中国农业大学

硕士学位论文

支持向量机多类分类算法的研究

姓名：袁玉萍

申请学位级别：硕士

专业：应用数学

指导教师：周志坚

20070501

摘 要

关于支持向量机多类分类问题的模型和算法的研究是当今研究的热点之一。无论是最近提出的“一对一对余”结构的算法，还是通常用的“一对一”结构的算法，对于K类分类问题，都需要解决 K(K-1)/2 个二次规划问题，使得支持向量机在求解大规模问题中就会产生速度很慢的缺陷。因此，研究高效的求解算法是很有意义并且急需解决的问题。本文主要研究多类分类问题，从最优化理论和算法的角度研究支持向量的最优化问题，并建立了高效的求解算法。

本文所做的主要研究工作如下：

1. 构造了基于线性规划的“一对一”三类结构支持向量分类器。由 Chih-Wei H 等人将几个常用的算法，如：“一对多”算法，“一对一”算法，“有向无环图”算法，“纠错输出编码”算法以及两种聚集算法进行了数据试验的比较，试验结果表明“一对一”算法更适合于解决多类分类问题。但其也存在一定的缺点，由于在构造子分类器的时候，只有两类数据参与训练，容易造成由其它数据的信息缺失而带来的错误分类的问题。由 Cecilio A 等人提出的基于二次规划的“一对一”三类结构支持向量机，与传统的“一对一”结构相比较，其优势在于在分解的过程中，除了需要计算被区分的两类训练点外，其它类别中训练点的信息也被充分利用，在一定程度上可以防止由信息的不完全带来的分类误差，同时，也减少了参数的个数。但由于增加了模型的复杂性，限制了其应用。本文构造了基于线性规划的“一对一”三类结构支持向量分类器，可以直接利用比较成熟的线性规划算法—预测-校正原对偶内点法，并在此基础上提出了基于预测-校正原对偶内点法的支持向量机的多类分类学习算法，这种算法可用于比较庞大的多类别识别问题。数值试验表明，本文提出的算法训练速度快，而且保持良好的分类精度。

2. 在 K-SVCR 算法结构的基础上，构造了新的模型。由 Angulo C 等人提出的 K-SVCR 算法，作者只给出了 K－SVCR 模型，并没有提供相应的求解算法，在一定程度上限制了 K－SVCR 算法的推广使用，并且其对偶目标函数为凸函数，而不是严格凸函数。 本文构造了新模型，该模型的特点是它的一阶最优化条件可以转化为一个线性互补问题，通过 Lagrangian 隐函数，可以将其进一步转化成一个严格凸的无约束优化问题。利用 Sherman-Moodbury-identity 等式减小相应优化问题的规模。并在此基础上利用了快速的 Armijo 步长的有限牛顿法和解决大型问题的共轭梯度法来求解无约束优化问题，理论和数值试验都表明有限牛顿法、共轭梯度法速度快、容易实现。另外，支持向量机的模型中含有多个参数，参数的取值直接影响分类的精确度，针对支持向量机结构参数的选取在没有理论支持的情况下，本文利用基本的遗传算法来解决优化问题，并有效估计未知参数。将上述算法应用于 benchmark 数据集的测试，实验表明了此算法的有效性。

关键词： 支持向量机，多类分类，原对偶内点法，牛顿法，共轭梯度法

Abstract

The study of Multi-class classification's models and algorithms in support vector machine(SVM) is an important and on-going research subject. Whatever the recently proposed one-against-one-against -rest and the most popular one-against-one methods, for a K-class classification problem, K(K-1)/2 quadratic programs need to be solved to assign a new pattern to a proper class. So it is extremely important to develop an algorithm that can solve these quadratic programs efficiently. This paper mainly aims at multi-class problems, we do some researches on the SVM by the optimization theory and algorithm and built up algorithm efficiently.

The main works in the paper are follows:

1、A one-versus-one tri-class SVM classifier based on linear programming is proposed. Chih-Wei H compared these experiments on benchmark datasets of these diffent algorithms, such as one-versus-all、one-versus-one 、Directed acidic graph 、Error-correcting-output-code and two all-together algorithms. The result turned out that one-versus-one algorithm adapt to even more settling the multi-class problems. Cecilio A proposed a one-versus-one tri-class SVM classifier based on quadratic programming. The algorithm'virtues is that datasets are utilized enough.However,its formulation is more complicated, which confines its applications.We present a separating hyperplane base on linear programming and a algorithm of Multi-class classification base on predictor-corrector primal dual interior point method.This algorithm is applied to many more huge identify problems of Multi-class. This article introduce this algorithm in detail . This algorithm is tested on benchmark datasets and obtain a better result.

2、We put forward a new formulation which is proposed based on the K-SVCR method. Angulo C proposed the K-SVCR method.They presented the model only,so confines its applications.The tatget value is protruding. In this paper, we start with a new formulation which is proposed based on the K-SVCR method. Then transform it as a complementarity problem and further a strongly convex unconstrained optimization problem by using the implicit Lagrangian function. Make use of Sherman-Moodbury-identity equation to correspond excellent turn the scale of problem .This indicates that the algorithm can be implemented efficiently in practice. Then the Newton algorithm and the conjugate gradient algorithm with global and finite termination properties is established for solving the resulting optimization problem. Make use of genetic algorithm solve the resulting optimization problem and estimate the unbeknown parameter. The procedure of fitness functions is established by Matlab language.This indicates that the algorithm can be implemented efficiently in practice. Preliminary numerical experiments on benchmark datasets show that the algorithm has good performance.

Key words: Support vector machine, Multi-class classification, Primal dual interior point method, Newton method, conjugate gradient method

独 创 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的地方外，论文中不包含其它人已经发表或撰写过的研究成果，也不包含为获得中国农业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：袁玉萍　　时间：2007 年 6 月 16 日

关于论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国农业大学可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

研究生签名：袁玉萍　　时间：2007 年 6 月 16 日

导师签名：[illegible]　　时间：2007 年 6 月 16 日

第一章 绪论

支持向量机(Support Vector Machine，SVM)是数据挖掘中的一项新技术，是基于统计学习理论[1]的结构风险最小化原理基础上提出来的一种学习算法，是借助于最优化方法解决机器学习问题的新工具。它最初于 20 世纪 90 年代由 Vapnik 提出，近年来在其理论研究和算法实现方面都取得了突破性进展，开始成为克服“维数灾难”和“过学习”等传统困难的有力手段。在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，并且能够推广应用到函数拟合等其它机器学习问题中。虽然统计学习理论和支持向量机方法中尚有很多问题需要进一步研究，但很多学者认为，他们正在成为继模式识别和神经网络研究之后机器学习领域中新的研究热点，并将推动机器学习理论和技术的发展。

由于支持向量机是基于统计学习理论的，本章首先介绍统计学习理论的基本核心内容，给出支持向量机算法的理论背景。

1.1 机器学习的基本问题和方法

传统统计模式识别的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，而在多数实际应用中，样本数目通常是有限的，很多方法都难以取得理想的效果。而统计学习理论是一种小样本统计理论，为研究有限样本情况下的统计模式识别和更广泛的机器学习问题建立了一个较好的理论框架，同时也发展了一种新的模式识别方法——支持向量机，能够较好地解决小样本学习问题。

统计学习理论就是研究小样本统计和预测的理论，核心内容包括：基于经验风险最小化准则的统计学习一致性条件；统计学习方法推广性的界；在推广界的基础上建立的小样本归纳推理准则；实现新的准则的实际方法。

1.1.1 机器学习问题的表示

机器学习问题的基本模型，可以用图（1-1）表示。其中，系统 S 是我们研究的对象，它在给定输入 x 下得到一定的输出 y，LM 是所求的学习机，输出为 y'。机器学习的目的是根据给定的已知训练样本求出对系统输入输出之间依赖关系的估计，使它能够对未知输出作出尽可能准确的预测。

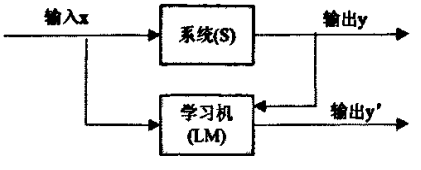


图 1-1 机器学习的基本模型

机器学习问题可以形式化地表示为：已知变量y与输出x之间存在一定的未知依赖关系，即存在一个未知的联合概率$F(x,y)$，机器学习就是根据l个独立同分布观测样本

$$(x_1, y_1),(x_2, y_2),\cdots,(x_l, y_l), \tag{1-1}$$

在一组函数$f(x, w)$中求一个最优的函数$f(x, w_0)$，使预测的期望风险

$$R(w)=\int L(y, f(x, w))dF(x, y) \tag{1-2}$$

最小。其中，$f(x, w)$称为预测函数集，$w\in\Omega$为函数的广义参数，故$f(x, w)$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，为损失函数，不同类型的学习问题有不同的损失函数。

在上面的表述中，学习的目标在于使期望风险最小化，但要使（1-2）式期望风险最小化，必须依赖于联合概率$F(x, y)$的信息，但是，联合概率$F(x, y)$是未知的，在实际的机器学习问题中，我们只能利用已知样本$(x_1, y_1),(x_2, y_2),\cdots,(x_l, y_l)$的信息，无法直接计算和最小化期望风险。因此传统方法采用经验风险最小化的准则。

1.1.2 经验风险最小化

经验风险最小化的准则，即用经验风险

$$R_{emp}[f]=\frac{1}{l}\sum_{i=1}^{l}L(y_i, f(x_i, w)) \tag{1-3}$$

逼近期望风险，由于$R_{emp}[f]$是用已知的训练样本定义的，因此称作经验风险。最小化经验风险在多年的学习方法研究中占据了主要地位，但是，仔细研究经验风险最小化原则和机器学习问题中的期望风险最小化要求，可以发现，从经验风险最小化准则代替期望风险最小化准则没有经过充分的理论论证，只是直观上合理的想当然做法。经验风险最小并不一定意味着期望风险最小，而且会出现“过学习”问题，训练误差小，并不能导致好的预测效果。另外，学习机器的复杂性不但与所研究的系统有关，而且要和有限的学习样本相适应。

1.1.3 复杂性和推广能力

在早期的研究中，人们总是把注意力集中在如何使$R_{emp}[f]$更小，但很快便发现，一味追求训练误差小并不是总能达到好的预测效果。人们将学习机器对未来输出进行正确预测的能力称作推广能力。某些情况下，当训练误差过小反而会导致推广能力的下降，这就是所谓的“过学习”(overfitting)问题。之所以出现“过学习”现象，一是因为学习样本不充分，二是学习机器设计不合理，这两个问题是互相关联的。只要设想一个简单的例子，假设我们有一组训练样本(x, y)，x分布在实数范围内，而y取值在$[0,1]$之间。那么不论这些样本是依据什么函数模型产生的，只要我们用一个函数$f(x, a)=\sin(ax)$去拟合这些样点，其中a是待定参数，总能够找到一个a使训练误差为零，显然得到的这个“最优函数”不能正确代表原来的函数模型。出现这种现象的原因，就是试图用一个复杂的模型去拟合有限的样本，结果导致丧失了推广能力，这就是有限样本下学习机器的复杂性与推广性之间的矛盾。

在有限样本情况下，我们可得出以下基本结论：

(1) 经验风险最小并不一定意味着期望风险最小；

(2) 学习机器的复杂性不但与所研究的系统有关，而且要和有限的学习样本相适应。在有限样本情况下学习精度和推广性之间的矛盾似乎是不可调和的，采用复杂的学习机器容易使学习误差更小，但却往往丧失推广性。因此，人们研究了很多弥补办法，比如在训练误差中对学习函数的复杂性进行惩罚；或者通过交叉验证等方法进行模型选择以控制复杂度等等，使一些原有方法得到了改进。但是，这些方法多带有经验性质，缺乏完善的理论基础。在模式识别中，人们更趋向于采用线性或分段线性等较简单的分类器模型。

1.2 统计学习理论的核心内容

统计学习理论被认为是目前针对小样本统计估计和预测学习的最佳理论。它从理论上较系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险的关系以及如何利用这些理论找到新的学习原则和方法等问题。其主要内容包括四个方面：

(1) 经验风险最小化原则下统计学习一致性的条件；

(2) 在这些条件下关于统计学习方法推广性的界的结论；

(3) 在这些界的基础上建立的小样本归纳推理原则；

(4) 实现这些新的原则的实际方法（算法）。

1.2.1 学习过程一致性的条件

所谓学习过程的一致性（consistency），就是指当训练样本数目趋于无穷大时，经验风险的最优值能够收敛到真实风险的最优值。只有满足一致性条件，才能保证在经验风险最小化原则下得到的最优方法在样本无穷大时趋近于使期望风险最小的最优结果。

学习过程的一致性：记 $f\left(x,\ w^*\right)$ 为在式（1-1）的 l 个独立同分布样本下在函数集中使经验风险取得最小的预测函数，由它带来的损失函数为 $L\left(y,\ f\left(x,\ w^*\middle|l\right)\right)$，相应的最小经验风险值为 $R_{emp}\left(w^*\middle|l\right)$。记 $R\left(w^*\middle|l\right)$ 为在 $L\left(y,\ f\left(x,\ w^*\middle|l\right)\right)$ 下的式（1-2）所取得的真实风险值（期望风险值）。当下面两式成立时称这个经验风险最小化学习过程是一致的：

$$R\left(w^*\middle|l\right)\xrightarrow{l\to\infty}R\left(w_0\right), \tag{1-4a}$$

$$R_{emp}\left(w^*\middle|l\right)\xrightarrow{l\to\infty}R\left(w_0\right). \tag{1-4b}$$

其中，$R\left(w_0\right)=\inf R\left(w\right)$ 为实际可能的最小风险，即式（1-2）的下确界或最小值。

定理 1.2.1（学习理论关键定理） 对于有界的损失函数，经验风险最小化学习一致的充分必要条件是经验风险在如下意义上一致地收敛于真实风险：

$$\lim_{n\to\infty}P\left[\sup_{w}\left|R\left(w\right)-R_{emp}\left(w\right)\right|>\varepsilon\right]=0,\forall\varepsilon>0 \tag{1-5}$$

其中，P 表示概率，$R_{emp}\left(w\right)$ 和 $R\left(w\right)$ 分别表示在 n 个样本下的经验风险和对于同一 w 的真实风

险。因为这一定理在统计学习理论中的重要性，被叫做学习理论的关键定理（Key Theorem of Learning），它把学习一致性的问题转化为式（1-5）的一致收敛问题。

我们的目的不是用经验风险去逼近期望风险，而是通过求使经验风险最小化的函数来逼近能使期望风险最小化的函数，因此其一致性条件比传统统计学中的一致条件更严格。虽然学习理论关键定理给出了经验风险最小化原则成立的充分必要条件，但这一条件并没有给出什么样的学习方法能够满足这些条件。为此，统计学习理论定义了一些指标来衡量函数集的性能，其中最重要的是 VC 维（Vapnik-Chervonenkis Dimension）。

1.2.2 函数集的学习性能与VC维

为了研究函数集在经验风险最小化原则下的学习一致性和一致性收敛的速度，统计学习理论定义了一些指标来衡量函数集的性能，其中最重要的是 VC 维，它是统计学习理论中的一个核心概念，是目前为止对函数集学习性能的最好描述指标。

