代　号　10701　　学　号　0222421153

分类号　TP183；Q812　　密　级　公开

西安電子科技大學

硕士学位论文

题（中、英文）目　基于 Boosting 的人工神经网络集成及其模式分类

Neural Network Ensemble Based on Boosting for Classification

作 者 姓 名　林存炜　指导教师姓名、职务　张军英 教授

学 科 门 类　工学　学科、专业　计算机应用技术

提交论文日期　二〇〇五年五月

摘 要

神经网络集成通过训练多个神经网络并将其结论进行合成，可以显著地提高学习系统的推广能力。它不仅有助于科学家对机器学习和神经计算的深入研究，还有助于普通工程技术人员利用神经网络技术来解决真实世界中的问题。因此，它被视为一种有广阔应用前景的工程化神经计算技术，已经成为机器学习和神经计算领域的研究热点。Boosting 算法是用来提高学习算法准确度的方法，它通过构造一个预测函数系列，然后以一定的方式将它们组合成一个预测函数。本文在对原AdaBoost算法的研究基础上，提出了优化训练集和用最小二乘优化弱分类器权系数的两种改进算法，理论和实验表明改进的算法具有训练误差更小，推广能力更好的优点。本文中，利用AdaBoost算法来生成神经网络集成中的个体网络，并用来做基因数据的模式分类。实验证明，该方法有效地解决极少样本、超高维的基因数据模式分类。

关键字：神经网络　神经网络集成　Boosting　AdaBoost　模式分类

Abstract

Neural network ensemble can significantly improve the generalization ability of learning systems through training a finite number of neural networks and then combining their results. It is not only helpful for scientists to investigate machine learning and neural computing but also helpful for common engineers to solve real-world problems using neural network techniques. Therefore neural network ensemble has been regarded as an engineering neural computing technology that has great application prospect. Also it has become a hot topic in both machine learning and neural computing communities. Boosting is a method for improving the accuracy of any given learning algorithm, which generate multiple versions of a hypothesis and combine them to create an aggregate hypothesis. Based on the detailed analysis of the adaboost algorithm, this paper presents two new adaboost algorithms which are the method based on optimizing training dataset and the method based on optimizing the weights of weak classifiers using LMS. The theories and experimental results prove that the two new algorithms have less training error and better generalization ability. In this paper, we use adaboost algorithm to generate single neural network for neural network ensemble, then do gene pattern recognition. The experimental results have verified the feasibility and validity of neural network ensemble based on Boosting for gene classification which often has few samples and high dimension.

Key words: neural network, neural network ensemble, boosting, adaboost, pattern recognition,

创新性声明

本人声明所呈交的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中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技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做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：林育娟　　日期 2006.2.24

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本人签名：林育娟　　日期 2006.2.24

导师签名：　　日期 2006.2.24

第一章 绪论

1.1 引言

我们先接触一个现实中有趣的问题。一个买足彩的彩民希望能够最大希望赢得赌注，于是他决定借助计算机程序通过有用的信息（各支球队的实力、状态、及最近的表现和成绩等等）来正确地判断各场比赛的结果。为了编写这样的一个程序，他请教了有名的博采专家下赌策略。不奇怪，专家也不能很清楚地对猜测某一场比赛输赢定出许多规则。但是，当提供一些跟某场比赛有关的数据时，博采专家就不难针对这场比赛凭以前的经验提出一个粗略的预测方法。比如，球队那些队员将上场，主教练会采取什么样的战术来应付对手，以及球队最近的状态怎样，等等，根据这些，博采专家就能粗略估计球队这场比赛将会取得一个怎样的结果。很明显，这样一个凭经验的方法 是粗糙的和不太准确的。但是，彩民还是要请教博采专家，是因为它应该至少能够提供一个比随机猜测略好的预测比赛方法 。进一步，彩民通过对不同比赛多次询问博采专家的观点，他就能够积累许多有经验的预测方法，这样，他就能够最大希望地赢得赌注。但是，有个问题是他该怎样合理地利用这些凭经验的预测方法。

为了利用这些凭经验的方法，彩民必须面对两个问题，第一，他该选那些对比赛最有用的数据给博采专家，以得到最有用的经验。第二，收集了许多凭经验的方法之后，他该怎样把它们组成一个精确的预测方法。

解决上述有趣的彩民买足彩问题，我们把它引入到机器学习领域。我们可以把这个问题看作是一个模式识别的决策问题。在现实中有许多类似问题，都需要凭一些以前的经验来做出决策判断。比如，天气预报，要预测是下雨还是晴天，就要根据以前的经验来判断，可是以前的经验判断又是千变万化的，所以我们就必须结合以前多次的经验来做判断。本文引出的 Boosting 通过结合如上所述粗糙的、不太正确的、单凭经验的预测方法，能构造出一个正确的预测方法，有效解决一些诸如此类原本没有一个准确预测方法的问题。当然，本文引用的 Boosting 及其改进加强的算法最终是用在人工神经网络集成上面的，并应用到基因数据的模式分类上的。因为本文工作是在基因数据上展开的，接下来就对基因数据做一下阐述。

1.2 基因数据综述

基因研究[1]是最近几十年发展起来的对人类生命科学研究最尖端、最前沿的科学研究，目前已成为是生物医学、生物化学、电子学、计算机科学、信息科学等学科的一个研究热点。当前国内外许多专家学者对此问题进行了深入研究，并取得相当的进展，例如DNA测序工程已经完成、后续的工作已经开展、而且有了相

当的基因级别的医学医疗应用，并取得较大的成功。关于基因的信息科学方面研究主要集中在基因DNA芯片技术的研究上，主要包括如何高精度的制造DNA芯片以及从DNA芯片提取出来的微阵列数据的后续处理、应用研究。本节对基因微阵列数据相关的技术进行一个概要而简单的介绍。

1、基因芯片技术简介

基因芯片(Gene Chip)又称DNA芯片（DNA Chip)、DNA阵列（DNA Arrays)、寡核苷酸微芯片（Oligonucleotide Micro-chip)，如图1.1所示，是指将许多特定的寡核苷酸片段或基因片段作为探针，有规律地排列固定于支持物上，然后与待测的标记样品的基因按碱基配对原理进行杂交，再通过激光共聚焦荧光检测系统等对芯片进行扫描，并配以计算机系统对每一探针上的荧光信号作出比较和检测，从而迅速得出所要的信息，其过程可用图1.2示出。上世纪90年代初，由美国Affymetrix公司的Fodor博士提出并开始基因芯片技术的研究，至今，基因芯片技术在医学各个领域中的应用均已取得巨大突破。1998年底，美国科学促进会将基因芯片列为1998年度自然科学领域十大进展之一。

图 1.1 DNA Chip

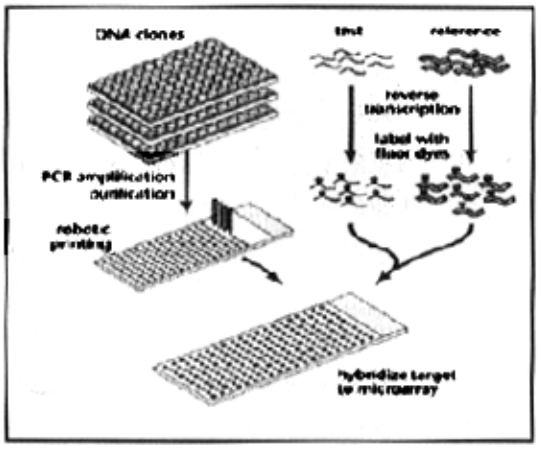

图1.2 DNA微阵列的提取过程

2、DNA微阵列提取

在整个基因芯片研究应用中，DNA微阵列的提取是最基本的也是最关键的过程，其基本过程大致分以下几步：

（1）基因芯片的制备

基因芯片的实质是高度集成的寡核苷酸阵列，制造基因芯片首先要解决的技术是如何在芯片片基上定位合成高密度的核酸探针。目前基因芯片的制备[2]主要采用三种方法：即光蚀刻合成法、压电印刷法、点样法。

（2）样品的制备

生物样品成分往往比较复杂，所以在与芯片接触前，必须对样品先进行处理。

为了提高结果的准确性，来自血液或组织中的DNA/mRNA样本须先行扩增，然后再被荧光素或同位素标记成为探针。

（3）杂交

影响杂交的因素很多，但主要是时间，温度及缓冲液的盐浓度。如果是表达检测，需要长历时、低温和高盐条件的较严谨性杂交。而如果是突变检测，需要短历时、高温和低盐条件高严谨性杂交。总之，杂交条件的选择要根据芯片上核酸片段的长短及其本身的用途来定。

（4）芯片的检测分析

对于荧光标记芯片应用荧光显微镜或激光共聚焦扫描仪等采集各杂交点荧光信号如位置和强度；同位素标记多采用放射自显影检测杂交信号；最近发展的纳米金标记，通过银放大后可直接在肉眼或普通光学显微镜下观察。最终再用相关软件进行信号的分析处理，得出待测样品的核酸信息。

3、基因芯片技术的应用

医学诊断。基因芯片技术[3]为临床疾病的诊断提供了一种全新的概念，它不仅使实验室检测的高通量、高自动化、微量化得以实现，而且使临床上对一些疑难疾病的准确诊断成为可能，从而给临床诊断工作带来革命性的进展。人类的许多疾病，如癌症、遗传性疾病等都与基因有关，因此利用基因芯片技术可寻找与该疾病有关的基因，实现对疾病的快速、简便、高效的诊断。Affmetix 公司把P53的全长序列和已知突变的探针集成在芯片上，制成基因芯片用于癌症的早期诊断。细胞色素P450芯片用于诊断有无药物代谢缺陷。华盛顿大学分子生物学系与病理系将5766个基因探针固定于芯片上，其中5376个分别选自卵巢癌、卵巢表面上皮细胞、正常卵巢的cDNA 文库，另外还有342个来自EST(表达序列标记) 克隆，包括一些已知确定的看家基因、细胞因子和因子受体基因、生长因子和受体基因、与细胞分裂相关的基因以及新近确定的肿瘤相关基因，用于卵巢癌中基因表达变化的监测。Lopez2Crapez 等应用基因芯片技术对75 例色素瘤患者的DNA 进行检测,并与直接DNA 测序法相比较，结果应用芯片法可以精确地确定所有的基因型。Drobyshev 等开发的β2寡核苷酸微集芯片用于地中海贫血患者红细胞中β2珠蛋白基因中三个突变位点的检测。Heller 等构建的96 个基因的cDNA 微集芯片用于检测风湿性关节炎(RA) 的相关基因。

寻找新基因。定量检测大量基因表达水平在阐述基因功能、探索疾病原因及机理、发现可能的诊断及治疗等方面是很有价值的。基因芯片技术在发现新基因及分析各个基因在不同时空表达方面是一项十分有用的技术，它具有样品用量极少，自动化程度高等优点，便于大量筛选新基因。目前，大量人类ESTs给cDNA微阵列提供了丰富的资源，数据库中400000个ESTs代表了所有人类基因，成千上万的ESTs微阵列将为人类基因表达研究提供强有力的分析工具。这将大大地加速人

类基因组的功能分析。

后基因组研究。基因组测序完成后，未知基因的功能研究是一个十分诱人的后基因组研究课题。斯坦福大学的Davis研究小组的研究提示DNA芯片技术将来可能应用于人类基因组测序完成后阐明开放读码框架ORF生物学功能的研究，可能会对深刻认识生命现象及药物设计带来重大影响。

此外，基因芯片还应用于药物研究，通过基因芯片技术可以将药物的生物效应和基因变化密切相联系，从而为药物的研究和开发注入了新的生机和活力；环境监测，基因芯片可以快速大规模的检测污染源，同时帮助寻找保护基因及能够治理污染源的基因产品；法律、军事医学及反恐，基因芯片可用于开发生物战病原体检测系统，研制生物战保护剂，进行血型、亲子鉴定及DNA 指纹图谱分析等，从而广泛应用于法律、军事途径，特别是在对反恐及防生物战的研究中。可以这样说基因芯片技术的出现，大大造福于人类社会。

4、基因表达分析和处理

生物机体组织在不同的时间条件下有不同的功能基因表达[4]，通过从该组织样品中提取出的mRNA与含有代表该组织的功能基因的芯片进行杂交，就可获得该样品的功能基因表达情况。例如，人类基因组中大约编码100000个不同的基因，通过DNA芯片技术进行检测，只需一次就几乎可以获得全部的表达情况。Lockhart等首先采用了光刻合成的20-mer观核苷酸阵列对鼠B细胞和T细胞在接受药物刺激后不同时段的IL-2，IL-3，IL-4，IL-10，GM-CSF及TNF-α等一些细胞因子的表达情况进行了检测。Chambers等人也首次利用DNA芯片技术对HCMV基因表达情况进行了研究。

