分类号　TP391

密　级　GK

学位代码　307
学校代码　10298
学　　号　3150116

南京林業大學
NANJING FORESTRY UNIVERSITY

硕士学位论文

论文题目：基于 L2P 范数距离度量的算法鲁棒性与稀疏性研究

作　　者：马　旭

专　　业：计算机应用技术

研究方向：模式识别

指导教师：刘应安　教　授

二〇一八年六月

诚樸雄偉
樹木樹人

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下进行的研究工作所取得的成果。尽我所知，除文中已经特别注明引用的内容和致谢的地方外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明并表示感谢。本人完全意识到本声明的法律结果由本人承担。

学位论文作者（本人签名）：马旭　　2018年 6 月20日

学位论文出版授权书

本人及导师完全同意《中国博士学位论文全文数据库出版章程》、《中国优秀硕士学位论文全文数据库出版章程》(以下简称"章程"，见 **www.cnki.net**)，愿意将本人的学位论文提交"中国学术期刊（光盘版）电子杂志社"在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表和以电子、网络形式公开出版，并同意编入 **CNKI** 《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按"章程"规定享受相关权益。

论文密级：

☑公开　□保密（___年___月至___年___月）（保密的学位论文在解密后应遵守此协议）

作者签名：马旭　　导师签名：[illegible]

2018年 6 月 20 日　　2018 年 6 月 25 日

稿酬领取通知

学位论文出版后，杂志社向被录用论文作者支付稿酬，稿酬支付标准为博士论文作者一次性获得价值 400 元人民币的"CNKI 网络数据库通用检索阅读卡"和 100 元人民币的现金稿酬；硕士论文作者一次性获得价值 300 元人民币的"CNKI 网络数据库通用检索阅读卡"和 60 元人民币现金稿酬。

请作者直接与杂志社联系领取学位论文发表证书和稿酬。联系方式如下：

联系人：吴老师　　**电话**：010 - 62791817（兼传真）、62793176、62701179（兼传真）

通讯地址：北京 清华大学邮局 84-48 信箱 采编中心　　**邮编**：100084

致　谢

三年的研究生学习生涯转瞬即逝，在这三年里，我收获颇丰，从刚开始的基础知识学习，毕业论文的选题以及最后大半年的毕业论文编纂每个环节都让我学习到了很多，也懂得了很多，同时也得到了很多老师和同学的关怀和帮助，值此之际，向他们表达我最诚挚的感谢！

首先，我要深深感谢我的导师刘应安老师。刘应安老师为人谦和，对学生认真负责。本科阶段刘老师就是我的班主任，一直苦口婆心劝我们读研读博，非常感谢刘老师给我们人生道路上方向的指引。在论文的选题，搜集资料和编写的过程中，刘老师都给了我很多宝贵的意见，一旦我论文实验过程中遇到一些瓶颈的时候，只要我去询问，他都会抽出他宝贵的时间来细心的为我分析问题的原因并运用他丰富的科研经验为我提供未来的研究方向，他严谨负责的治学作风和细心坚韧的科研态度时刻激励着我，这些都让我受益终生。

其次，我要感谢一直指导我的业巧林老师，是他让我了解并对模式识别机器学习产生了浓厚的兴趣，每当我在科研实验中遇到一些不懂得问题，他都会给我一些他的想法，这些都将为我后面的顺利完成实验提供了巨大的帮助。

最后，我要感谢南京林业大学信息科学技术学院的所有老师们，他们严谨的科学态度和对学生的无私奉献，带动了信息院整体师生的科研氛围，让我也受益匪浅。同时，我还要感谢计算机实验室的小伙伴们，他们在我心情低落的时候能够积极开导我，科研不顺的时候主动帮助我，这些小小的感动，一点一点的累积我都会铭记于心。以及我的师兄闫贺，卢棪，学弟陈向宇，黄捧等同学，他们在我学习道路上也给予了巨大的帮助。还有我的舍友张志华，渠堰墅，王宇阳和朱文等人，和他们一起生活的三年会是我人生中的一笔财富。感谢我的家人能够理解，鼓励和支持我的求学之路，没有他们默默的付出，也就没有我三年充实的研究生生活。

我相信今后的人生道路上，我都会将现在获得科研方法和为人处世的态度继续保持下去，以期更好的提升自己，为社会国家做贡献，再次感谢他们，祝他们每天开心，健康！

作者：马 旭

二〇一八年六月

摘　要

传统模式识别算法中距离度量往往是基于平方 L2 范数距离度量。而在实际应用中平方 L2 范数距离度量往往会放大噪声数据在整体数据距离中占比，导致算法的鲁棒性较差。基于平方 L2 范数距离度量鲁棒性缺陷，本文在研究分类和特征选择问题时分别采用 L2P 范数距离和 L21 范数距离度量来提高算法的鲁棒性。

孪生支持向量机（Twin Support Vector Machine）是一种特别适用于异或数据的有效分类器，通常基于平方 L2 范数距离度量来研究该分类器算法。由于平方 L2 范数距离度量容易受到异常值的影响，因此 TWSVM 需要一个更加有效的、鲁棒性强的距离度量。由于 L2P 范数距离度量比 L1 范数距离度量或平方 L2 范数距离度量能够更好地抑制异常值的影响，因此本文提出了一种基于 L2P 范数距离度量且鲁棒性强的孪生支持向量机。由于目标函数不光滑性和非凸性，基于 L2P 范数距离度量导致目标问题解决难度大。本文系统地给出一种有效的迭代算法，解决了基于 L2P 范数距离度量的目标最小化问题。理论研究证明了这个迭代算法基于 L2P 范数距离度量取代平方 L2 范数距离来改进 TWSVM 是有效的。实验表明，基于 L2P 范数距离度量的孪生支持向量机（pTWSVM）可以有效处理噪声数据，且具有较好的精度。

数据降维主要分为特征选择和特征抽取，通常它们总是被分开进行讨论研究。特征抽取旨在寻找新的特征子空间，而特征选择则致力于选择原始特征集的子集。为了获得更好的降维方法，本文提出了一种基于 L21 范数线性判别分析（LDA）的新特征选择方法，且算法具有更好的鲁棒性。然而算法求解非常具有挑战性，因为它需要同时最小化和最大化非光滑的 L21 范数项。针对这个问题，本文提出了一种迭代算法来解决基于 L21 范数距离度量的优化问题。理论研究证明了该算法的收敛性和计算效率，实验结果表明了该方法的有效性。

关键词： 鲁棒性;孪生支持向量机;特征选择;特征抽取;L2P 范数;LDA

Research on Robustness and Sparsity of L2P Norm Distance Metrics

Abstract

In traditional pattern recognition algorithms, the distance metric is often based on the square L2-norm distance. In practical applications, the squared L2-norm distance often amplifies the distance of the noise data in the overall data distance, resulting in the algorithm being not robust. Due to the non-robust defects of the square L2-norm distance, the paper improves the robustness of traitional algorithms by using the L2P-norm distance and the L21-norm distance respectively on the classification methods and feature selection methods.

Twin Support Vector Machine is an efficient classifier especially suitable for XOR data. It is usually studied based on the square L2 norm distance metric. Since the squared L2 norm distance is susceptible to outliers, TWSVM requires a more robust distance metric. In this paper, we propose a new robust twin support vector machine based on the L2P-norm distance, because the L2P-norm distance can better suppress the influence of outliers than the L1-norm distance or the squared L2-norm distance. However, the new objective function is non-smooth and non-convex, this makes it difficult to solve the objective function. As an important work of this paper, we systematically deduced an effective iterative algorithm to minimize the p-th order of L2-norm distance. Theoretical studies have proved that this iterative algorithm based on L2P-norm distance metric is effective in improving TWSVM. A large number of experiments show that the L2P-norm distance support vector machine (pTWSVM) can effectively deal with noise data and has better accuracy.

Data dimension reduction is mainly divided into feature selection and feature extraction. However, they are always discussed separately. Feature extraction aims at finding new feature subspaces, while feature selection focuses on selecting a subset of the original feature set. In order to obtain a better dimension reduction method, this paper proposes a new feature selection method based on L21-norm linear discriminant analysis (LDA). The new objective function provides better robustness. However, solving this objective function is very challenging because it requires minimizing and maximizing non-smooth L21 norm terms at the same time. To solve this problem, we propose an iterative algorithm. A series of theories proves the convergence and computational efficiency of the algorithm. The experimental results on various datasets shows the effectiveness of our new method.

Keywords： Robust; TWSVM; Feature selection; Feature extraction; L2P-norm; LDA

目 录

第一章　前　言

1.1 研究背景及意义

随着社会与科学技术的发展，越来越多的传统的行业将模式识别的相关算法应用到相关专业，如生物信息学，人脸识别，车牌识别，行人检测等等[1-3]，并且都取得了很好的效果，提高了人们的工作效率。但是在实际生活应用中，不论是图像，声音，视频等等数据，都存在少许噪声数据。而噪声数据往往会影响算法的效果，造成不必要的损失。因此在模式识别算法中如何抑制噪声数据对算法产生的影响,一直是一个值得我们探讨学习的课题。

模式识别就是通过计算机用数学的方法来对获取的数据样本进行处理与判读，来得到原始数据中的内在本质。

支持向量机（Support Vector Machine）是模式识别中的一个重要分类算法[4-7]，在机器学习等各领域中应用广泛。基于统计学习的支持向量机，由于统一了结构风险与经验风险，不仅具有很好的学习能力，还拥有很好的泛化能力。这一优点使得支持向量机算法在众多的分类算法中脱颖而出。

作为一个二分类监督学习算法，支持向量机也可以扩展到多分类算法中[8]。对于传统线性可分的情况，支持向量机通过最大间隔的原理，最终求解一个二次凸规划的问题。对于线性不可分的情况，支持向量机可以通过核函数的技巧，先将原始样本数据升维至高维空间再进行分类。理论证明，只要升高至合适的维度，样本最终都会被一个超平面可分。为了避免高维空间带来的维度灾难，支持向量机可以通过内积的形式避开对高维数据的计算。而且由于支持向量机的实际运算只需要支持向量点的参与，使得支持向量机具有极高的运算效率。

维度约减是数据预处理的一个重要步骤[9]。纬度约减目标是使用较少的特征来表示原本的高维特征，便于算法的计算运行。随着科技的发展，样本数据包含的特征也越来越多，传统的模式识别算法已经无法高效的进行计算。因此，对数据维度的预处理成了一个重要的步骤。纬度约减可以主要分为两个方面：特征选择和特征抽取。特征选择是从原本的样本特征中选取最具代表性的特征子集。特征抽取是将原市的样本数据特征迁移到一个低纬的特征空间中。而特征选择相比较特征抽取，则具有保留原始特征语义的优点。因此，学者们对特征选择进行了深入广泛的研究。

特征选择[10-12]是指从原始的数据集中选取出对后续分类等数据处理最有效果的特征子集的操作。特征选择是提高算法的学习性能，算法运行能力的一个重要的步骤，是模式识别机器学习等领域中对数据预处理的一个重要方法。特征选择主要包括产生过程，子集评价，停止准则，结果验证四个步骤。在相关领域的学习中，由于产生过程，子集评价是特征选择算法的核心，因此成为学者们重点关注的步骤。

然而不论是分类算法支持向量机还是特征选择算法，如何提高算法的泛化能力才是算法的核心问题。本文旨在针对支持向量机以及特征选择模型的不足之处，基于 L2P 范数距离来改进算法[13]，提高算法的泛化能力以及鲁棒性，为今后的实际应用提供切实有用的帮助。

1.2 国内外研究现状

由于传统的平方 L2 范数距离度量方法已经不能满足现实应用的需求，因此，我们必须寻找到能够取代平方 L2 范数距离度量的新的度量方法来未算法提供更好的鲁棒性或是稀疏性，提高算法的泛化能力。

2008 年，Nojun Kwak 对主成分分析法(principal component analysis, PCA)[14]通过 L1 范数距离进行了研究改进[15]。他所提出的 L1 范数优化方法[16]非常的直观简单，易于实现，并且还具有旋转不变性的特点。该方法原理是通过迭代算法寻找到一个局部最优解作为目标函数的解。

同样对于 L1 范数距离[17]，在 2016 年，闫贺，刘应安，业巧林等人将其应用在广义特征值支持向量机（proximal support vector machine via generalized eigenvalues， GEPSVM）上以提高分类算法的鲁棒性[18, 19]。GEPSVM 中点到平面的距离是用平方 L2 范数距离测量的，它会夸大平方运算中异常值的作用。为了优化这一点，他们提出了一个基于 L1 范数距离度量的强大而有效的 GEPSVM 分类器，称为 L1-GEPSVM。优化目标是尽量减少类内距离散度，同时最大化类内距离散度。众所周知，L1 范数距离的应用往往被认为是一种简单而有效的方法来减少异常值的影响，从而提高了模型的泛化能力和灵活性。另外，他们设计了一种有效的迭代算法来解决 L1 范数优化问题，该算法易于实现，并且在理论上保证了其收敛于逻辑上的局部最优解。因此，L1-GEPSVM 的分类性能更稳健。最后，通过在 UCI 数据集和人工数据集上的广泛实验结果证明了 L1-GEPSVM 的可行性和有效性。

2017 年，闫贺等人提出了基于 L1 范数距离的边界最小二乘支持向量机，以改进孪生支持向量机的鲁棒性，提高其算法的泛化性能。在文章中，学者们构造了最近提出的用于二分类的双边界支持向量机（TBSVM）的最小二乘版本，即通过最小二乘规划问题应用至孪生支持向量机，加以 L1 范数的应用。作为一种有效的分类工具，TBSVM 试图通过求解一对二次规划问题（QPPs）来寻找两个非平行平面，但这是非常耗时的，其时间复杂度非常高。在这里，作者们通过解决两个线性方程组而来取代解决两个 QPPs 来避免这个缺陷。此外，最小二乘法 TBSVM（LSTBSVM）的距离是通过 L2 范数来衡量的，但 L1 范数距离通常被认为是 L2 范数的一种替代方法，以提高存在异常值时的模型鲁棒性。受最小二乘孪生支持向量机（LSTWSVM），TBSVM 和 L1 范数距离的优点的启发，学者们提出了一种基于二元分类的 L1 范数距离度量的 LSTBSVM，称为 L1-LSTBSVM，它专门用于抑制异常值的负面影响以及提高大型数据集的计算效率。然后，作者设计了一个强大的迭代算法来解决 L1 范数优化问题，并且易于实现，从理论上保证了它对最优解的收敛性。

不同于 L1 范数，L21 范数由于不仅能够提供鲁棒性的优点，也能够提高算法稀疏性的特点。因此，近年来也被学者们广泛研究。

2015 年，Liang Du 等人通过 L21 范数改进了无监督的 K-means 算法。k-means 算法是数据聚类最常用的无监督方法之一。但是，标准 k-means 只能应用于原始特征空间。将 k-均值扩展到内核空间的核 k-means 可用于得到非线性结构并识别任意形状的数据。由于标准 k-means 和核 k-means 都应用平方 L2 范数残差来测量数据点与聚类中心之间的距离，因此一些异常值将导致较大的误差并支配学习函数。此外，内核方法的性能很大程度上取决于内核的选择。不幸的是，最适合特定任务的内核通常是未知的。在这篇文章中，作者首

先提出一个在特征空间中使用 L21 范数来提出了一种稳健的 k-means，然后将其扩展到内核空间。为了利用核方法的强大性，作者进一步提出了一种新的鲁棒多核 k-means（RMKKM）算法，该算法同时找到最佳聚类标签，聚类隶属度和多个内核的最优组合。

2016 年，Feiping Nie 等人提出了基于 L21 范数距离的 PCA 算法[20-22]。由于其主成分分析法中特征分解的高计算复杂度，难以将 PCA 应用于具有高维度的大规模数据。同时，由于基于平方 L2 范数的规划，目标函数对数据异常值很敏感。上文提到基于 L1 范数最大化的 PCA 方法被提出用于高效计算并对异常值具有鲁棒性。然而，这项工作使用了一个贪婪算法来解决特征向量。此外，基于 L1 范数最大化的目标可能不是正确的鲁棒的 PCA 公式，因为它失去了与最小化数据重构误差的理论联系，这是 PCA 最重要的理论基础和目标之一。在这篇文章中，作者建议最大化基于 L21 范数的鲁棒 PCA 目标函数，这在理论上与最小化重构误差理论相关。更重要的是，作者提出了高效的非贪婪优化算法来解决其目标函数。现实世界数据集上的实验结果表明了所提出的主成分分析方法的有效性。

同样通过 L21 范数，D kong 等人利用其改进了非负矩阵分解算法以提高算法的鲁棒性。非负矩阵分解（NMF）广泛用于数据挖掘和机器学习领域。 但是，在实际应用中许多数据包含噪音和异常值。因此需要一个更加强大的 NMF 改进算法。 在这篇文章中，作者提出了一个使用 L21 范数损失函数的 NMF 的鲁棒公式，还推导了一个严格收敛分析的计算算法。这个鲁棒的 NMF 方法有以下几个优点：（1）可以更好的处理噪声和异常值; （2）具有非常有效和简洁的目标函数; （3）与传统的标准 NMF 相比，改进的基于 L21 范数的 NMF 算法与其具有几乎相同的计算成本，因此可以用于更多现实世界的应用任务。

相比较 L21 范数距离度量，Hua Wang 等人提出了一种基于 L2P 范数距离的局部保留投影算法（Locality preserving Projection， LPP）[23]。局部保持投影（LPP）是一种基于流形学习的有效降维方法[24-27]，该方法定义在投影子空间中图的加权平方 L2 范数距离上。由于平方的 L2 范数距离容易对异常值敏感，所以需要一个更加鲁棒的 LPP 方法。在该文章中，基于现有关于 L1 范数或非平方 L2 范数距离提高统计学习模型的鲁棒性的基础上，作者提出了一个 L2P 范数的鲁棒的 LPP（rLPP）算法来最小化 l2 范数的 p 阶距离，它可以更好地容忍野值数据，因为它可以比 L1 范数和非平方 L2 范数距离更好的抑制野值噪声的影响。