VC 维直观定义是：假如存在一个有 l 个样本的样本集能够被一个函数集中的函数按照所有可能的 2^l 种形式分为两类，则称函数集能够把样本数为 l 的样本集打散(shatering)。指示函数集的 VC 维就是用这个函数集中的函数所能够打散的最大样本集的样本数目。若对任意数目的样本都有函数能将它们打散，则函数集的 VC 维是无穷大，有界实函数的 VC 维可以通过用一定的阈值将它转化成指示函数来定义。VC 维反映了函数集的学习能力，VC 维越大，则学习机器越复杂。

VC 维是统计学习理论中的一个核心概念，它是目前为止对函数集学习性能的最好描述指标。但是遗憾的是，目前尚没有通用的关于如何计算任意函数集的 VC 维的理论，只有对一些特殊的函数集的 VC 维可以准确知道，而对于一些比较复杂的学习机器，其 VC 维除了与函数集选择有关外，通常也受学习算法等的影响，因此其确定将更加困难。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。

1.2.3 推广性的界

统计学习理论中推广性的界是关于经验风险和期望风险之间的关系的重要结论，它们是分析学习机器性能和发展新的学习算法的重要基础。对于两类分类问题，对指示函数集 $f(x, w)$ 中的所有函数（包括使经验风险最小的函数），经验风险和期望风险之间至少以概率 $1-\eta(\eta \in (0,1])$ 满足如下关系：

$$R[f] \le R_{emp}[f] + \sqrt{\frac{8}{l}\left(h\left(\ln\frac{2l}{h}+1\right)+\ln\frac{4}{\eta}\right)} \tag{1-6}$$

其中 h 是函数集的VC维，l 是样本数。这一结论说明经验风险最小化原则下学习机器的实际风险是由两部分组成的，其中第一部分为训练样本的经验风险，另一部分称作置信范围。置信范围不但受置信水平 $1-\eta$ 的影响，而且更是函数集的VC维和训练样本数目的函数，且随着它的增加而单调减少，可将（1-6）式改写为

$$R(w) \le R_{emp}(w) + \Phi\left(\frac{l}{h}\right) \tag{1-7}$$

（1-6）式给出的是关于经验风险和期望风险之间差距的上界，它们反映了根据经验风险最小化原则得到的学习机器的推广能力，因此称作推广性的界。

1.2.4 结构风险最小化原则

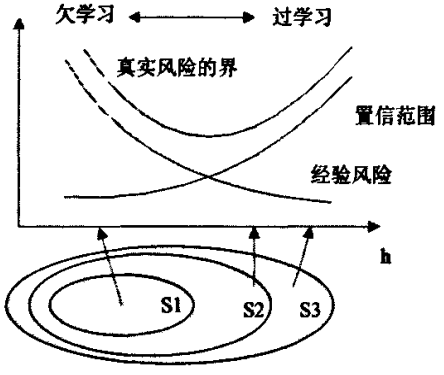
为了最小化期望风险，我们需要同时最小化经验风险和置信范围，根据（1-6）式的理论依据，可以用另一种策略来解决这个问题，首先把函数集$S=\{f(x,w),\ w\in\Omega\}$分解为一个函数子集序列

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots S, \tag{1-8}$$

使各个子集能够按照VC维的大小排列，即

$$h_1 \le h_2 \le \cdots \le h_k \le \cdots, \tag{1-9}$$

选择最小经验风险与置信范围之和最小的子集，就可以达到期望风险的最小，这个子集中使经验风险最小的函数就是要求的最优函数，这种思想称作结构风险最小化（Structural Risk Minimization）原则，简称SRM原则。如图1—2所示。



函数集子集:$S_1 \subset S_2 \subset S_3$ VC维: $h_1 \le h_2 \le h_3$

图 1-2 结构风险最小化示意图

在结构风险最小化原则下，一个分类器的设计过程包括以下两方面任务：
一方面选择一个适当的函数子集（使之对问题来说有最优的分类能力）；另一方面从这个子集中选择一个判别函数（使经验风险最小）。第一步相当于模型选择，而第二步相当于在确定了函数形式后的参数估计。与传统方法不同的是，在这里模型的选择是通过对它的推广性的界的估计进行的。 结构风险最小化原则为我们提供了一种不同于经验风险最小化的更科学的学习机器设计原则，但是由于其最终目的在于式（1-7）的两个求和项之间进行折衷，因此实际上实施这一原则并不容易。支持向量机就是一种比较好地实现了结构风险最小化思想的方法。

1.2.5 核函数

在利用支持向量机解决分类问题时，核函数是支持向量机的重要组成部分，对于非线性分类问题，首先要选择适当的核函数$K(\cdot,\cdot)$，或者说需要选择一个映射$\Phi(\cdot)$，把训练点所在的输入空间X映射到某一个高维的空间中去，然后在这个高维空间中求解优化问题。在映射过程中，并不需要显式计算样本点的象$\Phi(x_i)$，由于支持向量分类机的最终决策函数仅依赖于变换后的Hilbert空间中的内积$(\Phi(x_i)\cdot\Phi(x_j))$，并不需要知道具体的映射是什么，只要选定核函数$K(\cdot,\cdot)$就够了。由于多种不同的特征空间会导致不同的核函数$K(\cdot,\cdot)$，所以核函数应该有较大的选择范围，但是它必须满足如下定义。

定义 1.2.2（核函数（核或正定核）） 设X是R^n中的一个子集，称定义在$X\times X$上的函数$K(x,\ x')$是核函数（核或正定核），如果存在着从X到某个 Hilbert 空间 H 的映射$\Phi(\cdot)$，使得$K(x,\ x')=(\Phi(x)\cdot\Phi(x'))$，其中$(\cdot)$表示 H 中的内积。

常用的几种核函数：

(1) Gauss 径向基核（或 RBF 核）：

$$K(x,\ x')=\exp\left(-\frac{\|x-x'\|^2}{\sigma}\right) \tag{1-10}$$

(2) 多项式核：

$$K(x,\ x')=(x\cdot x')^d,\ d=1,2,\cdots,$$

$$K(x,\ x')=((x\cdot x')+1)^d,\ d=1,2,\cdots, \tag{1-11}$$

(3) 多层感知器，又称 Sigmoid 核

$$K(x,\ x')=\tanh(\kappa(x\cdot x')+\upsilon)^d \tag{1-12}$$

其中$\kappa>0,\ \upsilon>0$。

(4) B-样条核

$$K(x,\ x')=B_{2p+1}(x-x') \tag{1-13}$$

其中$B_{2p+1}(x)$是$2p+1$阶 B-样条函数。

1.3 论文的研究内容

支持向量机由于其特有的准确性和全局解而成为求解分类问题和回归问题的一个非常有效的工具，能较好地解决了小样本、非线性、高维数和局部极小点等问题，但是由于其要求解二次规划，对于 K 类分类问题，都需要解决 K(K-1)/2 个二次规划问题，使得支持向量机在求解大规模问题中就会产生速度很慢的缺陷。目前，建立高效的求解支持向量机中的最优化问题算法，是

支持向量机理论研究中一个急需解决且很有意义的一个问题，因此，本文针对以下几个方面对支持向量机的算法进行了研究和探讨。本文所做的主要研究工作如下：

（1） 构造了基于线性规划的“一对一”三类结构支持向量分类器。由 Chih-Wei H 等人将“一对多”算法，“一对一”算法，“有向无环图”算法，“纠错输出编码”算法，以及两种“聚集”算法进行了数据试验的比较，试验结果表明“一对一”算法更适合于解决多类分类问题。但其也存在一定的缺陷，由于在构造子分类器的时候，只有两类数据参与训练，容易造成由其它数据的信息缺失而带来的错误分类的问题。由 Cecilio A 等人提出的基于二次规划的“一对一”三类结构支持向量机，与传统的“一对一”结构相比较，其优势在于在分解的过程中，除了需要计算被区分的两类训练点外，其它类别中训练点的信息也被充分利用，在一定程度上可以防止由信息的不完全带来的分类误差。但由于增加了模型的复杂性，限制了其应用。本文构造了基于线性规划的“一对一”三类结构支持向量分类器，可以直接利用比较成熟的线性规划算法—预测-校正原对偶内点法。并在此基础上提出了基于预测-校正原对偶内点法的支持向量机的多类分类学习算法。这种算法可用于比较庞大的多类别识别问题。数值试验表明，本文提出的算法的训练速度快，而且保持良好的分类精度。

（2）在 K-SVCR 算法结构的基础上，构造了新的模型。由 Angulo C 等提出的 K-SVCR 算法，作者只给出了 K－SVCR 模型，并没有提供相应的求解算法，在一定程度上限制了 K－SVCR 算法的推广使用，并且其对偶目标函数为凸函数，而不是严格凸的。 本文构造了新模型，该模型的特点是它的一阶最优化条件可以转化为一个线性互补问题，通过 Lagrangian 隐函数，可以将其进一步转化成一个严格凸的无约束优化问题。利用 Sherman-Moodbury-identity 等式减小相应优化问题的规模。并在此基础上利用了快速的有限牛顿算法、解决大型问题的共轭梯度技术来求解无约束优化问题，理论和数值试验都表明有限牛顿算法、共轭梯度技术算法快速、容易实现；支持向量机的模型中含有多个参数，参数的取值直接影响分类的精确度，本文利用基本的遗传算法来解决优化问题，并有效估计未知参数。将上述算法应用于 benchmark 数据集的测试，实验表明了此算法的有效性。

1.4 论文的组织结构

第一章：首先介绍了机器学习的基本问题和方法，给出统计学习理论的核心内容。

第二章：给出几种常用的支持向量机模型，本文主要研究多类分类问题，详细介绍了目前对于多类分类问题国内外研究现状，并且总结了每个算法的优缺点。

第三章：介绍了目前解决线性规划的有效算法：Mehrotra 预测—校正原对偶内点算法理论；基于一种“一对一”三类结构多类分类问题，本章给出其线性规划支持向量机，并用 Mehrotra 预测—校正原对偶内点算法求解。

第四章：针对多类分类问题，本章提出一个新模型，将支持向量分类机和回归机结合在一起，增加了 b^2 项，保证了最优化问题是严格凸的二次规划问题。

第五章：利用有限牛顿法、共轭梯度技术的有限步终止的性质来求无约束优化问题；利用基本的遗传算法来解决优化问题，对优化问题中的参数进行有效估计。

第六章：结论与展望

第二章 支持向量分类机模型

支持向量机的理论目标是找到一个最优的超平面，使其能够尽量多的将每类数据点正确分开，同时要使分开的每类数据点距离分类面最远。其解决的方法是构造一个二次规划问题，然后求解该优化问题，得到分类器。支持向量机涉及到两个最优化问题—原始问题和对偶问题，算法是通过求解对偶问题的解得到原始问题的解，再根据原始问题的解来确定决策函数。本章的主要内容是介绍几种不同支持向量机分类模型[3]。

2.1 两类分类支持向量机的算法

两类分类问题 根据给定的训练集：

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in\{X\times Y\}^l, \tag{2-1}$$

其中 $x_i\in X\subset R^n$，$y_i\in Y=\{+1,-1\}$，$i=1,\cdots,l$，寻找一个从输入空间 X 到输出空间 Y 上的一个实值函数 $g(x)$，以便用决策函数 $f(x)=\mathrm{sgn}(g(x))$ 推断任一模式 x 相对应的 y 值。其中，$f(x)=\mathrm{sgn}(\cdot)$ 是符号函数

$$\mathrm{sgn}(a)=\begin{cases}+1\ ,\ a\geq 0;\\ -1,\ a<0.\end{cases} \tag{2-2}$$

l 个样本点组成的集合称为训练集，样本点称为训练点，x_i 是输入指标向量，或称输入模式，其分量称为特征，或输入指标；y_i 是输出指标，或称输出，$y_i=+1$ 表示输入 x_i 属于正类，$y_i=-1$ 表示输入 x_i 属于负类 。由此可见，分类问题，实质上就是对任意给定的一个新的模式 x，根据训练集，推断它所对应的输出 y 是+1还是−1。

分类问题可以分成两类分类问题和多类分类问题，上述分类问题是两类分类问题。多类分类问题的定义将在下面内容中介绍，它们的不同之处在于前者的输出只取两个值，而后者的输出取多个值。

2.1.1 支持向量分类机算法

支持向量机最初来自两类模式识别问题，其基本思想是要够造一个分类超平面作为决策平面，将两类分开，而且使两类之间的间隔最大[4,5]。

线性支持向量机可以分成线性可分和线性不可分两种情况。对于解决线性可分问题的基本途径是在正确分划训练集的超平面中，根据最大间隔原则，找出最终的决策超平面。

算法 2.1.1(线性可分支持向量分类机)

(1) 设已知训练集： $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in\{X\times Y\}^l$

其中 $x_i\in X\subset R^n$，$y_i\in Y=\{-1,1\}$，$i=1,\cdots,l$；

(2) 构造并求解最优化问题

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \left(x_i \cdot x_j\right) - \sum_{j=1}^{l} \alpha_j \\ s.t. \quad & \sum_{i=1}^{l} y_i \alpha_i = 0, \\ & \alpha_i \geq 0, \ i = 1, \cdots, \ l. \end{aligned} \tag{2-3}$$

得最优解 $\alpha^* = \left(\alpha_1^*, \cdots, \ \alpha_l^*\right)^T$；

(3) 计算 $w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i$；选择 α^* 的一个正分量 α_j^*，并据此计算

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

(4) 构造分划超平面 $\left(w^* \cdot x\right) + b^* = 0$，由此求得决策函数 $f(x) = \mathrm{sgn}\left(\left(w^* \cdot x\right) + b^*\right)$，或

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^* \left(x \cdot x_i\right) + b^*\right)$$

有时也称算法 2.1.1 为线性硬间隔分类机。如图 2-1

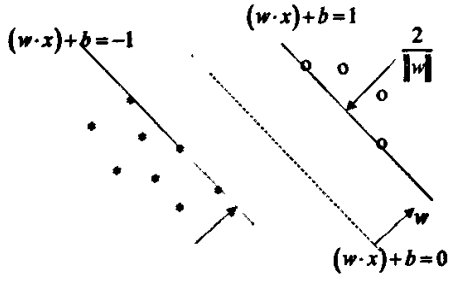


图 2-1 线性可分支持向量分类机最优分类面示意图

对于解决线性不可分问题，如果仍坚持用超平面进行分划，那么必须“软化”对间隔的要求，通过引入松弛变量 $\xi_i \left(\xi_i \geq 0\right)$，$i = 1, \cdots l$，可“软化”约束条件：

$$y_i \left(\left(w \cdot x_i\right) + b\right) \geq 1 - \xi_i, \ i = 1, \cdots, \ l.$$

显然应该设法避免 ξ_i 取太大的值，因此我们在目标函数中加入了惩罚参数 $C\left(C > 0\right)$，用于控制对错分样本的惩罚程度，C 越大，对错误的惩罚越重。

算法 2.1.2(线性支持向量分类机)