在生物信息论中，机体组织的信息可逐级分为：DNA、mRNA、蛋白质、细胞组织。其中DNA提供的信息最多，从更微观地角度去分析病类间的异同，相对来说，细胞组织提供的信息最少。随着人类基因组计划的顺利进行，基因组研究的重心转到了功能基因组学以及后续的基因数据的处理研究上。表达谱基因表达与传统的NorthernBlot相比有许多优点：系统微型化，样品需量极小；同时研究上万个基因的表达，研究效率明显提高；能更多地揭示基因之间表达变化的相互关系，从而研究基因与基因之间内在的作用关系；检测基因表达变化的灵敏度高，可检测到相差几个数量级的表达情况；节约费用和时间。当然基因数据的分析和处理也有很多困难：由于基因芯片成本高，因此目前所获得的基因微阵列数据是极少样本超高维的数据；盲的、非正则的数据。基于上述特点，目前关于基因微阵列数据的处理有以下几个过程：正则化；基因部分体积修正；基因选择；基因病类的发现和聚类；基因模式识别；基因预测以及基因调控等等。

DNA芯片技术成功地实现了人们对DNA微阵列数据的自动获取，这对了解疾病在基因级别产生的原因，基因级别的药物研制以及疾病的基因诊断和基因治疗具

有重要的理论和实际意义，但同时基于微阵列数据的基因研究工作才刚刚开始，要进行更深入的研究必须先进行数据信号的处理。经过基因选择、基因模式分类等几个关键性环节，才能进行基因诊断、药物实验等应用。

1.3 本文工作与结构

本论文主要的工作是对基因微阵列数据的分析和处理。文章引入并详细分析了统计机器学习方法Boosting算法及其加强改进AdaBoost算法。然后论述了人工神经网络集成，并阐述了基于Boosting的神经网络集成，特别做了减少神经网络集成推广误差的研究，使得基于Boosting的神经网络集成有更广泛的用途。最后，我们把训练集成的神经网络用于基因微阵列数据的分类上，试验证明了理论的有效性和可行性，同时也解决了实际的基因模式识别问题。

具体章节的内容安排如下：

第一章介绍了本文所要应用的基因芯片技术及基因微阵列数据；

第二章介绍了人工神经网络及人工神经网络集成，并从实现方法、理论分析和应用成果等三个方面论述神经网络集成；

第三章介绍了 Boosting 算法及其性能分析，首先介绍机器学习和 Boosting 的起源与发展，然后分析 Boosting 算法的训练误差与推广性误差；

第四章研究了怎样减少集成的推广性误差，并在前人基础上提出了两种改进的 AdaBoost 算法，从理论和实验证明方法的有效性；

第五章把理论研究应用到基因数据的分类上，并对试验结果加以分析，以验证理论的可行性和有效性；

第六章总结本论文的主要研究成果及意义，同时也指出了研究工作中存在的不足和进一步的解决思路，并展望下一步的工作。

本论文的研究工作在国家自然科学基金资助项目(No. 60371044)、国家留学回国人员科研基金项目和十五国防预研项目(No. 413070501)资助。

第二章 人工神经网络集成

2.1 引言

本章首先是人工神经网络简介，分神经网络的基本原理和神经网络的分类两方面进行介绍。其中神经网络的分类里面介绍了几种常用的人工神经网络如感知器和多层感知器（MLP），论文后面主要引用这两种神经网络。然后从单个的神经网络存在的不足引出人工神经网络集成，分实现方法、理论分析、应用成果三方面论述，主要作了集成的推广性能优于单个神经网络的分析，这是本文将神经网络集成用于分类器所最关心的。

2.2 人工神经网络简介

神经网络（Neural Network）是近年来再度兴起的一个高科技研究领域，是信息科学、脑科学、神经心理学等多种学科近年来研究的一个热点。科学家预言 21 世纪生物学的研究和系统理论将进入空前繁荣的时代。神经网络系统理论受其影响将得到飞速发展。经过各国科学家多年潜心研究，在神经网络的基础理论、方法、系统的综合与应用等方面取得了重大的成果，尤其是现金大规模集成技术和电子计算机软、硬件技术的发展，为神经网络的实现和应用提供了光明的前景。国外，神经网络的研究已成为自动控制领域中的一个热点，并已渗透到智能控制、模式识别、计算机视觉的自适应滤波、信号处理、非线形优化、自动目标识别、连续语言识别等各方面，并取得了令人鼓舞的进展，我国神经网络的研究正在加快步伐，接近国外先进水平。

2.2.1 神经网络的基本原理

国际著名的神经网络专家、第一个计算机公司的创始人和神经网络实现技术的研究领导人hecht-nielson给神经网络的定义[5]是："神经网络是一个以有向图为拓扑结构的动态系统，它通过对连续或断续式的输入作状态响应而进行信息处理"。

神经网络系统是由大量的、同时也是很简单的处理单元（或称神经元），通过广泛地互相连接而形成的复杂网络系统。虽然每个神经元的结构和功能十分简单，但由大量神经元构成的网络系统的行为确是丰富多彩和十分复杂的。神经网络系统是一个高度复杂的非线性动力学系统，不但具有一般非线性系统的共性，更主要的是它还具有自己的特点，比如高维性、神经元之间的广泛互连性以及自适应性或自组织性等。

1、生物神经元

典型的神经元由以下 4 个部分组成：①胞体；②树突；③轴突；④突触。神经元的基本工作机制是：一个神经元有两种状态——兴奋和抑制。平时处于抑制

状态的神经元，其树突和胞体接受其它神经元经由突触传来的兴奋电位，多个输入在神经元中以代数和的方式叠加；如果输入兴奋总量超过某个阈值，神经元就会被激发进入兴奋状态，发出输出脉冲，并由轴突的突触传递给其它神经元。神经元被触发之后有一个不应期，在此期间内不能被触发，然后阈值逐渐下降，恢复兴奋性。神经元是按照“全或无”的原则工作的，只有兴奋和抑制两种状态，但也不能认为神经元只能表达或传递二值逻辑信号。因为神经元兴奋时往往不是只发出一个脉冲，而是发出一串脉冲，如果把这一串脉冲看成是一个调频信号，脉冲的密度是可以表达连续量的。

2、人工神经元和神经网络

就目前的理论水平、制造水平和应用水平，人工神经元尚不可能是对人脑神经网络的全部的真实模拟，只能是对人脑神经网络有选择的、单一的、简化的构造和性能的模拟，从而形成了不同功能的，多种类型的，不同层次的神经网络模型。这里只介绍我们常用的最简单的模型：如图 2.1 所示。

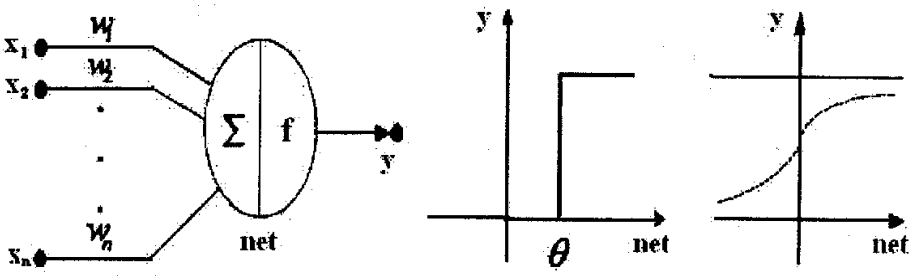

图 2.1 人工神经元模型与两种常见的输出函数

图中的 n 个输入 $x_i \in R$，相当于其它神经元的输出值，n 个权值 $w_i \in R$，相当于突触的连接强度，f 是一个非线性函数，θ 是阈值。神经元的动作如下：

$$net = \sum_{i=1}^{n} w_i x_i \tag{2-1}$$

$$y = f(net) \tag{2-2}$$

当 f 为阈值函数时，其输出为：

$$y = \mathrm{sgn}\left(\sum_{i=1}^{n} w_i x_i - \theta\right) \tag{2-3}$$

为使式子更为简约，可设阈值为：

$$\theta = -w_0 \tag{2-4}$$

$$W = (w_0, w_1, w_2, \cdots, w_n)^T \tag{2-5}$$

$$X = (1, x_1, x_2, \cdots x_n)^T \tag{2-6}$$

则

$$y = \mathrm{sgn}(W^T X) \tag{2-7}$$

或

$$y = f(W^T X) \tag{2-8}$$

这样的表达式可以将阈值合并到权值向量中处理。

其中，当f取各种不同形式的函数，如阶跃函数、单调上升有界函数、双曲正切函数等等，就产生不同的神经元模型。例如取f为阶跃函数时：

$$f(x)=\begin{cases}1 & (x>0)\\ 0 & (x\le 0)\end{cases} \tag{2-9}$$

如图 2.1（b），则得到 MP 模型。

某些重要的学习算法要求输出函数f可微，此时通常选用 Sigmoid 函数：

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-10}$$

见图 2.1(c)，选择 Sigmoid 函数作为输出函数是由于它具有以下有益的特性：①非线性，单调性。②无限次可微。③当权值很大时可近似阈值函数。④当权值很小时可近似线性函数。

神经网络的工作方式，由两个阶段组成：

1、学习期：神经元之间的连接权值可由学习规则进行修改，以使目标函数达到最小。

2、工作期：连接权值不变，由网络的输入得到相应的输出。

神经网络的性质主要决定于两个因素：一个是网络的拓扑结构；另一个是网络的学习、工作规则。二者结合起来就可以构成一个网络的主要特征。

神经网络按学习方式分为：有监督学习、无监督学习和再励学习三种：

（1）有监督学习（SL. Supervised Learning）即在学习过程中，网络根据实际输出与期望输出的比较，进行连接权系的调整，将期望输出称为教师信号，它是评价学习的标准。

（2）无监督学习（NSL. Nonsupervised Learning）即无导师信号提供给网络，网络能根据其特有的结构和学习规则，进行连接权系的调整，此时，网络的学习评价标准隐含于其内部。

（3）再励学习（RL. Reinforcement Learning ）把学习看作为试探评价（奖或惩）过程，学习机选择一个动作（输出）作用于环境之后，使环境的状态改变，并产生一个再励信号（奖或惩）反馈给学习机，学习机依据再励信号与环境当前的状态，再选择下一动作作用于环境，选择的原则，是使受到奖励的可能性增大。

随着网络结构和功能的不同，学习方法是多种多样的，但它们都遵循一些基本的、通用的学习规则，这些规则主要有：

（1） Hebb 学习规则

它是一类相关学习，其内容为：如果两个神经元同时兴奋，则他们之间的突触连接加强。用o_i表示神经元i的激活值，o_j表示神经元j的激活值，w_{ij}表示两个神经元之间的连接权，则 Hebb 学习规则可以表示为：

$$\Delta w_{ij} = \alpha \cdot o_i o_j \tag{2-11}$$

α 表示学习速率。

（2）δ 学习规则

它是用已知样本作为教师信号对网络进行学习。将式（2-11）中的o_i用网络期望输出d_i，与实际输出o_i之差的函数来代替，则权值的调整量为：

$$\Delta w_{ij} = \alpha \delta_i o_j \tag{2-12}$$

$$\delta_i = F(d_i - o_i) \tag{2-13}$$

（3）相近学习规则

设w_{ij}为神经元i到神经元j的连接权，o_i为神经元i的输出，则连接权的调整为：

$$\Delta w_{ij} = \alpha(o_i - w_{ij}) \tag{2-14}$$

在这种学习中，是使w_{ij}趋近于o_i的值。在 ART 等自组织竞争性网络中就采用了这种学习规则。

2.2.2 神经网路的分类

（1）单层感知器神经网络

感知器(Perceptron)是由美国学者F.Rosenblatt于1957年在M-P模型和Hebb学习律的基础上提出来的，它是一个具有单层计算神经元的神经网络，网络的传递函数是线性阈值单元。原始的感知器神经网络只具有一个神经元，主要用来模拟人脑的感知特征，由于采用阈值单元作为传递函数，所以感知器神经元只能输出两个值，即只有两个状态。感知器特别适用于简单的模式分类问题。当它用于两类模式的分类时，相当于在高维样本空间，用一个超平面将它们分开。F.Rosenblatt已经证明，如果两类模式是线性可分的（指存在一个超平面将它们分开），则算法一定是收敛的。但是，单层感知器网络只能用来解决线性可分问题，而对于非线性或线性不可分问题则无能为力。

