. A. Essa, A. P. Pentland. Coding, analysis, interpretation and recognition of facial expressions. In Proc Of IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 757-763.

[47] M. Bartlett. Face image analysis by unsupervised leanring and redundancy reduction. Ph. D.

Thesis, University of Caliofnria, San Diego, 1998.

[48] Y. Tian，T. Kanade，J. Cohn. Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach.Intell. 2001, 23(2): 97-115.

[49] C. Lisetti，D. Rumelhart. Facial expression recognition using a neural network. Proceedings of the 11th International Flairs Conefrence，AAAI Press, New York, 1998.

[50] Z. Zhang，M. Lyons，M. Schuste，S. Akamatsu. Comparison between geometry-based and Gabo-wavelets-based facial expression recognition using multi-layer perceptron. IEEE Proceedings of the Second Intenrational Conefrence on Automatic Face and Gesutre Recognition(FG'98)，Nara，JaPan，1998: 454-459.

[51] Maja Pantic，Leon J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images, IEEE Trans. On Systems,Man,and Cybernetics-Part B: Cybenretics, 2004, 34(3): 1449-1461.

[52] S. Ioannou, A. Raouzaiou, K. Kaprouzis, et al. Adaptive Rule-Based Facial Expression Recognition. G. Vouros, T. Panayiotopoulos(Eds.), Lecture Notes in Artificial Intelligence, Vol. 3025, Springer-Verlag, 2004: 466-475.

[53] P. Michel，R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. Proceedings of the 5th international conference on Multimodal interfaces, Vancouver, British Columbia, Canada, ACM Press New York, NY, USA, 2003: 258-264.

[54] I. Buciu, C. Kotropoulos, I. Pitas. ICA and Gabor representation for facial expression recognition. IEEE Intenrational Conefrence on Image Processing, Vol. 2, Barcelona, Spain, 2003: 855-858.

[55] T. M. Cover, P. E. Hart. Nearest neighbor pattern classification. IEEE Transactions on information theory, 1967, 13(1): 21-27.

[56] B. Moghaddam, T. Jebara, A. Pentland. Bayesian face recognition Pattern recognition, 2000, 33: 1171-1782.

[57] A. J. Calder, A. M. Burton, P. Miler, et al. A principal component analysis of facial expression. Journal of Vision Research, 2001, 41(9): 1179-1208.

[58] X. W. Chen, T. S. Huang. Facial expression recognition:a clustering-based approach. Pattern Recognition Letter, 2003, 24(9-10): 1295-1302.

[59] T. M. Mitchell. Machine Learning. McGRAW-HILL, 1997.

[60] B. E. Boser, I. M. Guyon, V. Vapnik. A Training Algorithm for Optimal Margin Classifiers.

In D. Haussler, editor, Proceedings of the Annual Conference on Computational Learning Theory, Pittsburgh, PA, ACM Press, July 1992: 144-152,.

[61] C. Cortes and V. Vapnik. Support Vector Networks. Machine Learning, 1995, 20:273-297.

[62] V. Vapnik. (张学工译). 统计学习理论本质[M]. 北京: 清华大学出版社, 2000.

[63] 袁亚湘, 孙文瑜. 最优化理论与方法. 科学出版社, 北京, 2001.

[64] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK, 2000.

[65] CHRISTOPHER, J. C. BURGES. A Tutorial on Support Vector Machines for Pattern Rescognition[J]. Data Mining and Knowledge Discovery, 1998, 2: 121-167.

[66] K. Crammer, Y. Singer. On the Algorithmic Implementation of Multi-class SVMs, JMLR, 2001.

[67] Krebel. Pairwise classification and support vector machines. In: Advances in Kernel Methods: Support Vector Machine.MIT Press, Cambrige, IMA, 1998: 255-268.

[68] J. Friedman. Another Approach to Polychotomous Classification. Dept. Statist., Stanford Univ., Stanford, CA, 1996.

[69] J. Platt, N. Crisitianini and J. Shawe-Taylor. Large Margin DAGs for multiclass classification[A]. Advances in Neural Information Processing Systems 12[C]. The MIT Press: S. A. Solla, T. K. Leen and K. R. Mulle, Cambridge, IMA, 2000: 547-553.

[70] T. G. Dietterich, G. Bakiri. Solving Multiclass Leaming Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Researeh, 1995, (2):263-286.

[71] I. Santhosh kumar, G. Manimaran, C. Siva ram Murthy. A pre-run-time scheduling algorithm for object-based distributed real-time systems[J]. Journal of Systems Architecture, 1999, 45(14): 1169-1188.

[72] R. Bastide. Approaches in unifying Petri nets and the object-oriented approach[C]. In: the lst workshop on object-oriented programming and models of Concurrency, Torino, Italy, 1995.

[73] Yeong-Jia, Daniel Mosses, Shi-Kuo Chang. An object-based model for dependable real-time distributed systems[C]. In: The 2nd IEEE work-shop on Object-Oriented Real-Time Dependable Systems, Laguna Beach, California, 1996.

[74] J. Weston and C. Watkins, Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, 1998.

[75] Vladimir N. Vapnik著，许建华，张学工译. 统计学习理论. 北京：电子工业出版社，2004.

[76] L. Kuncheva. Clustering-and-selection model for classifier combination [C]. Proceedings of the 4th International Conference on Knowledge-based Intelligent Engineering Systems (KES'2000), Brighton, UK, 2000, 3: 1275-1280.

[77] S. F. Dimitrios, S. Andreas. A multi-SVM classification system [C]. MCS，LNCS 2096, 2001: 198-207.

[78] 马笑潇，黄席樾，柴毅. 基于SVM的二叉树多类分类算法及其在故障诊断中的应用. 控制与决策，2003，18(3): 272-276.

[79] 宋辛科. 基于支持矢量机和决策树的多值分类器. 计算机工程，2005，31(14): 174-175.

[80] D. Landgrebe. Signal Theory Methods in Multispectral Remote Sensing. New Jersey: John Wiley and Sons Inc. 2003.

[81] Fumitake Takahashi, Shigeo Abe. Decision-tree-based multiclass support vector machine[A]. Proceeding of ICONIP'02[C], 2002.

[82] 杨淑莹. 图像模式识别：VC++技术实现. 北京：清华大学出版社；北京交通大学出版社，2005.

[83] 史朝辉， 王晓丹. 一种改进的DDAGSVM决策算法. 2005年中国模糊逻辑与计算智能联合学术会议，NLCFCI，2005：382-387.

[84] L. S. Franck, C. Miaskowski. Measurement of neonatal responses to painful stimuli: a research review. Journal of Pain and Symptom Management, 1997, 14(6): 343-378.

[85] S. Merkel, T. Voepel-Lewis, S. Malviya. Pain assessment in infants and young children: the FLACC scale. The American Journal of Nursing, 2002, 102(10): 55-58.

[86] 卢官明，邹婵洁，李晓南，郭旻. 新生儿疼痛面部表情的特征提取. 南京邮电大学学报，2008，第 5 期发表.

攻读硕士学位期间发表的论文

1．卢官明，郭旻，李晓南，邹婵洁. 基于SVM的新生儿疼痛表情识别. 南京邮电大学学报，2008，第6期发表。