(1) 设已知训练集：

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in \{X \times Y\}^l$$

其中 $x_i \in X \subset R^n$，$y_i \in Y = \{-1,1\}$，$i=1,\cdots,\ l$；

(2) 选择适当的惩罚参数 $C(C>0)$，构造并求解最优化问题

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j, \\ s.t. \quad & \sum_{i=1}^{l} y_i \alpha_i = 0, \\ & 0 \le \alpha_i \le C,\ i=1,\cdots,\ l. \end{aligned} \tag{2-4}$$

得最优解 $\alpha^* = (\alpha_1^*,\cdots,\ \alpha_l^*)^T$；

(3) 计算 $w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i$；选择 α^* 的一个正分量 $0 < \alpha_j^* < C$，并据此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j);$$

(4) 构造分划超平面 $(w^* \cdot x) + b^* = 0$，由此求得决策函数 $f(x) = \mathrm{sgn}((w^* \cdot x) + b^*)$。

有时也称算法 2.1.2 为线性软间隔分类机。

在解决线性不可分问题时，我们也可以采取另一途径“核函数”，把寻找超曲面的问题转化为寻找超平面的问题，就是引进从输入空间 R^n 到一个高维 Hilbert 空间的变换 Φ，利用这个变换，由原来的对应于输入空间映射到高维数的空间后，由原来的线性不可分则增加了线性可分的可能性。

算法 2.1.3(可分支持向量分类机)

(1) 设已知训练集：

$$T = \{(x_1, y_1),\cdots,(x_l, y_l)\} \in \{X \times Y\}^l$$

其中 $x_i \in X \subset R^n$，$y_i \in Y = \{-1,1\}$，$i=1,\cdots,\ l$；

(2) 选取适当的核函数 $K(x,\ x')$，构造并求解最优化问题

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i,\ x_j) - \sum_{j=1}^{l} \alpha_j, \\ s.t. \quad & \sum_{i=1}^{l} y_i \alpha_i = 0, \\ & \alpha_i \ge 0,\ i=1,\cdots,\ l. \end{aligned} \tag{2-5}$$

得最优解 $\alpha^* = (\alpha_1^*,\cdots,\ \alpha_l^*)^T$；

(3) 选择α^*的一个正分量α_j^*，并据此计算 $b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_j, x_i)$；

(4) 构造决策函数 $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^* K(x \cdot x_i) + b^*\right)$。

有时也称算法 2.1.3 为非线性硬间隔分类机。

2.1.2 两种常用的支持向量机算法

C-支持向量分类机和ν-支持向量分类机是支持向量机理论中最基本、最常用的方法。将线性软间隔分类机和非线性硬间隔分类机进行综合，便得到C-支持向量分类机。

算法 2.1.4（C-支持向量分类机——C-SVC）

（1）设已知训练集:

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in \{X \times Y\}^l$$

其中$x_i \in X \subset R^n$，$y_i \in Y = \{-1, 1\}$，$i = 1, \cdots, l$；

（2）选择适当的核函数$K(x, x')$和适当的参数$C(C > 0)$，构造并求解最优化问题

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j, \\ s.t. \quad & \sum_{i=1}^{l} y_i \alpha_i = 0, \\ & 0 \le \alpha_i \le C, \quad i = 1, \cdots, l. \end{aligned} \tag{2-6}$$

得最优解$\alpha^* = (\alpha_1^*, \cdots, \alpha_l^*)^T$；

（3）选取α^*的一个正分量$0 < \alpha_j^* < C$，并据此计算阈值

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

（4）构造决策函数 $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^* K(x, x_i) + b^*\right)$。

有时也称算法 2.1.4 为非线性硬间隔分类机。

在C-支持向量分类机中，将$\sum_{i=1}^{l} \xi_i^2$用$\sum_{i=1}^{l} \xi_i$来代替，便得到C-支持向量分类机的一种变形：

算法 2.1.5（C-支持向量分类机的变形）

（1）设已知训练集:

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in\{X\times Y\}^l$$

其中 $x_i\in X\subset R^n$，$y_i\in Y=\{-1,1\}$，$i=1,\cdots,l$；

（2）选择适当的核函数 $K(x,\ x')$ 和适当的参数 $C(C>0)$，构造并求解最优化问题

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j\left(K(x_i,\ x_j)+\frac{1}{C}\delta_{ij}\right)-\sum_{j=1}^{l}\alpha_j,\\&s.t.\ \ \sum_{i=1}^{l}y_i\alpha_i=0,\\&\qquad\alpha_i\geq 0,\ i=1,\cdots,\ l.\end{aligned}\tag{2-7}$$

其中 $\delta_{ij}=\begin{cases}1,\ i=j,\\0,\ i\neq j.\end{cases}$ 得最优解 $\alpha^*=(\alpha_1^*,\cdots,\ \alpha_l^*)^T$；

（3）选取 α^* 的一个正分量 $\alpha_j^*>0$，并据此计算阈值

$$b^*=y_j\left(1-\frac{\alpha_j^*}{C}\right)-\sum_{i=1}^{l}y_i\alpha_i^*K(x_i,\ x_j);$$

（4）构造决策函数 $f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}y_i\alpha_i^*K(x,\ x_i)+b^*\right)$。

由于C-支持向量分类机存在一定的缺点，选取C值比较困难，因此人们提出了一个改进的方法——ν-支持向量分类机（简称ν-SVC），它用另一个参数ν代替参数C，而参数ν又有直观上的意义，可以控制支持向量机的数目和误差，选取比较自然，有利于参数选择。

算法 2.1.6（ν-支持向量分类机——ν-SVC）

(1) 设已知训练集:

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in\{X\times Y\}^l$$

其中 $x_i\in X\subset R^n$，$y_i\in Y=\{-1,1\}$，$i=1,\cdots l$；

(2) 选择适当的参数ν和核函数 $K(x,\ x')$，构造并求解最优化问题

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i,\ x_j),\\&s.t.\ \ \sum_{i=1}^{l}y_i\alpha_i=0,\\&\qquad 0\leq\alpha_i\leq\frac{1}{l},\ i=1,\cdots,\ l,\\&\qquad\sum_{i=1}^{l}\alpha_i\geq\nu\end{aligned}\tag{2-8}$$

得最优解 $\alpha^* = (\alpha_1^*, \cdots, \alpha_l^*)^T$；

(3) 选取 $j \in S_+ = \{i | \alpha_i^* \in (0, 1/l),\ y_i = 1\}$，$k \in S_- = \{i | \alpha_i^* \in (0, 1/l),\ y_i = -1\}$

计算　　$b^* = -\frac{1}{2}\sum_{i=1}^{l} \alpha_i^* y_i \left(K(x_j, x_i) + K(x_k, x_i)\right)$；

(4) 构造决策函数 $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^* K(x, x_i) + b^*\right)$

2.2 多类分类问题

多类分类问题 根据给定的训练集：

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in \{X \times Y\}^l,$$

其中 $x_i \in X \subset R^n, y_i \in Y = \{\theta_1, \theta_2, \cdots, \theta_K\}, (K > 2), i = 1, 2, \cdots, l$，寻找一个从输入空间 $X \subset R^n$ 上的一个实值函数 $g(x)$，以便用决策函数 $f(x)$ 推断任一模式 x 相对应的 y 值。由此可见，求解多类分类问题，实质上就是找到一个把 R^n 上的点分成 K 部分的规则。

2.2.1 目前解决多类分类问题的国内外研究现状

目前已经提出了一些基于 SVM 的多类分类学习算法。从结构上分，主要有两大类方法："分解-重建（decomposition-reconstruction）"法和"聚集法（all-together）"。前者首先是先将多类分类问题转化为两类分类问题来构造模型，然后再把得到的各个分类器进行重新组合来得到多类分类问题的分类器，该类方法包括"一对多"算法[6]、"一对一"算法[7, 8]、"有向无环图(DAG)"算法[9]、"纠错输出编码"算法[10-12]、"一对一对余"算法[14-15]，"层树分类"算法[16,17]等。后者是通过直接构造一个基于支持向量机的模型来解决多类分类问题，通常需要解决大规模的优化问题，该类方法包括"K 类 SVM"算法[18]、"QP-MC-SV"算法[19]、"LP-MC-SV"算法[20]、"球结构分类"算法[21]等。

(1) "一对多"算法（One-versus-the rest Method），针对不同的 K 个类别，构造 K 个支持向量机子分类器，第 k 个子分类器是将第 k 类与其余的类别分开。在构造第 k 个子分类器的时候，将属于第 k 类别的样本数据标记为正类，将不属于 k 类别的样本数据标记为负类。测试时，对测试数据分别计算各个子分类器的函数值，并选取函数值最大对应的类别为测试数据的类别。完成这个过程需要计算 K 个二次规划。该方法简单，容易实现，但是不足之处在于建立的两类问题其正负类的规模是比较不对称的（特别是在类别数较多时），容易产生没有被分类的点和属于多类别的点。

(2) "一对一"算法（One-versus-one Method）是由 Knerr 提出来的，对于 K 个类别中的任意两个类别，构造一个二分类器，结果共构造 K(K-1)/2 个子分类器。在构造类 i 和类 j 的子分类器时，在样本数据集中选取属于类 i 和类 j 的样本数据作为训练样本数据，并将属于类 i 的数据标记

为正，属于类 j 的数据标记为负。最后采用投票法决定测试数据的类别。完成这个过程需要计算 K(K-1)/2 个二次规划。该方法优点在于每个判别函数的支持向量数少，而不足之处在于计算量非常庞大，并且存在不同的子分类器的输出并不具有可比性的问题。同时由于在构造子分类器的时候，只有两类数据参与训练，容易造成由其它数据的信息缺失而带来的错误分类的问题。

(3)“有向无环图算法”(Directed acidic graph，DAG)，该算法在 K 类训练样本中构造所有不同类别的二分类器，共需构造 K(K-1)/2 个两类分类器，每个分类器对应两类。由于利用了图论中的有向无环图的思想，其拓扑结构如图 2-2。测试时，将测试点输入根节点，每次判别时排除掉最不可能的一个类别，经过 K-1 次判别后剩下的最后一个类别即为该点所属的类别。该算法的优点是推广误差只取决于类数 K 和节点上的类间间距，而与输入空间的维数无关.

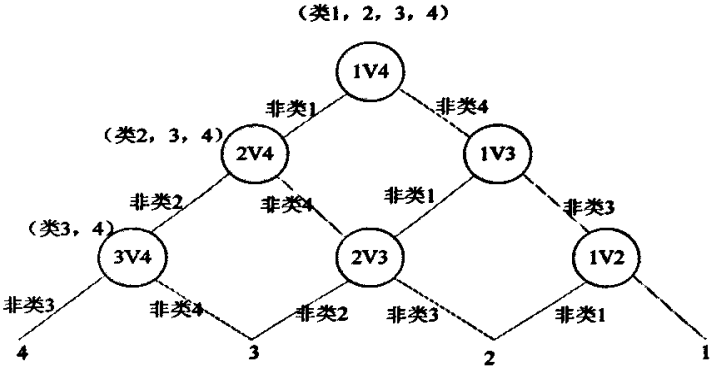


图 2-2 四类问题有向无环图结构图

(4)“纠错输出编码”算法(error-correcting-output-code)，对于 K 类分类问题，可以根据不同方法构造一系列的两类分类问题，对于每个两类分类问题可以建立一决策函数，共得到 L 个决策函数，如果这些决策函数完全正确，K 类中的每一类都对应一个元素为-1 或+1 的长度为 L 的数列，按照 K 类中的第一类、第二类、……、第 K 类的顺序，把这些数列排列起来，便可得到一个 K 行 L 列的编码矩阵，若要判断一个测试输入点的归属，首先用所得到的 L 个决策函数，得到一个元素为-1 或 1 的长度为 L 的数列，然后将此数列与先前得到矩阵比较，相应于矩阵中有一行且仅有一行向与此数列相同，这个行数就是输入点的归属类；若矩阵中没有一行与该数列相同，可以通过计算汉明距离找出最接近的一行，该行对应的类别即为该点的类别。

(5)“一对一对余”算法(one-versus-one-versus-rest)，它们结合了“一对一”和“一对多”的结构，把输入分为三类来构造决策函数 $f_{j,k}(\cdot)$，即它们在将两类 (θ_j,θ_k) 分开的同时也把其余类与这两类分开，共需构造 K(K-1)/2 个两类分类器，然后采用投票法，得票最多的类为此点所属的类。

(6)“层树分类”算法，是对“一对一”算法的改进，将 K 个分类合并为两个大类，每类里面再分为两个子类，如此下去，直到最基本的 K 个分类。给定一个测试样本，从根节点开始根据子分类器的输出值决定其到下一个节点，一直如此，到叶子节点为止，得到测试样本的归属类别。这种方法提高了判别的准确率，但是每级的根节点直接影响分类结果，不同的根节点选取方式可能产生分类的不确定性。

(7)“K 类 SVM”算法是由 Bennett 等人提出的，是将所有的样本点放入一个二次规划中，只需要一次就可以决定分类。这种方法的局限在于由于一次需要处理所有的数据，约束条件大量增加，进行分类的二次规划规模相当庞大。即使转化成线性规划，数据的规模依然受限。

(8)“QP-MC-SV”算法， Westton 提出了两种新的 K 类 SVM 算法，他在构造决策函数是很自然地同时考虑所有的类，将原始问题推广为相应的决策函数。

(9)“LP-MC-SV”算法，此算法与(2)相似，这两种方法的缺点是计算量都比较大，优点是得到的决策分类面的支持向量机的数据均比常规少。

(10)“球结构分类”算法，将同一类数据用超球来界定，数据空间变为由若干个超球组成，在三维上面像是很多肥皂泡的集合. 此算法在复杂性、扩充性和数据规模上都比较占优势。

以上方法都各有千秋，但都没有脱离支持向量机的最优超平面，形象地说，支持向量在多分类问题上相当于多个超平面将数据空间分割，每一类数据都被若干超平面围在一个区域里。
本论文主要研究多类分类问题的算法，以下章节以多类分类问题为主要内容。

2.3 小结

本章主要阐述常用的几种支持向量机分类机的模型，包括：线性可分支持向量分类机、线性支持向量分类机、可分支持向量分类机、C-支持向量分类机，ν-支持向量分类机。无论哪种支持向量机都涉及到两个最优化问题—原始问题和对偶问题，算法是通过求解对偶问题的解得到原始问题的解，再根据原始问题的解来确定决策函数，这是支持向量机很关键的主要内容。本章还介绍了目前解决多类分类问题的国内外研究现状。

第三章　基于预测-校正原对偶内点法的多类分类支持向量机

3.1 引言

由于支持向量机是一种新型的有监督的机器学习方法，具有坚实的理论基础和良好的推广能力，比较好地解决了小样本、高维数、非线性等问题，已被成功地应用于字体识别[22]、文本自动分类[23]、人脸识别[24]等问题。但是，作为一种新兴的学习机器，支持向量机也存在一些有待完善的地方，例如：支持向量机一般采用多个两类分类支持向量机来求解，这就需要求解多个二次规划问题，从而导致算法的计算复杂性，使得它在求解大规模数据上具有一定的局限性，尤其是对于多类分类问题；支持向量机算法由于变量数目过多，因此只能在小型问题的求解中使用；支持向量机是通过求解对偶问题来解原问题，求解对偶问题的困难之处在于 Hessian 阵的稠密性，当数据规模很大时，将面临维数灾难或者由于内存限制导致无法训练。因此，建立高效的求解支持向量机中的最优化问题的算法，是支持向量机理论研究中一个急需解决的问题。