（2）多层感知器（MLP）神经网络

多层感知器（Multilayer Perceptrons，MLPs）典型地由三部分组成：一组感知单元（源节点）组成输入层，一层或多层计算节点的隐藏层，还有一层计算节点的输出层，输入信号在层层递进基础上前向传播通过网络。如果只有一层隐藏层，我们称之为单隐层的MLP神经网络。通常，在监督学习的方式下使用误差反向传播算法训练多层感知器。将在下面的BP神经网络里介绍误差反向传播算法。

（3）BP神经网络

BP(Back Propagation)神经网络是指采用误差反向传播算法的多层前向神经网络，它是D.E.Rumelhatt和J.L.McCelland及其研究小组在1986年研究并设计出来的。BP算法是最小均方差算法的一般化，用梯度搜索技术，使均方差的代价函数最小。

算法的学习过程，由正向传播和反向传播组成。在BP正向传播过程中，输入信息从输入层经过隐含层，再传向输出层，每一层的神经元的状态值只影响下一层的神经元的状态值；如果在输出层不能得到期望的输出值，则转入反向传播，将误

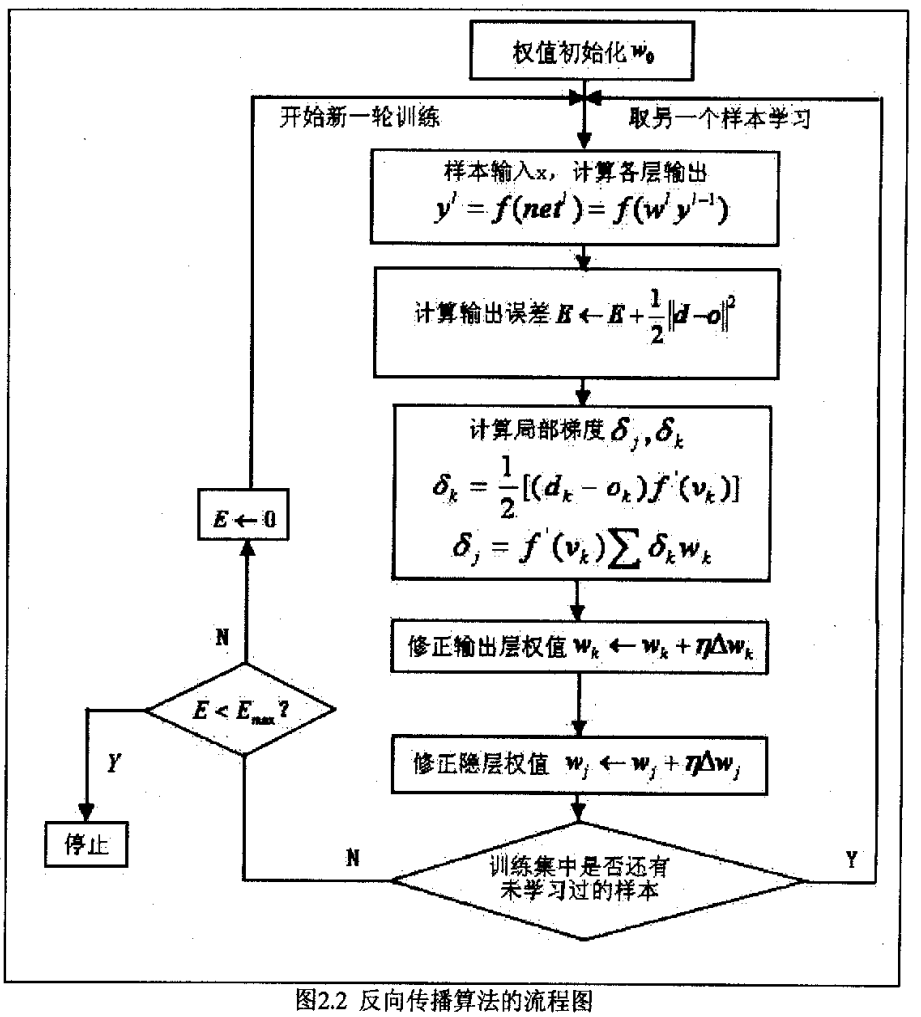

图2.2 反向传播算法的流程图

差信号沿逆向通路返回，通过修正各层神经元的权值，使得网络的总误差值收敛到极小。反向传播算法BP算法的整个学习过程的流程图如图2.2所示。(其中 E_{max} 为样本学习过程中所要求达到的误差给定值)。与感知器和线性神经网络不同的是，BP网络的神经元采用的传递函数通常是Sigmoid函数，所以可以实现输入和输出间的任意非线性映射，这使得它在诸如函数逼近、模式识别、数据压缩等领域有着更加广泛的应用。

（4）径向基函数网络

径向基函数（RBF）网络是以函数逼近理论为基础而构造的一类前向网络，这类网络的学习等价于在多维空间中寻找训练数据的最佳拟合平面。径向基函数网络的每个隐层神经元传递函数都构成了拟合平面的一个基函数，网络也由此而得

名。径向基函数网络是一种局部逼近网络，即对于输入空间的某一个局部区域只存在少数的神经元用于决定网络的输出。而 BP 网络则是典型的全局逼近网络，即对每一个输入/输出数据对，网络的所有参数均要调整。二者构造有本质的不同，径向基函数网络与 BP 网络相比规模通常较大，但学习速度较快，并且网络的函数逼近能力、模式识别与分类能力都优于后者。

另外，还有自组织网络，反馈网络（如 Elman 和 Hopfield 网络），限于本论文的应用就不一一作介绍。

2.3 人工神经网络集成

神经网络已经在很多领域得到了成功的应用，由于缺乏严密理论体系的指导，其应用效果完全取决于使用者的经验。虽然 Hornik 等人[6]证明，仅一个非线性隐层的前馈网络就能以任意精度逼近任意复杂度的函数，但一些研究者指出，对网的配置和训练是 NP 问题。在实际应用中，由于缺乏问题的先验知识，往往需要经过大量费力耗时的实验摸索才能确定合适的神经网络模型、算法以及数值设置，其应用效果完全取决于使用者的经验。即采用同样的方法解决同样的问题，由于操作者不同其结果也很可能大相径庭。在实际应用中，操作者往往是缺乏神经计算经验的普通工程技术人员，如果没有易于使用的工程化神经计算方法，神经网络技术的应用效果将很难得到保证。

1990 年，Hansen 和 Salamon[7]开创性地提出了网络集成(neural network ensemble)方法。它们证明，可以简单地通过训练多个神经网络并将其结论进行合成，可以显著地提高学习系统的推广能力。它不仅有助于科学家对机器学习和神经计算的深入研究，还有助于普通工程技术人员利用神经网络技术来解决真实世界中的问题。因此，它被视为一种有广阔应用前景的工程化神经计算技术，已经成为机器学习和神经计算领域的研究热点。1996 年，Sollich 和 Krogh[8]为神经网络集成下了一个定义，即“神经网络集成是用有限个神经网络对同一个问题进行学习，集成在某输入示例下的输出由构成集成的各神经网络在该示例下的输出共同决定”。目前这个定义已被广泛接受。下面从实现方法、理论分析和应用成果等三个方面论述神经网络集成。

2.3.1 神经网络集成研究

在神经网络集成的研究中，始终存在着两方面的内容。一方面，研究者们试图设计出更有效的神经网络集成实现方法，以直接用于解决问题。另一方面，研究者们试图对神经网络集成进行理论分析，以探明这种方法为何有效、在何种情况下有效，从而为实现方法的设计提供指导。

1、实现方法

对神经网络集成实现方法的研究主要集中在两个方面，即怎样将多个神经网络的输出结论进行结合以及如何生成集成中的个体网络。

（1）结论生成方法

当神经网络集成用于分类器时，集成的输出通常由个体网络的输出投票产生。通常采用绝对多数投票法(某分类成为最终结果当且仅当有超过半数的神经网络输出结果为该分类)或相对多数投票法(某分类成为最终结果当且仅当输出结果为该分类的神经网络的数目最多)。理论分析和大量试验表明，后者优于前者。因此，在对分类器进行集成时，目前大多采用相对多数投票法。

当神经网络集成用于回归估计时，集成的输出通常由各网络的输出通过简单平均或加权平均产生。Perrone 等人[9]认为，采用加权平均可以得到比简单平均更好的推广能力。但是，也有一些研究者认为，对权值进行优化将会导致过配(overfitting)，从而使得集成的推广能力降低，因此，他们建议使用简单平均。

此外还存在多种结合方式。例如，有的研究者利用神经网络这样的学习系统，通过学习来对多个预测进行结合；有的研究者通过对一组子网进行优化，使各子网都可以较好地处理一个输入子空间，从而一步步地进行结合；有的研究者不使用线性结合方法，而是使用一些随个体网络输出的确定程度而变化的动态权值来产生最终的分类；有的研究者以最小化分类误差为标准选择出相对于每一个输出分类的最佳网络，然后估计出最优线性权以将个体网络集成起来形成理想分类器。

2、个体生成方法

在生成集成中个体网络方面，最重要的技术是 Boosting[10] 和 Bagging(BootstrapAggregating)[11]，这在下一章节中将重点介绍。此外还存在多种个体生成方法。例如，有些研究者使用不同的目标函数、隐层神经元数、权空间初始点等来训练不同的网络，从而获得神经网络集成中的个体；有的研究者使用交叉验证技术来产生神经网络集成中的个体；有的研究者利用遗传算法进化出的神经网络种群作为集成中的个体；本论文作者使用 Boosting 算法对网络群体进行选择以获得集成中的个体。

3、理论分析

对神经网络集成的理论分析与对其实现方法的研究类似，也分为两个方面，即对结论生成方法的分析以及对网络个体生成方法的分析。

（1） 结论生成方法分析

1990 年，Hansen 和 Salamon[7]证明，对神经网络分类器来说，采用集成方法能有效提高系统的推广能力。假设集成由 N 个独立的神经网络分类器构成，采用绝对多数投票法，再假设每个网络以 $1-p$ 的概率给出正确的分类结果，并且网络之间错误不相关，则该神经网络集成发生错误的概率 p_{err} 为：

$$p_{err} = \sum_{k>N/2}^{N} \binom{N}{k} p^k (1-p)^{N-k} \qquad (2\text{-}15)$$

在 $p<1/2$ 时，p_{err} 随 N 的增大而单调递减。因此，如果每个神经网络的预测精度都高于 50%，并且各网络之间错误不相关，则神经网络集成中的网络数目越多，集成的精度就越高。当 N 趋向于无穷时，集成的错误率趋向于 0。采用相对多数投票法时，神经网络集成的错误率比式(1)复杂得多，但是 Hansen 和 Salamon[7]的分析表明，采用相对多数投票法在多数情况下能够得到比绝对多数投票法更好的结果。

在实际应用中，由于各个独立的神经网络并不能保证错误不相关，因此，神经网络集成的效果与理想值相比有一定的差距，但其提高推广能力的作用仍相当明显。1993 年，Perrone 和 Cooper[9]证明，在将神经网络集成用于回归估计时，如果采用简单平均，且各网络的误差是期望为 0 且互相独立的随机变量，则集成的推广性误差为各网络推广性误差平均值的 1/N，其中 N 为集成中网络的数目；如果采用加权平均，通过适当选取各网络的权值，能够得到比采用简单平均法更好的推广能力。

1996 年,Sollich 和 Krogh[8]指出，在神经网络集成的规模较大，即个体网络较多时，对结论的权进行优化没有好处，适于使用简单平均等结论合成方法；而在神经网络集成的规模较小，即个体网络较少，或者数据集中噪音较多时，对结论的权进行优化将提高学习系统的推广能力。

常用的一些神经网络模型在学习过程中容易陷入局部极小，这通常被认为是神经网络的主要缺点之一。然而，Perrone 和 Cooper[9]却认为，这一特性对神经网络集成推广能力的提高起到了重要作用。这是因为，如果各神经网络互不相关，则它们在学习中很可能会陷入不同的局部极小，这样神经网络集成的差异度(variance)就会很大，从而减小了推广性误差。换句话说，各局部极小的负作用相互抵消了。