1.3 传统算法的不足

如上文所述，在传统的模式识别算法中，无论是分类算法还是特征选择算法，算法的目标函数为了便于求解，都是基于平方 L2 范数距离。平方 L2 范数距离能够为目标函数提供很好的凸性，但是同时也更容易受到噪声数据以及野值的影响，造成算法的不鲁棒的特性。由于基于平方 L2 范数距离的算法不鲁棒，导致传统算法在实际应用中往往不能得到令人满意的结果。

针对上述研究的不足，本文基于 L2P 范数距离度量进行算法研究，着重对分类算法中的 TWSVM 算法以及特征选择算法中的 DFS 算法进行研究。

1.4 本文主要研究创新工作

本文针对传统算法的不足，提出了分类算法 TWSVM 和特征选择算法 DFS 的改进，并

进行了系列研究，主要研究工作如下：

（1）由于传统算法的学习函数都是基于平方 L2 范数距离，基于平方 L2 范数距离度量算法容易受到噪声数据以及离群数据的影响，使得算法鲁棒性差，因此，本文从算法的公式推导中，提出基于 L2P 范数（包括 L21 范数）的算法改进。

（2）改进后的目标函数是一个非凸的目标函数，因此非常难于直接求解。在吸取国内外学者研究经验的基础上，本文提出了一种迭代算法来求解目标函数。

（3）针对每一个迭代算法，本文都对其收敛性以及时间复杂度在理论上进行了研究。理论证明迭代算法严格收敛，并且时间复杂度可接受。

（4）通过大量的实验，本文从不同方面对算法的精确度，运算时间，收敛效率等性能进行了实验，实验表明本所所提出的算法相比较目前的相关工作都表现出了更好的性能。

1.5 本文内容安排

第一章，介绍了本文的研究背景、研究的目的和意义，系统分析了国内外模式识别算法研究现状以及算法研究的不足，给出了本文研究的创新及研究内容框架。

第二章，介绍了支持向量机的一些发展，从寻找单个分类面到寻找两个分类面，重点对算法模型进行了分析。通过对算法的分析，发现其中的不足之处，为后续对其进行算法改进奠定基础。

第三章，介绍了特征选择的相关工作，从降维工程的分类到特征选择和特征抽取的区别；介绍了一些降维工程中常用的算法，分析了当前特征选择中的不足之处，为后续的改进指明了方向。

第四章，重点展开对 L2P 范数距离 TWSVM 工作的介绍。从理论上推导出基于 L2P 范数距离度量的 TWSVM，并设计算法求解这个非凸的目标。理论研究证明了该算法的收敛性以及时间复杂度，从而说明了算法的可行性。实验结果表明相比其它分类算法，本文所提出的分类算法确实具有较好的性能。

第五章，详细介绍了基于 L21 范数距离的特征选择。它将线性判别分析特征抽取工作通过 L21 范数距离融合到特征抽取中，使得新的算法具有更好的稀疏性和可解释性。由于目标函数的非凸性，本文设计了一个有效的迭代算法，使得能够在极少次数的迭代过程后求得目标函数的解。理论研究证明了算法严格收敛且时间复杂度较低，实验结果表明相比较目前流行的特征选择算法，本文提出的算法具有更好的鲁棒性。

第六章，简要概括了本文的主要工作，并对后续工作进行了展望。

第二章　支持向量机概述

支持向量机是模式识别中一个重要的分类算法，它通过寻找最大间隔，同时引入软间隔的概念，利用拉格朗日函数进行求解计算。本章节将重点介绍传统支持向量机，广义特征值支持向量机以及孪生支持向量机。

2.1 传统支持向量机

1995 年，Vaprink 根据统计学习理论提出如果数据服从独立同分布原则，要使得机器学习得到输出与实际输出差距尽可能小，算法应该遵循结构风险最小化而不是经验风险最小化的原则[28-30]。依据这一理论，Vaprink 提出了支持向量机。

假设有包含n个点的数据集$\{x_1，x_2，\cdots，x_n\}$，该数据集可以记为$X \in R^{n\times d}$，其中n为样本个数，d为样本维度。如果第i个点x_i属于正类，那么标记该点为$+1$，如果其为负类，那么标记该点为-1。第i个点x_i的标记可以表示为$y_i \in \{+1,-1\}$。y_i为第i个样本的标签。支持向量机寻找的不是一个能分类的平面，而是基于最大间隔原理来寻找最优的分类平面。这个平面的方程可以表示为

$$\boldsymbol{w}^T\boldsymbol{x}+\boldsymbol{b}=0 \tag{2-1}$$

其中w表示平面。为了使得平面到两类样本的距离最大化，可以得到如下的目标函数

$$\max\frac{1}{||\boldsymbol{w}||} \tag{2-2}$$

求公式(2-2)的最大值问题可以转化为如下的最小值问题

$$\min||\boldsymbol{w}|| \tag{2-3}$$

然而公式(2-3)中默认假定两类样本是线性可分，既能够找到一个平面能够完全的将数据区分。但是在现实数据中，两类样本往往是无法用一个平面完全分开。因此，支持向量机引入了松弛变量的概念。引入松弛变量的支持向量机公式如下：

$$\begin{aligned}&\min_w\frac{1}{2}||w||^2+C\sum_{i=1}^{n}\zeta_i\\&s.t.\ y_i\left(w^Tx_i+b\right)\geq 1-\zeta_i,i=1,\cdots,n\\&\quad\zeta_i\geq 0,i=1,\cdots,n\end{aligned} \tag{2-4}$$

其中ζ_i为第i个样本的松弛变量，C为平衡系数。对公式（2-4）进行到拉格朗日函数运算，可得如下公式：

$$L(w,b,\zeta,\alpha,\beta)=\frac{1}{2}||w||^2+C\sum_{i=1}^{n}\zeta_i-\sum_{i=1}^{n}\alpha_i\left(y_i\left(w^Tx_i+b\right)-1+\zeta_i\right)-\sum_{i=1}^{n}\beta_i\zeta_i \tag{2-5}$$

将拉格朗日函数对w,b,ζ,α,β分别求偏导，可以得到如下公式

$$\frac{\partial L}{\partial \boldsymbol{w}}=0\Rightarrow \boldsymbol{w}=\sum_{i=1}^{n}\alpha_i y_i x_i \tag{2-6}$$

$$\frac{\partial L}{\partial \boldsymbol{b}}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0 \tag{2-7}$$

$$\frac{\partial L}{\partial \zeta_i}=0\Rightarrow C-\alpha_i-\beta_i=0,\ i=1,\cdots,n \tag{2-8}$$

将公式(2-6)(2-7)(2-8)带入拉格朗日函数，可以得到整个对偶目标函数

$$\begin{aligned}&\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle\\&s.t.\quad 0\le\alpha_i\le C,i=1,\cdots,n\\&\qquad\sum_{i=1}^{n}\alpha_i y_i=0\end{aligned} \tag{2-9}$$

然而大部分的时候，数据并非线性可分,这时候能够区分数据的超平面就不存在。对于非线性的数据，SVM 通过核函数的方法，将数据映射到高维空间中，来解决在低维空间中不可分的问题。常见的核函数包括多项式核函数，高斯核函数，和线性核函数。通过核函数映射的目标问题可以写成如下形式

$$\begin{aligned}&\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k\left(x_i,x_j\right)\\&s.t.\quad 0\le\alpha_i\le C,i=1,\cdots,n\\&\qquad\sum_{i=1}^{n}\alpha_i y_i=0\end{aligned} \tag{2-10}$$

其中$k(a,b)$表示核函数的映射。

2.2 广义特征值支持向量机

2005 年，Olvi L.Mangasarian 等人提出了基于传统支持向量机改进的广义特征值支持向量机(Generalized Eigenvalues Proximal Support Vector Machine,GEPSVM)[6, 31-33]。不同于传统的支持向量机寻找一个分类平面，GEPSVM 旨在寻找两个不平行的分类平面，并且每一个分类平面离相应的样本最近，离相对的样本最远[34]。求解这两个分类平面只需要求解两个简单的广义特征值问题，因此相比较传统支持向量机求解二次图规划问题，GEPSVM 拥有更低的时间复杂度。

在传统的线性支持向量机中，对于异或问题(XOR problem)，传统的线性支持向量机并不能有效的区分两个样本。而 GEPSVM 通过两个分类平面，则很好的解决了这个问题。对于一个严格的异或样本，GEPSVM 能够达到 100%的分类精度，而传统线性支持向量机只能有一半的分类精度。

假设正类样本A $\in R^{n1\times d}$对应的平面法向量为w_1，偏差为b_1，负类样本B $\in R^{n2\times d}$对应的平面法向量为w_2，偏差为b_2。对于平面 1，要求平面距离正类样本尽可能的近，距离负类样本尽可能的远；对于平面 2，要求平面距离负类样本尽可能的近，距离正类样本尽可能的远；这可以引入如下的优化目标：

$$\min_{w_1,b_1\neq 0}\frac{\left\|Aw_1+eb_1\right\|^2/\left\|\begin{bmatrix}w_1\\b_1\end{bmatrix}\right\|^2}{\left\|Bw_1+eb_1\right\|^2/\left\|\begin{bmatrix}w_1\\b_1\end{bmatrix}\right\|^2} \tag{2-11}$$

$$\min_{w_2,b_2\neq 0}\frac{\left\|Bw_2+eb_2\right\|^2/\left\|\begin{bmatrix}w_2\\b_2\end{bmatrix}\right\|^2}{\left\|Aw_2+eb_2\right\|^2/\left\|\begin{bmatrix}w_2\\b_2\end{bmatrix}\right\|^2} \tag{2-12}$$

公式(2-11)(2-12)即 GEPSVM 需要求解的两个平面的优化目标函数。对于平面 1 的优化目标，公式可以简写为

$$\min_{w_1,b_1\neq 0}\frac{\left\|Aw_1+eb_1\right\|^2}{\left\|Bw_1+eb_1\right\|^2} \tag{2-13}$$

为了防止过拟合问题，这里给 GEPSVM 的目标加入一个 L2 正则项

$$\min_{w_1,b_1\neq 0}\frac{\left\|Aw_1+eb_1\right\|^2+\delta\left\|\begin{bmatrix}w_1\\b_1\end{bmatrix}\right\|^2}{\left\|Bw_1+eb_1\right\|^2} \tag{2-14}$$

其中δ是一个非负的参数。公式(2-14)的几何解释即正类样本离目标平面尽可能近，负类样本离目标平面尽可能的远。对于另一个平面，可以通过同样的方式获得。我们定义

$$\begin{aligned} G&=\begin{bmatrix}\mathbf{A} & e\end{bmatrix}^T\begin{bmatrix}\mathbf{A} & e\end{bmatrix}+\delta\mathbf{I}\\ H&=\begin{bmatrix}\mathbf{B} & e\end{bmatrix}^T\begin{bmatrix}\mathbf{B} & e\end{bmatrix}\end{aligned} \tag{2-15}$$

$$z=\begin{bmatrix}w_1\\b_1\end{bmatrix} \tag{2-16}$$

其中e为维度合适的单位列向量，I为维度合适的单位对角阵。为了方便，公式(2-14)可以简写为

$$\min_{z\neq 0}\frac{z^T\mathbf{G}z}{z^T\mathbf{H}z} \tag{2-17}$$

公式(17)就是瑞利商问题。求解公式(17)等价于求解以下问题

$$\mathbf{G}z=\lambda\mathbf{H}z,\ z\neq\mathbf{0} \tag{2-18}$$

求解的目标z即公式(2-18)的特征值问题中的最小特征值所对应的特征向量。同理，另

一个平面也可以通过同样的方式求解。

GEPSVM 每一个平面都只需求解一个广义特征值问题，因此 GEPSVM 的效率相比较传统SVM得到了很大的提高。而且由于GEPSVM求解两个不平行的平面，这使得GEPSVM相比较传统 SVM 在交叉数据上具有更加明显的优势。

2.3 孪生支持向量机

与 GEPSVM 相似，孪生支持向量机(Twin support vector machine) 也是寻找两个不平行的分类平面[35-37]，但是寻找这两个分类平面的方法完全不同[30]。GEPSVM 求解的是一对广义特征值问题，而 TWSVM 求解的是一对凸二次规划问题。在传统支持向量机中，所有的数据点都参与凸二次规划问题的求解。而在 TWSVM 中，对于每一个平面，只有相应的类别的数据点参与问题求解，而其他的数据点存在于约束中。由于 TWSVM 求解的是较小规模的凸二次规划，这使得 TWSVM 的运算效率能够比传统的支持向量机高出许多。

假设有n个数据点可以表示为一个矩阵$\mathrm{X}=\{\mathrm{A}_1,\mathrm{A}_2,\cdots\mathrm{A}_n\}$，A属于一个$d$维的实值空间$\mathrm{X}\in R^{n\times d}$.同样，$y_i\in\{+1,-1\}$表示对应的第i个样本属于正类或事负类。假设正类样本为A,负类样本为B，那么可以通过求解以下的两个问题来得到分类平面法向量w^1,w^2和偏差b^1,b^2:

$$\begin{aligned}(\mathrm{TWSVM1})\quad &\min_{w^1,b^1,q}\frac{1}{2}\left\|\mathbf{A}w^1+e_2b^1\right\|^2+c_1e_2^Tq\\ &\text{s.t.}\ -\left(\mathbf{B}w^1+e_2b^1\right)+q\geq e_2,\ q\geq 0\end{aligned}\tag{2-19}$$

$$\begin{aligned}(\mathrm{TWSVM2})\quad &\min_{w^2,b^2,q}\frac{1}{2}\left\|\mathbf{B}w^2+e_2b^2\right\|^2+c_2e_1^Tq\\ &\text{s.t.}\ -\left(\mathbf{A}w^2+e_1b^2\right)+q\geq e_1,\ q\geq 0\end{aligned}\tag{2-20}$$

其中$c_1,c_2>0$是非负的平衡参数，e_1,e_2是维度合适的单位列向量，q是松弛变量。

TWSVM 寻找的两个平面，每一个平面要求离相应类别的数据点尽可能的近。因此，最小化公式(2-19)和(2-20)能够使得相应的平面到相应的数据点距离最小化。同时，公式(2-19)和(2-20)要求所求的平面与对立的数据点要有一个函数间隔最小为 1 的距离。同时，一系列的松弛变量使得目标函数允许部分点存在错分，而目标函数中第二项就是松弛变量的总和。

对于求解 TWSVM，我们需要像传统 SVM 一样求解一个凸二次规划问题。公式(2-19)对应的拉格朗日函数可以表达为如下形式：

$$\underset{w^1,b^1,q,\alpha,\beta}{L}=\frac{1}{2}\left(\mathbf{A}w^1+e_1b^1\right)^T\left(\mathbf{A}w^1+e_1b^1\right)+c_1e_2^Tq-\alpha^T\left(-\left(\mathbf{B}w^1+e_2b^1\right)+q-e_2\right)-\beta^Tq\tag{2-21}$$

其中α,β为拉格朗日乘子向量。通过 KKT 条件和对每一个变量求导，可以得到如下的公式：

$$\mathbf{A}^T\left(\mathbf{A}w^1+e_1b^1\right)+\mathbf{B}^T\alpha=0\tag{2-22}$$

$$e_1^T\left(\mathbf{A}w^1+e_1b^1\right)+e_2^T\alpha=0\tag{2-23}$$

$$c_1 e_2 - \alpha - \beta = 0 \tag{2-24}$$

$$-\left(\mathbf{B}w^1 + e2b^1\right) + q \geq e_2 \tag{2-25}$$

$$\alpha^T\left(-\left(\mathbf{B}w^1 + e_2 b^1\right) + q - e_2\right) = 0 \tag{2-26}$$

$$\beta^T q = 0 \tag{2-27}$$

$$\alpha \geq 0,\ \beta \geq 0,\ \mathrm{q} \geq 0 \tag{2-28}$$

结合(2-24)和(2-28)，可以得到

$$0 \leq \alpha \leq c_1 \tag{2-29}$$

接下来，通过将(2-22)与(2-23)相加可以得到

$$\left[\mathbf{A}^T\ \ e_1^T\right]\left[\mathbf{A}\ \ e_1\right]\left[w^1, b^1\right]^T + \left[\mathbf{B}^T\ \ e_2^T\right]\alpha = 0 \tag{2-30}$$

定义

$$\mathbf{H} = \left[\mathbf{A}\ \ e_1\right],\ \ \mathbf{G} = \left[\mathbf{B}\ \ e_2\right],\ \ u = \left[w^1, b^1\right]^T \tag{2-31}$$

通过这些定义，可以重写公式(2-30)为

$$\mathbf{H}^T\mathbf{H}u + \mathbf{G}^T\alpha = 0$$

$$u = -\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{G}^T\alpha \tag{2-32}$$

通过公式(2-32)可以发现，寻求的第一个分类平面的法向量与偏差可以表达为样本与拉格朗日乘子积的形式。由于需要对H^TH进行求逆运算，虽然H^TH是一个半正定矩阵，但是仍有可能在某些情况下奇异。因此，添加一个正则项$\epsilon \mathrm{I}, \epsilon > 0$，其中I是一个任意维度的单位对角阵。因此，修正过后的公式(32)可以重写为

$$u = -\left(\mathbf{H}^T\mathbf{H} + \varepsilon\mathbf{I}\right)^{-1}\mathbf{G}^T\alpha \tag{2-33}$$

但是在后续的工作中为了方便，仍然使用公式(32)来进行计算。如果有必要可以用公式(33)来代替公式(32)。

通过拉格朗日公式(21)和上述的 KKT 条件，可以得到第一个 TWSVM 平面的对偶形式如下：

$$\begin{aligned} &\max_{\alpha}\ e_2^T\alpha - \frac{1}{2}\alpha^T\mathbf{G}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{G}^T\alpha \\ &s.t.\quad 0 \leq \alpha \leq c_1 \end{aligned} \tag{2-34}$$

通过 SMO 算法[38-40]利用凸二次规划求解公式(34)，可以得到最优的α值，带入公式(32)或者公式(33)可以得到所求平面 1 的法向量w^1和偏差b^1。同理，可以通过同样的方法来得到平面 2 的法向量w^2和偏差b^2.