由 Chih-Wei H 等人[20]将一些常用的解决多类分类问题的算法进行了比较，如："一对多（1-v-r）"算法，"一对一（1-v-1）"算法，"有向无环图"算法，"纠错输出编码"算法，以及两种聚集算法进行了数据试验的比较，试验结果表明"一对一"算法更适合于解决多类分类问题。

由 Cecilio A 等人[25]提出的基于二次规划的"一对一"三类结构支持向量机，彻底减小了优化问题的规模。与传统的"一对一"结构相比较，其优势在于在分解的过程中，除了需要计算被区分的两类训练点外，其它类别中训练点的信息也被充分利用，在一定程度上可以防止由信息的不完全带来的分类误差。但由于增加了模型的复杂性，限制了其应用。

由于有比较成熟的处理大规模的线性规划算法和软件，本文采取了"一对一"的三类结构的思想，建立了一种基于线性规划的"一对一"三类结构支持向量分类器，并且给出了基于预测-校正原对偶内点法的支持向量机学习算法。这种算法可用于比较庞大的多类别识别问题。数值试验表明，本文提出的算法的训练速度快，而且保持良好的分类准确率。

3.2 求解 1-v-1 三类结构 SVM 方法的线性规划

原-对偶内点法是以对数障碍函数为基础，每次迭代要么原问题的约束等式不成立，要么对偶问题的约束等式不成立，但是原问题和对偶问题中的变量必须满足正值性。关键之处是把搜索点控制在可行域的内部，并在可行域的边界上设置一道"障碍"，当迭代点靠近可行域边界时，使目标函数值迅速增大，并在迭代中适当控制步长，从而使迭代点始终留在可行域内部，随着障碍因子的减少，障碍函数的作用将逐渐降低，最后将收敛于原问题的最优解。

由于在原-对偶内点法中，求解牛顿方程是内点法中最主要的计算量，包括系数矩阵的因子化、前代和回代，而因子化的计算量要远远高于前代和回代。Mehrotra 预测-校正算法就是将牛顿方向分成两部分：一部分是预测方向，用于消减原-对偶的不可行性和动态估计障碍参数；另一部分是校正方向，用于保持当前迭代点远离可行域的边界。因此，预测-校正算法在每一步迭代都需要求解两次同样规模、同样稀疏的方程，虽然增加了每次迭代的计算量，但只需一次因子化，使

总的迭代次数和计算时间都显著减少，大幅度的改进了原-对偶内点法的性能。Mehrotra 预测-校正算法能够很好的估计障碍因子，能获得较大的迭代步长，鲁棒性较强，改善了算法的收敛性能。

3.2.1 Mehrotra 预测-校正原对偶内点法理论

Mehrotra 预测-校正法(Predictor-corrector method)[26]是直接应用原-对偶法和级数展开法，不断地改正牛顿搜索方向以加快对偶空隙（或对偶松弛条件不符值）的减小。

首先，考虑如下标准形式的原问题和对偶问题

线性规划原问题：　　对偶问题：

$$\begin{aligned} P:\min_{x}\ & c^Tx \\ s.t.\ & Ax=b, \\ & x\geq 0. \end{aligned} \qquad \begin{aligned} D:\max_{y,z}\ & b^Ty \\ s.t.\ & A^Ty+z=c, \\ & z\geq 0. \end{aligned} \tag{3-1}$$

其中 $x\in R^n$，$A\in R^{m\times n}\left(m\leq n\right)$，$b\in R^m$，$c\in R^n$，$y\in R^m$ 是对偶变量，$z\in R^n$ 是对偶问题中加入的松弛变量。

对对偶问题引进对数障碍函数，则问题转换为

$$\begin{aligned} \max_{y,z}\ & b^Ty+\mu\sum_{j=1}^{n}\ln z_j \\ s.t.\ & A^Ty+z=c \end{aligned} \tag{3-2}$$

其中 μ 为障碍参数$\left(\mu>0\right)$，显然 x 为上式的约束条件的拉格朗日乘数，因此相应的拉格朗日函数为

$$L\left(x,\ y,\ z,\ \mu\right)=b^Ty+\mu\sum_{j=1}^{n}\ln z_j-x^T\left(A^Ty+z-c\right) \tag{3-3}$$

其中 下标 j 表示向量第 j 个元素。

由库恩—图克条件（Karush-Kuhn-Tucker，KKT）第一阶最优条件，即其分别对变量 y，x，z 求偏导，并令其等于零，得下列方程组

$$Ax-b=0 \tag{3-4}$$

$$A^Ty+z-c=0 \tag{3-5}$$

$$DGe-\mu e=0 \tag{3-6}$$

其中，$D=diag(x_1,x_2\cdots x_n)$，$G=diag(z_1,z_2,\cdots,z_n)$，$e$ 为各元素均为1的 n 维列向量。式(3-4)为原始可行条件，式（3-5）为对偶可行条件，式（3-6）为互补松弛条件，因此，原对偶内点法在寻优过程中既考虑了问题的原始可行性，同时也考虑了相应对偶问题的可行性。

以$\left(x^k,y^k,z^k\right)$为当前迭代点，$\left(\Delta x,\Delta y,\Delta z\right)$表示牛顿修正方向，由 Newton 方法求解式（3-4）—（3-6）的 KKT 最优性条件：

$$\begin{bmatrix} -G & 0 & -D \\ -A & 0 & 0 \\ 0 & -A^T & -I \end{bmatrix}\begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}=\begin{bmatrix} DGe-\mu e \\ Ax^k-b \\ A^Ty^k+z^k-c \end{bmatrix} \tag{3-7}$$

用阻尼牛顿法求解该方程组得修正方向：

$$\begin{aligned} \Delta y &= -\left(AG^{-1}DA^T\right)^{-1}\left[G^{-1}\left(A^Ty^k+z^k+\mu e-DGe-Ax^k+b\right)\right] \\ \Delta z &= -A^Ty^k-z^k+c-DA^T\Delta y \\ \Delta x &= G^{-1}\left(\mu e-DGe\right)-DG^{-1}\Delta z \end{aligned} \tag{3-8}$$

然后更新原对偶变量：$x^{k+1}=x^k+\alpha_p^k\Delta x^k$，$y^{k+1}=y^k+\alpha_D^k\Delta y^k$，$z^{k+1}=z^k+\alpha_D^k\Delta z^k$　　(3-9)

式中　k 表示迭代次数；α_p，α_D 分别为原变量和对偶变量的迭代步长：

$$\bar{a}_p=\min_{1\le j\le n}\left\{\frac{-x_j}{\Delta x_j}\middle|\Delta x_j<0\right\},\ \bar{a}_D=\min_{1\le j\le n}\left\{\frac{-z_j}{\Delta z_j}\middle|\Delta z_j<0\right\}$$

取步长　　$a_p=\gamma\bar{a}_p,\ a_D=\gamma\bar{a}_D,\left(0.99<\gamma<0.9995\right)$　　(3-10)

γ 为安全因子，通常取 $\gamma=0.9995$，以保证 $x>0,\ z>0$。

原-对偶内点法以互补间隙 $\rho=(x^{k+1})^Tz^{k+1}$ 作为收敛依据（若 $\rho\le10^{-6}$ 则认为算法已收敛），若不满足收敛条件则修正障碍参数：　$\mu=\left(\frac{\rho}{\left(x^k\right)^Tz^k}\right)^2\left(\frac{\rho}{n}\right)$　　(3-11)

然后进入下一次迭代，直到算法收敛。

Mehrotra 提出的预测-校正内点法将修正方向 Δ 分为两部分：$\Delta=\Delta_{af}+\Delta_{co}$，一部分是预测方向 Δ_{af}，用于消减原对偶的不可行性和动态估计障碍参数；另一部分是校正方向 Δ_{co}，用于保持当前迭代点远离可行域的边界。从而使牛顿方向可以高阶近似逼近中心路径[27]。

首先解下列“仿射”方程组，令式(3-7)中的 $\mu=0$，可得仿射方向 Δ_{af}：

$$\begin{bmatrix} -G & 0 & -D \\ -A & 0 & 0 \\ 0 & -A^T & -I \end{bmatrix}\begin{bmatrix} \Delta x_{af} \\ \Delta y_{af} \\ \Delta z_{af} \end{bmatrix}=\begin{bmatrix} DGe \\ Ax^k-b \\ A^Ty^k+z^k-c \end{bmatrix} \tag{3-12}$$

这是预测-校正法的预测过程。计算仿射迭代步长 α_{af}，$\alpha_{af}=\min\{\alpha_p,\alpha_D\}$，式中 α_p，α_D 由式（3-10）确定。然后计算仿射补偿间隙：　$\rho_{af}=\left(x^k+\alpha_{af}\Delta x^k{}_{af}\right)^T\left(z^k+\alpha_{af}\Delta z^k{}_{af}\right)$

计算仿射障碍参数：令　$\mu=\left(\frac{\rho_{af}}{\left(x^k\right)^T z^k}\right)^2\left(\frac{\rho_{af}}{n}\right)$

根据求得的预测方向 Δ_{af} 和估计的障碍参数 μ，则行成校正方程：

$$\begin{bmatrix}-G & 0 & -D\\ -A & 0 & 0\\ 0 & -A^T & -I\end{bmatrix}\begin{bmatrix}\Delta x_c\\ \Delta y_c\\ \Delta z_c\end{bmatrix}=\begin{bmatrix}-\Delta G\Delta De-\mu e\\ 0\\ 0\end{bmatrix} \tag{3-13}$$

式中　$\Delta D=diag(\Delta x_1,\Delta x_2,\cdots,\Delta x_n)$，$\Delta G=diag(\Delta z_1,\Delta z_2,\cdots,\Delta z_n)$。求解此式可获得校正方向 Δ_{co}，即可得总的牛顿方向 $\Delta=\Delta_{af}+\Delta_{co}$。

因此,新的迭代点为：

$$\begin{aligned}x^{k+1}&=x^k+\alpha_p^k\Delta x^k\\ y^{k+1}&=y^k+\alpha_D^k\Delta y^k\\ z^{k+1}&=z^k+\alpha_D^k\Delta z^k\end{aligned} \tag{3-14}$$

3.3 1-v-1 三类结构线性规划支持向量机

3.3.1 基于二次规划的 1-v-1 三类结构支持向量机

“一对一” 三类结构 SVM 方法[25] (One-versus-one tri-class SVM Method)是在已建立的多类分类算法的基础之上，提出将所有的数据参与构造支持向量机子分类器。针对 K 个类别，构造 K（K-1）/2 个支持向量机子分类器。在构造类 i 和类 j 的分类器时，不仅选取属于类 i 和类 j 的样本数据作为训练样本数据，并将属于类 i 的数据标记为 +1，属于类 j 的数据标记为 −1，而且区别于这两类的其它类别的样本数据作为一类数据同时参与到训练过程中，并标记为 0。在训练过程中，结合 SVM 分类器中的规范化超平面和固定间隔的有序回归[28]，使所有的样本点同时参与训练，可以有效地防止在“一对一”方法的训练过程中信息缺失问题，以及在“一对多”方法中的数据不对称问题。

对 K 类问题，数学表示如下：令 Z 是有 K 个类别 $\{\theta_1,\cdots,\theta_K\}$ 的样本集，其中类 k 为 $Z_k=\{z_i=(x_i,y_i):y_i=\theta_k\}$，样本数 $n_k=|Z_k|$，共有样本数为 $n=n_1+n_2+\cdots n_K$。在构造类 θ_i 和 θ_j 子分类器时，不妨以 Z_1 和 Z_3 分别记标号为 −1 和 +1 的两类，以 Z_2 记标号为 0 的类，即除去 θ_i 和 θ_j 的其它类。该方法的二次规划模型为

$$\min_{w\in F,b_1,b_2\in R}\ \frac{1}{2}\|w\|^2+C\sum_i\left(\xi_i^1+\xi_i^2+\xi_i^{*2}+\xi_i^{*3}\right) \tag{3-15}$$

$$s.t.\ \ \left\langle x_i,w\right\rangle-b_1\le-1+\xi_i^1,\qquad z_i\in Z_1, \tag{3-16}$$

$$\left\langle x_i,w\right\rangle-b_1\ge1-\xi_i^{*2},\qquad z_i\in Z_2, \tag{3-17}$$

$$\langle x_i, w\rangle - b_2 \le -1 + \xi_i^2, \quad z_i \in Z_2, \tag{3-18}$$

$$\langle x_i, w\rangle - b_2 \ge 1 - \xi_i^{*3}, \quad z_i \in Z_3, \tag{3-19}$$

$$\xi_i^1,\ \xi_i^2,\ \xi_i^{*2},\ \xi_i^{*3} \ge 0. \tag{3-20}$$

3.3.2 基于线性规划的 1-v-1 三类结构支持向量机

3.3.2.1 线性核

引入对偶变量 $\alpha_i^1(z_i \in Z_1)$，$\alpha_i^{*2}(z_i \in Z_2)$，$\alpha_i^2(z_i \in Z_2)$，$\alpha_i^{*3}(z_i \in Z_3) \ge 0$，根据二次规划问题（3-15）—（3-20）的 Karush-Kuhn-Tucker (KKT) 条件，可以得到

$$w = -\sum_{z_1} \alpha_i^1 x_i + \sum_{z_2} \alpha_i^{*2} x_i - \sum_{z_2} \alpha_i^2 x_i + \sum_{z_3} \alpha_i^{*3} x_i$$

令

$$v_i = \begin{cases} -\alpha_i^1, & z_i \in Z_1, \\ \alpha_i^{*2} - \alpha_i^2, & z_i \in Z_2, \\ \alpha_i^{*3}, & z_i \in Z_3. \end{cases}$$

则 w 可以改写为 $w = \sum_i v_i x_i$ 。用线性规划代替二次规划，对问题（3-15）—（3-20）需要进行如下改动：

（ i ）将目标函数式（3-15）的 l_2 模 $\|w\|^2$ 用 l_1 模 $\sum_{z_1} \alpha_i^1 + \sum_{z_2} \alpha_i^{*2} + \sum_{z_2} \alpha_i^2 + \sum_{z_3} \alpha_i^{*3}$ 代替；