1995 年，Krogh 和 Vedelsby[12]给出了神经网络集成推广误差计算公式。假设学习任务是利用 N 个神经网络组成的集成对 f: R^n? R 进行近似，集成采用加权平均，各网络分别被赋以权值 w_α，并满足式(2-16)：

$$\sum_{\alpha} w_\alpha = 1; \quad w_\alpha > 0 \qquad (2\text{-}16)$$

再假设训练集按分布 p(x)随机抽取，网络 α 对输入 X 的输出为 $V^\alpha(X)$，则神经网络集成的输出为：

$$\bar{V}(X) = \sum_{\alpha} w_\alpha V^\alpha(X) \qquad (2\text{-}17)$$

神经网络 α 的推广性误差 E^{α} 和神经网络集成的推广性误差 E 分别为：

$$E^{\alpha} = \int dxp(x)(f(x) - V^{\alpha}(x))^2 \tag{2-18}$$

$$E = \int dxp(x)(f(x) - \bar{V}(x))^2 \tag{2-19}$$

各网络推广性误差加权平均为：

$$\bar{E} = \sum_{\alpha} w_{\alpha} E^{\alpha} \tag{2-20}$$

神经网络 α 的差异度 A^{α} 和神经网络集成的差异度 $\bar{A}$ 分别为：

$$A^{\alpha} = \int dxp(x)(V(x) - \bar{V}(x))^2 \tag{2-21}$$

$$\bar{A} = \sum_{\alpha} w_{\alpha} A^{\alpha} \tag{2-22}$$

则神经网络集成的推广性误差为：

$$E = \bar{E} - \bar{A} \tag{2-23}$$

式(2-22)中的 $\bar{A}$ 度量了神经网络集成中各网络的相关程度。若集成是高度偏向(biased)的，即对于相同的输入，集成中所有网络都给出相同或相近的输出，此时集成的差异度接近于 0，其推广性误差接近于各网络推广性误差的加权平均。反之，若集成中各网络是相互独立的，则集成的差异度较大，其推广性误差将远小于各网络推广性误差的加权平均。因此，要增强神经网络集成的推广能力，就应该尽可能地使集成中各网络的误差互不相关。

（2）个体生成方法分析

1997 年，Freund 和 Schapire[13]以 AdaBoost 为代表，对 Boosting 类方法进行了分析。在下一章节中也将对此作重点分析。

2.3.2 神经网络集成的应用成果

由于神经网络集成方法操作简单且效果明显，因此，该技术已在很多领域中得到了成功的应用。Hansen 等人[14]利用由相对多数投票法结合的神经网络集成进行手写体数字识别，实验结果表明，集成的识别率比最好的单一神经网络识别率高出 20%—25%。此后，Schwenk 和 Bengio[15]将 AdaBoost 与神经网络结合进行手写体字符识别，系统对由 200 多个人的手写字符所组成的数据库能达到 1.4%的错误率，而对 UCI 字符数据集则能达到 1.5%的错误率。

Gutta 和 Wechsler[16]将神经网络集成和决策树相结合进行正面人脸识别，其集成由 RBF 网络采用相对多数投票法构成，实验结果表明，使用神经网络集成不仅增加了系统的健壮性，还提高了识别率。此后，Gutta 等人[17]用 RBF 网络的集成进行多性别、多人种和多姿态的正面人脸识别。他们使用了“分而治之

(divideandconquer)”的模块化方法，并利用决策树和支持向量机来实现挑选个体网络以确定输出分类的选通(gating)功能。CarnegieMellon 大学、微软中国研究院的合作者一起，将神经网络集成用于图像在深度方向上发生偏转的多姿态人脸识别，在省去了偏转角度估计预处理的情况下，系统的识别精度甚至高于多个单一神经网络在理想偏转角度估计预处理协助之下所能取得的最佳识别精度，除此之外，系统还能在进行识别的同时给出一定的角度估计信息。

Cherkauer[18]用简单平均法集成具有不同隐层神经元数的 BP 网络，并用它代替 NASA 的喷射推进实验室研制的 JARTOOL 中的 Gauss 分类器，对 Magellan 空间探测器收集到的关于金星的合成孔径雷达图像进行分析，在火山检测方面达到了行星地质专家的水平。Shimshoni 和 Intrator[19]利用神经网络集成进行地震波分类。他们采用了二级集成方式，地震波信号的三种不同表示分别被输入到采用不同网络结构的三个集成中，每个集成都被赋予一个可信度，第二级集成就以该可信度为权值，通过加权平均对第一级的三个集成进行结合。

此外，神经网络集成还在语音识别、文本过滤、遥感信息处理、疾病诊断等多个领域成功地得到了应用，限于篇幅，这里就不再一一详细介绍了。

2.3.3 神经网络集成的发展与探讨

目前，在神经网络集成的研究中仍然存在着很多有待解决的问题。

(1)关于神经网络集成的研究目前基本上是针对分类和回归估计这两种情况分别进行的，这就导致了多种理论分析以及随之而来的多种不同解释的产生。如果能为神经网络集成建立一个统一的理论框架，不仅可以为集成技术的理论研究提供方便，还有利于促进其应用层面的发展。

(2)神经网络集成中的个体网络差异较大时，集成的效果较好，但如何获得差异较大的个体网络以及如何评价多个网络之间的差异度，目前仍没有较好的方法。如果能找到这样的方法，将极大地促进神经网络集成技术在应用领域的发展。在这方面，西安交大傅向华[20]作了关于增量构造负相关异构神经网络集成的方法对此取得了一定的进展。

(3)在使用神经网络集成，尤其是 Boosting 类方法时，训练样本的有限性是一个很大的问题。如何尽可能地充分利用训练数据，也是一个很值得研究的重要课题。

(4)神经网络的一大缺陷是其“黑箱性”，即网络学到的知识难以被人理解，而神经网络集成则加深了这一缺陷。目前，从神经网络中抽取规则的研究已成为研究热点。从神经网络中抽取规则可以改善 Boosting 系统的可理解性，可以在一定程度上缓解集成的不可理解性。

2.4 本章小结与讨论

本章首先介绍了人工神经网络及其基本原理，并介绍了几种常用的神经网络，将在文章后面引用。然后从神经网络技术的应用效果出发，介绍了易于使用的工程化神经计算方法的人工神经网络集成，并从实现方法、理论分析和应用成果等三个方面论述神经网络集成，对神经网络集成用于分类器或回归分析时的重要性能指标推广性误差作了详细分析，最后对神经网络集成的发展做了进一步探讨。在本章中，引出了神经网络集成的个体生成方法：Boosting类方法。本论文后面章节，将对神经网络集成的Boosting类方法进行重点讨论。

第三章 Boosting 算法及其性能分析

3.1 引言

本章首先从模式识别的机器学习领域引出了该领域的一种有效的分类决策方法：Boosting，接着介绍了 Boosting 的起源与发展和 AdaBoost 算法与模型。Boosting 算法是用来提高学习算法准确度的方法；AdaBoost 算法是 Boosting 最初提出发展到能够实际应用的通用算法，它通过在训练集的不同子集上，多次调用简单学习算法，最终按加权投票方式融合多次简单学习算法的预测结果得到最终学习结果。接下来部分是 Boosting 算法分析与研究，先是 Boosting 算法的研究和应用，然后是 AdaBoost 算法的训练误差及其收敛性分析，最后，也是最重要的部分，作了 AdaBoost 算法的推广性误差的分析，这也是我们将其应用到提升弱分类器为强分类器所关心的一项性能指标。

3.2 Boosting 简介及其模型

3.2.1 模式识别的机器学习

机器学习研究是如何使机器从过去已有的现象中学习做出准确预测的自动技术。自从计算机问世以来，人们就想知道它们能不能自我学习。如果我们理解了它们学习的内在机制，即怎样使它们根据经验来自动提高，那么影响将是空前的。想象一下，在未来，计算机能从医疗记录中学习，获取治疗新疾病最有效的方法；住宅管理系统分析住户的用电模式，以降低能源消耗；个人软件助理跟踪用户的兴趣，并为其选择最感兴趣的在线早间新闻。对计算机学习的成功理解将开辟出许多全新的应用领域，并使其计算能力和可定制性上升到新的层次。同时，透彻理解机器学习的信息处理算法，也会有助于更好地理解人类的学习能力及缺陷。目前，我们还不知道怎样使计算机具备和人类一样强大的学习能力。然而，一些针对特定学习任务的算法已经产生。关于学习的理论认识已开始逐步形成。人们开发出很多实践性的计算机程序来实现不同类型的学习，一些商业化的应用也已经出现。例如，对于语音识别这样的课题，迄今为止，基于机器学习的算法明显胜过其他的方法。在数据挖掘领域，机器学习算法理所当然地被用来从包含设备维护记录、借贷申请、金融交易、医疗记录等此类信息的大型数据库中发现有价值的信息。随着对计算机认识的日益成熟，机器学习必将在计算机科学和技术中扮演越来越重要的角色。

通过一些专项成果可以看到机器学习[21]这门技术的现状：计算机已经能够成功地识别人类的讲话、预测肺炎患者的康复率、检测信用卡的欺诈、在高速公路

上自动驾驶汽车、以接近人类世界冠军的水平对弈西洋双陆棋这样的游戏等。已有很多理论成果能够对训练样例数量、假设空间大小和已知假设中的预期错误这三者间的基本关系进行刻画。人类正开始获取人类和动物学习的原始模型，用以理解它们和计算机的学习算法间的关系。在过去的十年中，无论是应用、算法、理论，还是生物系统的研究，都取得了令人瞩目的进步。

机器学习的目的[22]是根据给定的训练样本获得对某系统输入输出之间依赖关系的估计，使它能够对未知输出做出近可能准确的预测。一般地表示为：变量y与x存在一定的未知依赖关系，即遵循某一未知的联合概率$F(x,y)$，(x,y) 之间的确定性关系可以看作是其特例，机器学习问题就是根据n个独立同分布观测样本：

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) \tag{3-1}$$

一组函数$\{f(x,w)\}$中求一个最优的函数$f(x,w_0)$对依赖关系进行估计，使期望风险最小。

$$R(w)=\int L(y,f(x,w))\,df(x,y) \tag{3-2}$$

其中，$\{f(x,w)\}$称作预测函数集，w 为函数的广义参数，$\{f(x,w)\}$可以表示任何函数集；$L(y,f(x,w))$为由于用$f(x,w)$对y进行预测而造成的损失，不同类型的学习问题有不同形式的损失函数，预测函数也称作学习函数、学习模型或学习机器。

有三类基本的机器学习问题，即模式识别、函数逼近和概率密度估计。1）对模式识别问题，输出y是类别标号，两类情况下y= {0,1}或{-1,1}，预测函数称作指示函数，损失函数可以定义为：

$$L(y,f(x,w))=\begin{cases}0, & if \quad y=f(x,w)\\ 1, & if \quad y\neq f(x,w)\end{cases} \tag{3-3}$$

使风险最小就是Bayes 决策中使错误率最小。2）在函数逼近问题中，y是连续变量（这里假设为单值函数），损失函数可定义为：

$$L(y,f(x,w))=(y-f(x,w))^2 \tag{3-4}$$

即采用最小平方误差准则。3）对概率密度估计问题，学习的目的是根据训练样本确定 x 的概率密度，记估计的密度函数为 $p(x,w)$，则损失函数可以定义为：

$$L(p(x,w))=-\log p(x,w) \tag{3-5}$$

机器学习可以通过评估损失函数来评估预测函数。在机器学习模式识别中如何能够更好地分类，如本文引言中提到的分类决策问题，是一个值得考虑的问题，下面引入用来解决这一类问题有效的方法：Boosting算法。

3.2.2 Boosting 的起源与发展

Kearns和Valiant[24]指出，在PAC学习模型中，若存在一个多项式级学习算法来辨别一组概念，并且辨别正确率很高，那么这组概念是强可学习的；而如果学

习算法辨别一组概念的正确率仅比随机猜测略好，那么这组概念是弱可学习的。Kearns和Valiant[24]提出了弱学习算法与强学习算法的等价性问题，即是否可以将弱学习算法提升成强学习算法。如果两者等价，那么在学习概念时，只需找到一个比随机猜测略好的弱学习算法，就可以将其提升为强学习算法，而不必直接去找通常情况下很难获得的强学习算法。也即，能否把在PAC模型中表现稍好于随机猜想的弱学习机算法推进（Boost）成非常准确的强学习机算法的问题。

1989年，Schapire[10]通过一个构造方法提出了可证明的多项式时间的算法，其构造过程就是最初的Boosting算法，一年后，Freund提出了一个效率更高的Boosting算法。但这种算法尽管在某些情况下是最佳的，却存在一些实际的问题：它们都要求事先知道弱学习算法正确率下限，而这一点在现实应用问题中是很难满足。所以当时Boosting算法并未能应用于具体的模式识别问题。直到1995年， Freund和Schapire[13]合作提出了新的Adpative Boosting（AdaBoost）算法，该算法的效率与最初的Boosting算法效率几乎一样，却没有正确率下限的限制，因而可以很容易的应用到实际问题中。逐渐形成成熟的理论，并推广应用于具体的实际应用中。所谓“Ada”就是Adaptive，即可以根据上一轮训练错误率自适应于每种弱假设。本论文以后出现的Boosting如无特别说明，均指AdaBoost。