一旦两个平面确定，便可以确定一个新的点的类别。假设一个新的点$x \in R^d$,那么可以

通过这个点到两个平面的距离来判断新的点的类别：

$$\min_{l=1,2}\left|x^T w^l + b^l\right| \tag{2-35}$$

如果x点离平面 1 近并且离平面 2 远，那么x点属于平面 1 对应的正类样本。如果x点离平面 1 远离平面 2 近，那么x点属于平面 2 对应的负类样本。

2.4 本章小结

本章简单介绍了支持向量机的几种改进，包括传统支持向量机，广义特征值支持向量机，孪生支持向量机。传统支持向量机秉持最大间隔的思想，求解一个凸二次规划的问题。对于传统支持向量机，原目标问题已经可以求解，但是由于时间复杂度过高，通过拉格朗日函数采用对偶形式来求解。GEPSVM 寻求两个分类平面，要求每一个平面离相应的类别数据尽可能近，离对应类别数据尽可能远。通过求解两个对立的广义特征值问题，GEPSVM 能够得到两个不平行的分类平面。GEPSVM 相比较传统的支持向量机，由于是求解广义特征值问题而非凸二次规划问题，运算时间有了极大的提高。并且，GEPSVM 很好的解决了异或问题。TWSVM 同样是求解两个分平行分类平面，但是从根本上与 GEPSVM 不同。TWSVM 是通过求解两个较小规模的凸二次规划问题来得到平面。由于每一个问题的规模较小，因此 TWSVM 的时间复杂度约为传统支持向量机的时间复杂度的四分之一。

但是，上述的算法都是基于平方 L2 范数距离(欧式距离)进行求解。平方 L2 范数具有凸函数的性质，因此便于求解目标函数。但是由于样本中往往包含一些噪声和野值，平方 L2 范数则往往会放大野值的影响，使得算法不具有鲁棒性。为了缓和野值造成的 SVM 算法的鲁棒性缺陷问题，在下一章中，将通过 L2P 范数距离来重新定义目标函数，提高算法的鲁棒性与算法的泛化能力。

第三章　特征选择概述

特征选择是特征工程的一种方法，ji 即从原始样本空间中寻找到最有利于后续分类工作的的特征子集。本章节详细介绍特征选择的相关算法以及理论分析。

3.1 特征选择与特征提取

特征是决定样本之间的相似性和区别性的重要属性，因此特征成为了模式识别分类器设计的关键[2, 46, 47]。一个样本数据往往包含不同的数据特征，有些特征能够对分类器起到积极的正作用，而有些特征则对分类器分类毫无帮助，甚至会影响分类器的分类性能。如何找到合适的特征来代表样本数据是模式识别的一个核心问题。

然而，在实际问题中，常常无法找到那些最具有代表性的特征，或者受限于各种条件限制而无法对其进行测量。这使得样本数据的特征工程任务复杂化。在模式识别中，样本的特征主要包括三大基本特征：物理，结构和数字特征。物理特征和结构特征易于为人所感知，但是往往会难于定量的描述，因此，在模式识别中，这两类特征并不是很好的选择。而数字特征则往往易于机器学习的描述和判别，可以通过统计，概率等方式来进行分类器的分类学习。

在一般情况下，人们普遍认为增加特征的维度(特征数目)将有助于分类器算法的分类进度提高。但是随着科技的发展，维度已经不再是限制分类器性能的条件。相反，在实际应用中，过高的维度反而会对分类器算法产生负效应。首先，过高的维度会导致算法的时间复杂度过高，大大提高了算法的运算成本。其次，过高的维度需要更大的存储空间。最后，过高的维度甚至会降低分类器的分类新精度，因为部分特征是冗余的甚至是噪声特征。基于以上考虑，对于模式识别算法，降低特征维数，选出最有代表性的特征是设计有效分类器的重要一步。

特征选择和特征抽取是模式识别中数据降维[48-50]的两种不同方法。特征抽取后的特征是原本特征在一个映射空间中形成的新的特征集。特征选择是选择原本特征中最具有代表性的特征子集。然而，特征选择和特征抽取有许多的相同点。首先，这两者能够达到的效果是相同的，即减少原样本的维度并最大可能保留样本的内在本质。其次，两者都是可以通过学习函数得到，而不是随意抽取或选择。但是特征选择和特征抽取所采用的方式却大不相同。特征抽取方法主要是通过属性之间的关系来得到新的特征，如组合不同的特征属性得到新的特征，但是这样却改变了原始的特征空间。而特征选择是从原始的特征空间中，通过某种评价函数，选择最具有代表性的特征子集，而没有改变其原始的特征空间。特征选择和特征抽取的基本任务是从原始的特征中获取最有效的信息。

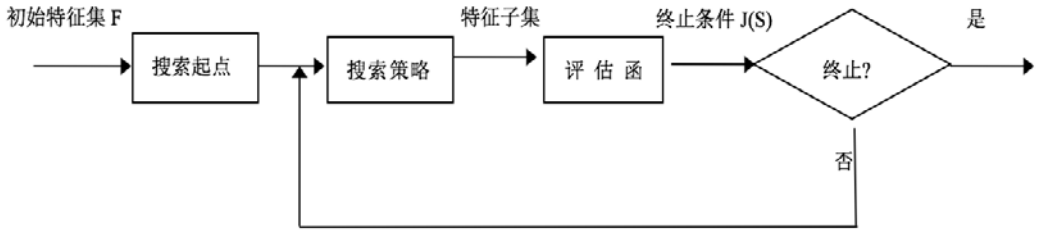
目前，模式识别中还没有特征提取和特征抽取的一般方法，因为降维工程一般是面向问题的，不具有普适性，很难有一个统一的比较与评价。

3.2 特征选择分类

特征选择[51-53]又称特征子集选择或者属性选择，指从全部特征中选择出一个特征子集，能够使得后续构造分类器模型性能更加优秀。

在模式识别的实际应用中，特征数量往往较多，其中可能往往包含与分类无关的特征，或者有噪声的特征，特征之间往往也会存在相互依赖或者冗余关系。特征选择致力于剔除与分类无关或者冗余的特征，保留最具有代表性的特征子集，从而提高后续分类算法精度，减少分类算法的运行时间，节省内存空间开销。

对于一般的情况下，特征选择过程可以分为四个部分，包括初始子集设定，搜索策略，子集评价和终止条件。



通过特征选择的方式，可以将特征选择分为三类：过滤式(Filter)，包装式(Wrapper)和嵌入式(Embedded)。

过滤式的特征选择特征子集搜索与评价模型的训练过程并不重合，往往将过滤得到的特征用于训练中。换言之，即现对输入数据集进行特征选择，然后在训练学习分类器，使得 特征选择的过程和后续的学习方法无关。这就相当于先用特征选择方法对原始特征进行过滤，在用过滤后的特征来进行模型训练。

包裹式特征选择与过滤式特征选择不同，包裹式的特征选择直接把最终的学习器的分类精度当作特征子集好坏的评价标准。包裹式的特征选择的目标就是为了给学习器选择最有利于其性能的特征子集。

从传统意义上而言，由于包裹式的特征选择方法直接依附于给定的学习器而进行优化，往往会拥有一个很好的分类性能，然而由于包裹式特征选择在特征选择过程中需要多次的训练学习器，因此包裹式的特征选择的计算开销相比较过滤式的特征选择往往会大上很多。

结合于包裹式特征选择与过滤式特征选择，嵌入式特征选择将特征选择过程与学习器训练的过程融为一体，这两个步骤在同一个优化过程中完成，即在学习器训练的过程中自动的进行了特征选择。

3.3 纬度约减算法

纬度约减算法主要包括特征选择算法和特征抽取方法。本小节主要介绍一些常见的特征抽取方法如主成分分析法(Principal Component Analysis, PCA)，线性判别分析法(Linear Discriminant Analysis, LDA),还有一些常见的特征选择方法如决策树等。

3.3.1 主成分分析法

主成分分析法(PCA)是一种最常用的维度约减方法[54]，它的原理是最大可分性，即样本点在这个超平面上的投影点尽可能的分开。假定数据样本 X 包含了数据点x_i,那么样本点x_i在超平面W上的投影点即$W^T x_i$,如果要使样本点的投影尽可能的分开，那么则应该使得样本的投影后的数据点的方差尽可能的大，即离样本中心点$\bar{x_i}$最可能的分散。于是优化目标

可以写为：

$$\max \sum_{i=1}^{n} \left\| \mathbf{W}^T x_i - \mathbf{W}^T \overline{x_i} \right\|_2^2 \quad s.t. \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \tag{3-1}$$

对于公式(4-1)，我们将投影向量提出，可以得到

$$\max \mathbf{W}^T \sum_{i=1}^{n} \left(x_i - \overline{x_i} \right)^T \left(x_i - \overline{x_i} \right) \mathbf{W} \quad s.t. \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \tag{3-2}$$

定义求和项为全局散度矩阵S_t，那么公式(3-2)可以简写为

$$\max \mathbf{W}^T \mathbf{S}_t \mathbf{W} \quad s.t. \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \tag{3-3}$$

对公式(3-3)使用拉格朗日橙子法可得

$$\mathbf{S}_t \mathbf{W} = \lambda \mathbf{W} \tag{3-4}$$

于是，只需要对散度矩阵S_t进行特征值分解，将所有的特征值按照降序排序，再取前n个特征值对应的特征向量，组合成

$$\mathbf{W} = \left(w_1, w_2, \cdots w_n \right) \tag{3-5}$$

即主成分分析的解。

传统的降维方法降维后的维数n往往是由用户事先指定的。但是对于主成分分析法，可以从重构的角度来选取降维后的维数，即设置一个重构阈值t。

$$t = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{3-6}$$

其中λ_i是第i个特征值。

3.3.2 线性判别分析法

线性判别分析(LDA)是一种经典的降维算法[41, 55, 56]，其核心思想非常朴素，即找到一个投影平面，使得相同类别的点距离尽可能的尽，不同类别的点距离尽可能的远。假设给定数据集合D $= \{(x_i, y_i)\}_{i=1}^{m}$,令$X_i$，$\mu_i$分别表示对应类别的样本数据点集合以及对应样本类别的平均值，即中心点。

欲使得同类样本的投影点距离尽可能的小，即使得同类样本投影点的协方差尽可能的小；而欲使得异类样本的投影点距离尽可能的大，可以让类中心的距离尽可能的大。我们同时考虑这两个方面，则可以得到如下的最大化优化目标：

$$\max \frac{\mathbf{W}^T\left(\sum_{i=1}^{c} n_i\left(\mu_i-\mu\right)^T\left(\mu_i-\mu\right)\right)\mathbf{W}}{\mathbf{W}^T\left(\sum_{i=1}^{c}\sum_{j}^{c_i}\left(x_k-\mu_i\right)^T\left(x_k-\mu_i\right)\right)\mathbf{W}} \tag{3-7}$$

我们定义类内散度矩阵为S_w,内间散度矩阵为S_b,则公式(4-7)可以简化为

$$\max \frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_w\mathbf{W}} \tag{3-8}$$

这就是 LDA 最大化目标函数，即内间散度矩阵和类内散度矩阵的广义瑞丽商(Generalized Rayleigh Quotient)。

对于公式(3-8),注意到分子分母都是关于W的二次项。因此，公式的解与W的长度无关，只与其方向有关。为了便于求解，将公式(4-8)转换成以下形式：

$$\begin{aligned} &\min \ -\mathbf{W}^T\mathbf{S}_b\mathbf{W} \\ &s.t.\ \mathbf{W}^T\mathbf{S}_w\mathbf{W}=\mathbf{I} \end{aligned} \tag{3-9}$$

对公式(3-9)使用拉格朗日函数，可以得到如下公式：

$$\mathbf{S}_b\mathbf{W}=\lambda\mathbf{S}_w\mathbf{W} \tag{3-10}$$

可以求得对应的最小特征值对应的特征向量组合成求解的W。只需将原始数据样本投影到已求解的低维超平面中，即可得到降维后的数据。

3.3.3 决策树

决策树[57-59]是在已知各种情况发生概率的基础上，通过构建树模型，实现取经线值大于等于零的概率的决策方法。在特征选择中，决策树的构建过程是非常重要的一步，也是实现特征选择的主要步骤。对于决策树而言，选择的特征可以将原始特征空间划分成块状的特征子空间。 特征选择是要选取出对原始输入样本数据最具有分类能力的特征，这样可以帮助决策树的大大提高其学习效率。如果决策树选择的某一个特征在进行分类后的结果与随机分类的结果没有太大差异，那么认为这个特征是没有分类能力的，也就是说这样的特征可以丢弃。对于决策树而言，信息增益和信息增益比是常用的选择特征的准则。

信息增益是熵的一种增益变化情况。熵是无序度的度量，在信息论和统计中，熵表示随机变量不确定性的度量。假设X是一个取有限值的离散型随机变量，那么对于此随机变量的熵的定义如下：

$$H(p)=-\sum_{i=1}^{n} p_i \log p_i \tag{3-11}$$

从公式(3-11)中可以发现,熵只依赖于样本的分布,而与样本的取值没有关系。熵越大，随机变量的不确定性就越大。

信息增益表示得知特征 X 的信息而使得类 Y 的信息不确定性减少的程度。假定特征 A 对训练数据集 D 的信息增益为 g(D,A)，定义为集合 D 的经验熵 H(D)与特征 A 给定条件下 D 的经验条件熵 H(D|A)之差：

$$g(D,A)=H(D)-H(D|A) \tag{3-12}$$

信息增益大的特征具有更强的分类能力,即算法目标所寻取的目标特征。根据信息增益准则进行特征选择的方法是：对训练数据集 D，计算其每个特征的信息增益，并比较它们的大小，选择最大的特征。

然而通过信息增益选取特征的时候，存在偏向于选择取值较多的特征的问题。使用信息增益比可以纠正这一问题。假定特征 A 对训练数据集 D 的信息增益比 gR(D,A)定义为其信息增益 g(D,A)与训练数据集 D 关于特征 A 的值的熵 HA(D)之比，即:

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)} \tag{3-13}$$

$$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|} \tag{3-14}$$

其中 n 是特征 A 取值的个数。

3.4 本章小结

本章介绍了降维工程中相关的一些原理以及常见的算法。无论是特征抽取还是特征选择，都能够将原始的高维样本数据降维到较低的维度。但是传统算法中，往往是特征选择与特征抽取相分离的，而且缺乏对样本鲁棒性以及特征鲁棒性的研究。基于此问题，本文在第五章中提出了一种新型的特征选择方法，将特征抽取融合到特征选择中去，并通过 L21 范数距离，提高算法的鲁棒性。

第四章　基于 L2P 范数距离度量的 TWSVM

通过对传统支持向量机以及其相关推广算法的理论分析，本文在总结了传统 TWSVM 算法不足的问题上，提出了基于 L2P 范数距离度量的 TWSVM，以提高算法的鲁棒性。

4.1 范数定义

范数是具有长度概念的一种函数。[21, 41]它常常用来度量空间中某个向量空间或者矩阵中向量的长度或者大小。假设我们规定‖·‖是矩阵 X 的一个范数函数，那么‖·‖必须满足如下的条件：

正定性：

$$\left\|X\right\|\geq 0 \tag{4-1}$$

正齐次性:

$$\left\|cX\right\|=\left|c\right|\left\|X\right\| \tag{4-2}$$

三角不等式:

$$\left\|X+Y\right\|\leq\left\|X\right\|+\left\|Y\right\| \tag{4-3}$$

对于任意向量x，常见的向量范式包括 L1 范数，L2 范数，无穷范数，Lp 范数。

向量的 L1 范数即向量中所有元素的绝对值之和，可以表达为：

$$\left\|x\right\|_1=\sum_{i=1}^{n}\left|x_i\right| \tag{4-4}$$

向量的 L2 范数即传统的欧里几德距离(欧式距离)，即向量各元素的绝对值的平方和再开方，可以表达为：

$$\left\|x\right\|_2=\left(\sum_{i=1}^{n}\left|x_i\right|^2\right)^{\frac{1}{2}} \tag{4-5}$$

向量的无穷范数即向量中所有元素绝对值中最大值，可以表达为如下形式：

$$\left\|x\right\|_\infty=\max_i\left|x_i\right| \tag{4-6}$$

向量的 Lp 范数相当于 L2 范数的推广，即向量中蒜素绝对值的 p 次方之和的 1/p 次幂。当 p=2 时，向量的 Lp 范数即向量的 L2 范数。向量的 Lp 范数可以表达为如下的形式：

$$\left\|x\right\|_p=\left(\sum_{i=1}^{n}\left|x_i\right|^p\right)^{\frac{1}{p}} \tag{4-7}$$

假设有矩阵X $=[x_1,x_1,\cdots,x_n]^T$包含n个数据点，并且每一个数据点$x_i\in R^d$,那么矩阵X则是一个n × d大小的矩阵。$x_{i,j}$代表矩阵X第i行中的第j列所对应的元素。对于矩阵X，常见

的矩阵范式包括 L1 范数，L2 范数，无穷范数，Lp 范数和 F 范数。

矩阵的 L1 范数又称列和范数，即所有矩阵的列向量绝对值之和的最大值，可以表达为

$$\left\|\mathbf{X}\right\|_1 = \max_i \sum_{j=1}^{d} \left|x_{i,j}\right| \tag{4-8}$$

矩阵的 L2 范数又称矩阵谱范数，即矩阵平方的最大特征值的开方，可以表达为：

$$\left\|\mathbf{X}\right\|_2 = \sqrt{\lambda} \tag{4-9}$$

其中λ为$X^T X$的最大特征值。矩阵的无穷范数又称作行和范数，即所有矩阵的行向量绝对值之和的最大值，可以表达为如下形式：

$$\left\|\mathbf{X}\right\|_\infty = \max_i \sum_{j=1}^{n} \left|x_{i,j}\right| \tag{4-10}$$

矩阵的 F 范式即矩阵每一个元素的平方和在开平方，可以表达为：

$$\left\|\mathbf{X}\right\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{d} \left|a_{i,j}\right|^2 \right)^{\frac{1}{2}} \tag{4-11}$$