（ ii ）用 $w = \sum_i v_i x_i$ 代替约束（3-16）—（3-19）中的 w；

（iii）增加约束 $\alpha_i^1,\ \alpha_i^{*2},\ \alpha_i^2,\ \alpha_i^{*3} \ge 0$。

这样便得到线性规划模型为：

$$\min \quad \frac{1}{n_1}\sum_{z_1} \alpha_i^1 + \frac{1}{n_2}\sum_{z_2}\left(\alpha_i^{*2} + \alpha_i^2\right) + \frac{1}{n_3}\sum_{z_3} \alpha_i^{*3} + C\sum_i \left(\xi_i^1 + \xi_i^2 + \xi_i^{*2} + \xi_i^{*3}\right) \tag{3-21}$$

$$s.t. \quad \sum_i v_i \langle x_i, x_j\rangle - b_1 \le -1 + \xi_j^1, \quad z_j \in Z_1, \tag{3-22}$$

$$\sum_i v_i \langle x_i, x_j\rangle - b_1 \ge 1 - \xi_j^{*2}, \quad z_j \in Z_2, \tag{3-23}$$

$$\sum_i v_i \langle x_i, x_j\rangle - b_2 \le -1 + \xi_j^2, \quad z_j \in Z_2, \tag{3-24}$$

$$\sum_i v_i \langle x_i, x_j \rangle - b_2 \geq 1 - \xi_j^{*3}, \quad z_j \in Z_3, \tag{3-25}$$

$$\alpha_i^1,\ \alpha_i^{*2},\ \alpha_i^2,\ \alpha_i^{*3},\ \xi_i^1,\ \xi_i^{*2},\ \xi_i^2,\ \xi_i^{*3} \geq 0. \tag{3-26}$$

求解线性规划问题（3-21）—（3-26）得到分类器

$$f_1(x) = \mathrm{sgn}\left(\sum_i v_i \langle x_i, x \rangle - b_1\right) \tag{3-27}$$

$$f_2(x) = \mathrm{sgn}\left(\sum_i v_i \langle x_i, x \rangle - b_2\right) \tag{3-28}$$

3.3.2.2 非线性核

利用核函数，不难把它推广到非线性支持向量机。在选定核函数 $K(\cdot,\cdot)$ 后，用 $K(x_i, x_j)$ 和 $K(x_i, x)$ 分别代替约束（3-22）－（3-25）中的 $\langle x_i, x_j \rangle$ 和式（3-27）－（3-28）中的 $\langle x_i, x \rangle$，这样便得到优化问题

$$\min \quad \frac{1}{n_1}\sum_{z_1}\alpha_i^1 + \frac{1}{n_2}\sum_{z_2}\left(\alpha_i^{*2} + \alpha_i^2\right) + \frac{1}{n_3}\sum_{z_3}\alpha_i^{*3} + C\sum_i\left(\xi_i^1 + \xi_i^2 + \xi_i^{*2} + \xi_i^{*3}\right) \tag{3-29}$$

$$s.t. \quad \sum_i v_i K(x_i, x_j) - b_1 \leq -1 + \xi_j^1, \quad z_j \in Z_1, \tag{3-30}$$

$$\sum_i v_i K(x_i, x_j) - b_1 \geq 1 - \xi_j^{*2}, \quad z_j \in Z_2, \tag{3-31}$$

$$\sum_i v_i K(x_i, x_j) - b_2 \leq -1 + \xi_j^2, \quad z_j \in Z_2, \tag{3-32}$$

$$\sum_i v_i K(x_i, x_j) - b_2 \geq 1 - \xi_j^{*3}, \quad z_j \in Z_3, \tag{3-33}$$

$$\alpha_i^1,\ \alpha_i^{*2},\ \alpha_i^2,\ \alpha_i^{*3},\ \xi_i^1,\ \xi_i^{*2},\ \xi_i^2,\ \xi_i^{*3} \geq 0. \tag{3-34}$$

求解问题（3-29）—（3-34）得到分类器

$$f_1(x) = \mathrm{sgn}\left(\sum_i v_i k(x_i, x) - b_1\right) \tag{3-35}$$

$$f_2(x) = \mathrm{sgn}\left(\sum_i v_i k(x_i, x) - b_2\right) \tag{3-36}$$

3.3.3 求解 1-v-1 三类结构 SVM 方法的线性规划

近年来，原-对偶内点法，以及预测-校正原对偶内点法因具有计算速度快，鲁棒性好等优点而被广泛地应用于求解线性和非线性优化问题，并且在求解大规模问题上有着显著的优势。在原-对偶内点法中，系数矩阵的因子化和前代、回代计算量是最主要的计算量。由于系数矩阵因子化的计算量要远远高于前代、回代运算，因此预测—校正技术的主要目的[29]是将系数矩阵因子化次数减少到最少，甚至以增加单步迭代的时间为代价，原-对偶内点法的性能将大幅度得以提高。本文中以预测-校正原对偶内点法为基础对优化问题（3-29）－（3-34）进行求解。

记 $m=n_1+n_2+n_2+n_3$，其中，n_1,n_2,n_3 分别为类 Z_1,Z_2,Z_3 中的样本点数，$e_{n_1},e_{n_2},e_{n_3},e_m$ 分别为维数为 n_1,n_2,n_3,m 的各个分量均为 1 的行向量。记

$\lambda=\left(\frac{1}{n_1}e_{n_1},\frac{1}{n_2}e_{n_2},\frac{1}{n_2}e_{n_2},\frac{1}{n_3}e_{n_3},Ce_m,0_{(m+2)}\right)$，$I_{m\times m}$ 为 m 阶单位矩阵，

$$H=\begin{pmatrix}-(k(x_i^1,x_j^1)) & (k(x_i^2,x_j^1)) & -(k(x_i^2,x_j^1)) & (k(x_i^3,x_j^1))\\ (k(x_i^1,x_j^2)) & -(k(x_i^2,x_j^2)) & (k(x_i^2,x_j^2)) & -(k(x_i^3,x_j^2))\\ -(k(x_i^1,x_j^2)) & (k(x_i^2,x_j^2)) & -(k(x_i^2,x_j^2)) & (k(x_i^3,x_j^2))\\ (k(x_i^1,x_j^3)) & -(k(x_i^2,x_j^3)) & (k(x_i^2,x_j^3)) & -(k(x_i^3,x_j^3))\end{pmatrix}\in R^{m\times m},$$

$$B^T=\begin{pmatrix}-e_{n_1} & e_{n_2} & 0 & 0\\ 0 & 0 & -e_{n_2} & e_{n_3}\end{pmatrix}\in R^{2\times m},\ \beta^T=\begin{pmatrix}-e_{n_1} & -e_{n_2} & -e_{n_2} & -e_{n_3}\end{pmatrix}。$$

引入松弛变量 $\eta_j^1,\eta_j^{*2},\eta_j^2,\eta_j^{*3}\geq 0$，模型（3-29）－（3-34）可以简单地化成如下形式：

$$\begin{aligned}\min\ & \lambda r\\ s.t.\ & Ar=\beta\\ & r\geq 0\end{aligned}\tag{3-37}$$

其中 $A=[H,-I,B,I]$。预测-校正原对偶内点法需要求解仿射方程：

$$\begin{bmatrix}A & 0 & 0\\ 0 & A^T & I\\ G & 0 & D\end{bmatrix}\begin{bmatrix}\Delta r^{Aff}\\ \Delta\tau^{Aff}\\ \Delta u^{Aff}\end{bmatrix}=-\begin{bmatrix}Ar^k-\beta\\ A^T\tau^k+u^k-\lambda^T\\ \Delta D\Delta Ge\end{bmatrix},\tag{3-38}$$

其中 τ 为对偶变量和 u 为对偶松弛变量，$G=diag(r^k)$，$D=diag(u^k)$，$\Delta G=diag(\Delta r^k)$，$\Delta D=diag(\Delta u^k)$。其计算步骤如下：

（1）初始化：给出原变量 r 与对偶变量 τ 的初始值，并保证原变量 r 与对偶松弛变量 u 的正值性，给出障碍因子 $\mu\geq 0$，置迭代次数 $k=0$，收敛精度 $\varepsilon_1=\varepsilon_2=10^{-6}$，对偶间隙 $Gap=100\varepsilon_1$。

（2）如果对偶间隙 $Gap<\varepsilon_1$ 和 KKT 条件的最大范数 $\left\{\left\|\beta-Ar^k\right\|,\left\|\lambda^T-A^T\tau^k-u^k\right\|\right\}<\varepsilon_2$，则输出最优结果并终止计算。否则转（3）。

（3）形成并求解仿射方程（3-38）得到仿射方向 $\Delta r^{aff},\Delta\tau^{aff},\Delta u^{aff}$，障碍因子 μ。

（4）根据求得的仿射方向 Δ_{aff} 和估计的障碍因子 μ，计算校正方向 Δ_{co}；

（5）计算 $\Delta=\Delta_{aff}+\Delta_{co}$，得到牛顿方向 $\Delta r,\Delta\tau,\Delta u$，根据迭代公式（3-14），更新 r,τ,u，测试

对偶间隙值Gap（互余松弛条件不符值），转 Step 2，置$k=k+1$。

3.4 数值实验

本章报告了在 UCI 数据库中的 Iris，Wine，Glass，Vehicle，Vowel 五个数据集上进行实验的结果。实验数据如表 3-1 所示。可以看到表 3-1 数据的类别数，样本数和特征向量的维数均不同，保证了实验数据具有一定的代表性。由于这五个数据集没有测试数据，将训练数据分为 10 份，尽量使每 1 份中包含的各个类别样本数相同，依次将每 1 份作为测试数据，其余 9 份作为训练数据进行测试，共测试 10 次，结果取平均值。

实验中用到了多项式核(ploy)函数$k(x,y)=((x,y)+1)^d$ $(d\in N)$和 RBF 径向基核函数$k(x,y)=\exp(-\sigma\|x-y\|^2)$ $(\sigma\in R)$。测试中对正则化因子C和核的宽度参数σ进行了打网格式赋值，其中$C=[2^9,2^8,\cdots,2^{-9}]$，$\sigma=[2^6,2^5,\cdots,2^{-6}]$。实验平台 Intel Pentium IV，2.00GHz, 512M RAM PC 机, 用 Matlab7.0 编程。识别时间和识别率见表 3-2。

表 3-1　测试数据

测试数据	Iris	Wine	Glass	Vehicle	Vowel
类别数	3	3	6	4	11
特征向量维数	4	13	9	18	10
训练集	150	178	214	846	528

表 3-2　试验结果

测试数据	Iris	Wine	Glass	Vehicle	Vowel
核/C/σ或d	RBF/2^{-1}/2^{-2}	P/2^8/1	RBF/2^7/2^0	RBF/2^8/2^3	RBF/2^{-2}/2^3
测试精度	[97.33, 99.33]	[90.00, 99.41]	[67.14, 76.19]	[85.12, 87.50]	96.70
测试时间	5.55s	6.72s	111.35s	2.11s	2416.3s

RBF: RBF 径向基核函数; P: 多项式核函数.

[·,·]表示当有测试点不仅在它应该归属的类别上得到最多票，同时在其他类别上也得到相同最多票时的准确率情况。括号里的前一个数据是把这种测试点都归为分类错误时的测试精度，后一个数据是把这种测试点都归为分类正确时的测试精度。

从实验结果可以看出，使用本文提出的基于预测-校正原对偶内点法的支持向量机的多类分类学习算法是非常有效的，不仅能将准确率保持在较高的水平，而且计算时间短。

3.5 小结

本章首先介绍了预测-校正原始对偶内点法的理论内容，针对多类分类问题的支持向量机的研究，基于二次规划的“一对一”三类结构支持向量基础上，提出了一种基于线性规划的分类器，并且给出了基于预测-校正原对偶内点法的支持向量机的多类分类学习算法。这种算法可用于比较庞大的多类别识别问题。数值试验表明这种支持向量机模型训练速度快，推广能力好。

第四章　建立多类分类支持向量机新模型

4.1 引言

关于支持向量机多类分类问题的模型和算法的研究是当今研究的热点之一。最近对该问题提出了一种具有新型结构的K-SVCR算法[14]，与其它算法相比较，此算法最大的优点在于在训练的过程中，能够利用训练数据的所有信息。但是作者只给出了K-SVCR模型，并没有提供相应的求解算法，在一定程度上限制了K-SVCR算法的推广使用，并且其对偶目标函数为凸函数，而不是严格凸函数。本章的主要内容是在K-SVCR算法结构的基础上，构造出一个新模型。该模型将支持向量分类机和回归机结合在一起，增加了b^2项，保证了最优化问题是严格凸的二次规划问题。该模型的特点是：它的一阶最优化条件可以转化为一个线性互补问题，通过Lagrangian隐含数[30]，可以将其进一步转化成严格凸的无约束优化问题，利用 Sherman-Moodbury-identity 等式减小相应优化问题的规模。文中将给出一些相关理论的推导和证明。

4.2 解决多类分类问题建立的新模型

4.2.1 线性核公式

对任意的一个类别对$(\theta_j,\theta_k)\in Y\times Y$，我们希望建立一个决策函数$f(x)$，使其能够将$\theta_j,\theta_k$和剩余的其它类别区分开来。不失一般性，假设训练点$x_i$，$i=1,\cdots,l_1$属于$\theta_j$类，并且将它们标号为+1；训练点$x_i$，$i=l_1+1,\cdots,l_1+l_2$属于$\theta_k$类，并且将它们标号为-1；其它的训练点看作一类，以0来标记。具体地说，我们希望建立如下的决策函数：

$$f(x_i)=\begin{cases}+1, & i=1,\cdots,l_1,\\ -1, & i=l_1+1,\cdots,l_1+l_2,\\ 0, & i=l_1+l_2+1,\cdots,l.\end{cases} \tag{4-1}$$

为了方便，以下我们记 $l_{12}=l_1+l_2, l_3=l-l_{12}$.