3.2.3 AdaBoost 算法及其模型

Freund和Schapire引入的AdaBoost的主要思想：给定一个弱的学习算法和训练集。在训练集的不同子集上，多次调用简单学习算法，最终按加权投票方式融合多次简单学习算法的预测结果得到最终学习结果。其伪代码如图3.1所示：

算法的输入：训练样本集$S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$，其中$x_i\in X$，X表示领域或实例空间。$y_i\in$某一标签集Y。学习器接受的例子$(x_i, y_i)$是从分布为P的$X\times Y$上随机的选择。假定是两类问题，$Y=\{-1,+1\}$，它是多类问题的基础。 AdaBoost反复地调用给定的弱或基学习算法，其主要思想之一是在训练集中维护一套权重分布。初始时，所有例子的权重都设为相等(即1/m)。但是每一回错分的实例其权重将增加，以使弱学习器被迫集中在训练集中的难点(实例)上。弱学习的任务就是根据分布D_t找到合适的弱假设h_t：$X\to Y$最简单情况下每个h_t的范围是二值的： $\{-1,+1\}$。于是该学习器的任务就是最小化错误$\varepsilon_t=\Pr_{i\sim D_t}[h_t(x_i)\neq y_i]$。$\varepsilon_t$表示第$t$轮训练时，所有错分样本的出现概率和。一旦得到$h_t$，AdaBboost算法选择一个参数$\alpha_t\in R$，该参数直观地测量$h_t$的重要程度。最终假设H是T次循环后，用加权多数投票把T个弱假设的输出联合起来得到的。对二值h_t，典型地，设：

$$\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t}) \qquad (3\text{-}6)$$

给定样本集 $(x_1, y_1), \cdots, (x_m, y_m)$ 其中 $x_i \in X, y_i \in \{-1, +1\}$

初始化，设数据分布为均匀分布：$D_1(i) = \frac{1}{m}$

For $t = 1, 2, \cdots T$：

根据分布 D_t 训练弱学习算法

得到该轮的预测结果 $h_t : X \to \{-1, +1\}$，并且有误差 $\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$

令 $\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$

更新 D_t：

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if\ h_t(x_i) = y_i \\ e^{\alpha_t} & if\ h_t(x_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i) \times \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

其中 Z_t 是为使 D_t 为概率分布的归一化因子

最终的预测输出为：$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

图 3.1 AdaBoost 算法

AdaBoost算法的演算我们用图3.2来直观说明，对两类数据集：波折号表示单个分类器产生的边界（到第五轮），实线表示合成分类器的边界。（数据来自[23]）

3.3 Boosting 算法性能分析

3.3.1 Boosting 算法的研究和应用

目前，研究者们对Boosting算法的研究主要集中在算法改进的两个方面：

1）如何更合理地选出多个弱学习的训练集。在Boosting算法中，下一轮训练的样本主要来自上一轮分错的样本中间，这样就会出现这样的问题：如何使下一轮参与训练的样本能更好地集中在容易出错的样本中间。例如，Breiman和Ratsch[25]针对以上情况提出了对 α_t 的修改：

$$\alpha_t = \arg\min_{\alpha \in R} \sum d_n^{(t)} \exp\{-e \cdot (\varphi - y_n h_t(x_n))\} \qquad (3\text{-}7)$$

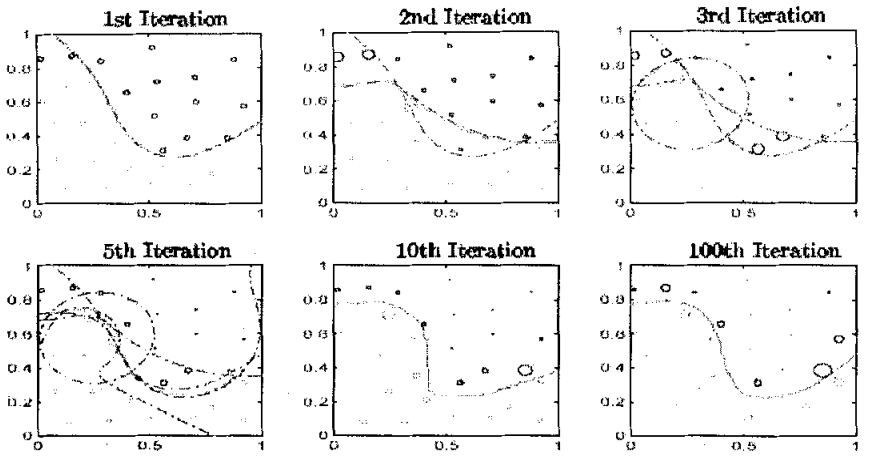

图 3.2 单个分类器与合成分类器的训练边界

其中，e为该轮训练所有错分样本的权值和。φ为一可调参数。还有，Schapire等利用Hamming Loss来代替e。总之关于这方面的研究已取得一些进展，详细内容可以参考文献[26]。

2）如何能高效、合理地设计融合规则。目前，较为常见的融合规则有乘积规则联合，求和规则联合、极小规则联合、极大规则联合、中值规则联合以及多数投票联合和加权求和等。本文介绍基于Boosting的人工神经网络集成就是在算法稳健性详细分析的基础上对多个人工神经网络融合规则的改进。

目前，基于弱学习算法的Boosting方法已成功应用于[27]很多方面：Fround和Schapire将Boosting应用于UCI benchmark数据集，他们还应用到文本识别、OCR等。还能被应用于文本过滤、语音识别、ranking problem等领域。

Boosting算法有许多优点：快速、简单、易于实现而且无需参数调节，也无须任何关于弱学习算法的先验知识，而且理论能灵活的应用于任何分类方法，这一点是其他任何机器学习不可能实现的。这样使我们设计分类器的精力不用过分集中于寻找精度高的算法，只需找出比随机猜测好的弱分类器即可。同时，Boosting也有其固有的缺点，对于具体问题，其过度依赖于取样后的数据集和弱学习算法，而且对噪声比较敏感。

总之，理论上Boosting算法是一种可以集成任何弱分类算法的算法框架，它有比较完整的数学理论基础，而且也有实验表明该算法对少样本、高维数据具有很好的适用性，和其他的分类算法相比，Boosting具有适应性强、精度高的优点。

3.3.2 AdaBoost 算法的训练误差及其收敛性分析

AdaBoost算法能有效的降低训练集的误差，这一点Schapire和Singe[13]已经证明，且最终预测的训练误差有上界：

$$\frac{1}{m}\left|\{i:H(x_i)\neq y_i\}\right|\leq\frac{1}{m}\sum_i\exp(-y_i f(x_i))=\prod_t z_t \qquad (3\text{-}8)$$

其中 $f(x)=\sum_t \alpha_t h_t(x)$。

应用优化理论，从式(3-8)中我们可以得出，通过选择 α，h_t可以使得 z_t 最小化：

$$z_t=\sum D_t(i)\exp(-\alpha_t y_i h_t(x_i)) \qquad (3\text{-}9)$$

也就是说使其训练错误方法降到最低。对于2 类问题， 跟据 $\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$，可以得到：

$$\prod_t z_t=\prod_t[2\sqrt{\varepsilon_t(1-\varepsilon_t)}]=\prod_t 1-4\gamma_t^2\leq\exp(-2\sum_t\gamma_t^2) \qquad (3\text{-}10)$$

式(3-10)中 ε_t 为h_t的训练误差。令 $\gamma_t=\frac{1}{2}-\varepsilon_t$。

从式(3-10)可以看出，当弱学习算法略好于随机猜测时(两类问题 $\varepsilon_t<0.5$)，训练误差按指数函数下降，并最终能下降到零。AdaBoost算法以前的Boosting算法也有相似的性质。然而，以前的算法在学习前需要得到已知的下界γ。实践中关于这样的边界的知识是很难得到的。而AdaBoost算法可以调整单个弱假设的错误率，所以说是自适应的。

一个模式识别训练算法是否可行，首先要从理论上考察其收敛性，否则，如果算法本身就不可行，更谈不上算法的有效性、精度。一般认为，训练算法的收敛性是指随着训练学习次数的增加，训练误差会逐渐下降，最终达到设计者要求的精度，这样的系统才能作为一个学习系统。试想一下，一个识别系统在训练过程中会出现振荡或过学习，即随着学习次数的增加训练误差反而上升或者一直处于上下波动状态，这时系统肯定不可行的。因此，如何能保证设计出一个可行的系统才是设计模式识别系统最基本的目标。从上面对AdaBoost的训练误差分析可以看出当弱学习算法略好于随机猜测时(两类问题 $\varepsilon_t<0.5$)，训练误差按指数函数下降，并最终能下降到零，这说明AdaBoost算法具有收敛性。

3.3.3 AdaBoost 的推广性误差的分析

上面谈论了AdaBoost的训练误差，在实际的机器学习应用系统中，推广性误差才是大家所真正关心的。理论上，为了产生最好的分类器，我们期望AdaBoost的训练误差降到零。Freund和Schapire[26]用基于样本个数、弱学习分类空间的VC维来衡量推广性误差。根据Baum和Hausslev的理论，Freund和Schapire推导出：AdaBoost的最大推广性误差为：

$$\hat{P}_r[H(x)\neq y]+\tilde{o}\sqrt{\frac{Td}{m}} \qquad (3\text{-}11)$$

式（3-11）中 $\hat{P}_r[\cdot]$表示在训练集上的经验概率，从式（3-11）中我们可以看出，

当T变大时，该式也增大，也就是当训练轮数过多时，容易产生过学习（Overfit）。但是在早期的大量实验中，没有发生过过学习现象，更有甚者，当训练误差降到零后，其推广性误差还在继续下降，这似乎有悖于式(3-11)。基于这个实验现象，Schapire等人在Bartlett[28]的基础上，用边界的概念来分析这种训练误差为零时推广性误差还在下降的现象：

$$I = m\arg in_f(x,y) = \frac{yf(x)}{\sum_t |\alpha_t|} = \frac{y\sum \alpha_t h_t(x)}{\sum_t |\alpha_t|} \qquad (3\text{-}12)$$

$I \in [-1,+1]$，当且仅当测试的最终预测正确时I 为+1。

在公式（3-12）中，Schapire等人[29]认为，较大的正边界表示可信度高的正确的预测，较大的负边界表示可信度较高的错误的预测，同理，较小的边界表示可信度低的预测。当训练误差降低为零后，在训练空间上的边界仍能提高，由以下式子表示：

$$\hat{P}_r[m\arg in_f(x,y) \le \theta] + \tilde{o}(\sqrt{\frac{Td}{m\theta^2}}) \qquad (3\text{-}13)$$

也可以用下面的模型（图3.3）来解释上述理论：

图3.3中，假设存在两个不同的分类，若以超平面A为边界，则最小边界为h_1；以超平面B为边界，则最小边界为h_2，$h_2 > h_1$。当训练误差降到零后，继续增大边界使A变成B，从而增大了最小边界，使可靠性增加，使推广性误差得以继续降低。

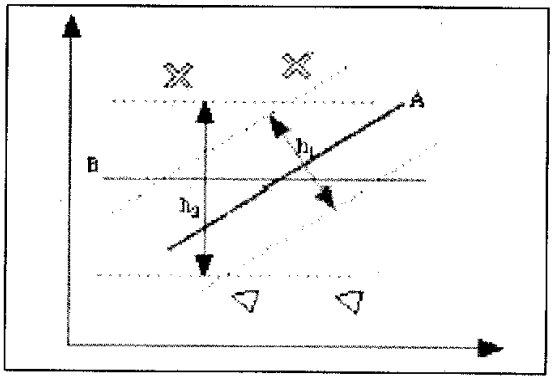

图 3.3 边界模型

3.4 本章小结与讨论

本章首先介绍了 Boosting 类方法的 AdaBoost 算法。然后分析了 AdaBoost 算法的训练误差，得出多轮弱学习算法的训练误差都有一个上界，并且当弱学习算法好于随机猜测时，最终假设的训练误差随着训练轮数呈指数趋势下降，并最终

能趋于零，证明了该算法具有收敛性。然后分析了分类器最关心的问题：推广性能力。介绍了基于样本个数、弱学习分类空间的 VC 维来衡量推广性误差，用边界的概念来分析当训练误差趋于零时，推广性误差还在下降的现象；另外还作了 Boosting 算法的研究，提出两个问题 1）如何更合理地选出多个弱学习的训练集；2）如何能高效、合理地设计融合规则。论文下一章节将就此展开讨论。