不同的范数由于对向量或矩阵的度量方式不同，使得不同的范数具有不同的性质。如 L1 范数更加具有稀疏性，Lp 范数更加具有鲁棒性等等。因此，不同的范数在模式识别等算法中具有不用的应用，下文将应用 L2P 范数来优化改进一些模式识别算法，使得算法具有更好的泛化性和更好的效果。

4.2 相关工作

在数据挖掘和模式识别的许多应用中，支持向量机（SVM）在过去的几十年中一直是模式识别中的重要分类方法。它已被成功应用于广泛的领域。标准的 SVM 致力于获得一个最优分类超平面，该平面在两个数据集之间具有最大间隔，以减少泛化误差。 SVM 的一个优点是调节结构复杂性和经验风险之间的折中。

但是，SVM 可能不满足现实世界的应用的需求。由于需要解决二次凸规划问题（QPPs），SVM 的计算复杂性将成为一个问题。另外，在处理某些特殊数据集时，SVM 是不适用的，如交叉的异或数据集，不平衡数据集等。因此，许多学者对 SVM 进行了改进研究。

在 2001 年，G. Fungand 等人提出了一种近似支持向量机算法 (PSVM)。PSVM 将两个平行平面尽可能分开以对点进行分类。 不同于传统的支持向量机，PSVM 只需要求解单个线性方程组，而不是求出二次方程或线性方程。 PSVM 使得支持向量机的解变得快速并且高效。 2006 年，O.L.Mangasarian 和 E.W.Wild 通过广义特征值提出了一个非平行分类平面的支持向量机算法(GEPSVM)。GEPSVM 去除了 SVM 产生的边界在输入空间中平行的必要条件。

与 PSVM[42]和 GEPSVM 不同，2007 年 Jayadeva 提出了一种新的非平行分类平面的支持向量机 Twin Support Vector Machine（TWSVM）。 它需要解决一对二次凸规划问题。 两

个二次凸规划问题中的每一个都是一个典型 SVM 的表示，但不是所有的数据点都同时用于两个问题的约束。

尽管如此，上述相关工作都是基于平方二范数距离度量，这很容易导致样本野值对样本数据产生影响。为了能够提供一个鲁棒的方法，基于 L1 范数距离度量的方法已经在许多论文中引入。L1 范数度量的公式可以提供更好的鲁棒性，并且是 L0 范数的最优凸逼近。L1 范数比 L0 范数更适合于优化[43]，因为 L0 范数优化是一个 NP 难问题的优化问题。

大量研究表明，使用 L1 范数最小化和非平方 L2 范数(L2P 范数，$0 < p \leq 2$)最小化可以为目标函数提供鲁棒性[23, 44, 45]，可以更好地容忍噪声造成的偏差，特别是那些离正常样本数据群特别远的野值。 因此，许多研究通过 L2P 范数距离改进了各种模型。 受上述启发，本文中，我们主要针对带有异常值数据样本的数据集上 TWSVM 的鲁棒性问题。 在经典的 TWSVM 中，它的学习函数是将样本距离的平方最小化。 如我们所知，平方后的样本距离更加扩大了由噪声野值引起的样本的误差距离。 基于这一点，我们认为低阶 L2 范数距离可以强调正常点距离占整体样本距离的百分比。 对于 L2P 范数距离，p 值应该低于 2，才可以用于改进 TWSVM。

本小节主要介绍一些向量的定义。在本大章节中，向量都是列向量。行向量将通过列向量经由一个上标转置符号来定义。 假设A表示正类的样本矩阵并且B表示负类的样本矩阵。m1和 m2 分别表示正类样本的数量和负类样本的数量。所有的样本点都属于R^n的实值空间。因此，所有矩阵A和矩阵B的大小分别为$m1 \times n$和$m2 \times n$。 对于一个矩阵A，A_i表示矩阵在实值空间R^n中的第i行。A_i的平方 L2 范数可以表示为$\|A_i\|_2^2$。 因此，矩阵的平方 L2 范数定义如下：

$$\left\|\mathbf{A}\right\|_2^2 = \sum_{i=1} \left\|\mathbf{A}_i\right\|_2^2 \tag{4-12}$$

平方 L2 范数的公式表达可以推广至 p 序 L2 范数（L2P 范数）：

$$\left\|\mathbf{A}\right\|_2^p = \sum_{i=1} \left\|\mathbf{A}_i\right\|_2^p \tag{4-13}$$

另外，为了后文的公式书写，我们定义e1为行数与正类样本个数相同的单位向量。同样，e2为行数与负类样本个数相同的单位向量。I则表示维度合适的单位对角阵。

4.3 L2P-TWSVM 模型推导

4.3.1 模型推导

从上文 TWSVM 的公式(19)和公式(20)中可以清楚地看出学习函数中的平方 L2 范数距离。 它可能不能很好的满足样本存在噪声数据情况下对分类精度的要求。 通过 TWSVM 获得的分类结果可能会被异常值所明显地影响。 也就是说，P 阶 L2 范数距离度量是一种取代平方 L2 范数距离度量的很好的方法。 如果能找到合适的 p 值，算法将强调正常数据的距离并能够最好地忽略异常值距离产生的影响。 假设平方 L2 范数距离是一个基准，如果$p < 2$，数据的距离将缩短，并且异常数据样本造成的影响将被减轻。 本文认为 p 值的确定取决于异常值占整体样本的百分比。

TWSVM 的改进可以通过解决以下问题来表示：

$$\min_{w^1,b^1,q} \frac{1}{2}\left\|\mathbf{A}w^1+e_1b^1\right\|_2^p + c_1e_2^Tq$$
$$\text{s.t.} -\left(\mathbf{B}w^1+e_2b^1\right)+q\geq e_2,\ q\geq 0 \tag{4-14}$$

$$\min_{w^2,b^2,q} \frac{1}{2}\left\|\mathbf{B}w^2+e_2b^2\right\|_2^p + c_2e_1^Tq$$
$$\text{s.t.} -\left(\mathbf{A}w^2+e_1b^2\right)+q\geq e_1,\ q\geq 0 \tag{4-15}$$

公式(4-14)的拉格朗日函数为

$$L\left(w^1,b^1,q,\alpha,\beta\right)=\frac{1}{2}\left\|\mathbf{A}w^1+e_1b^1\right\|_2^p + c_1e_2^Tq+\alpha^T\left[\mathbf{B}w^1+e_2b^1-q+e_2\right]-\beta^Tq \tag{4-16}$$

其中α，β为拉格朗日乘子。

注意到公式(4-16)涉及到 L2P 范式，因此这个函数很难直接求解。针对这样的问题，将含有 L2P 范数的项拆分为平方 L2 范数和$(p-2)$次方的 L2 范数的乘积：

$$\left\|\mathbf{A}w^1+e_1b^1\right\|_2^p=\left\|\mathbf{A}w^1+e_1b^1\right\|_2^{p-2}\left\|\mathbf{A}w^1+e_1b^1\right\|_2^2 \tag{4-17}$$

定义

$$\varsigma=\left\|\mathbf{A}w^1+e_1b^1\right\|_2^{p-2} \tag{4-18}$$

那么拉格朗日函数(4-16)可以重写为以下形式

$$L\left(w^1,b^1,q,\alpha,\beta\right)=\frac{1}{2}\varsigma\left\|\mathbf{A}w^1+e_1b^1\right\|_2^2 + c_1e_2^Tq+\alpha^T\left[\mathbf{B}w^1+e_2b^1-q+e_2\right]-\beta^Tq \tag{4-19}$$

对每一个参数进行求导计算，加上 KKT 条件，可以得到下列的公式

$$\frac{\partial L}{\partial w^1}=\varsigma\mathbf{A}^T\left(\mathbf{A}w^1+e_1b^1\right)+B^T\alpha=0 \tag{4-20}$$

$$\frac{\partial L}{\partial b^1}=\varsigma e_1^T\left(\mathbf{A}w^1+e_1b^1\right)+e_2^T\alpha=0 \tag{4-21}$$

$$\frac{\partial L}{\partial q}=c_1e_2-\alpha-\beta=0 \tag{4-22}$$

$$\alpha\geq 0,\ \beta\geq 0 \tag{4-23}$$

通过公式(4-22)和（4-23）可以得到

$$0\leq\alpha\leq c_1 \tag{4-24}$$

为了简化公式，定义

$$\mathbf{H}=\left[\mathbf{A}\ e_1\right],\ \mathbf{G}=\left[\mathbf{B}\ e_2\right],\ u=\left[w^1,b^1\right]^T \tag{4-25}$$

因此，公式(4-18)可以表达为

$$\varsigma=\left\|\mathbf{H}u\right\|_2^{p-2} \tag{4-26}$$

将公式(4-20)和公式(4-21)相加，可以得到

$$\varsigma\left[\mathbf{A}^T\ e_1^T\right]\left[\mathbf{A}\ e_1\right]\left[w^1, b^1\right]^T+\left[\mathbf{B}^T\ e_2^T\right]\alpha=0 \tag{4-27}$$

这个可以简化表达为

$$\varsigma\mathbf{H}^T\mathbf{H}u+\mathbf{G}^T\alpha=0 \tag{4-28}$$

由公式（4-28）可以得到u的解析解为

$$u=-\left(\frac{1}{\varsigma}\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{G}^T\alpha \tag{4-29}$$

尽管$\mathbf{H}^T\mathbf{H}$是一个半正定矩阵，但是在某些情况下仍有可能存在奇异。因此，给公式(3-29)添加一个正则项，如下所示：

$$u=-\left(\frac{1}{\varsigma}\mathbf{H}^T\mathbf{H}+\varepsilon\mathbf{I}\right)^{-1}\mathbf{G}^T\alpha \tag{4-30}$$

其中$\varepsilon>0$并且I是一个合适维度的对角矩阵。

通过拉格朗日函数和 KKT 条件可以得到原 L2P 范数 TWSVM 的对偶问题的最小化形式，即

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\alpha^T\mathbf{G}\left(\frac{1}{\varsigma}\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{G}^T\alpha-e_2^T\alpha\\&s.t.\qquad 0\le\alpha\le c_1\end{aligned} \tag{4-31}$$

通过公式(4-31),可以求解一个凸二次规划问题来得到最优的α，然后带入公式(4-29)即可得到第一个分类平面的法向量和偏差的值。同理，可以通过同样的方式来求解第二个平面。

4.3.2 迭代算法

由于ς是关于 u 的未知变量，因此不能直接求解出目标值。为此，本文提出了一个有效的迭代算法来解决这个问题，使得u和ς在每次迭代中自动变化直至迭代收敛。

算法 4-1 一个迭代算法解决 L2P 范数距离 TWSVM 问题

Algorithm.4-1 An iterative algorithm for L2P norm distance TWSVM

输入： 训练数据集$A \in R^{m1\times n}$，$B \in R^{m2\times n}$，参数$p, c1, c2$；

步骤一：计算$H \in R^{m1\times(n+1)}$ $G \in R^{m2\times(n+1)}$ $I \in R^{(n+1)\times(n+1)}$；

步骤二：初始化向量$u \in R^{(n+1)\times 1}$；

循环至收敛

步骤三：计算$\varsigma = \left\|Hu^{T}\right\|_2^{p-2}$；

步骤四：通过公式(66)计算拉格朗日乘子α；

步骤五：更新u，如果需要添加正则项；

结束循环

输出：$u \in R^{(n+1)\times 1}$；

同样，另一个平面向量可以通过同样的程序求得。

4.3.3 收敛性证明

为了证明这个算法的收敛性，需要借助以下的一个定理：

定理 1： 对于任意非零向量u, v,当$0 < p \le 2$,下列不等式成立：

$$\left\|\left\|u\right\|_2^p - \frac{p}{2}\left\|v\right\|_2^{p-2}\left\|u\right\|_2^2 \le \left\|v\right\|_2^p - \frac{p}{2}\left\|v\right\|_2^{p-2}\left\|v\right\|_2^2 \right. \tag{4-32}$$

证明： 假设函数$f(x) = 2x^{\frac{p}{2}} - px + p - 2$,我们对此函数进行求导，可以得到

$$f'(x) = p\left(x^{\frac{p-2}{2}} - 1\right) \tag{4-33}$$

$$f''(x) = \frac{p(p-2)}{2}x^{\frac{p-4}{2}} \tag{4-34}$$

很明显，当$x > 0$并且$0 < p \le 2$ $f''(x) \le 0$，并且$x = 1$是可以使得$f'(x) = 0$的唯一解，注意 $f(1) = 0$。因此， $2|x|^p - px^2 + p - 2 \le 0$。因此，可以得到以下公式：

$$2\left(\frac{\left\|u\right\|_2}{\left\|v\right\|_2}\right)^p - p\left(\frac{\left\|u\right\|_2}{\left\|v\right\|_2}\right)^2 + p - 2 \le 0 \tag{4-35}$$

$$\Rightarrow \left\|\left\|u\right\|_2^p - \frac{p}{2}\left\|v\right\|_2^{p-2}\left\|u\right\|_2^2 \le \left\|v\right\|_2^p - \frac{p}{2}\left\|v\right\|_2^{p-2}\left\|v\right\|_2^2 \right.$$

理论 1： 该算法可以在每次迭代中单调地减小问题（4-14）的目标函数值，并使目标函数值收敛到局部最优。

证明： 将公式(4-14)用G, H表达，可以改写为

$$J(u)=\min_{w^1,b^1,q}\frac{1}{2}\left\|\mathbf{H}u\right\|^p+c_1e_2^Tq$$
$$s.t.\ \ -(\mathbf{G}u)+q\geq e2,\ q\geq 0 \tag{4-36}$$

公式(4-14)和公式（4-36）等价，这里用J来表示这个目标函数的值。 假设ũ是下一次迭代的的u的值，那么

$$\begin{aligned}\tilde{u}&=\arg\min_{w^1,b^1,q}\frac{1}{2}\left\|\mathbf{H}u\right\|^p+c_1e_2^Tq\\&=\arg\min_{w^1,b^1,q}\frac{1}{2}\varsigma\left\|\mathbf{H}u\right\|^2+c_1e_2^Tq\end{aligned} \tag{4-37}$$

结合公式(4-35)可以得到

$$\frac{1}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}\tilde{u}\right\|_2^2+c_1e_2^Tq\leq\frac{1}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}u\right\|_2^2+c_1e_2^Tq$$
$$\Rightarrow\frac{1}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}\tilde{u}\right\|_2^2\leq\frac{1}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}u\right\|_2^2$$
$$\Rightarrow\frac{p}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}\tilde{u}\right\|_2^2\leq\frac{p}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}u\right\|_2^2 \tag{4-38}$$

根据定理 1 有

$$\left\|\mathbf{H}\tilde{u}\right\|_2^p-\frac{p}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}\tilde{u}\right\|_2^2\leq\left\|\mathbf{H}u\right\|_2^p-\frac{p}{2}\left\|\mathbf{H}u\right\|_2^{p-2}\left\|\mathbf{H}u\right\|_2^2 \tag{4-39}$$

结合公式(4-38)和公式(4-39)，可以得到

$$\left\|\mathbf{H}\tilde{u}\right\|_2^p\leq\left\|\mathbf{H}u\right\|_2^p$$
$$\Rightarrow\frac{1}{2}\left\|\mathbf{H}\tilde{u}\right\|^p+c_1e_2^Tq\leq\frac{1}{2}\left\|\mathbf{H}u\right\|^p+c_1e_2^Tq$$
$$\Rightarrow J(\tilde{u})\leq J(u) \tag{4-40}$$

因此，在每一次的迭代过程中，这个算法的目标函数都会单调递减。由于这个目标函数具有值为 0 的下界，所以算法能够单调递减直至收敛。

4.3.4 核函数 L2P-TWSVM

为了将L2P范数距离TWSVM推广至非线性分类，通过核函数来修改算法的目标函数。对于 TWSVM，核函数的分类平面分别是

$$K(x^T,\mathbf{C}^T)w^1+b^1=0,K(x^T,\mathbf{C}^T)w^2+b^2=0 \tag{4-41}$$

其中$\mathrm{C^T}=[\mathrm{A^T};\mathrm{B^T}]$,$\mathrm{K}(\cdot)$表示任意选择的核函数。如果$\mathrm{K}(\cdot)$是一个线性核函数如$\mathrm{K}(\mathrm{x^T},\mathrm{C^T})=\mathrm{x^TC}$，那么核函数 L2P 范数距离 TWSVM 就会退化为原始的 L2P 范数距离 TWSVM。

构建如下最优化的核函数 L2P 范数距离 TWSVM 的目标函数：

$$\begin{aligned}&\min_{w^1,b^1,q}\frac{1}{2}\left\|K\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^p+c_1e_2^Tq\\&\text{s.t.}-\left(K\left(\mathbf{B},\mathbf{C}\right)w^1+e_2b^1\right)+q\geq e_2,\ q\geq 0\end{aligned}\tag{4-42}$$

$$\begin{aligned}&\min_{w^2,b^2,q}\frac{1}{2}\left\|K\left(\mathbf{B},\mathbf{C}\right)w^2+e_2b^2\right\|_2^p+c_2e_1^Tq\\&\text{s.t.}-\left(K\left(\mathbf{A},\mathbf{C}\right)w^2+e_1b^2\right)+q\geq e_1,\ q\geq 0\end{aligned}\tag{4-43}$$

公式(3-42)对应的拉格朗日函数可以表达为：

$$L\left(w^1,b^1,q,\alpha,\beta\right)=\frac{1}{2}\left\|k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^p+c_1e_2^Tq+\alpha^T\left[k\left(\mathbf{B},\mathbf{C}\right)w^1+e_2b^1-q+e_2\right]-\beta^Tq\tag{4-44}$$

为了便于求解这个拉格朗日函数，将包含 L2P 范数距离的项拆分为如下形式：

$$\left\|k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^p=\left\|k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^{p-2}\left\|k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^2\tag{4-45}$$

公式(4-45)中，可以将乘积的前一项用ς来表示，那么新的拉格朗日函数可以表达为：

$$L\left(w^1,b^1,q,\alpha,\beta\right)=\frac{1}{2}\varsigma\left\|k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right\|_2^2+c_1e_2^Tq+\alpha^T\left[k\left(\mathbf{B},\mathbf{C}\right)w^1+e_2b^1-q+e_2\right]-\beta^Tq\tag{4-46}$$