本文提出了一个新模型：

$$\min_{(w,b,\xi,\eta,\tilde{\eta})\in R^{n+1+l_{12}+l_3+l_3}} \frac{1}{2}\|w\|^2+\frac{\delta_1}{2}\xi^T\xi+\frac{\delta_2}{2}(\eta^T\eta+\tilde{\eta}^T\tilde{\eta})+\frac{1}{2}b^2 \tag{4-2}$$

$$s.t. \quad D(Aw+be)\geq e-\xi, \tag{4-3}$$

$$Bw+be\leq \varepsilon e+\eta, \tag{4-4}$$

$$-Bw-be\leq \varepsilon e+\tilde{\eta}. \tag{4-5}$$

其中，$\delta_1,\delta_2,\varepsilon$是事先选定的正数，且$\varepsilon<1$，此外，$A\in R^{l_{12}\times n}$，$A_{i\cdot}$表示$A$的第$i$行向量，对应于类$\theta_j$或$\theta_k$内的某一个训练点；$D$是一个对角矩阵，其对角上的元素为+1或-1，由相应的$A_{i\cdot}$的类标号决定；$B\in R^{l_3\times n}$，$B_{i\cdot}$表示B的第i行向量，对应于除类θ_j和θ_k外，剩余类中某一个训练点。e表示单位向量。若记$\mu=\delta_2$，$\tau=\sqrt{\delta_1/\delta_2}$，可以将式(4-2)-(4-5)转化成下面的规划问题：

$$\min_{(w,b,\xi,\eta,\tilde{\eta})\in R^{m_1+k_2+k_3}} \frac{1}{2}\|w\|^2+\frac{\mu}{2}(\xi^T\xi+\eta^T\eta+\tilde{\eta}^T\tilde{\eta})+\frac{1}{2}b^2 \tag{4-6}$$

$$s.t. \quad \tau D(Aw+be)\geq \tau e-\xi, \tag{4-7}$$

$$Bw+be\leq \varepsilon e+\eta, \tag{4-8}$$

$$-Bw-be\leq \varepsilon e+\tilde{\eta}. \tag{4-9}$$

通过引入 Lagrangian 乘子，可以得到式(4-6)- (4-9)的对偶问题:

$$\min_{0\leq\gamma\in R^{k_1+k_2+k_3}} \frac{1}{2}\gamma^T\left(\bar{D}(SS^T+ee^T)\bar{D}+\frac{I}{\mu}\right)\gamma-t^T\gamma \tag{4-10}$$

其中 I 表示单位矩阵且

$$\bar{D}=diag(\tau D,-I,I),\ S=\left[A^T,B^T,B^T\right]^T,\ t=\left[\tau e^T,-\varepsilon e^T,-\varepsilon e^T\right]^T. \tag{4-11}$$

同时，还可以得到 $w=S^T\bar{D}\gamma,\quad b=e^T\bar{D}\gamma$ (4-12)

这样我们就得到了分离超平面为

$$X^Tw+b=X^TS^T\bar{D}\gamma+b=0 \tag{4-13}$$

定义矩阵 $G:=\bar{D}[S,-e]$，$Q:=GG^T+\frac{I}{\mu}$ (4-14)

则可以将公式(4-10)变形为: $\min_{0\leq\gamma\in R^{k_1+k_2+k_3}} \frac{1}{2}\gamma^TQ\gamma-t^T\gamma$ (4-15)

引理 4.2.1: 矩阵 $Q=GG^T+\frac{I}{\mu}$ 是对称的正定矩阵。

证明:因为 $\mu>0,\frac{I}{\mu}$ 是正定的, GG^T 是半正定，由于对称正定矩阵和对称半正定矩阵的和是正定矩阵,因此 Q 是对称的正定矩阵。

由引理 4.2.1 我们可以知道式(4-15)是一个只带有非负约束的正定二次规划问题。

4.2.2 非线性核公式

在这一节里，给出非线性核的一般形式，分离超平面式(4-13)可以转换成下面的非线性核超平面:

$$K\left(x^T,\ S^T\right)\bar{D}\gamma+b=0 \tag{4-16}$$

其中，$K(\cdot,\cdot)$ 是在上述的常用核中任意选择的核函数，S、$\bar{D}$ 在式(4-11)已定义，根据式(4-6)-(4-9)，我们考虑下面的广义支持向量机[31]

$$\min_{\gamma,\zeta,b} \frac{1}{2}\gamma^TM\gamma+\frac{\mu}{2}\zeta^T\zeta+\frac{1}{2}b^2 \tag{4-17}$$

$$s.t. \quad \bar{D}K\left(S,\ S^T\right)\bar{D}\gamma+b\bar{D}e\geq t-\eta. \tag{4-18}$$

其中 $M \in R^{(l_{12}+l_3+l_4)\times(l_{12}+l_3+l_4)}$ 是一个对称半正定矩阵，t 在公式(4-11)中定义

$$\zeta=\left[\xi^T,\ \eta^T,\ \tilde{\eta}^T\right]^T$$

且
$$K\left(S,\ S^T\right)=\begin{bmatrix} K\left(A,\ A^T\right) & K\left(A,\ B^T\right) & K\left(A,\ B^T\right) \\ K\left(B,\ A^T\right) & K\left(B,\ B^T\right) & K\left(B,\ B^T\right) \\ K\left(B,\ A^T\right) & K\left(B,\ B^T\right) & K\left(B,\ B^T\right) \end{bmatrix} \tag{4-19}$$

相似于[32]，可以任选对称半正定矩阵 M。当 $M=\bar{D}^2$，$\gamma=\bar{D}^{-1}K\left(S,\ S^T\right)^T\bar{D}\lambda$ 时，则得到式(4-17)-(4-18)的对偶问题为：

$$\min_{0\le\lambda\in R^{l_{12}+l_3+l_4}}\frac{1}{2}\lambda^T\left[\bar{D}\left(K\left(S,\ S^T\right)K\left(S,\ S^T\right)^T+ee^T\right)\bar{D}+\frac{I}{\mu}\right]\lambda-t^T\lambda \tag{4-20}$$

对于式(4-20)可以看作是在式(4-10)线性核中的 SS^T 由非线性核 $K\left(S,S^T\right)K\left(S,S^T\right)^T$ 所代替，定义矩阵：

$$G=\bar{D}\left[K\left(S,\ S^T\right),e\right]\ ,\qquad Q=GG^T+\frac{I}{\mu} \tag{4-21}$$

4.3 将有约束问题转化为无约束问题

下面将式(4-15)(或(4-20))转化成无约束问题。

由 Karush-Kuhn-Tucker（KKT）条件可以得到式(4-15)的一阶最优化条件为

$$\gamma^T\left(Q\gamma-t\right)=0,\qquad Q\gamma-t\ge 0,\qquad \gamma\ge 0\ . \tag{4-22}$$

对于任意的正常数 v，式(4-22)的等价形式为

$$Q\gamma-t=\left(Q\gamma-t-v\gamma\right)_+, \tag{4-23}$$

其中，函数 $\left(\alpha_i\right)_+:=\max\left\{\alpha_i,0\right\}$。当 $v\ge\|Q\|$ 时，利用 Lagrangian 隐含数[30]将式(4-15)等价地转化为

$$\min_{\gamma\in R^{l_{12}+l_3+l_4}}L(\gamma):=\frac{1}{2}\gamma^TQ\gamma-t^T\gamma+\frac{1}{2v}\left(\left\|(Q\gamma-t-v\gamma)_+\right\|^2-\left\|Q\gamma-t\right\|^2\right) \tag{4-24}$$

$L\left(\gamma\right)$ 的梯度为：

$$\begin{aligned}\nabla L(\gamma)&=\left(Q\gamma-t\right)+\frac{1}{v}(Q-vI)(Q\gamma-t-v\gamma)_+-\frac{1}{v}Q(Q\gamma-t)\\&=\frac{(vI-Q)}{v}\left((Q\gamma-t)-\left((Q-vI)\gamma-t\right)_+\right)\end{aligned} \tag{4-25}$$

$L(\gamma)$的通常意义下的海色阵不存在，但是它的广义的海色阵(Hessian)存在[32,33]。

$$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\left(Q+diag\left((Q-\nu I)\gamma-\mu\right)_*(\nu I-Q)\right) \tag{4-26}$$

其中 $\left(diag\left((Q-\nu I)\gamma-\mu\right)_*\right)_{jj}=\begin{cases}1, & if\ (Q-\nu I)_{j\cdot}\gamma-\mu_j>0,\\ a, & if\ (Q-\nu I)_{j\cdot}\gamma-\mu_j=0,\\ 0, & if\ (Q-\nu I)_{j\cdot}\gamma-\mu_j<0.\end{cases}$ (4-27)

$a\in[0,1]$的常数。为简单起见,我们定义：

$$N(\gamma):=(Q\gamma-\mu)-\left((Q-\nu I)\gamma-\mu\right)_+, \tag{4-28}$$

得：　$\partial N(\gamma)=Q+diag\left((Q-\nu I)\gamma-\mu\right)_*(\nu I-Q)$ (4-29)

因此，　$\nabla L(\gamma)=\frac{(\nu I-Q)}{\nu}N(\gamma)$ ，$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\partial N(\gamma)$ (4-30)

命题 4.3.1 若另$E(\gamma)=diag\left((Q-\nu I)\gamma-\mu\right)_*$，则：

(1) 当$\nu>\|Q\|$时，$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\left(Q+diag\left((Q-\nu I)\gamma-\mu\right)_*(\nu I-Q)\right)$是正定的。

(2) $\partial N(\gamma)=Q+diag\left((Q-\nu I)\gamma-\mu\right)_*(\nu I-Q)$, $\partial N(\gamma)$是非退化的，且

$$\partial N(\gamma)^{-1}=\left(I-PG\left(I+G^TPG\right)^{-1}G^T\right)C^{-1} \tag{4-31}$$

其中：$C=\nu E(\gamma)+\frac{I-E(\gamma)}{\mu}$，$P=C^{-1}\left(I-E(\gamma)\right)$

证明：(1) 已知：$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\left(Q+diag\left((Q-\nu I)\gamma-\mu\right)_*(\nu I-Q)\right)$，由引理 4.2.1 知，

矩阵$Q:=GG^T+\frac{I}{\mu}$是对称正定的，且在$\nu\geq\|Q\|$时，$(\nu I-Q)$是对称正定的。又因为：

$$\left((\nu I-Q)Q\right)^T=Q^T\left(\nu I-Q^T\right)=Q(\nu I-Q)=\nu Q-Q^2=(\nu I-Q)Q$$

可知，$(\nu I-Q)Q$是对称正定的，由$(\cdot)_*$函数的定义，$E(\gamma)$中的所有元素都在$[0,1]$内，

由$E(\gamma)=\left(E(\gamma)^{\frac{1}{2}}\right)^2$，因此

$$(\nu I-Q)E(\gamma)(\nu I-Q)=\left(E(\gamma)^{\frac{1}{2}}(\nu I-Q)\right)^{T}\left(E(\gamma)^{\frac{1}{2}}(\nu I-Q)\right)$$

此式表明，$(\nu I-Q)E(\gamma)(\nu I-Q)$是对称半正定的，$\partial^2 L(\gamma)$可表示为：

$$\partial^2 L(\gamma)=\frac{1}{\nu}\left((\nu I-Q)Q+(\nu I-Q)E(\gamma)(\nu I-Q)\right)$$

即$\partial^2 L(\gamma)$是对称正定的。

(2) 因为$(\nu I-Q)$和$\partial^2 L(\gamma)$是正定的，由式（4-30）：$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\partial N(\gamma)$，知

$\partial N(\gamma)$是非退化的，　$\partial N(\gamma)=\nu(\nu I-Q)^{-1}\partial^2 L(\gamma)$　　(4-32)

将式(4-26)代入到式(4-32)中，且由式(4-14)，可以得到：

$$\begin{aligned}\partial N(\gamma)&=Q+E(\gamma)(\nu I-Q)=\nu E(\gamma)+(I-E(\gamma))Q\\&=\nu E(\gamma)+(I-E(\gamma))\left(\frac{I}{\mu}+GG^{T}\right)\\&=\nu E(\gamma)+\frac{I-E(\gamma)}{\mu}+(I-E(\gamma))GG^{T}\\&=C+(I-E(\gamma))GG^{T}\end{aligned}\qquad(4\text{-}33)$$

其中：$C=\nu E(\gamma)+\frac{I-E(\gamma)}{\mu}$

因为$E(\gamma)$中的所有元素都在$[0,1]$内，C是非退化的，很容易证明$(I+G^{T}PG)$是非退化的，由 Sherman-Moodbury-identity 等式，可得：

$$\begin{aligned}\partial N(\gamma)^{-1}&=C^{-1}-PG(I+G^{T}PG)^{-1}G^{T}C^{-1}\\&=\left(I-PG(I+G^{T}PG)^{-1}G^{T}\right)C^{-1}\end{aligned}\qquad(4\text{-}34)$$

其中：$P=C^{-1}(I-E(\gamma))$　　得证

由(4-34)式可以知道，可以由$(n+1)\times(n+1)$阶矩阵$(I+G^{T}PG)$来代替大型的$(l_{12}+l_3+l_3)\times(l_{12}+l_3+l_3)$阶矩阵$(I+PGG^{T})$。

求解式(4-15)(或式(4-20))得最优解$\bar{\gamma}$（或$\bar{\lambda}$）。

对于线性核的超平面决策函数可以表示为：

$$f(x)=\begin{cases}+1, & X^T S^T \bar{D}\bar{\gamma}+e^T \bar{D}\bar{\gamma}\geq \varepsilon,\\ -1, & X^T S^T \bar{D}\bar{\gamma}+e^T \bar{D}\bar{\gamma}\leq -\varepsilon,\\ 0, & \text{其他.}\end{cases} \tag{4-35}$$

其中，$\bar{\gamma}$ 是式(4-15)的解。

对于非线性核的超平面决策函数可以表示为：

$$f(x)=\begin{cases}+1, & K\left(X^T,S^T\right)\bar{D}\bar{\lambda}+e^T \bar{D}\bar{\lambda}\geq \varepsilon,\\ -1, & K\left(X^T,S^T\right)\bar{D}\bar{\lambda}+e^T \bar{D}\bar{\lambda}\leq -\varepsilon,\\ 0, & \text{其他.}\end{cases} \tag{4-36}$$

其中，$\bar{\lambda}$ 是(4-20)式的解。

4.4 小结

本章在基于最近提出的一个多类分类的新算法K-SVCR的基础上，构造了新的模型。分别就两种情况：线性核和非线性核，分别给出了其原始问题和对偶问题的表达式。通过Lagrangian隐含数，可以将其进一步转化成一个严格凸的无约束优化问题，分别得到了各自的超平面决策函数。同时，利用Sherman-Moodbury-identity等式，将算法中$(n+1)\times(n+1)$阶矩阵$\left(I+G^T PG\right)$代替了大型的$(l_{12}+l_3+l_3)\times(l_{12}+l_3+l_3)$阶矩阵$\left(I+PGG^T\right)$，对于很多多类分类问题，$n$（数据的维数）远远小于$l$（样本点的个数），大大减小了问题的规模，提高算法的效率。

第五章　求支持向量机中最优化问题的算法

5.1 引言

支持向量机将机器学习问题转化为求解最优化问题，并应用最优化理论来构造算法。最优化理论是支持向量机的重要理论基础之一，本章主要从最优化理论和方法的角度对支持向量机中的最优化问题进行研究。提出了利用快速的 Armijo 步长有限牛顿算法、解决大型问题的共轭梯度技术来求解无约束优化问题，理论和数值试验都表明有限牛顿算法、共轭梯度技术算法快速、容易实现。支持向量机的模型中含有多个参数，参数的取值直接影响分类的精确度，本文利用基本的遗传算法来解决优化问题，并有效估计未知参数。将上述算法应用于 benchmark 数据集的测试，实验表明了此算法的有效性。

5.2 利用 Armijo 步长的有限牛顿法求解

利用有限牛顿算法对式(4-15)、(4-20)进行求解。有限牛顿算法具有全局收敛和有限步终止的性质。

5.2.1 有限牛顿算法

在这一节里，对于式(4-15)、式(4-20)我们建立了牛顿法，然后证明此算法在有限步内终止。

由上章可知：　$N(\gamma):=(Q\gamma-\mu)-((Q-\nu I)\gamma-\mu)_+$

得：　$\partial N(\gamma)=Q+diag((Q-\nu I)\gamma-\mu)_*(\nu I-Q)$

因此，　$\nabla L(\gamma)=\frac{(\nu I-Q)}{\nu}N(\gamma)$ ，　$\partial^2 L(\gamma)=\frac{(\nu I-Q)}{\nu}\partial N(\gamma)$

由于$\nabla L(\gamma)$和$\partial^2 L(\gamma)$的表达式中都有$(\nu I-Q)/\nu$项，且他们都是正定，牛顿迭代可简化为：

$$\partial N(\gamma_i)(\gamma_{i+1}-\gamma_i)+N(\gamma_i)=0 \qquad (5\text{-}1)$$

下面给出牛顿算法

算法 5.2.1

Step 1　取初始点$\gamma_0,(\gamma_0\in R^{l_1+l_2+l_3})$，且置$i=0$；

Step 2　若$N(\gamma_i)=0$，则停止计算；

Step 3　计算搜索方向$d_i=-\partial N(\gamma_i)^{-1}N(\gamma_i)$；

Step 4　计算$\gamma_{i+1}=\gamma_i+\alpha_i d_i$，$\alpha_i$是*Armijo*步长，$\alpha_i=\max\left\{1,\frac{1}{2},\frac{1}{4},\cdots\right\}$

$$L(\gamma_i)-L(\gamma_i+\alpha_i d_i)\geq-\delta\alpha_i\nabla L^T(\gamma_i)d_i,\ \ \delta\in\left(0,\frac{1}{2}\right);\tag{5-2}$$

Step5　置$i=i+1$转 Step 1.