第四章 Boosting 算法的改进

4.1 引言

Boosting算法能够有效提升弱分类器为强分类器。但是人们总是追求一些更合理，更有效的算法来解决实际问题。前文提到，目前，研究者们对Boosting算法的研究主要集中在算法改进的两个方面：1）如何更合理地选出多个弱学习的训练集。2）如何能高效、合理地设计融合规则。本章在这两方面进行了初步研究，分别提出了两种改进的AdaBoost算法。其后，对两种改进的AdaBoost算法进行性能分析，理论证明其性能均优于原始AdaBoost算法。接着，我们进行实验证明，采用Iris数据，共三类样本，每类均为50个样本，每个样本是4维的向量（第一维是花萼的长度，第二维是花萼的宽度，第三维是花瓣的长度，第四维是花瓣的宽度），我们取其中两类，每类前30个样本为训练样本，后20个为测试样本，实验结果得出两种改该进的算法对数据分类有更小的训练误差和测试误差，也表明其性能比原始的AdaBoost算法更优。

4.2 Boosting 算法改进（一）

（1）基本思想

针对如何更合理地选出多个弱学习的训练集，在一般AdaBoost算法中，下一轮训练的样本主要来自上一轮分错的样本中间。这样就会出现这样的问题：如何使下一轮参与训练的样本能更好地集中在容易出错的样本中间。对此，我们对原始AdaBoost算法简单地提出了一个改进的设想，我们选择下一轮的训练样本不再来自上一轮分错的样本中间，而是来自前面T轮组合判决（$\sum \alpha_t h_t$）之后分错的样本中间。这样，前T轮组合判决（$\sum \alpha_t h_t$）的分类结果比第T轮单轮的预测结果h_t更准确，错分的样本相对也少，因此下一轮参与训练的样本就能更好地集中在那些难分的样本中间，从而得到最终分类更好的结果。

改进的算法与原始算法的不同之处在于更新D_t不同：原始算法根据第t轮弱分类器h_t的分类错误率ε_t更新D_t；改进算法根据前面T轮组合判决$\sum \alpha_t h_t$分类错误率$\varepsilon\varepsilon_t$更形D_t。

（2） 误差分析

根据公式（3-10），原先的AdaBoost算法ε_t为h_t的训练误差，改进的AdaBoost算法ε_t为前t轮弱分类器组合$\sum \alpha_t h_t$的训练误差，显然改进的AdaBoost训练误差ε_t更小，也就更容易收敛。同样，根据公式（3-11），训练集更集中在少数难分的样本中间，它们在训练集中占的比例较少，则改进的AdaBoost算法的$\hat{P}_t[\cdot]$在训练集

上的经验概率要小，其推广性误差也小。因此，从理论上说，此改进的AdaBoost算法是有效的，而且比原始AdaBoost算法性能优越。

经过修改的AdaBoost_revise1的算法如图4.1。

给定样本集$(x_1,y_1),\cdots,(x_m,y_m)$，其中样本集$x_i \in X$，$y_i=\{-1,+1\}$

初始化，设数据分布为均匀分布：$D_t(i)=\frac{1}{m}$

For $t=1,2,..T$:

根据分布D_t训练弱学习算法

得到该轮的预测结果h_t：$X \to \{-1,+1\}$，并且有误差$\varepsilon_t=\Pr_{i\sim D_t}[h_t(x_i)\neq y_i]$

令$\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$

组合$h=sign(\sum_{k=1}^{t}\alpha_k h_k)$，得到经前$t$轮组合判决后的预测结果$h$：$X \to \{-1,+1\}$

重新计算误差$\varepsilon\varepsilon_t=\Pr_{i\sim D_t}[h(x_i)\neq y_i]$，并令$\alpha\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon\varepsilon_t}{\varepsilon\varepsilon_t})$

根据$\alpha\alpha_t$更新D_t：

$$D_{t+1(i)}=\frac{D_t(i)}{Z_t}\times\begin{cases}e^{-\alpha\alpha_t} & if \quad h(x_i)=y_i\\ e^{\alpha\alpha_t} & if \quad h(x_i)=y_i\end{cases}$$

其中Z_t是使D_t为概率分布的归一化因子

最终的预测输出为：$H(x)=sign(\sum_{t=1}^{T}\alpha_t h_t(x))$

图4.1 AdaBoost_revise1算法

（3） 实验分析

为了验证改进AdaBoost_revise1算法的可行性和有效性，我们进行了分类性能分析实验：数据来源于美国加利福尼亚大学的机器学习数据集[4]（UCI Datasets）中的测试数据Iris,选用其中的Setosa和Versicolour两类100个样本，每个样本是个四维向量（第一维是花萼的长度，第二维是花萼的宽度，第三维是花瓣的长度，第四维是花瓣的宽度），用每类的前30个训练后20个做测试；视单层感知器网络为弱分类器，对boosting单层感知器网络进行实验。实验结果如表4.1所示。

从表4.1中，可以看出，在两种算法中无论训练误差还是测试误差随着训练轮数T的增加呈下降趋势，但当T大于3时，误差明显下降，当T大于10时其下降趋势略有小幅振荡并趋于平缓。同时，改进AdaBoost_revise1算法比原始的AdaBoost

算法有更小的训练误差和测试误差，说明前者比后者具有更好的分类识别能力，同时具有更好的推广能力。我们还可以从图4.2直观地看出。

表4.1 AdaBoost与AdaBoost_revise1的试验结果比较

误识率 / 轮数 T	原始 Adaboost 算法		Adaboost_reives1 算法	
	训练	测试	训练	测试
1	0.41667	0.475	0.41667	0.475
2	0.41667	0.475	0.31667	0.375
3	0.3	0.35	0.033333	0.125
4	0.083333	0.15	0.1	0.175
5	0.1	0.2	0.016667	0.05
10	0.05	0.15	0.033333	0.05
20	0.0167	0.075	0.016667	0.05
50	0.0167	0.075	0.016667	0.05

图4.2中，横轴表示轮数，竖轴表示误识率，曲线为原始AdaBoost和改进AdaBoost_revise1算法50轮训练误差和测试误差走势图。可以看出改进的AdaBoost_revise1算法无论训练误差还是测试误差都要比原始AdaBoost更低，随着训练轮数的增加都呈下降的趋势并趋于平稳，而且改进AdaBoost_revise1算法的测试误差下降的幅度更大一些说明其具有更好的推广性，这一点正是每个理想模式分类器最希望的。

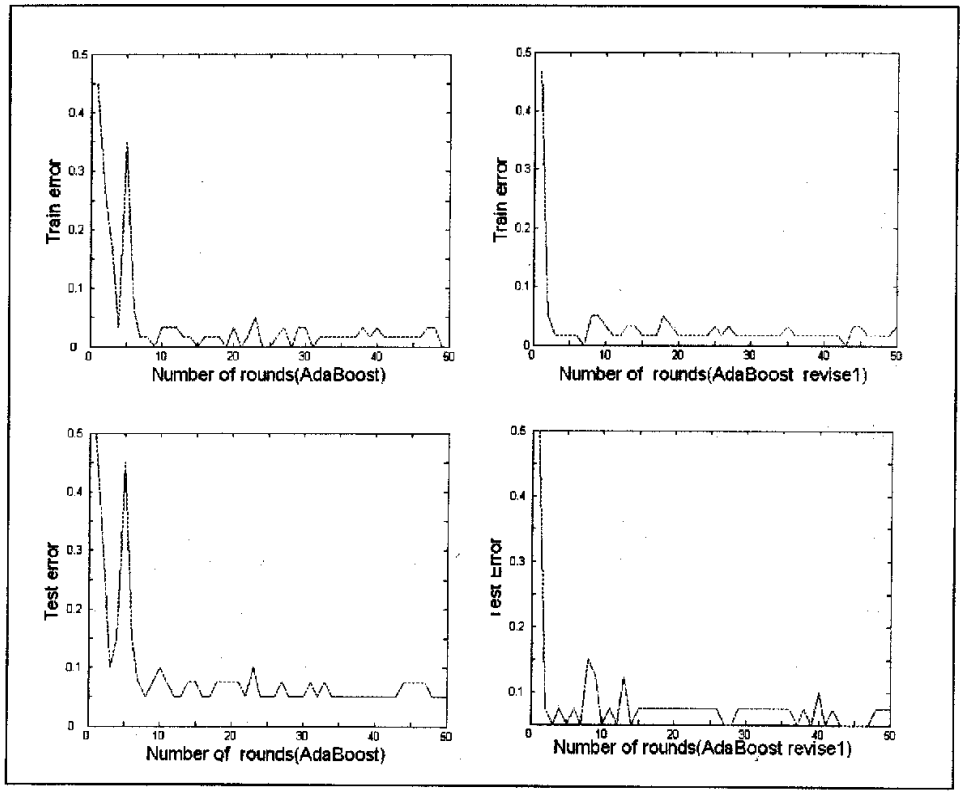

图 4.2 AdaBoost 和 AdaBoost_revise1 训练和测试误差曲线

4.2 Boosting 算法改进（二）

如何能高效、合理地设计融合规则，我们在前人研究的基础上，基于AdaBoost的原理和训练误差最小原则，从数学优化的角度出发，提出了用最小二乘[30]的方法来优化每个弱分类器的投票权，随后的实验及其结果证明了该方法的有效性。

（1） 基本思想

在 AdaBoost 算法中，1)弱分类器组合为 $h=\sum_{t=1}^{K}\alpha_t h_t$，上一节已证明该分类器随 K 是收敛的；2)对两类问题，每一个弱分类器 h_t 对于任一样本其值仅为 $\{-1,+1\}$；3)组合系数为 α_t（ε_t 误差越小，对应的 h_t 权越大)； 4)对多个弱分类器，任意样本 X 经这些分类器有 $h_1, h_2, \ldots, h_k$，即将样本 X 影射到 K 维隐空间的正超立方体的 2^K 个顶点之一上去（或靠近顶点），而组合分类器则是对这个超立方体的顶点的一种线性分类，分类面为 $\sum_{t=1}^{K}\alpha_t h_t=0$（过空间原点），$K$ 越大，空间维数越大，则对样本就越易正确分类。为了直观地说明，2 次分类映射到 2 维空间的 4 个顶点，可用图 4.3 示出：

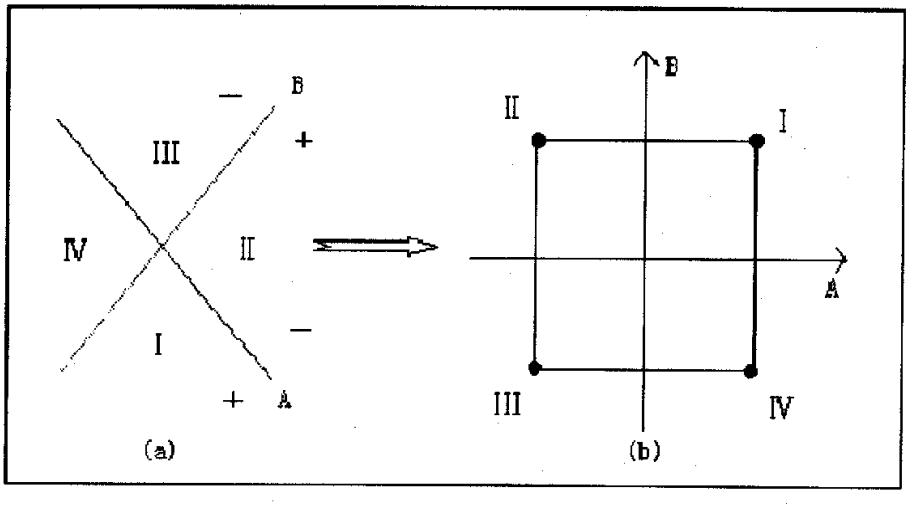

图4.3 2次分类映射到2维空间的4个顶点

在图4.3中，(a)图中第一轮、第二轮分类器分类线A、B将整个样本空间分为Ⅰ、Ⅱ、Ⅲ、Ⅳ四个区域。两次分类后，四个区域分别被映射到以A、B分类线为2维的空间的4个顶点。

实际上，存在如下问题:a)样本不能保证在隐空间中一定线形可分，即使K很大时，也是如此。b)若线形可分，也不一定用过原点的以 α_t 为系数的分类面的错分率最低。由此，我们对弱分类器组合作了改进，可以简单地改进设计AdaBoost算法的多个弱分类器耦合为线性组合：

$$\sum \alpha_t h_t + b \tag{4-1}$$

我们用ω_t代替α_t，θ代替b，Z_i为输出，则：

$$Z_i = sign(\sum_{t=1}^{k} \omega_t h_t(x) + \theta) \tag{4-2}$$

有研究者[31]用感知器网络迭代算出ω_t和θ。在这里，我们用最小二乘法估计出ω_t和θ，即对如下公式作变换：

$$\begin{pmatrix} h_1(x_1) & h_2(x_1)\ldots h_k(x_1) & 1 \\ \vdots & \ddots & \vdots \\ h_1(x_N) & h_2(x_N)\cdots h_k(x_N) & 1 \end{pmatrix} \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_k \\ \theta \end{bmatrix} = \begin{bmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_N \end{bmatrix} \tag{4-3}$$