可以通过求导以及 KKT 条件得到下列条件：

$$\frac{\partial L}{\partial w^1}=\varsigma k\left(\mathbf{A},\mathbf{C}\right)^T\left(k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right)+k\left(\mathbf{B},\mathbf{C}\right)^T\alpha=0\tag{4-47}$$

$$\frac{\partial L}{\partial b^1}=\varsigma e_1^T\left(k\left(\mathbf{A},\mathbf{C}\right)w^1+e_1b^1\right)+e_2^T\alpha=0\tag{4-48}$$

$$\frac{\partial L}{\partial q}=c_1e_2-\alpha-\beta=0\tag{4-49}$$

$$\alpha\geq 0,\ \beta\geq 0\tag{4-50}$$

将公式(4-47)和公式(4-48)相结合，可以得到

$$\varsigma\left[k\left(\mathbf{A},\mathbf{C}\right)^T\ e_1^T\right]\left[k\left(\mathbf{A},\mathbf{C}\right)\ e_1\right]\left[w^1,\ b^1\right]^T+\left[k\left(\mathbf{B},\mathbf{C}\right)^T\ e_2^T\right]\alpha=0\tag{4-51}$$

为了简化公式，定义

$$\mathbf{E}=\left[k\left(\mathbf{A},\mathbf{C}\right)\ e_1\right],\mathbf{R}=\left[k\left(\mathbf{B},\mathbf{C}\right)\ e_2\right]\tag{4-52}$$

并且用向量u = $[w^1,b^1]^T$来表示此分类平面。因此，公式(4-51)可以改写为：

$$\varsigma\mathbf{E}^T\mathbf{E}u+\mathbf{R}^T\alpha=0\tag{4-53}$$

由公式(4-51)可以得到关于超平面向量u的解析解：

$$u=-\frac{1}{\varsigma}\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{R}^T\alpha\tag{4-54}$$

这样核函数的 L2P 范数距离 TWSVM 的最小化对偶形式为

$$\min_{\alpha} \frac{1}{2\varsigma}\alpha^T \mathbf{R}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{R}^T\alpha + e_2^T\alpha$$
$$s.t. \quad 0 \le \alpha \le c_1 \tag{4-55}$$

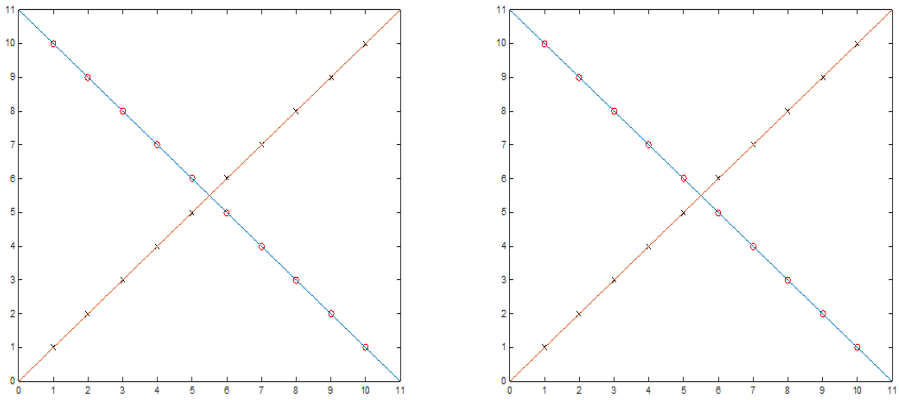
通过同样的方法，可以得到另一个超平面的关于核函数的目标函数最小化对偶形式。一旦这两个核函数的 L2P 范数距离 TWSVM 问题解决了，一个新的点就可以通过和线性 L2P 范数距离 TWSVM 的相似方式来分类。

在实际的实验中，如果样本数量规模很大，那么核技巧可以用来降低 L2P 范数距离 TWSVM 的维数。 在线性情况下，正则化项往往能提高算法的性能。

4.4 L2P-TWSVM 算法实验

4.4.1 二进制数据

为了直接比较 TWSVM 和 L2P 范数距离 TWSVM 之间的差异，本文对人造数据集进行了一个小实验。构建的数据集包含两类数据，分别严格分布在 y=x and y=-x+10 这两条直线上。 这两类点是严格的交叉异或数据。在二维笛卡尔坐标系中，数据集严格分布在两条线上，没有噪音。尽管 L2P 范数距离 TWSVM 致力于提高 TWSVM 的鲁棒性，但它在没有噪声的情况下应该具有与 TWSVM 相同的精度。并且，由于没有噪声，算法只需要迭代一次即可获得最终的收敛结果。 图 1 的两份图像分别显示了 TWSVM 和 L2P 范数距离 TWSVM 的分类平面。 此外，二元异或数据集显示为图像中的点。



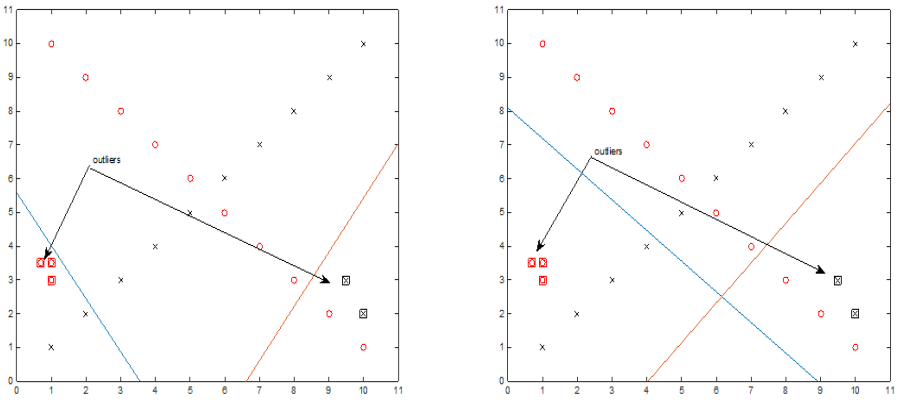
(传统 TWSVM)　　(L2P 范数距离 TWSVM)

图 4-1： 异或数据实验分类平面

Fig.4-1 Classification surfaces on XOR data

图 4-1 表明这两种算法对二元异或数据集具有良好的分类效果，图 4-1 中分类平面几乎相同，结果符合预期猜想。

为了引入野值，本文模拟了一些数据点，这些数据点改变了它们原始的分布并且被用方框表示出来。接下来，再次进行相同的实验以观察两个算法获得的分类平面之间的差异。图 4-2 显示了新数据集和两种方法的得到的分类平面。



(传统 TWSVM)　　(L2P 范数距离 TWSVM)

图 4-2： 存在野值的异或数据实验分类平面

Fig.4-2 classfication surfaces on XOR data with noise data

从图 4-2 中我们可以发现 TWSVM 和 L2P 范数距离 TWSVM 的分类表平面在结构上是相似的，并且 pTWSVM 提供了更好的分类效果。 这证明 pTWSVM 比 TWSVM 更不易受异常值影响，并且具有良好的鲁棒性。

4.4.2 精度比较

在本节中，本文收集了几种不同的公共数据集，以比较不同分类算法的性能。 表 3-1 给出了数据集的描述。

表格 4-1 数据集描述

Tab.4-1 Datasets Description

数据集名称	样本个数	样本维度
heart	270	13
australian	690	14
pima	768	8
monk1	561	6
sonar	208	60
spect	267	44
cancer	683	9
ionodata	351	34
haberman	306	3
monk3	554	6
wpbc	194	33
bupa	345	6
checkdata	297	13

为了公平起见，每个比较的算法都使用线性核。本文将提出的新算法与一些广泛使用的算法进行比较，包括原始 TWSVM，SVM，GEPSVM 和最新的 L1GEPSVM 。本文使用十折交叉验证方法来获得每种算法的最佳参数和 L2P 范数距离 TWSVM 的 p 值。在表 2

中给出了每个算法的平均精度，平均运算时间和十次精度的标准差。不同数据集上的最佳性能以粗体显示。为了比较新方法的统计性检验，本文进行了配对 T 检验，将这些方法与本文的新方法进行比较。当配对 T 检验中的 p-value<0.05 时，认为该算法与本文提出的新算法存在显著性差异，P-value<0.05 表明两个算法分类精度之间的存在很大差异。

表格 4-2 算法分类精度比较(平均精度 ± 标准差，时间：秒，p-value 值)

Tab.4-2 Methods Comparision(Average±STD, time:s, p-value)

	L2PTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value
	0.84±2.77	0.78±5.56	0.82±3.95	0.82±3.00	0.79±4.38	0.67±2.48
heart	0.1623	0.0132	0.0071	0.9383	0.7859	0.0427
	—	0.0720	0.5675	0.4698	0.1113	5.97e-5
	0.84±2.52	0.67±4.96	0.84±4.00	**0.85±1.65**	0.66±4.66	0.57±3.06
australian	1.2176	0.0211	0.1180	8.1210	1.0614	0.8684
	—	8.62e-6	0.8613	0.6857	1.65e-6	2.55e-7
	0.76±3.82	0.75±4.05	0.75±2.30	0.75±3.43	0.74±4.24	0.74±4.26
pima	1.1706	0.0137	0.0412	1.8497	0.9329	0.9378
	—	0.5455	0.5268	0.5713	0.3572	0.3558
	0.70±7.07	**0.79±3.98**	0.70±3.18	0.55±9.29	0.76±2.29	0.66±4.55
monk1	0.3543	0.0125	0.0934	0.1614	0.8432	0.0777
	—	5.06e-7	0.7047	5.10e-7	0.0515	0.1641
	0.68±10.04	0.71±4.88	0.68±5.55	**0.74±3.57**	0.72±9.52	0.72±6.15
sonar	0.3965	0.0158	0.0079	1.5948	4.2953	0.0257
	—	0.0816	0.8062	0.0293	0.0184	0.2800
	0.79±1.50	0.58±4.83	0.79±5.49	0.71±4.40	0.78±5.09	0.79±5.49
spect	0.1442	0.0187	0.0062	1.5253	2.7397	0.0253
	—	2.02e-6	0.9740	0.0041	0.6591	0.9951
	0.96±1.28	0.91±7.14	0.96±1.63	**0.97±1.16**	0.95±2.26	0.95±1.80
cancer	1.4262	0.0159	0.0925	0.2452	1.0705	0.3123
	—	0.0033	0.9934	0.6237	0.4251	0.3608
	0.90±1.90	0.82±4.49	0.85±5.65	0.86±3.17	0.79±4.40	0.86±5.68
ionodata	0.2017	0.0140	0.0094	1.4361	2.1234	0.3446
	—	3.19e-4	0.0121	0.0204	6.43e-4	0.2272
	0.63±19.51	**0.75±4.78**	0.73±5.17	0.64±21.10	0.74±5.02	0.73±5.28
haberman	0.1335	0.0123	0.0079	0.2823	0.7074	0.0204
	—	1.80e-4	0.0014	0.6761	6.57e-4	0.0105

	L2PTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value	平均精度 时间(s) p-value
monk3	0.82±5.97 0.6786 —	**0.87±2.17** 0.0142 0.0908	0.78±2.78 0.0361 0.0240	0.48±3.51 0.1020 8.39e-9	0.79±3.63 0.8342 0.0619	0.77±3.37 0.5351 0.0323
wpbc	**0.78±5.84** 0.1236 —	0.72±7.38 0.0131 0.0042	0.76±7.06 0.0060 0.1583	0.73±6.79 1.8354 0.0298	0.76±6.56 1.4888 0.1706	0.76±7.59 0.0634 0.5226
bupa	**0.69±3.35** 0.2546 —	0.54±5.15 0.0118 3.99e-5	0.67±4.72 0.0091 0.1921	0.66±6.25 0.8766 0.0232	0.5391±4.03 0.7477 4.03e-6	0.62±7.25 0.0975 0.1008
checkdata	0.53±4.87 1.3981 —	**0.57±5.84** 0.0197 0.0134	0.50±4.76 0.0785 0.0703	0.51±4.74 0.6881 0.1096	0.52±5.78 0.9978 0.4931	0.51±3.90 0.5537 0.4245

实验的结果表明，当由 L2P 范数距离 TWSVM 获得的超平面与 TWSVM 获得的超平面相同时，只有一次循环。理论上，当参数 p 值不固定为 2 时，L2P 范数距离 TWSVM 提供更多的参数选择来优化算法。另外，从表二可以看出，对于大多数数据集，新方法的标准偏差总是小于其他方法的标准偏差。这意味着本文提出的新方法具有更好的鲁棒性，并且算法具有更高的稳定性，这符合预期的效果。

表二中许多 p-value 值小于 0.05，即在大多数数据集上基于 L2P 范数距离 TWSVM 的精度明显高于其他分类器的精度。 例如，在 ionodata 和 monk3 数据集上比较 L2P 范数距离 TWSVM 和 NLPTSVM 的 p-value 值分别为 0.0121 和 0.0240，因此 pTWSVM 在两个数据集上明显优于 TWSVM。这种情况也出现在 SVM 中。此外，在一些数据集中，pTWSVM 并不具有最高的准确性。 例如，L1GEP 在 australian 和 cancer 数据机上的正确率。 然而，在这些数据集上比较 L2P 范数距离 TWSVM 与它们的 T 检验的 p-value 值分别为 0.6857 和 0.6237，它们之间在统计学上没有显著差异。 T 检验的 p-value 值也证明了四个数据集上 L2P 范数 TWSVM 和 NLPTSVM 之间存在显着差异。

关于算法运算的时间， NLPTSVM 总是比 L2P 范数距离 TWSVM 更快。 这可以从其公式来解释。 虽然它们都是基于传统 TWSVM 的迭代算法，但是 L2P 范数距离 TWSVM 解决的是一对凸二次规划问题（QPPs），而 NLPTSVM 解决的是线性规划问题（LPP）。

实验结果表明，L2P 范数距离 TWSVM 不仅有效，而且对大多数数据集也是更好的选择。

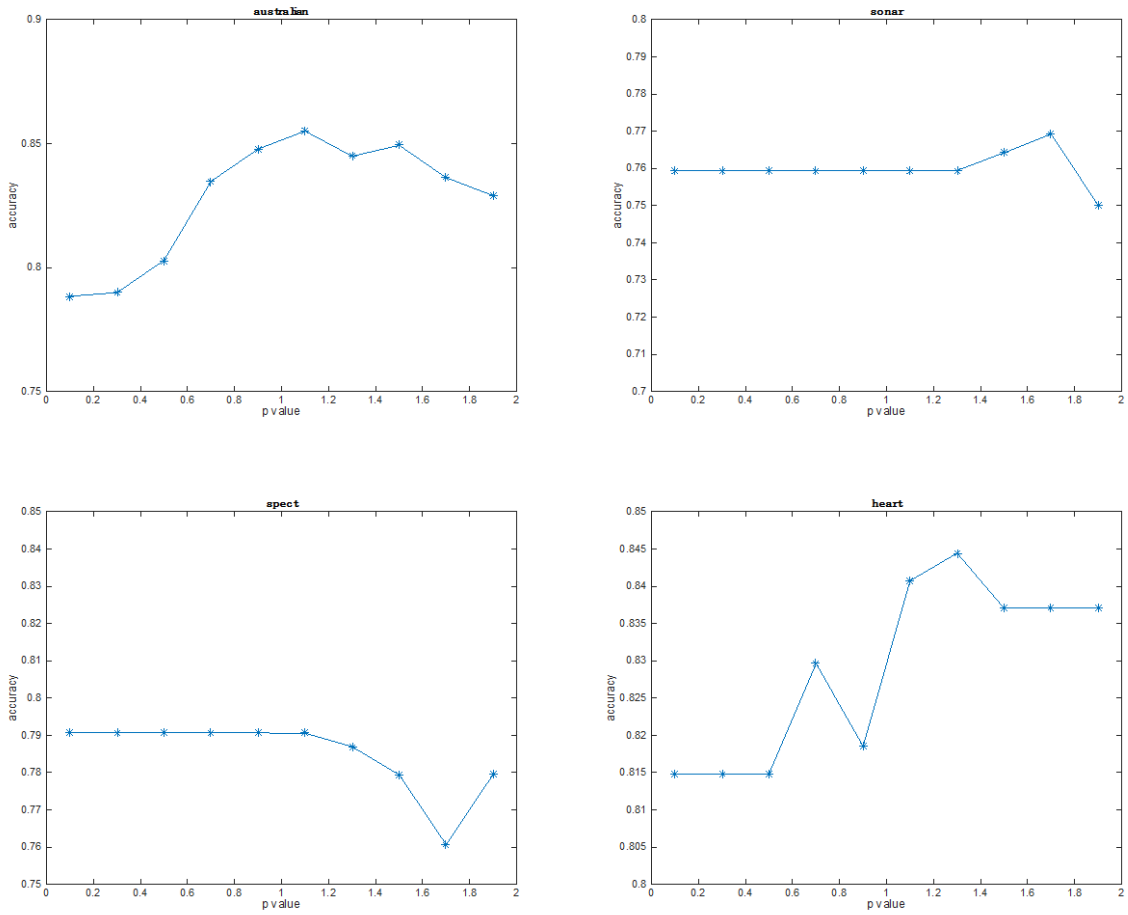
4.4.3 参数 p 值研究

本文新提出的方法存在一个关于如何确定 p 值的问题。一般情况下，针对具体的目标函数，p 值受异常值的影响。为了获得更高的精确度，噪声数据的比例越大，p 值越小，反

之亦然。公式(4-14)很明显的表明 p 值直接影响目标函数的结果。本文将公式分成两部分：异常值数据点的距离和正常数据点的距离。p 值的作用是强调这两部分的比例。因此认为参数 p 值可以直接影响实验精度。

本文对几个基准数据集进行实验。为了测量精度的影响，本文将其余参数设置为特定值$c1 = c2 = 1$。然后记录不同 p 值下算法的正确率。为了研究其对分类性能的影响，本文将 p 值的目标范围固定在 0.1 到 2 之间变化。通过实验数据，本文模拟了相应的正确率曲线。所有的记录如图 4-3 所示。