根据下面的理论可知：算法 5.2.1 是有限终止的。

命题 5.2.2:当$\nu>\|Q\|$时,由算法 5.2.1 产生的序列$\{\gamma_i\}$在有限步内最终收敛于式(4-24)的全局最优解$\bar{\gamma}$。

证明：由于式(4-24)是严格凸的无约束求最小值问题,根据文献[34]，由算法 5.2.1 产生的序列$\{\gamma_i\}$收敛于全局最优解$\bar{\gamma}$,且$N(\bar{\gamma})=0$,利用文献[35,36]相似的方法，来证明此方法有限步内终止，我们想要说明：当γ_i充分接近$\bar{\gamma}$时，取$\alpha_i=1$，由算法 5.2.1 得到的下一个迭代点$\gamma_{i+1}=\bar{\gamma}$，因此，我们需要证明下列等式:

$$0=N(\gamma_{i+1})=(Q\gamma_{i+1}-\mu)-\left((Q-\nu I)\gamma_{i+1}-\mu\right)_+\tag{5-3}$$

由（4-28)、(4-29)、(5-1）式，可得：

$$(Q\gamma_{i+1}-\mu)-\left((Q-\nu I)\gamma_i-\mu\right)_+-diag\left((Q-\nu I)\gamma_i-\mu\right)_*(Q-\nu I)(\gamma_{i+1}-\gamma_i)=0\tag{5-4}$$

为了证明式(5-3)，比较式(5-3)与式(5-4)，只须证明下列等式：

$$\left((Q-\nu I)\gamma_{i+1}-\mu\right)_+-\left((Q-\nu I)\gamma_i-\mu\right)_+-diag\left((Q-\nu I)\gamma_i-\mu\right)_*(Q-\nu I)(\gamma_{i+1}-\gamma_i)=0\tag{5-5}$$

对于每一个j，$j=1,\cdots,\ l_{12}+l_3+l_3$，我们来证明上式是成立的，因此，需要考虑以下九种可能情况的组合：

(1) 当$(Q-\nu I)_{j\cdot}\gamma_{i+1}-\mu_j>0,\ \ (Q-\nu I)_{j\cdot}\gamma_i-\mu_j>0$时，

则 $(Q-\nu I)_{j\cdot}\gamma_{i+1}-\mu_j-(Q-\nu I)_{j\cdot}\gamma_i+\mu_j-1\times(Q-\nu I)_{j\cdot}(\gamma_{i+1}-\gamma_i)=0$

因此，式（5-5）成立。

(2) 当$(Q-\nu I)_{j\cdot}\gamma_{i+1}-\mu_j>0,\ \ (Q-\nu I)_{j\cdot}\gamma_i-\mu_j=0$时，

点γ_i充分接近于点$\bar{\gamma}$时，这种情况是不可能发生的。

(3) 当$(Q-\nu I)_{j\cdot}\gamma_{i+1}-\mu_j>0,\ \ (Q-\nu I)_{j\cdot}\gamma_i-\mu_j<0$时，

点γ_i充分接近于点$\bar{\gamma}$时，这种情况是不可能发生的。

(4) 当$(Q-\nu I)_{j\cdot}\gamma_{i+1}-\mu_j=0,\ \ (Q-\nu I)_{j\cdot}\gamma_i-\mu_j>0$时，

则 $0-(Q-\nu I)_{j\cdot}\gamma_i+\mu_j-1\times(Q-\nu I)_{j\cdot}(\gamma_{i+1}-\gamma_i)=-(Q-\nu I)_{j\cdot}\gamma_{i+1}+\mu_j=0$

因此，式(5-5)成立。

(5) 当$(Q-vI)_{j.}\gamma_{i+1}-\mu_j=0$，$(Q-vI)_{j.}\gamma_i-\mu_j=0$ 时，则

$$0-0-[0,1](Q-vI)_{j.}(\gamma_{i+1}-\gamma_i)=-[0,1]\left\{(Q-vI)_{j.}\gamma_{i+1}-\mu_j-(Q-vI)_{j.}\gamma_i+\mu_j\right\}=0$$

因此，式(5-5)成立。

(6) 当$(Q-vI)_{j.}\gamma_{i+1}-\mu_j=0$，$(Q-vI)_{j.}\gamma_i-\mu_j<0$时，

则　$0-0-0\times(Q-vI)_{j.}(\gamma_{i+1}-\gamma_i)=0$

因此，式(5-5)成立。

(7) 当$(Q-vI)_{j.}\gamma_{i+1}-\mu_j<0$，$(Q-vI)_{j.}\gamma_i-\mu_j>0$时，

点γ_i充分接近于点$\bar{\gamma}$时，这种情况是不可能发生的。

(8) 当$(Q-vI)_{j.}\gamma_{i+1}-\mu_j<0$，$(Q-vI)_{j.}\gamma_i-\mu_j=0$时，

点γ_i充分接近于点$\bar{\gamma}$时，这种情况是不可能发生的。

(9) 当$(Q-vI)_{j.}\gamma_{i+1}-\mu_j<0$，$(Q-vI)_{j.}\gamma_i-\mu_j<0$时，

则　$0-0-0\times(Q-vI)_{j.}(\gamma_{i+1}-\gamma_i)=0$

因此，式(5-5)成立。

从以上证明可知：γ_i是充分接近于$\bar{\gamma}$，即$\gamma_{i+1}=\bar{\gamma}$，表明牛顿法在有限步内迭代终止的。

5.1.2 数据实验

在K类多类分类问题中，考虑其中任意的两类(θ_j, θ_k)，都可以构造一个决策函数$f_{j,k}(\cdot)$把相应的两类θ_j，θ_k分开，同时也把其余类别于这两类分开。共有 K(K-1)/2 个决策函数。对于一个新的训练点x_p，可以得到 K(K-1)/2 个输入，那么如何判断其真正的类别呢？我们采取投票的方式：当$f_{j,k}(x_p)=+1$时，就给θ_j类加上+1票，其它类为0票；当$f_{j,k}(x_p)=-1$时，就给θ_k类加上+1票，其它类为0票；当$f_{j,k}(x_p)=0$时，就给θ_j、θ_k类加上-1票，其它类为0票。当我们检验完所有的K(K-1)/2个分类器后，每一类都会获得一个总票数，最后，x_p属于获得票最多的那一类。

为了检验此算法的有效性，将上述算法用 Matlab 语言编程实现，并用三类公开数据集：Iris、Wine、Glass 来进行测试算法 5.2.1。我们使用的电脑：英特尔奔腾 IV，内存是 512MB，CPU 为 2.00GHz，在 Matlab7.0 数学软件上进行的试验.数据集的统计信息如表 5-1。

表 5-1　数据集的统计信息

数据集	训练数据个数	测试数据个数	类别数	属性个数
Iris	135	15	3	4
Wine	161	17	3	13
Glass	193	21	6	9

评价一个分类器算法的最重要的指标是分类准确率，本次试验我们采取10-折交叉确认的方法，在试验过程中，每个数据集被随机等分成 10 个子集，任取其中一个子集作为测试集，其余的并在一起作为训练集，这样的试验总共做十次，将分类正确的点数占十倍测试集中点数的百分比值作为准确率，以此来评价此算法的性能.

本次实验对于数据集：Iris，Wine 采取的是多项式核 $K(x, x')=((x\cdot x')+1)^d$，$d=1$；对于 Glass 数据集采取的 Gauss 径向基核（或 rbf 核）：$K(x, x')=\exp\left(-\|x-x'\|^2/\sigma\right)$，在算法 5.2.1 中的参数 δ 值固定，均取 $\delta=10^{-4}$，在公式(4-6)-(4-9)中的参数 ε，μ，τ，对于不同的数据集取不同的数值。对 Iris 测试集，选取的参数为 $\varepsilon=10^{-4}$，$\mu=10^{-3}$，$\tau=10$；对 Wine 测试集，选取的参数为 $\varepsilon=10^{-4}$，$\mu=10^{-3}$，$\tau=2$；对 Glass 测试集，选取的参数为 $\varepsilon=10^{-4}$，$\mu=10^{-2}$，$\tau=1$，$\sigma=2^{-6}$。为了比较算法的性能，我们将算法 5.2.1 的试验结果与“一对多”算法(1-a-a)，“一对一”算法(1-a-1)，“二次多分类支持向量机(qp-mc-sv)”算法和“线性多分类支持向量机(lp-mc-sv)”算法[17]，以及“K-SVCR”算法，“ν-K-SVCR”算法[13]的试验结果进行比较，比较结果见表 5-2。

表 5-2 算法 5.2.1 和算法 1-a-a，1-a-1，qp-mc-sv，lp-mc-sv，K-SVCR，ν-K-SVCR 的错误率

数据集	1-a-a	1-a-1	qp-mc-sv	lp-mc-sv	K-SVCR	ν-K-SVCR	算法 5.2.1
Iris	1.33	1.33	1.33	2.0	[1.93, 3.0]	1.33	1.33
Wine	5.6	5.6	3.6	10.8	[2.29, 4.29]	3.3	2.35
Glass	35.2	36.4	35.6	37.2	[30.47, 36.35]	[32.38, 36.19]	33.33

说明：

$[\cdot,\cdot]^*$：当有测试点不仅在它应该归属的类别上得到最多票，同时在其他类别上也得到相同最多票时的错误率情况。括号里的前一个数据是认为这种测试点都分类正确时的错误率，而后一个数据是认为这种测试点都分类错误时的错误率。

从表二进行对比可以看到：算法 5.2.1 在各数据集上的错误率与其他算法的错误率相当，甚至比有的算法精确度有所提高。另外，一个算法的效率也是非常关键的，对于 K 类问题，需要解决 K(K-1)/2 个二次规划问题。例如：Glass 数据集中 K=6，我们需要解决 15 个二次规划函数，因此训练速度也是一个很重要的问题，我们于 K-SVCR 算法执行的时间进行比较，所记录的时间为十次训练平均一次的执行时间（秒）。

表 5-3 记录的时间为十次训练平均一次的执行时间（秒）

数据	K-SVCR(s)	算法 5.2.1(s)
Iris	15.420	0.5673
Wine	17.811	2.7421
Glass	257.72	32.4690

从表 5-3 中可以看出新算法是非常有效的，比K-SVCR算法训练速度快得多。由于利用 Sherman-Moodbury-identity 等式时，将算法中大型的 $(l_{12}+l_3+l_3)\times(l_{12}+l_3+l_3)$ 阶矩阵 $(I+PGG^T)$ 由 $(n+1)\times(n+1)$ 阶矩阵 $(I+G^TPG)$ 代替，减小了优化规模。

5.3 利用共轭梯度法求解

共轭梯度法是用于求解大规模无约束非线性规划的一类有效算法，它是最优化中最常用的方法之一，它具有算法简单，存储需求小，易于实现等优点，十分适合于大规模优化问题。因此，本文采取共轭梯度法对式(4-15)、式(4-20)进行求解。

算法 5.3.1

step1:取初始点 $\gamma_0=0$，输入矩阵 $A=\partial^2 L(\gamma_0)$ 和向量 $b=-\nabla L(\gamma_0)$，置精度要求 $\varepsilon>0$，$i=0$，置 $r_0=b-A\cdot\gamma_0$，$\rho_0=\|r_0\|_2^2$；

step2:若 $\rho_i\le\varepsilon\|b\|_2^2$，则停止计算，$\gamma_i$ 为式(4-15)、式(4-20)的解；否则转到 step4；

step3:当 $i=0$ 时，置 $p_i=r_i$，否则，置 $p_i=r_i+\dfrac{\rho_i}{\rho_{i-1}}p_{i-1}$；

step4:置 $\omega_i=Ap_i$，$\alpha_i=\rho_i/p_i^T\omega_i$，$\gamma_{i+1}=\gamma_i+\alpha_i p_i$，$r_{i+1}=r_i-\alpha_i\omega_i$

step5:计算 $\rho_{i+1}=\|r_{i+1}\|_2^2$；

step6: 置:$i=i+1$,转到 step2。

其中 α_i 为函数 $L(\gamma)$ 在点 γ_i 的步长，p_i 为搜索方向。

算法 5.3.1 是 Polak-Ribiere-Polyak(PRP)共轭梯度法，采取的是精确线搜索。对于正定二次规划问题，采取精确线搜索的共轭梯度法具有有限终止性，即在第一个搜索方向是最速下降方向时，可知不超过 n 次迭代就可找到最优点。利用文献[37]中的结果可得采取精确线搜索的 PRP 方法对正定函数全局收敛性。(PRP)共轭梯度法产生的点列 $\{\gamma_i\}$ 收敛于 $L(\gamma)$ 的唯一极小点 $\bar{\gamma}$。

5.3.1 数据实验

为了比较算法的性能,这里我们仍使用上述的三类公开数据集。模型评价准则与上述实验相同。实验环境：英特尔奔腾 IV，内存是 512MB，CPU 为 2.00GHz，在 Matlab7.0 数学软件上进行的试验.