上式，K为训练轮数，N为训练样本数，$\tilde{z}_i$为理想输出（即训练样本标签）求得：

$$\hat{\omega} = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_k \\ \theta \end{bmatrix} = (H^T H)^{-1} \cdot H^T \cdot \tilde{z} \tag{4-4}$$

这样，根据公式（4-4）求得ω_t和θ，则最终的预测输出为：

$$H(x) = \sum_{t=1}^{T} \omega_t h_t(x) + \theta \tag{4-5}$$

（2）误差分析

显然，改进的AdaBoost_revise2算法用最小二乘方法估计出的弱分类器权系数ω_t和θ代替α_t，从数学分析上，是对训练结果的一种优化，因此，对最终预测结果比原始AdaBoost算法有效小的训练误差；但是，是否有更小的测试误差，或者说是，更好的推广能力，下面就此展开分析。

1）原AdaBoost推广能力分析

简单起见，我们对2次分类映射到2维空间的4个顶点作分析，一般地，在图4.3(a)样本空间中有两种分类方法：Ⅰ区为一类，其余为另一类；或者，Ⅲ区为一类，其余为另一类。因此，当（a）图映射到（b）图时，其线性分类判决函数只有不过原点的线性函数才能将Ⅰ区和Ⅱ、Ⅲ、Ⅳ区分开。基于上述原则，可以得出：AdaBoost算法中，多个弱分类的耦合预测函数 却是一个经过原点的线性函数。但该预测函数不能完全地将Ⅱ、Ⅳ区域分到同一类中。也就是说，AdaBoost的弱分类器耦合出的分类器不稳健。稳健性是指分类系统在一定的参数（结构、大小等）摄动下，仍能维持某些性能的特性，是衡量系统推广能力的一个重要指标。上述过程如图4.4所示，分类线a、b、c都不能满足上述两种分类规则。

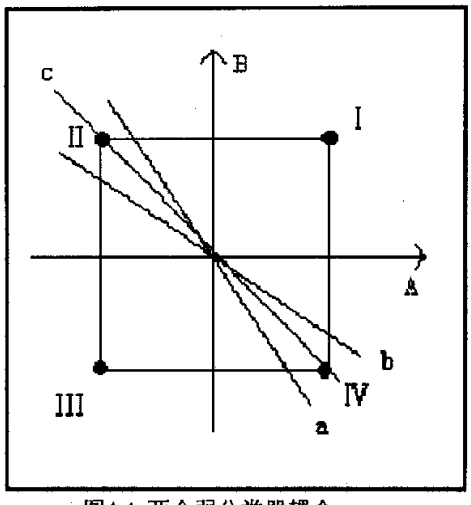

图4.4 两个弱分类器耦合

2）改进的AdaBoost_revise2的推广能力分析。

这里，我们讨论两类问题，因为两类问题是多类问题的基础，改进的AdaBoost将输入空间非线性分类样本经多个弱分类器映射成隐空间中2^K个顶点（或附近的点），并用线性分类器$\sum_{t=1}^{K}\omega_t h_t+\theta=0$实现分类，该线形分类器不必过原点，比原来过原点的分类面$\sum_{t=1}^{K}\alpha_t h_t=0$（过空间原点）更容易实现隐空间的分类。这样，用最小二乘估计出的ω_t和θ，基本保证了隐空间中的稳健分类。

（3）实验分析

为了验证改进AdaBoost_revise2算法的可行性和有效性，我们也进行了分类性能分析的实验：我们对两类问题进行实验，同样，我们采用Iris数据两类样本，选用其中的Setosa和Versicolour两类共100个样本，每个样本是一个四维向量（第一维是花萼的长度，第二维是花萼的宽度，第三维是花瓣的长度，第四维是花瓣的宽度），用每类的前30个训练后20个做测试；采用单层感知器为弱分类器，对boosting单层感知器网络进行实验。实验结果如表4.2所示：

表4.2 AdaBoost与AdaBoost_revise2的试验结果比较

误识率 / 轮数 T	原始 Adaboost 算法		Adaboost_reives2 算法	
	训练	测试	训练	测试
1	0.41667	0.475	0.41667	0.475
2	0.4	0.475	0.4	0.475
3	0.05	0.15	0.05	0.15
4	0.033333	0.075	0.033333	0.075
5	0.016667	0.05	0.016667	0.05
10	0.016667	0.075	0	0.075
20	0.033333	0.1	0.016667	0.05
50	0.016667	0.05	0	0.05

从表4.2中，可以看出，在两种算法中无论训练误差还是测试误差随着训练轮数的增加都呈下降趋势并趋于稳定的很小误差，但当T较大时，改进AdaBoost_revise2算法比原始的AdaBoost算法有更小的训练误差和测试误差，这说明前者比后者具有更好的分类识别能力，同时具有更好的推广能力。

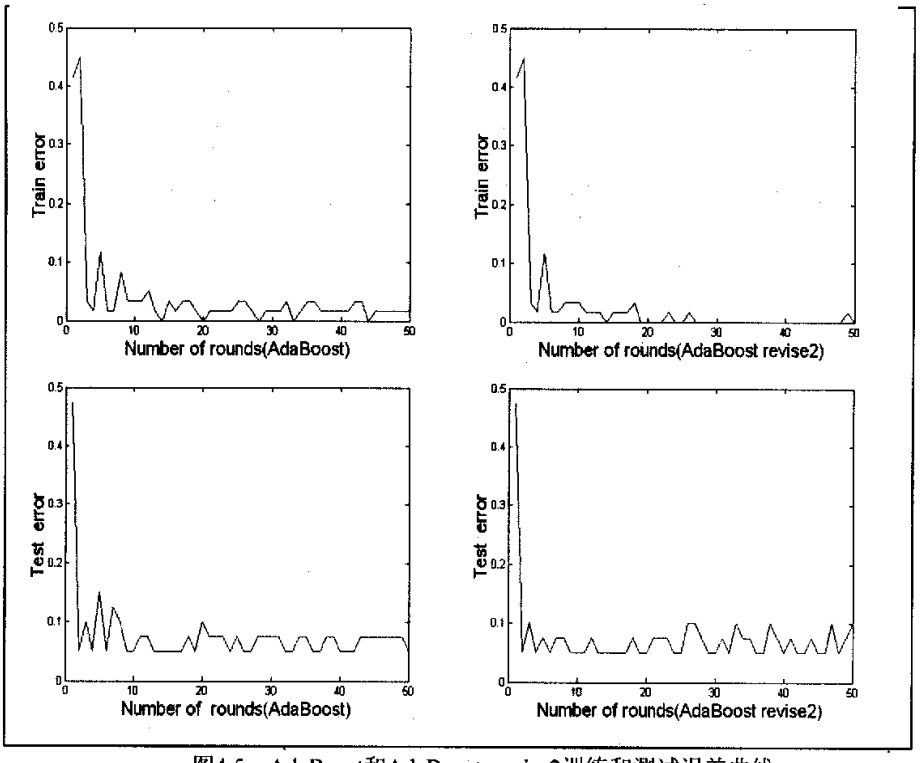

图4.5 AdaBoost和AdaBoost_revise2训练和测试误差曲线

为了直观表示，我们绘制了AdaBoost和AdaBoost_revise2训练和测试误差曲线。在图4.2中，横轴表示轮数，竖轴表示误识率，为原始AdaBoost和改进AdaBoost_revise2算法50轮训练误差和测试误差走势图。明显可以看出改进的AdaBoost_revise2算法的训练误差比原始AdaBoost走势更趋于零，说明训练结果好，随着训练轮数的增加训练误差和测试误差都呈下降的趋势，有小幅震荡趋于平稳，改进AdaBoost_revise2算法的测试误差略有改进，说明其推广性能力稍好，这一点是每个理想模式分类器所希望的。

4.3 本章小结与讨论

本章我们在前人基础上进一步研究Boosting算法，从1）如何更合理地选出多个弱学习的训练集2）如何能高效、合理地设计融合规则这两条原则出发，我们分别提出了AdaBoost_revise1算法：用前T轮组合判决之后分错的样本代替直接用第T轮分错的样本选择下一轮的训练集；AdaBoost_revise2算法：用最小二乘优化弱

分类器权系数的两种改进的AdaBoost算法。我们从理论上作了两种算法的训练误差和推广性能力分析，并用Iris数据从实验上证明了这两种方法的有效性和优越性。当然，Boosting类方法还有许多其它值得研究的地方，比如如何根据分布 D_t 来分配训练样本，以及如何从两类问题推广到多类问题，目前还都没有一个定论，值得探讨。

第五章 基于 Boosting 的人工神经网络集成的基因模式分类

5.1 引言

Boosting具有把弱分类器提升为强分类器的能力，感知器网络只能用于线性分类，可视为是分类能力最弱的弱分类器。我们设想提出，能不能用Boosting训练的感知器网络集成，这是易于实现的，来代替复杂的多层感知器（MLP）神经网络，本章第二节作了Boosting感知器网络集成与MLP神经网络的比较。研究、设计模式分类器的最终目的就是用该分类器来解决具体的应用问题。本章第三节我们将用基于原始的AdaBoost算法和两种改进算法的感知器网络集成做基因模式分类。由于基因数据维数高，样本少，我们先用前人的基因特征选择方法进行基因数据降维。AdaBoost算法思路是根据每次训练的情况确定下次训练的样本，这样多轮训练后可以理解为各轮所选的这些样本的并集覆盖了比原始样本更大的分布情况。因此，可用AdaBoost算法来弥补基因数据样本少的不足。我们采用(leave-one-out)交叉验证实验，依次选一个样本为测试样本，其余为训练样本，这样有效验证系统的性能。在这里，我们选用NCI数据，其基因数目亦即空间维数为2308维，经特征选择降为6维，4种病类64（23+8+12+21）个有效样本。实验结果表明基于Boosting的神经网络集成有效解决基因模式分类，而且基于改进的AdaBoost算法系统性能更优。

5.2 Boosting 感知器网络集成与 MLP 神经网络的比较

我们知道单个感知器由于其结构简单，只能用于线性分类，因此，对复杂的样本分类，它可视为是分类能力最弱的弱分类器。应用 Boosting 算法，我们可将其提升为强分类器。而多层感知器 MLP(Multilayer Perceptron) 神经网络其本身就是一个强分类器。那么，我们就想用 boosting 单层感知器网络能不能代替 MLP 神经网络，有没有更好的分类结果。在这里，我们考虑代替单隐层的 MLP 神经网络，它是网络规模最小的 MLP 神经网络。具体是，对同样的输入，假设为 $n \times r$ 矩阵，n 个样本，r 维矢量，Boosting 的弱学习算法采用感知器网络， boosting 算法共进行 t 轮，那么，每一轮训练一个感知器， t 轮共训练 t 个感知器，那么网络训练的规模就有 $r \times t$ 个的网络权值，这相当于对同样输入，有 t 个隐节点的单隐层 MLP 神经网络的训练的规模。

下面我们就用进行 t 轮的采用感知器网络为弱学习算法的 AdaBoost 算法与同等权值训练规模的 MLP 神经网络作比较，我们还是采用 Iris 数据两类样本（每类 50 个，共 100 个），每个样本是个四维向量，用每类的前 30 个训练后 20 个做

测试，结果如表 5.1 所示。

表 5.1 Boosting 感知器网络与单隐层 MLP 的实验比较

误识样本数 / 轮数 T (隐节点个数)	Adaboost 算法		单隐层 MLP	
	训练（共 60 个样本）	测试（共 40 个样本）	训练（共 60 个样本）	测试（共 40 个样本）
1	23	19	24	14
2	3	12	9	10
3	3	12	24	14
4	2	6	4	3
5	2	3	1	2
10	3	3	2	4
20	2	2	2	4
50	2	2	1	3

从表5.1中，可以看出，采用感知器为弱学习算法的AdaBoost算法，在不同的训练轮数下比同等训练规模的MLP神经网络训练误差差不多，但是测试误差明显要小，这是一个模式分类器所关心的，这说明前者比后者具有更好的分类识别能力，同时具有更好的推广能力。由此，我们可以用一些简单的神经网络，经AdaBoost算法集成，这是易于实现的，来代替大型、复杂的神经网络，往往它们更难于设计。