图 4-3 显示，参数 p 值的确定与特定数据集密切相关。由此可以达出两个结论：一是当参数 p 值太小时，分类精度不是很稳定;另一个是，当值在 1.0 到 1.2 之间时，L2P 范数距离 TWSVM 总是有非常好的性能。这两点以从一下三个方面来解释。首先，当数值较小时，S的值可能非常大以至于目标函数的值不准确。其次，正则化参数被设置为$1e-7$，它可能对奇异性问题的计算结果有影响。 最后，数据集的数据分布和数值大小会影响计算过程。但是，当参数 p 的值稍大时，这些问题将会大大缓解，分类性能会上升并稳定下来。为了获得更好的准确性，我们在后续实验中采用了一种通过十折交叉验证方式从 0.1,0.2 ... 2.0 中选择最合适的 p 值。



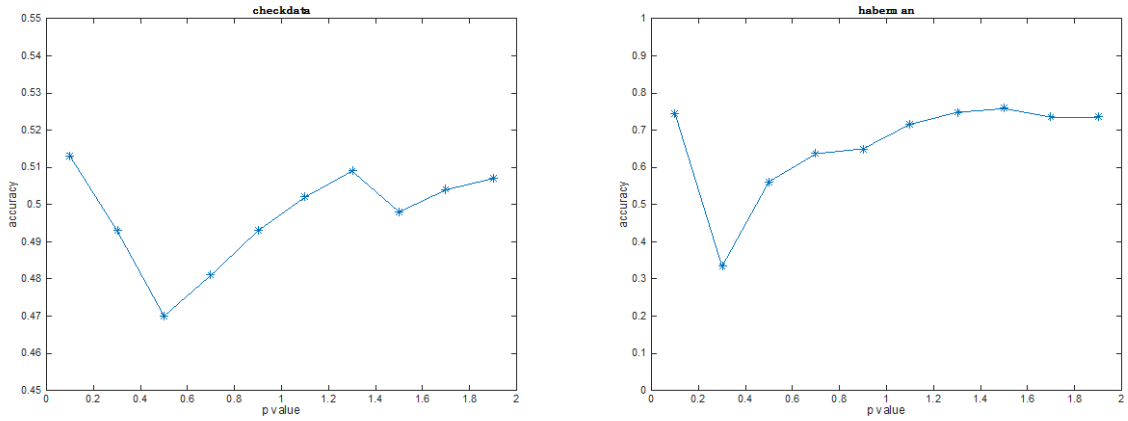


图 4-3：不同 p 值下算法正确率折线图

Fig.4-3: Accuracy line with different p

4.4.4 算法收敛性分析

由于该算法是一种迭代算法，因此算法的收敛性是一个重要的问题。 在前文中，从理论上严格证明了它的收敛性，现在从实验中研究它的收敛性。 本文用以下几个数据集进行试验，并且固定 p 值，算法在每次迭代中的目标值绘制在图 4-4 中。

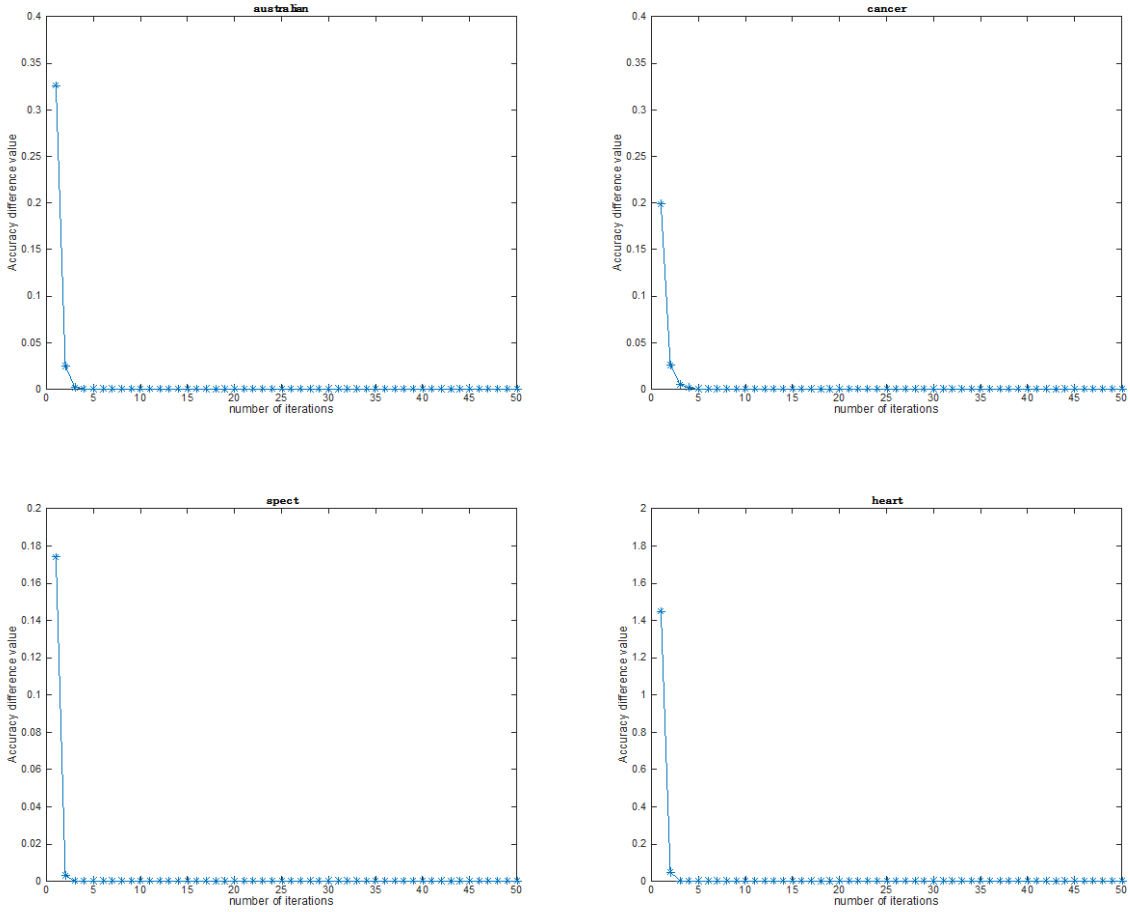


图 4-4 迭代次数 vs. 目标值差异

Fig.4-4 Iteration numbers vs. objective value difference

图 4-4 显示本文新提出的算法的目标值差异随迭代过程不断减少。 而且，对于每个数据集，该算法通常会在 5 次内收敛到渐近线，这表明了该算法在计算上和时间上的可行性。根据实验结果，本文在实验中设定了一个停止阈值为10^{-5}，这足以在收敛性方面取得令人满意的结果。

4.4.5.噪声数据实验

由于新提出的 L2P 范数距离 TWSVM 算法具有处理噪声样本的主要优点，因此本文将重点关注以下设定异常值的数据集实验。为了模拟异常值数据样本，给定输入数据集$X = [x_1, \ldots, x_n] \in R^{m\times n}$，给它加入一个噪声矩阵$\tilde{X} \in R^{d\times n}$，并且该噪声矩阵中的元素都满足标准独立同分布原则。然后，在$X + \sigma\tilde{X}$数据集上执行与原始数据相同的计算程序，其中$\delta = nf\frac{\|X\|_F}{\|\tilde{X}\|_F}$并且 nf是一个给定的噪音因子。 在所有的实验中，设定$nf = 0.1$。 本文的新方法与以前的其他方法进行比较，结果见表 4-3。

表 4-3 加入 20%噪声的分类效果 (平均精度 ±标准差，p-value 值)

Tab.4-3 Methods Comparision with 20% noise(Average±STD, time:s, p-value)

	pTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value
heart	**0.70±8.84**	0.67±6.68	0.68±1.17	0.70±4.53	0.65±5.92	0.66±2.15
	—	0.2522	0.4363	0.9993	0.0820	0.0922
australian	**0.68±2.97**	0.62±7.25	0.65±4.77	0.59±3.15	0.610±5.52	0.57±3.19
	—	0.0037	0.1955	3.39e-4	0.0068	7.08e-4
pima	**0.75±3.20**	0.72±3.18	0.74±5.07	0.74±3.31	0.72±2.73	0.71±5.53
	—	0.2118	0.7443	0.5015	0.2411	0.0258
monk1	0.68±4.16	0.80±4.21	0.65±2.93	0.54±6.60	**0.80±2.84**	0.66±4.55
	—	0.0010	0.1227	1.47e-5	3.41e-5	0.2543
sonar	**0.75±8.01**	0.70±9.19	0.68±8.60	0.74± 6.90	0.73±2.17	0.72±6.53
	—	0.0232	0.0127	0.9782	0.3558	0.1043
spect	0.76±4.35	0.55±5.25	0.79±5.01	0.72±5.43	0.77±3.67	**0.79±5.49**
	—	1.68e-6	0.2039	0.0331	0.7666	0.1662
cancer	0.96±1.93	0.95±0.59	0.96±1.55	**0.96±0.99**	0.95±1.50	0.95±1.95
	—	0.9173	0.7347	0.5161	0.5975	0.7156
ionodata	**0.90±2.80**	0.81±4.42	0.86±4.67	0.87±2.33	0.81±3.75	0.87±5.19
	—	1.26e-4	0.0593	0.0948	3.03e-5	0.0874
haberman	0.74±4.69	0.74±4.06	0.72±5.06	0.74±2.63	**0.75±4.66**	0.72±5.16
	—	0.8491	0.3094	0.8794	0.5593	0.3124
monk3	**0.86±5.00**	0.84±3.63	0.79±1.93	0.70±14.79	0.79±2.94	0.77±3.37
	—	0.0813	0.0010	1.75e-6	2.86e-4	2.29e-4
wpbc	**0.79±7.17**	0.68±7.52	0.74±5.26	0.60±3.72	0.76±5.71	0.76±7.59
	—	1.74e-5	0.0181	5.20e-7	0.0560	0.0166

	pTWSVM	L1GEP	TWSVM	SVM	GEPSVM	NLPTSVM
	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value	平均精度 p-value
bupa	**0.68±5.10**	0.61±3.73	0.64±10.58	0.64±4.35	0.51±4.88	0.63±6.83
	—	0.0065	0.1574	0.1724	1.24e-5	0.0771
checkdata	0.53±4.51	**0.57±4.69**	0.51±2.72	0.50±1.99	0.53±1.94	0.51±3.84
	—	0.0920	0.2867	0.1932	0.8108	0.4727

如表 4-3 所示，在添加相同噪声的情况下，新提出的 L2P 范数距离 TWSVM 证明了其强壮的鲁棒性。L2P 范数距离 TWSVM 在不同的数据集上基本表现出了最高的分类精度。与没有添加噪声时的分类结果相比较，实验结果表明每种算法的分类精度都有所降低，其中 L2P 范数距离 TWSVM 算法下降最少。此外，我们注意到，在这五个数据集中，pTWSVM 没有表现出最好的精度，相应的 p 值分别为 3.41e-5,0.1662,0.5161,0.5593,0.0920。 五个 p 值只有一个小于 0.05，这意味着其他四个在统计显着性上没有显着差异。

通过与原始数据和污染数据的算法分类精度对比，我们可以获得算法的在不同噪声污染情况下的精度差异。 为了深入研究，在实验中采取了不同的η值。 以下图片总结了不同算法在不同η值的基准数据集上的性能。

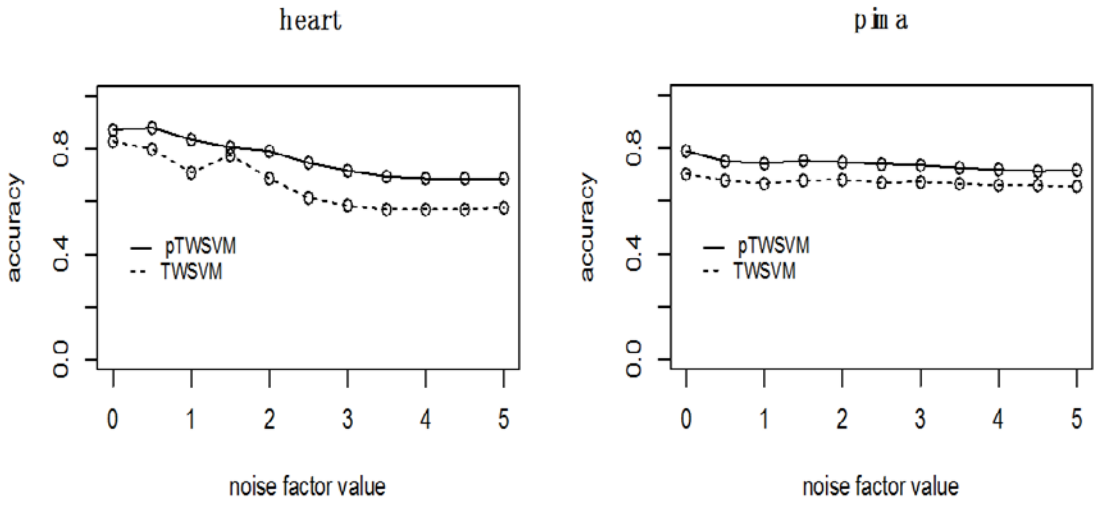


图 4-5　不同噪声程度下算法分类精度

Fig.4-5 accuracy with different noise factor

从图 4-5 中，可以得到以下几点结论：

（1）所提出的 L2P 范数距离 TWSVM 方法在实验数据集上一直优于传统的 TWSVM 方法，这证明了该方法能够有效提高噪声数据分类精度。同时，这也表明，新的给予 L2P 范数距离的 TWSVM 方法在实际应用中可以取得较好的效果。

（2）无论噪声系数值是多少，L2P 范数距离 TWSVM 的精度始终高于传统 TWSVM。尽管如表格二所示，L2P 范数距离 TWSVM 方法在无噪声的原始基准数据集上分类精度的提高相对不突出的，但在带有离群值数据样本的噪声数据中，新方法的分类精度的提升相当大。例如，对于具有异常值的 heart 数据集，不同η值下 L2P 范数距离 TWSVM 平均分类精度为 0.7481，而 TWSVM 的平均分类精度为 0.6633。所以本文提出的方法相比传统 TWSVM 方法分类精度提高了12.78% = (0.7481 − 0.6633)/0.6633.。相反，在无噪声条件下

相同数据集上的分类精度的提高为$4.47\% = (0.8667 - 0.8296)/0.8296$。这一现象普遍存在于其他数据集上，这表明所提出的方法具有更好的对噪声数据处理的能力。

（3）图 4-5 同时也显示 L2P 范数距离 TWSVM 的准确性变化是平坦的并且变化不大，这清楚地表明新提出的 L2P 范数距离 TWSVM 方法比原始 TWSVM 方法更快且更容易趋于稳定。这一特点证实了新方法对异常数据样本具有很好的鲁棒性。

4.5 本章总结

本文提出了一个基于 L2P 范数距离的鲁棒 TWSVM，它的目标函数是一个非光滑非凸的最小化问题。与平方 L2 范数距离相比，L2P 范数距离 TWSVM 具有更好的分类精度，并且对于远离的数据样本非常鲁棒。 与传统的 TWSVM 相比，新方法具有更多挑战性的优化问题。 为了解决这个问题，本文引入了一种高效的迭代算法，并对算法的收敛性进行了严格的理论分析。

该算法仍有几个改进的方向：（1）处理奇异性的问题，在上文的研究中，是通过正规化项解决的。（2）在每次迭代期间，如果 p 值太小，例如 0.1,0.2，则该值将变得非常大。 这会导致算法分类精度不准确。（3）决定参数 p 的值仍然是一个开放的问题，而这个问题在许多算法中也没有解决。

第五章　基于 L21 范数距离度量的判别特征选择

通过对特征选择的介绍，本文提出了一种更加鲁棒的判别特征选择算法，不仅对噪声数据鲁棒，而且还对噪声特征鲁棒。并且，由于是基于 L21 范数距离度量，这使得本算法具有更好的稀疏性，更加满足特征选择的要求，使得算法性能有了较大的提高。

5.1 相关工作

在数据挖掘和模式识别的许多应用中，数据往往具有超高维的特征。太多的特征增加了算法处理数据的计算时间和内存开销。此外，许多的特征是冗余的甚至和分类不相关的，这不利于算法分类。因此，降维工作一直是模式识别领域数据处理的重要组成部分。降维工程是致力于找到数据的“内在维度”。这使得我们致力于寻找一种去除无用特征或在较低维空间中能够表示原始输入数据的方法。

降维工程可以分为两种方式：特征抽取和特征选择。特征抽取方法将原始特征转换为具有较低维度的新特征空间。与特征抽取不同，特征选择试图消除不相关或多余的特征，并同时保留最具判别性的特征。因此，特征选择保留了特征的主要的原始语义，并为新特征提供了可解释性。因此，特征选择越来越受到欢迎，近年来许多研究集中在的特征选择上。在近期的研究中，越来越多的人关注特征抽取与特征选择的结合。

Fisher 线性判别分析（LDA）是最受欢迎的监督特征抽取方法之一。LDA 搜索一个新的特征空间，它可以最大化“类间散度”并同时最小化“类内散度”。这一约束允许不同类别的数据点尽可能分离，并且在新的投影空间中相同类别数据点尽可能多聚集。在过去的几十年中， LDA 算法有了许多的扩展使之转化为特征选择方法。Fisher Score 算法是一种基于线性判别分析的广泛使用的特征选择方法。该方法通过计算特征和相同类型样本之间的方差来分别对每个特征进行评估和排序，然后选择排名最高的特征作为目标特征。但是，这种方法忽略了特征之间的关系并且忽略了冗余特征的存在。因此，该方法不具备去除特征冗余的能力，并且不能处理特征之间的关系。为了克服这个缺点，M. Masaeli 等人提出了一种新的算法线性判别特征选择（LDFS）[14, 60]。这是一个受 LDA 的启发，基于过滤器的特征选择方法。 LDFS 为传统的 LDA 提供正则化项来约束寻求的投影平面。由于选择的特征是通过学习机制获得的，LDFS 可以同时去除冗余特征和不相关的特征。因此 LDFS 在特征选择中起着重要作用。然而，LDFS 的公式是有缺陷的，因为它忽略了投影矩阵的任意伸缩性的可能性。任意伸缩性可以导致全零的平凡解的存在。因此，当算法的解为平凡解时，LDFS 不能获得最具判别力的特征。