本次实验对于数据集：Iris，Wine 采取的是多项式核 $K(x,\ x')=\left((x\cdot x')+1\right)^d$，$d=1$；对于 Glass 数据集采取的 Gauss 径向基核（或 rbf 核）：$K(x,\ x')=\exp\left(-\|x-x'\|^2/\sigma\right)$，在公式(4-6)-(4-9)中的参数 ε，μ，τ，对于不同的数据集取不同的数值。对 Iris 测试集，选取的参数为 $\varepsilon=10^{-4}$，$\mu=10^{-3}$，$\tau=10$；对 Wine 测试集，选取的参数为

$\varepsilon=10^{-4}$，$\mu=1$，$\tau=\sqrt{2}/2$；对 Glass 测试集，选取的参数为 $\varepsilon=0.0001$，$\mu=1$，$\tau=1$，$\sigma=2^{-6}$。

表 5-4 K-SVCR 算法和牛顿法(算法 5.2.1)、共轭梯度法(算法 5.3.1)错误率和时间

	K-SVCR 算法		牛顿法(算法 5.2.1)		共轭梯度法(算法 5.3.1)	
数据	错误率	时间(秒)	错误率	时间(秒)	错误率	时间(秒)
Iris	[1.93, 3.0]	15.423s	1.33	0.5673s	0.67	0.1374s
Wine	[2.29, 4.29]	17.811s	2.35	2.7421s	2.94	0.3919s
Glass	[30.47, 36.35]	257.720s	33.33	32.4690s	31.75	9.8090s

从表中可以看出，我们的算法的对于数据集 Iris 和 Glass 的错误率比 K-SVCR 算法的错误率都要低，和数据集 Wine 的错误率基本持平，但是所训练的时间要远远少于 K-SVCR 算法。对于牛顿法(算法 5.2.1)与共轭梯度法(算法 5.3.1)相比较，共轭梯度法(算法 5.3.1)所用的时间比牛顿法(算法 5.2.1)还要少，可以看出共轭梯度法是非常有效的。

5.4 利用遗传算法求解

遗传算法提供了一种求解复杂系统优化问题的通用框架，是一种借鉴生物界自然选择和自然遗传机制的高度并行、随机、自适应搜索的方法。它不依赖于问题具体的领域，对问题的种类有很强的鲁棒性，所以广泛应用于许多学科[38,39]。

由于在有限牛顿算法、共轭梯度算法中存在很多待估的参数，参数取值的好坏，直接影响算法的精度和效率。本节就上述算法：有限牛顿法、共轭梯度法，利用遗传算法，求其目标函数（十折交叉的错误率）的最小值，同时可将有限牛顿法算法、共轭梯度法算法中的参数进行有效的估计。

5.4.1 遗传算法与传统方法的比较

求最优解或近似最优解的传统方法主要有解析法、随机法和穷举法。解析法主要包括爬山法和间接法；随机法主要包括导向随机方法和盲目随机方法；穷举法主要包括完全穷举法、回溯法、动态规划法和限界剪枝法。大多数古典的优化算法是基于一个单一的度量函数(评估函数)的梯度或较高次统计，以产生一个确定性的试验解序列；遗传算法不依赖于梯度信息，而是通过模拟自然进化过程来搜索最优解（Optimal Solution），它利用某种编码技术，作用于称为染色体的数字串，模拟由这些串组成的群体的进化过程。它通过有组织的、随机的信息交换来重新组合那些适应性好的串，生成新的串的群体[40]。

对于求解函数的最优化问题，遗传算法与一般传统方法有着本质的区别。如图 5-5 所示：

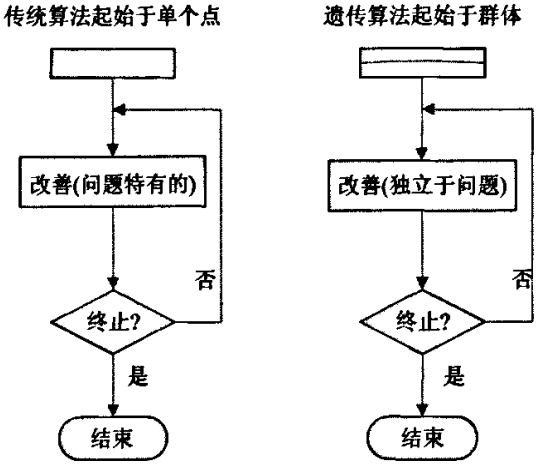


图 5-5 传统算法和遗传算法对比较

5.4.2 利用遗传算法解决优化问题

遗传算法的实现主要有六个主要因素： 参数的编码、初始群体的设定、适应度函数的设计、遗传操作、算法控制参数的设定和约束条件的处理。本节利用基本的遗传算法来解决优化问题，且同时对有限牛顿算法、共轭梯度算法中的参数进行有效的估计。除适应度函数的设计以外，其他因素都采取 Matlab 6.5 的默认值。适应度函数也称为评价函数，是根据目标函数确定的用于区分群体中个体好坏的标准，是自然选择的惟一依据。适应度函数总是非负的，因此须将目标函数映射成求最大值的形式且函数值非负。

这里我们仍然采取十折交叉确认误差来评价算法的数量标准，我们以十折交叉产生的正确率作为适应度函数。由于在一种算法中存在几个参数，我们要给这些参数在其取值范围内赋值，目前人们都是凭个人的经验来取不同的数值，算法的精度最高的认为是最优的参数值，这样要花很多的时间，那么利用遗传算法就可以对参数进行有效的估计，也节省了运算时间。

本节用两类公开数据集： Iris 来进行测试遗传算法，实验环境：英特尔奔腾 IV，内存是 512MB，CPU 为 2.00GHz，在 Matlab7.0 数学软件上进行的试验。

表 5-6 遗传算法与有限牛顿算法、共轭梯度算法的参数估计值

算法	数据集	δ	epsilon	μ	τ	正确率
Newton	Iris	0.0001	0.0001	0.001	10	98.67%
GA-Newton	Iris	0.21901	0.00013	0.16288	9.6444	99.33%
CG	Iris	——	0.0001	0.001	10	[99.37%, 99.37%]
GA-CG	Iris	——	0.01984	0.06176	6.2471	[99.37%, 100%]*

从图中可以看出，首先利用遗传算法估计出参数值，然后再用有限牛顿法、共轭梯度法求解得到的正确率比大量选参得到的精度要高一些，但是实际应用中遗传算法也存在一定的缺陷，主

要表现在算法的早熟现象、局部寻优能力差，收敛速度慢等问题。

5.5 小结

本章利用牛顿法、共轭梯度技术的有限步终止的性质来求优化问题，通过减小相应二次规划的矩阵维数，来提高算法的效率，数据试验的执行结果也表明此算法在正确率和速度上，都有很好的效果；由于在前面几个算法中，都涉及到不止一个参数最后利用基本的遗传算法来解决优化问题，且同时对有限牛顿算法、共轭梯度算法中的参数进行有效的估计，从表 5-6 中可以看到利用遗传算法对优化问题估计出参数值，比大量选参得到的错误率要低一些。但是由于遗传算法在解空间内进行充分的搜索，因此其搜索时间要长一些。

第六章 结论和展望

6.1 结论

统计学习理论和支持向量机方法之所以受到广大的重视，在于它们对有限样本情况下模式识别中的一些根本性问题进行了系统的理论研究，并且在此基础上建立了一种较好的通用学习方法。以往困扰机器学习的很多问题，如：模型选择与过学习问题、非线性和维数灾难问题、局部极小点问题等，在这里都得到了很多程度上的解决。

支持向量机在理论上和实际应用中都有很好的表型，但是由于其训练方法归结为要求解二次规划，对于大规模分类问题，都需要解决很多个二次规划问题，因此，需要占计算机很大的内存和很长的运行时间，使得支持向量机在求解大规模问题中就会产生速度很慢的缺陷。因此，建立合理的模型和研究高效的求解算法是很有意义且急需解决的问题。

（1）采用了基于线性规划的 1-v-1 三类结构支持向量的分类器，并且给出了基于预测-校正原对偶内点法的支持向量机的多类分类学习算法。将此算法用于比较庞大的多类别识别问题，如：benchmark 数据集的测试，均取得了较好的结果，算法训练速度快，而且保持良好的分类精度 。

（2）在 K-SVCR 算法结构的基础上，构造了一个新的模型。利用 Sherman-Moodbury-identity 等式减小相应优化问题的规模。继承了 K-SVCR 算法结构的优点。

（3）利用有限牛顿法、共轭梯度技术的有限步终止的性质来求解优化问题，通过 benchmark 数据集的测试，可以看出算法在训练速度和分类精度上均比 K-SVCR 算法有更好的表现。

（4）利用遗传算法的全局搜索特性得到 SVM 的最优参数值，有效提高了分类的精度和效率

6.2 展望

支持向量机虽然在分类和回归方面有较好的表现，但其作为一门新兴学科，在理论上和实际应用中都有非常大的研究空间，从理论到应用都还有很多尚未充分解决的问题，如：结构风险最小化原则中的函数子集结构的设计、支持向量机中的内积函数选择等。本文工作只是探讨了支持向量机的模型和算法。在以下几个方面还有待进一步研究。

（1）本文虽然提出了几个支持向量机算法，但其中都含有较多的参数，而参数对分类的准确性至关重要，因此，可以通过理论分析提出简单合理的无参数支持向量机模型进行研究；

（2）将本文中提出的算法可以扩大它的应用性范围，数据实验可以更加充实一些；

（3）虽然本文给出了用遗传算法来估计参数，但此算法的速度很慢，还需将此 GA-SVM 算法进行改进，有待进一步研究。

（4）对本文中提出的模型和算法有其优缺点，因为每一种方法都有其局限性。因此可以探讨对新方法理论和应用的研究。可以尝试利用其它更有效优化方法解决优化问题。

参考文献

[1] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

[2] 边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，2000，285-295

[3] 邓乃扬，田英杰. 数据挖掘中的新方法-支持向量机. 北京：科学出版社，2004

[4] 张学工. 关于统计学习理论与支持向量机. 自动化学报，2000，26(1)：32-42

[5] Cristianini N，Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge：Cambridge University Press，2000

[6] Bottou L，Cortes C，Vapnik V，etal. Comparison of classifier methods: a case study in handwriting digit recognition in: Proceedings of the International Conference on Pattern Recognition(ed.IAPR). IEEE Computer Society Press. 1994，77-82

[7] Hastie T J, Tibshirani R J. Classification by pairwise coupling. Jordan M I , Kearns M J and Solla S A (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA,1998, 10:507-513

[8] KreBel U. Pairwise classification and support vector machines in Advances in Kernel Methods: Support Vector Learning (eds.Scholkopf B, Burges C.J.C,and Smola A.J). Cambridge : MIT Press ,1999,255-268

[9] Platt J. Large margin DAGS for multiclass classification,in advances in neural information //Processing Systems 12,MIT Press,2000.547-553

[10] Allwein E L, Schapire R E and Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research,2001,1:113－141

[11] Crammer K and Singer Y. On the learnability and design of output codes for multiclass problems . Machine Learning ,2002,47:201-233

[12] Dietterich T G and Bakiri G. Solving multi-class learning problems vis error-correcting output codes. Journal of Artificial Intelligence Research, 1992, 263-286

[13] Fürnkranz J. Round robin classification. Journal of Machine Learning Research,2002, 2, 721-747

[14] Angulo C, Parra X, Catala A. K-SVCR A support vector machine for multi-class classifycati-On Neurocomputing,2003,55,57-77

[15] Zhong P, Fukushima M. A new multi-class support vector algorithm. Optimization Methods & Software, 2006, 21: 359-372

[16] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-basd vector machines. Journal of Machine Learning Research,2002,2,265-292.

[17] Bennett K P. Combining support vector and mathematical programming methods for classification in Advances in Kernel Methods: Support Vector Learning (eds. Scholkopf B, Burges C.J.C, Smola A J). Cambridge:MIT Press,1999,307-326

[18] Lee Y, Lin Y, Wahba G. Multicategory supp-ort vector machines. Computing Science and

Statistics,2001,33:498-512.

[19] Weston J,Watkins C. Multiclass support vector machine. TR CSDTR 9804, Department of Computer Science Egham,Survey TW 2003x, England, 1998

[20] Chih-Wei Hsu and Chih-Jen Lin. A Comparision of Methods for Multi-class Support Vector Machines. IEEE Transactions on Neural Networks. 2002.13(2): 415～425

[21] Rifkin R and Klautau A. In defense of one-vs-all classification. Journal of Machine Learning Research, 5 （2004），101-141。

[22] Scholkopf B，Sung K，Burges C et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Transactions On Signal Processing，1997，45：2758-2765

[23] 卢增祥，李衍达. 交互SVM学习算法及其在文本信息过滤中的应用. 清华大学学报，1999，39(7)：93-97

[24] Osuna E，Freund R，Girosi F. Training support vector machines：An application to face detection. In Proceeding of the 1997 Computer Vision and Pattern Recognition (CVPR'97)，Puerto Rico，June 1997. 130-136

[25] Angulo C, Ruiz F J, Gonalez L, Ortega J A. Multi-classification by using tri-class SVM. Neural Processing Letters, 2006, 23: 89-101.

[26] 胡清淮，魏一鸣. 线性规划及其应用. 科学出版社. 2004.

[27] 徐进东，丁晓群，谭振成，李晨. 基于非线性预报－校正内点法的电力系统无功优化研究. 电网技术，2005，29（9）：36-40.

[28] Shashua A, Levin A. Taxonomy of large margin principle algorithms for ordinal regression problems[J]. Neural Information Processing Systems, 2002, 16.

[29] Mehrotra S. On the implementation of a primal-dual interior point method. SIAM Journal Optimization,1992, 2(4):575-601

[30] Mangasarian O L, Solodov M V. Nonlinear complementarity as unconstrained and constrain-ed minimization. Mathematical Programming,Serise.B,1993,62: 277-297

[31] O.L.Mangasarian.Generalized support vector machine .Advances in Large Margin Classifiers,（2000）135-146.MIT Press,Cambrige,MA

[32] F. Facchinei, Minimization of SC^1 functions and the Maratos effect . Oper.Res.Lett.17 (1995) 131-137

[33] J.-B.Hiriart-Urruty,J.J.Strodiot.V.H.Nguyen. Generalized Hession matrix and second-order optimality condition for problems with C^{LI} data , Appl. Math.Optim.11（1984）43-56

[34] O.L.Mangasarian.Parallel gradient distribution in unconstrained optimization.SIAM Journal on Control and Optimization,1995,33,1916-1925.

[35] G.Fung , O.L. Mangasarian.Finite Newton method for Lagrangian support vector machine classification .Neurocomputing,2003,55,39-55.

[36] O.L. Mangasarian. A finite Newton method for classification .Optimization Methods and Software,2002,17,913-929.

[37] Powell M J D. Restart procedures of the conjugate gradient method. Math Program

1977,2:241-254.

[38] 周明, 孙树栋. 遗传算法原理及应用. 国防工业出版社, 1999

[39] Yu-Hsin Liu. Global maximum likelihood estimation procedure for multinomial probit model parameters. Transportation research part B 34, 2000-08

[40] 雷英杰, 张善文, 李续武等. 遗传算法工具箱及应用[M].西安电子科技大学出版社,2005, 3-9

致 谢

值此论文完成之际，衷心感谢我的导师周志坚教授对我的帮助和悉心指导，从论文的开题直至论文的最后完成阶段，周老师都给予了精心的指导。周老师严谨的治学态度、诲人不倦的师者风范以及对工作认真负责的态度是我学习的榜样。

深深的感谢钟萍老师，在我的整个论文过程中，钟老师都给了我精心指导。对我学习和生活上的无微不至的关怀使我终生难忘。她的渊博知识、严谨的治学态度是我在两年的学习生活中受益非浅，也将永远鞭策着我在未来的人生旅途中奋发努力。

衷心感谢关心、帮助过我的老师和同学们，与他们在一起渡过了难忘的两年时光，同他（她）们的讨论和交流给了我很大启发，使我受益良多。

最后，我要特别感谢我的爱人及家人为我完成学业所付出的一切，正是他们无私的奉献和全力支持使我顺利完成学业。

作者简历

袁玉萍，女，1970 年 10 月 23 日出生于黑龙江省密山市。1994 年 7 月毕业于哈尔滨师范大学数学系，获理学学士学位。1994 年 7 月至 2005 年 9 月，在黑龙江八一农垦大学理学院任教。2005 年至今，在中国农业大学理学院应用数学专业攻读硕士学位。