5.3 基于 Boosting 的神经网络集成的基因模式分类

基因微阵列数据研究的最终目的之一就是对疾病的早期检测、诊断以及最终的治疗提供最大信息量的帮助。因此对基因数据样本的分析，找出病变与正常、一种病变与另一种病变在基因级别上的异同进而区别它们，从信息科学的观点来讲，就是不同病类在基因级别上的模式分类。目前，对于基因模式的分类，很多学者进行了研究和实验，主要有Zhang et al[32](2001), Dudoit et al[33](2002), Furey et al[34](2000)等。传统的统计模式识别方法都是在样本数目足够多的前提下进行研究的，所提出的各种方法只有在样本数趋向无穷大时其性能才有理论上的保证。但由于现有基因数据的特点——样本极其有限但维数极高，很多方法都难以取得很理想的效果。因此，适合基因的模式分类的算法就必须具有解决极少样本超高维模式识别的能力。AdaBoost算法思路是根据每次训练的情况确定下次训练的样本，这样多轮训练后可以理解为各轮所选的这些样本的并集覆盖了比原始样本更大的分布情况。这样，可用AdaBoost算法来弥补基因数据维数高、样本少的不足。

但是基因数据可以通过基因特征选择从超高维降到低维空间，这样利用神经网络分类时可大大减少网络的训练开销和时间。在这里，我们采用刘利平[35]的利用SVM/MLP交叉验证的基因选择方法。这样，我们主要验证AdaBoost算法对极小样本模式的分类能力，并解决病类在基因级别上的模式分类问题。我们对真实的

基因微阵列数据进行了实验：基因数据是neuroblastoma神经细胞和非霍吉金氏淋巴细胞肿瘤的源基因表达数据（美国癌症研究院(NCI)提供），其基因数目亦即空间维数为2308维，4种病类64（23+8+12+21）个有效样本。利用SVM/MLP交叉验证的基因选择方法选出了能够完全识别该四种病类的六个基因向量（151，246，509，545，1389，1955）。我们对经过基因特征选择的NCI基因数据进行(leave-one-out)交叉验证实验。具体是，a)先在第一种病类中取一个样本作测试，其余的样本归为一类做训练样本与其它三种病类两两组合，共得到$C_4^2=6$种组合，依次取其中一种组合为训练样本，然后用Boosting感知器网络进行分类，共得到6种对测试样本的分类识别，对该六种结果采用相对多数投票，若果投票结果不为1，说明识别错误，则识别错误的样本数加1；b)依次在第一种病类中取一个样本作测试，重复a)的实验直到取完第一种病类的所有样本；c)依照对第一种病类的做法，依次完成对其它三种病类的分类识别，并统计误识的样本数。下面，分别基于原始的AdaBoost算法和改进的AdaBoost_revisel和AdaBoost_revise2算法进行分类识别实验（弱学习算法采用感知器网络），有以下表5.2的实验结果：

表5.2 NCI数据的三种算法的分类实验比较

误识样本数 / 训练轮数T	AdaBoost(原始)	AdaBoost_revisel	AdaBoost_revise2
10	6	8	4
20	4	2	4
50	5	2	4

显然，从表5.2中的实验结果可以看出，基于三种算法的感知器网络集成最终分类的误识样本数都远远少于样本总数，说明基于AdaBoost算法的神经网络集成对基因微阵列数据模式识别的有效性。而基于两种改进算法的网络集成的误识样本数要少于基于原始AdaBoost算法的进行实验的误识样本，这也说明了基于改进的AdaBoost_revisel和AdaBoost_revise2算法的神经网络集成比基于原始AdaBoost算法的神经网络集成有更好的基因模式分类能力。

5.4 本章小结与讨论

本章第一节比较了采用单层感知器为弱学习算法，经t轮AdaBoost算法运行与同等权值训练规模的单隐层MLP神经网络的误差，得出前者更有效的结果，由此设想用一些简单的神经网络，经AdaBoost算法集成，这是易于实现的，来代替大型、复杂的神经网络，往往它们更难于设计。

第二节，介绍了用基于Boosting的神经网络集成对基因数据的模式分类。DNA芯片的出现为基因诊断和基因治疗提供了很好的前提和可能性，DNA微阵列数据

的超高维空间和超小样本特性也给我们提出了新的挑战。基于Boosting的神经网络集成因其根据每次训练的情况确定下次训练的样本，这样多轮训练后所选的样本的并集覆盖了比原始样本更大的分布情况。这样，可用来弥补基因数据维数高、样本少的不足，有效地完成基因数据的模式分类。同时，实验也说明了基于改进AdaBoost_revise1和AdaBoost_revise2算法的神经网络集成比原始AdaBoost算法的神经网络集成有更好的基因模式分类能力。

第六章 总结与展望

在信息科学技术和生物医学技术飞速发展的今天，二者的结合导致了基因微阵列数据处理研究的高速发展。而人工神经网络是模仿生物系统行为机理而提出的，我们对于它的研究就更有可能产生行之有效的智能化方法。本论文是在智能信息处理方面的神经网络领域展开工作的。

文章首先阐述了人工神经网络和神经网络集成。然后引用了AdaBoost算法，并对AdaBoost算法的训练误差和推广性作了详尽的分析；在作了深入研究之后，分别从原AdaBoost算法1）如何更合理地选出多个弱学习的训练集；2）如何能高效、合理地设计融合规则两条原则出发，我们提出了两种新的AdaBoost_revise1和AdaBoost_revise2算法。其中：

1. AdaBoost_revise1算法是基于优化训练集的方法

2. AdaBoost_revise2算法是用最小二乘法优化弱分类器权系数的方法。

我们采用了单层感知器网络为弱学习算法，得到了基于boosting的神经网络集成。理论和实验表明改进的AdaBoost_revise1和AdaBoost_revise2算法比原AdaBoost算法有更好的结果，证明了我们新算法的可行性和有效性。

同时，我们采用单层感知器为弱学习算法，经t轮AdaBoost算法运行与同等权值训练规模的单隐层MLP神经网络的比较，得出了前者更有效的结论。

文章最后，用基于boosting的神经网络集成对基因数据进行模式分类，有效解决超高维、少样本的基因数据的分类。

由于时间和水平的限制，本文所达到的研究成果是有限的，一些工作仍需不断的改进和完善，一些工作可以进一步探讨与展开。比如，考虑如何更好地改进Adaboost的推广能力。而如何根据分布D_t来分配训练样本，以及如何从两类问题推广到多类问题，因为目前还都没有定论，值得探讨。另外还可以在如何用简单的神经网络，经AdaBoost算法集成，这是易于实现的，来代替大型、复杂的神经网络，往往它们更难于设计，这也值得在以后工作中探讨。

致 谢

本文的研究工作是在我的导师张军英教授的亲切关怀和悉心指导下完成的。张老师从我选课开始到课程的学习都给予了细心的指导，从论文的选题、论证、研究到最后完成，自始至终无不凝聚着导师的心血。在生活上，张老师也给予我亲切的关怀与帮助。衷心感谢我的导师张军英教授。

感谢张老师创造的宽松民主的学术氛围、仁爱和谐的工作环境；张老师渊博的学识、敏锐的学术洞察力、诲人不倦的师者风范、严谨缜密的治学作风和平易近人的态度，都使我受益终生。

感谢张老师每周一次召开的讨论班，给我们和高年级、低年级的同学有共同学习、进步的机会。

感谢和我一起从事研究工作的所有同学，大家共同创造了一个融洽的工作氛围，使得工作能舒心的开展。

感谢我的父母及家人，是他们默默的支持和关怀，使得我完成毕业论文。

感谢所有支持我和关心我的老师们和朋友们！

参考文献

[1] 孙啸，王睥等. 一种高密度基因芯片设计的新方法. 电子学报. 2001(03).

[2] 潘继红，韩金祥. 基因芯片的制备方法. 生命的化学. 2002(3)..

[3] 郭大东，毕爱莲. 基因芯片技术及其应用. 今日科技. 2002(2).

[4] 徐伟文，李文全. 表达谱基因芯片. 生物化学与生物物理进展. 2001(6).

[5] 巫 影，陈定方，唐小兵. 神经网络综述. 科技进步与对策. 2002(6).

[6] Hornik K M, Stinchcom be M, White H. Multilayer feed forward networks are universal pproximators. NeuraNetworks,1989,2(2):359-366

[7] Hansen L K,Salamon P. Neural network ensembles. IEEE Trans attern Analysis and Machine Intelligence,1990,12(10):993-1001

[8] Sollich P, Krogh A. Learning with ensembles: How over-fitting can be useful. In: Touretzky D, Mozer M, Hasselmo Meds. Advances in Neural Information Processing Systems 8,Cambridge,MA:MITPress,1996.190-196

[9] Perrone M P, Cooper L N. When networks disagree: Ensemble method for neural networks. In: Mamm one R J ed. Artificial Neural Networks for Speech and Vision, NewYork: Chapman&Hall,1993.126-142

[10] Schapire R E. The strength of weak learn ability. Machine Learning, 1990,5(2): 197-227

[11] Breiman L. Bagging predictors. Machine Learning,1996,24(2):123-140

[12]Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky D Leen T eds. Advances in Neural Information Processin Systems 7, Cambridge, MA:MITPress,1995.231-238

[13]Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997,55(1):119-139

[14]Hansen L K, Liisberg L, Salamon P. Ensemble methods for handwritten digit recognition. In: Procthe 1992 IEEE Workshop on Neural Networks for Signal Processing Copenhagen, Denmark, 1992.333-342

[15] Schwenk H, Bengio Y. Boosting neural networks. Neural Computation, 2000,12(8):1869-1887

[16]Gutta S, Wechsler H. Facere cognition using hybrid lassifier systems. In: Proc the IEEE International Conference on Neural Networks, Washington, DC, 1996. 1017 -1022

[17]Gutta S, Huang J R J, Jonathon P, Wechsler H. Mixture of experts for classification of gender, ethnicorigin and pose of human faces. IEEE Trans Neural Networks, 2000, 11(4): 948-960

[18]Cherkauer K J. Human expert level performance on ascientific image analysis task by a system using combined artificial neural networks. In: Proc the 13th AAA I Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, Portland, OR, 1996.15-21

[19]Shimshoni Y, Intrator N. Classification of seismic signals by integrating ensembles of neural networks. IEEE T rans Signal Processing, 1998,46(5):1194-1201

[20] 傅向华, 冯博琴, 马兆丰等. 增量构造负相关异构神经网络集成的方法. 西安交通大学学报. Vol.38,No.8,Aug.2004

[21] Tom M. Mitchell. 机器学习. 北京. 机械工业出版社, 2003.

[22] 边肇祺, 张学工. 模式识别. 北京. 清华大学出版社, 2000.

[23]G. Ratsch, S. Becker, and K-R. Muller. Soft margins for AdaBoost. Machine learning, 42(3):287-320, March 2001. also NeuroCOLT Technical Report NC-TR-1998-021.

[24] Valiant.L.G. A theory of the learnable. Communications of the ACM. 1984, 27(11). 1134~1142.

[25] Gunnar Ratsch. Robust Boosting via Convex Optimization:Theory and Applications. Potsdam:Machematisch-Naturwissenschaftlichen Fakultat der Universitat Postsdam. 2001.

[26] Rebert E. Schapire, Yoram Singer. Improved boosting algorithms using confidence-rated predictions. In Proceedings of the Eleventh Annual conference on Computatinal Learning Theory, Machine Learning. 1999. 80~91.

[27] Yoav Freund, Robert E. Schapire. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference. 1996. 148~156.

[28] Peter L, Bertlett. The sample comleity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Transactions on Information Theory. 1998,44(2). 525~536.

[29]Schapire R E, Freund Y, Bartlett Y, Lee W S. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 1998, 26(5):1651-1686

[30]Simon Haykin 著, 叶世伟, 史忠植 译. 神经网络原理（原书第二版）. 机械工业出版社. 2004.1

[31] 刘申岭, 刘利平等. 基于Boosting的癌症基因模式分类. 西安电子科技大学2003年研究生学术年会. 2003,11. 200~203.

[32] zhang H.,Yu C.etc. Recursive Partitioning for Tumor Classification with Gene

Expression Microarray Data. PNAS. 2001. 6730~6735.

[33] Dudoit S, Fridlyand J.etc. Comparison of Discrimination methods for the Classification of Tumors using Gene Expression Data. JASA. 2002. 77~87.

[34] Furey T, Cristianini N.etc. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression data. Bioinformatics. 2000. 531~537.

[35]刘申岭. 基于SVM的基因选择. 西安电子科技大学2004年研究生毕业论文. 2004.1.3

研 究 成 果

林存炜、李维勤，一种确定盲分离源信号个数的方法，西安电子科技大学计算机学院第一届学术年会录取，西安，2004. 11