2016 年，Hong Tao 等人提出了判别特征选择（Discriminative Feature Selection，DFS）的新方法[41]，它可以通过限制公式条件而使平凡解不存在。DFS 不再同时解决最小化项和最大化术项的问题，而是强制其中一项成为固定的约束条件。此外，L21 范数正则化在公式中加以引入。这些改进使得 DFS 不仅具有 LDFS 的优点，可以同时去除冗余和不相关的特征，而且还可以避免无效的平凡解。 虽然 DFS 是一种高效的和有创造性的特征选择方法，但它的学习函数是基于平方 L2 范数，这可能导致 DFS 容易出现异常值数据样本和异常值特征。 换句话说，DFS 的选择特征可能不是最具有判别力的，因为学习过程可能受

到噪声样本和噪声特征的影响。

本文重点针对特征选择中样本数据存在异常数据点和异常值特征的鲁棒性问题进行研究。许多以前的研究使用正则化项来提高模式识别方法的鲁棒性。到目前为止，众所众知，在用于特征选择的学习函数中使用 L21 范数距离度量的文章很少。受到 DFS 的启发，本文提出了一种鲁棒的基于 L21 范数距离度量的线性判别分析方法 L21FS。新的 L21FS 方法解决同时最小化和最大化问题。

在本节中，先介绍本章节中相关的符号和定义。LDA 是模式识别领域流行的降维方法。假设有n个属于c类的数据点$\{x_1, x_2 \cdots x_n\}$。为了方便起见，数据集可以用矩阵X表示。此外，X_i表示第i个数据点并且X_j表示第 j个特征。LDA 的目标是寻找一个投影平面，以便不同类别点之间的距离最大化，同一类别点之间的距离最小化。为了评估数据点的距离，引入了基于平方 L2 范数距离的散度矩阵：

$$\mathbf{S}_b = \sum_{k=1}^{c} n^k \left(\mu^k - \mu\right)\left(\mu^k - \mu\right)^T \tag{5-1}$$

$$\mathbf{S}_w = \sum_{k=1}^{c}\left(\sum_{i=1}^{n_k}\left(x_i^k - \mu^k\right)\left(x_i^k - \mu^k\right)^T\right) \tag{5-2}$$

$$\mathbf{S}_t = \sum_{i=1}^{n}\left(x_i - \mu\right)\left(x_i - \mu\right)^T \tag{5-3}$$

其中，S_b，S_w，S_t分别表示类间散度矩阵，类内散度矩阵和总散度矩阵。表 1 中给出了本章节中的相关符号。

表格 5-1 定义

Tab.5-1 Definitions

定义	描述
c	类别数
n	数据点数
μ	总体样本平均值
k	第 □个类别
S_b	类间散度矩阵
S_w	类内散度矩阵
S_t	总散度矩阵
W	投影矩阵

很明显，S_t是S_b和S_w的总和，可以写成：

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w \tag{5-4}$$

LDA 的目标函数可以写成如下形式：

$$J(\mathbf{W}) = \max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}} \tag{5-5}$$

其中W就是所寻找的投影矩阵。

由于类间散度矩阵，类内散度矩阵和总散度矩阵紧密相关，所以原始 LDA 可以导出许多变化。而且，将最大化问题转化为最小化问题极大地扩展了这种可能性。受到这种观点的启发，LDFS 重写了传统 LDA 的表达式：

$$J(\mathbf{W})=\min_{\mathbf{W}}-\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}} \tag{5-6}$$

对于投影矩阵W，它的每一行代表相应特征的重要性。 如果某一行主要由零组成，这意味着相应的特征对分类没有贡献。相反，与所选特征相对应的行至少必须具有一个非零项。 因此，为了实现特征选择的能力，LDFS 必须迫使投影矩阵包含更多的零行。因此，LDFS 引入了$l_{\infty,1}$范数正则化术项，这有助于缓解过度拟合并提高泛化性能。改进后的目标可以写成如下形式：

$$J(\mathbf{W})=\min_{\mathbf{W}}-\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}+\gamma\left\|\mathbf{W}\right\|_{\infty,1}=\min_{\mathbf{W}}-\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}+\gamma\sum_{j=1}^{d}\left\|\mathbf{W}_j\right\|_{\infty} \tag{5-7}$$

然而，LDFS 的公式则决定了它在达到平凡解时将失去特征选择的能力，这得到了证明。因此，一种基于 LDFS 的新方法被提出，称为判别特征选择（Discriminative Feature Selection，DFS）。为了避免任意缩放以及平凡解的存在，DFS 强制投影矩阵W独立于S_b。此外，L21 范数正则项用来代替$L_{\infty,1}$范数正则项。新公式可以写成如下形式：

$$\min_{\mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}}-tr\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)+\gamma\left\|\mathbf{W}\right\|_{2,1} \tag{5-8}$$

公式(5-8)致力于最大化类间散度迹，并且第二项能够调节解的稀疏性和学习函数的经验风险。由于$l_{\infty,1}$范数和$l_{2,1}$范数都是l_1范数的拓展，DFS 同样也利用了$L_{2,p}$范数正则项来代替$l_{\infty,1}$范数正则项。

如上所述，由于平凡解问题，DFS 是 LDFS 的更好选择。尽管如此，它仍然忽略了特征选择的鲁棒性。DFS 无法很好的处理存在异常数据点和异常值特征的鲁棒性问题。回顾 DFS 的表达式，它可以被重写为：

$$\begin{aligned}
&\min_{\mathbf{W}^T\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\mathbf{W}=\mathbf{I}} -tr\left(\mathbf{W}^T\left(\sum_{k=1}^{c}n^k\left(\mu^k-\mu\right)\left(\mu^k-\mu\right)^T\right)\mathbf{W}\right)+\gamma\left\|\mathbf{W}\right\|_{2,1}\\
&\Rightarrow \min_{\sum_{i=1}^{n}\mathbf{W}^T(x_i-\mu)(x_i-\mu)^T\mathbf{W}=\mathbf{I}} -tr\left(\sum_{k=1}^{c}n^k\mathbf{W}^T\left(\mu^k-\mu\right)\left(\mu^k-\mu\right)^T\mathbf{W}\right)+\gamma\left\|\mathbf{W}\right\|_{2,1}\\
&\Rightarrow \min_{\sum_{i=1}^{n}\left\|(x_i-\mu)^T\mathbf{W}\right\|_2^2=\mathbf{I}} -\sum_{k=1}^{c}n^k\left\|\left(\mu^k-\mu\right)^T\mathbf{W}\right\|_2^2+\gamma\left\|\mathbf{W}\right\|_{2,1}
\end{aligned} \tag{5-9}$$

如公式（5-9）所示，很明显，目标函数涉及平方 L2 范数项。 众所周知，平方 L2 范数对异常值的存在很敏感。距离的估计可能受到偏离数据样本和离群特征的影响。也就是说，这个目标函数在受污染的数据集上是不合适的，因为较大平方误差距离主宰了总和距离。

5.2 L21FS 模型推导

5.2.1 模型推导

首先使用*L21*范数距离推导出一个鲁棒的目标函数并且求解优化这个问题。这个目标函数很难求解，因为它涉及到一系列的*L21*范数项并且目标函数不是一个凸函数问题。然后，引入一个能够有效解决问题的迭代算法。接下来将证明该算法的收敛性。

如上所述，虽然 DFS 是在正则化项中引入*L21*范数，但 DFS 的学习函数仍是基于平方*L2*范数距离。由于噪声特征和噪声样本的存在，它可能会失去选择最具判别性特征的能力。在本文的新方法中，*L21*范数距离不仅用于正则化项中，还用于学习函数中。因此本文提出的新方法理论上是能够提供更好的鲁棒性和稀疏性的。

在本节中，首先给出了一些*L21*范数相关的概念和定义，然后提出新方法的目标函数。最终，将介绍相关的迭代算法和相关收敛性证明。

对于一个矩阵$X = [X_i^j] \in R^{m*n}$，定义X的第i行为X_i，X_i表示样本矩阵中的第i个数据点。同样，用X^j来表示矩阵的第j列，即矩阵的第j个特征。因此，X_i^j表示样本中第i个数据点的第j个特征。

对于矩阵X，传统的平方 L2 范数距离定义如下：

$$\left\|\mathbf{X}\right\|_2^2 = \sum_{i=1}^{m}\mathbf{X}_i^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{X}_i^j\right)^2 \tag{5-10}$$

相应的*L21*范数距离定义如下：

$$\left\|\mathbf{X}\right\|_{2,1} = \sum_{i=1}^{m}\left\|\mathbf{X}_i\right\|_2 = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}\left(\mathbf{X}_i^j\right)^2} \tag{5-11}$$

与传统的线性判别分析算法类似，L21FS 也需要通过*L21*范数距离来定义类间散度矩阵和类内散度矩阵。假设投影空间为W，那么新空间中类间散度的距离可表示为：

$$\sum_{i=1}^{c} n_i\left\|\overline{\mathbf{X}_i}\mathbf{W} - \overline{\mathbf{X}}\right\|_2 = \left\| \begin{array}{c} n_1\left(\overline{\mathbf{X}_1} - \overline{\mathbf{X}}\right)\mathbf{W} \\ \vdots \\ n_c\left(\overline{\mathbf{X}_c} - \overline{\mathbf{X}}\right)\mathbf{W} \end{array} \right\|_{2,1} \tag{5-12}$$

其中$\bar{x}$是矩阵 X的平均值，c是类别个数，$\bar{x}_i$是第i类样本的平均值，n_i是第i类样本的样本个数。那么类间数据点矩阵X_b可以定义为

$$\mathbf{X}_b = \left[\begin{array}{c} n_1\left(\overline{\mathbf{X}_1} - \overline{\mathbf{X}}\right)\mathbf{W} \\ \vdots \\ n_c\left(\overline{\mathbf{X}_c} - \overline{\mathbf{X}}\right)\mathbf{W} \end{array} \right] \tag{5-13}$$

同样，投影后的*L21*范数距离类内散度矩阵值可以表示为

$$\sum_{i=1}^{c}\sum_{j=1}^{ni}\left\|\mathbf{X}_{ij}\mathbf{W}-\overline{\mathbf{X}_i}\mathbf{W}\right\|_2=\left\|\begin{array}{c}\left(\mathbf{X}_{11}-\overline{\mathbf{X}_1}\right)\mathbf{W}\\ \vdots\\ \left(\mathbf{X}_{1n_1}-\overline{\mathbf{X}_1}\right)\mathbf{W}\\ \vdots\\ \left(\mathbf{X}_{cn_c}-\overline{\mathbf{X}_c}\right)\mathbf{W}\end{array}\right\|_{2,1} \tag{5-14}$$

其中类内数据点矩阵为

$$\sum_{i=1}^{c}\sum_{j=1}^{ni}\left\|\mathbf{X}_{ij}\mathbf{W}-\overline{\mathbf{X}_i}\mathbf{W}\right\|_2=\left\|\begin{array}{c}\left(\mathbf{X}_{11}-\overline{\mathbf{X}_1}\right)\mathbf{W}\\ \vdots\\ \left(\mathbf{X}_{1n_1}-\overline{\mathbf{X}_1}\right)\mathbf{W}\\ \vdots\\ \left(\mathbf{X}_{cn_c}-\overline{\mathbf{X}_c}\right)\mathbf{W}\end{array}\right\|_{2,1}。$$

回顾 LDA 的优化准则，它要求目标最小化类内散度矩阵值，同时最大化类间散度矩阵值。利用这个想法，可以通过以下目标函数实现最优□*21*范数投影矩阵：

$$\mathbf{W}^*=\arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\left\|\mathbf{X}_w\mathbf{W}\right\|_{2,1} \tag{5-15}$$

$$s.t.\ \left\|\mathbf{X}_b\mathbf{W}\right\|_{2,1}=cons$$

在公式（5-15）中，类内散度值将被固定为一个常数以便于计算。到目前为止，可以通过求解这个最小优化问题来获得*L21*范数距离的最优投影矩阵。对于矩阵W^*，它的每一行都对应一个特征。如果一行中的所有元素均为零，则意味着相应的特征对分类没有贡献。为了将*L21*范数距离 LDA 转换成特征选择方法，必须强制更多的行为零。因此，有必要引入*L21*范数距离正则项：

$$\mathbf{W}^*=\arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\left\|\mathbf{X}_w\mathbf{W}\right\|_{2,1}+\gamma\left\|\mathbf{W}\right\|_{2,1} \tag{5-16}$$

$$s.t.\ \left\|\mathbf{X}_b\mathbf{W}\right\|_{2,1}=cons$$

其中$\gamma>0$是可以调节投影矩阵的行稀疏程度的参数。越大的γ意味着更多的行被迫接近于零，反之亦然。目标函数（5-16）可以改写为

$$\mathbf{W}^*=\arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\frac{\left\|\mathbf{X}_w\mathbf{W}\right\|_{2,1}+\gamma\left\|\mathbf{W}\right\|_{2,1}}{\left\|\mathbf{X}_b\mathbf{W}\right\|_{2,1}} \tag{5-17}$$

尽管得出该目标函数的出发点是清晰而简单的，但这个目标函数并不是一个光滑的凸

优化问题，难以有效解决。因此，下面给出一个迭代算法来解决这个同时解决最小最大化问题。

在推导出新方法之前，先介绍以下的一些引理。

引理 1： 对于任何矩阵，当没有一行是零时，有以下等式：

$$\left\|\mathbf{X}\right\|_{2,1}=trace\left(\mathbf{X}^T\mathbf{D}_{\mathbf{X}}\mathbf{X}\right) \tag{5-18}$$

$$s.t.\ \mathbf{D}_{\mathbf{X}}=diag\left(\frac{1}{\left\|\mathbf{X}_1\right\|},\ \cdots,\ \frac{1}{\left\|\mathbf{X}_m\right\|}\right)$$

根据引理 1，类间散度值可以被重写为

$$\sum_{i=1}^{c}n_i\left\|\overline{\mathbf{X}_i}\mathbf{W}-\overline{\mathbf{X}}\mathbf{W}\right\|_2=\left\|\mathbf{X}_b\mathbf{W}\right\|_{2,1}=tr\left(\mathbf{W}^T\mathbf{X}_b\mathbf{D}_b\mathbf{X}_b^T\mathbf{W}\right) \tag{5-19}$$

其中D_b是一个对角矩阵，定义为

$$\mathbf{D}_b=diag\left(\frac{1}{\left\|n_1\left(\overline{\mathbf{X}_1}-\overline{\mathbf{X}}\right)\mathbf{W}\right\|_2},\ \cdots,\ \frac{1}{\left\|n_c\left(\overline{\mathbf{X}_c}-\overline{\mathbf{X}}\right)\mathbf{W}\right\|_2}\right). \tag{5-20}$$

那么类间散度矩阵可以表示为

$$\mathbf{S}_b=\mathbf{X}_b\mathbf{D}_b\mathbf{X}_b^T. \tag{5-21}$$

同样，类间散度值可以写为

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}\left\|\mathbf{X}_{ij}\mathbf{W}-\overline{\mathbf{X}_i}\mathbf{W}\right\|_2=\left\|\mathbf{X}_w\mathbf{W}\right\|_{2,1}=tr\left(\mathbf{W}^T\mathbf{X}_w\mathbf{D}_w\mathbf{X}_w^T\mathbf{W}\right) \tag{5-22}$$

其中，

$$\mathbf{D}_w=diag\left(\frac{1}{\left\|\left(\mathbf{X}_{11}-\overline{\mathbf{X}_1}\right)\mathbf{W}\right\|_2},\ \cdots\ \frac{1}{\left\|\left(\mathbf{X}_{1n_1}-\overline{\mathbf{X}_1}\right)\mathbf{W}\right\|_2},\ \cdots\ \frac{1}{\left\|\left(\mathbf{X}_{cn_1}-\overline{\mathbf{X}_c}\right)\mathbf{W}\right\|_2}\right) \tag{5-23}$$

那么类内散度矩阵可以表示为

$$\mathbf{S}_w=\mathbf{X}_w\mathbf{D}_w\mathbf{X}_w^T. \tag{5-24}$$

根据引理 1，可以得到

$$\left\|\mathbf{W}\right\|_{2,1}=trace\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right) \tag{5-25}$$

其中，

$$\mathbf{D}=diag\left(\frac{1}{\left\|\mathbf{W}_1\right\|_2},\cdots,\frac{1}{\left\|\mathbf{W}_i\right\|_2}\right). \tag{5-26}$$

回顾上述 L21FS 的公式，可以用传统的 L2 范数距离公式来解决：

$$\mathbf{W}^* = \arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{tr\left(\mathbf{W}^T\mathbf{X}_w\mathbf{D}_w\mathbf{X}_w^T\mathbf{W}\right)+\gamma\times tr\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)}{tr\left(\mathbf{W}^T\mathbf{X}_b\mathbf{D}_b\mathbf{X}_b^T\mathbf{W}\right)}$$

$$\mathbf{W}^* = \arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{tr\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)+\gamma\times tr\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)}{tr\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}. \tag{5-27}$$

因此，这个目标函数可以用特征值问题来解决。最佳投影矩阵W^*是对应于最小特征值的特征向量：

$$\left(\mathbf{S}_b\right)^{-1}\left(\mathbf{S}_w+\gamma\mathbf{D}\right)\mathbf{W}=\mathbf{W}\Lambda. \tag{5-28}$$

5.2.2 迭代算法

不过需要注意的是，S_w，S_b和D都依赖于投影矩阵W，因此它们也是未知的变量。本文提出了一种迭代算法来获得公式（5-16）和（5-17）的解，并且下文中将证明该算法的收敛性。

算法 5-1: 一种解决 L21FS 问题的迭代算法

Algorithm.5-1 An iterative algorithm for L21FS

输入：数据X，参数 γ；

初始化列正交矩阵 W，参数 γ；

计算类间数据点矩阵 X_b 和类内数据点矩阵 X_w；

循环至收敛

计算 D_b，D_w 和 D；

通过D_b，D_w构建散度矩阵 S_b，S_w；

通过特征值问题求解公式(5-28)；

通过获得的特征向量更新W；

结束循环

输出：列正交矩阵W；

备注 1：由于S_b的秩等于类别数减 1$(c-1)$，即S_b是不满秩的，所以对S_b进行求逆运算会存在奇异性问题。为了处理这个问题，在S_b的主对角元素上增加一个小的值ς。

备注 2：请注意，在实际问题中，W的某些行将是零，它会导致D_b，D_w和D的一些元素不存在。同样，替换$\|w_i\|_2$为$\sqrt{w_i^2+\varsigma}$。

5.2.3 收敛性证明

在这一小节中，将证明这个算法能够迫使目标函数值每次递减直到收敛。首先，引入以下引理。

引理 2：对于函数$f(a,b)=a-\frac{a^2}{2c}+\kappa b-\kappa\frac{b^2}{2d}$，给定任意非零值$,b,c,d\neq 0\in R^n$，并

且$\kappa \geq 0$, 下面的不等式成立：

$$f\left(a,b\right)\leq f\left(c,d\right). \tag{5-29}$$

因此，对于任意非零向量$v,u,\tilde{v},\tilde{u}$，有

$$v-\frac{v^2}{2\tilde{v}}+ku-k\frac{u^2}{2\tilde{u}}\leq \tilde{v}-\frac{\tilde{v}^2}{2\tilde{v}}+k\tilde{u}-k\frac{\tilde{u}^2}{2\tilde{u}}. \tag{5-30}$$

定理 1：在固定$tr(W^TS_bW)$值的情况下，该算法将在每次迭代中减小公式（5-16）的目标值，直到其收敛到局部最优。

证明： 首先，通过$\tilde{W}$表示更新的W。 在每次迭代中，都有

$$\tilde{\mathbf{W}}=\arg\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} tr\left(\mathbf{W}^T\mathbf{X}_w^T\mathbf{D}_w\mathbf{X}_w\mathbf{W}\right)+\gamma\times tr\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right) \tag{5-31}$$

$$s.t.\ tr\left(\mathbf{W}^T\mathbf{X}_b^T\mathbf{D}_b\mathbf{X}_b\mathbf{W}\right)=cons,$$

这表明

$$tr\left(\tilde{\mathbf{W}}^T\mathbf{X}_w^T\mathbf{D}_w\mathbf{X}_w\tilde{\mathbf{W}}\right)+\gamma\times tr\left(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}}\right) \tag{5-32}$$

$$\leq tr\left(\mathbf{W}^T\mathbf{X}_w^T\mathbf{D}_w\mathbf{X}_w\mathbf{W}\right)+\gamma\times tr\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right).$$

为了方便，定义矩阵W的第i行为w_i，那么

$$\sum_i\frac{\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{X}_w\mathbf{W}_i\right\|}+\gamma\sum_i\frac{\left\|\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|}\leq\sum_i\frac{\left\|\mathbf{X}_w\mathbf{W}_i\right\|^2}{2\left\|\mathbf{X}_w\mathbf{W}_i\right\|}+\gamma\sum_i\frac{\left\|\mathbf{W}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|}$$

$$\Rightarrow\sum_i\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|-\left(\sum_i\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|-\sum_i\frac{\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{X}_w\mathbf{W}_i\right\|}\right)$$

$$+\gamma\left[\sum_i\left\|\tilde{W}_i\right\|-\left(\sum_i\left\|\tilde{W}_i\right\|-\sum_i\frac{\left\|\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|}\right)\right]$$

$$\leq\sum_i\left\|\mathbf{X}_w\mathbf{W}_i\right\|-\left(\sum_i\left\|\mathbf{X}_w\mathbf{W}_i\right\|-\sum_i\frac{\left\|\mathbf{X}_w\mathbf{W}_i\right\|^2}{2\left\|\mathbf{X}_w\mathbf{W}_i\right\|}\right)$$

$$+\gamma\left[\sum_i\left\|W_i\right\|-\left(\sum_i\left\|W_i\right\|-\sum_i\frac{\left\|\mathbf{W}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|}\right)\right]. \tag{5-33}$$

根据公式(5-30),有

$$\sum_i\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|-\sum_i\frac{\left\|\mathbf{X}_w\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{X}_w\mathbf{W}_i\right\|}+\gamma\sum_i\left\|\tilde{\mathbf{W}}_i\right\|-\gamma\sum_i\frac{\left\|\tilde{\mathbf{W}}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|} \tag{5-34}$$

$$\leq \sum_i \left\|\mathbf{X}_w \mathbf{W}_i\right\| - \sum_i \frac{\left\|\mathbf{X}_w \mathbf{W}_i\right\|^2}{2\left\|\mathbf{X}_w \mathbf{W}_i\right\|} + \gamma \sum_i \left\|\mathbf{W}_i\right\| - \gamma \sum_i \frac{\left\|\mathbf{W}_i\right\|^2}{2\left\|\mathbf{W}_i\right\|}.$$

结合公式(5-33)和(5-34)，可以得到

$$\sum_i \left\|\mathbf{X}_w \tilde{\mathbf{W}}_i\right\| + \gamma \sum_i \left\|\tilde{\mathbf{W}}_i\right\| \leq \sum_i \left\|\mathbf{X}_w \mathbf{W}_i\right\| + \gamma \sum_i \left\|\mathbf{W}_i\right\|. \tag{5-35}$$

公式(5-35)可以重写为

$$\left\|\mathbf{X}_w \tilde{\mathbf{W}}\right\|_{2,1} + \gamma \left\|\tilde{\mathbf{W}}\right\|_{2,1} \leq \left\|\mathbf{X}_w \mathbf{W}\right\|_{2,1} + \gamma \left\|\mathbf{W}\right\|_{2,1} \tag{5-36}$$

因此，在$tr(W^T S_b W) = cons$约束条件下，该算法将在每次迭代中单调递减公式（5-16）的目标函数值。需要注意的是，目标函数（5-16）一定大于 0，这意味着它具有下限。因此，该算法将单调减小目标函数(5-16)的目标值，直到它收敛到问题的局部最优值W。

5.2.4 时间复杂度分析

在优化 L21FS 算法的过程中，最耗时的操作是解决步骤 5 中$(S_b)^{-1}(S_w + \gamma D)W = W\Lambda$的特征值问题。该操作的时间复杂度近似于$O(n^3)$。由于该算法是一种迭代算法，因此整个计算复杂度与该算法的迭代次数有关。经验上，实验结果表明该算法只需要几次迭代就能达到收敛。因此，所提出的方法在实践中表现良好。

5.2.5 评价标准

一旦找到了最优投影矩阵W^*，接下来就是确定特征重要性的评估原则。在本算法中，按照每行的欧几里德距离度量按降序排列特征。也就是说，如果$\|W_i\|_2$值越大，则相应的特征就越重要。有了这个原则，就可以得到排名靠前的特征。

5.3 L21FS 算法实验

在本节中，进行了大量实验来评估本文新方法的性能。所有的代码都写在 MATLAB_R2014b 中。实验环境：2.7GHz Intel Core i5 CPU，8 GB 1867 MHz DDR3 内存。采用 LIBSVM 算法对数据点进行分类。为了更加精确，本文的方法的测试精度是使用传统的十折交叉验证来计算的。几个比较算法中的参数也同样是通过十折交叉验证得到。

首先，进行两个小实验来展示本文的新算法能够找出最具判别特征的能力。然后，将本文的 L21FS 方法与几种相关的最先进的特征选择方法进行比较。之后，通过实验来显示参数□对 L21FS 的性能的影响。为了进一步研究本文新方法和其他方法之间的分类精度的差别，本文还采用了配对 T 检验方法。然后，分别对具有噪声数据和噪声特征的数据集进行实验。最后，通过收敛曲线图研究了新方法的收敛性。

5.3.1 数据集描述

本文的实验是使用了几个广泛使用的公开数据集，包括 ORL，USPS，MADELON，LUNG_DISCRETE，ISOLET5，ISOLET，COIL20 和 COLON。所有数据集的介绍如下：

ORL 包含 1992 年 4 月至 1994 年 4 月在实验室拍摄的一组人脸图像，共有 40 个不同

的人。每幅图像的大小为32×32。每个人有十个不同的图像。

USPS 是一个流行的手写体公开数据集，总共包含 9298 个大小为16×16手写数字图像，其中包括 7291 个训练图像和 2007 个测试图像。

MADELON 是一个人造数据集，它是 NIPS 2003 特征选择挑战的一个数据集。这是一个连续输入变量的两类分类问题。这个数据集的特点在于这个数据集的特征是多变量和高度非线性的。

ISOLET5 和 ISOLET 包含 150 个样本，每个字母的名字录入两次。 这些数据被分为五组，分别称为 isolet1 至 isolet5。

COIL20 包含 20 个对象。当物体在转盘上旋转时，每个物体的图像相差 5 度，每个物体有 72 个图像。每个图像的大小是 32x32 像素，每个像素有 256 个灰度级。因此，每个图像由 1024 维向量表示。

COLON 含有从结肠癌患者收集的 62 个样本。其中 40 例肿瘤样本来自肿瘤，22 例正常样本来自同一患者结肠的健康部位。基于测量的表达水平的置信度选择了约 6500 个基因中的两千个。

5.3.2 ORL 人脸数据集小实验

为了直观地展示本文的新方法能够选择最显着特征的能力，在 ORL 数据集上展示了一个小实验，该实验收集了 40 人的面部图片。随机选择两个人的照片数据作为训练数据。为了绘制图片，选择排名靠前的{*32,64,128,256,512,640,768,896,1024*}个特征。需要说明的是，未选择的特征由白色点表示，所选特征由原始值表示。在图 5-1 中，第一行是重绘的一个人，最后一行是重绘的另一个人。

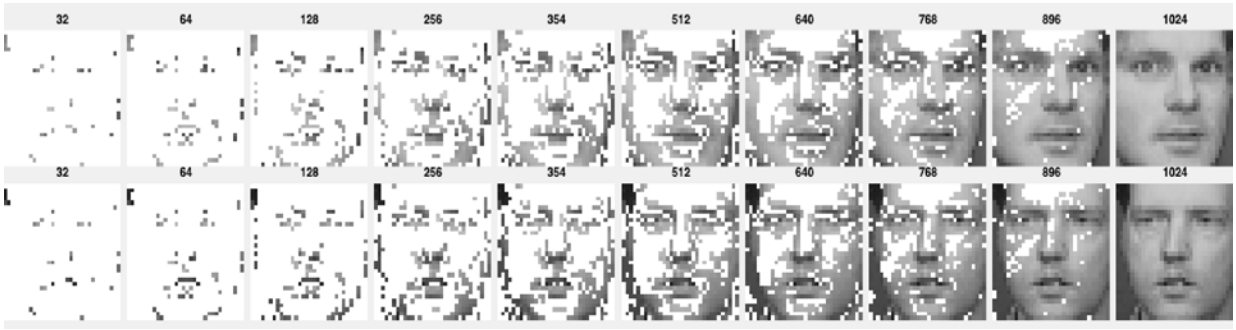


图 5-1 ORL 小实验

Fig.5-1 Toy experiment on ORL

从图 5-1 中可以发现，只需要 64 个特征，图片就足够清晰到识别一个人。需要注意的是，这 64 个特征清楚地显示了眼睛，鼻子和嘴巴，这正是脸部最有辨别力的部分。此外，这些特征不会聚集在一起或随机广泛分布，它们只是显示区别不同人的关键部分。这有力地证明了 L21FS 能够选择最强大和最具辨别性的特征，这与期望一致。

5.3.3 Iris 鸢尾花小实验

Iris 数据集是 UCI ML 数据库的一个流行数据集，包括 3 个类别，每个类别共有 50 个样本。每个样品包含四个特征，代表萼片长度，萼片宽度，花瓣长度和花瓣宽度。由于这个数据集的简单性和普及性，对其进行实验来直观地表达本文提出的方法的效果。

在这个实验中，从 Iris 数据集中选择两个特征，然后在直角二维坐标系中绘制了每个样本点。遍历了这四个特征的所有可能组合并将其画出。此外，还绘制了通过 L21FS 实现的选定特征所呈现的样本，所有图片如图 5-2 所示。

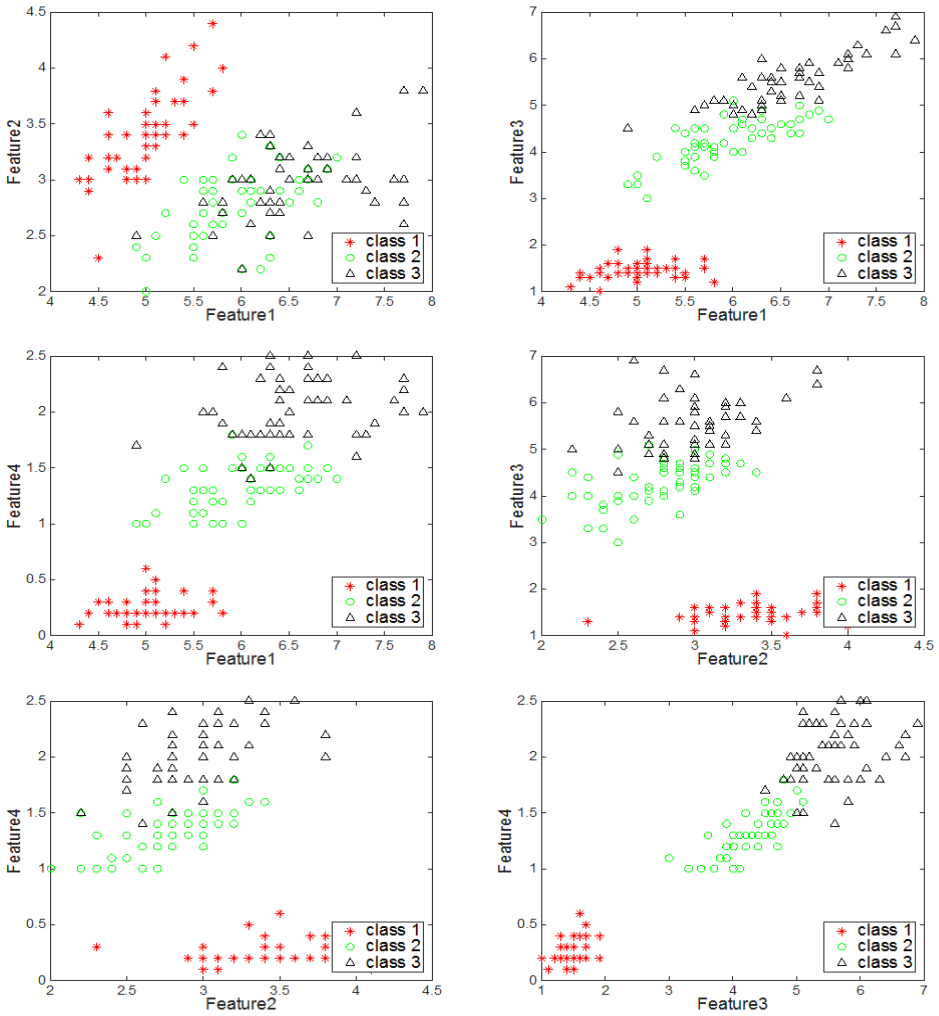


图 5-2-1 六种二维 Iris 图可能.

Fig.5-2-1 Six possibilitis of 2-D Iris

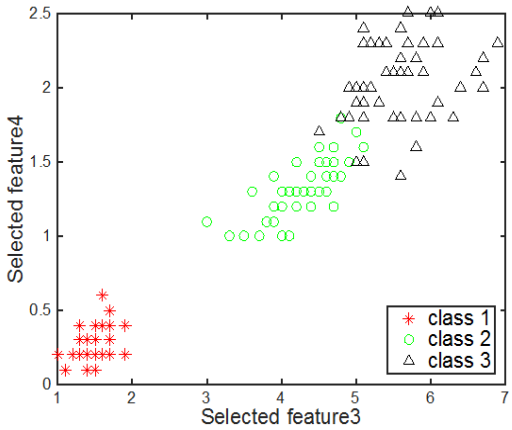


图 5-2-2　L21FS 选出的二维 Iris 图.

Fig.5-2-2 The 2-D Iris selected by L21FS

从图 5-2 中可以看出，L21FS 选择了这六种可能性中视觉区别最明显的两个特征。当

仔细观察图 5-2（2）时，可以发现同一类的点被组合在一起，不同类之间的点相距很远。这种现象与传统 LDA 的思想是一致的。 因此，Iris 数据集上的小实验很好地反映了 L21FS 选择最显着特征的能力。

5.3.4 算法比较

为了显示本文新方法的性能，在公开数据集上进行了试验，并将本文的方法与其他四种广泛使用的最新的特征选择方法进行比较：

Discriminative Feature Selection（DFS），它通过正则化来改善 LDFS 存在平凡解的问题，具有选择最具判别性特征并同时去除冗余特征的能力。

Laplace Score（LS）[53]，它评估每个特征对保持局部性的贡献的重要性，并选择排名最高的特征。

Multi-Cluster Feature Selection（MCFS），选择可以保留数据的多集群结构的特征[61]。与传统的排序方法不同，MCFS 在多种学习和 L1 正则化模型的帮助下选择了最佳特征。

Unsupervised Maximum Margin（UMM），它结合了特征选择和 K-Means 聚类方法来选择最具判别力的特征子集。

为了描述本文新方法的效果，采用了以下几个指标：

平均精度：采用十折交叉验证来评估每种方法的性能。在每个实验中，数据集被分成十个相同大小的子集进行训练和测试。10 个分类任务的平均精度将代表相应方法的分类精度。

运算时间：平均运行时间表示了算晕的时间开销。

方差：方差越小表明该算法具有更好的鲁棒性，受数据影响较小。

统计检验：执行配对 T 检验来比较 L1FS 和其他方法[31]。 T 检验的 p 值表示两个分类准确度值之间差别的概率。p 值越小，表示观察到的两种方法之间的差异越大。 典型的 p 值阈值为 0.05 。例如，如果 p 值小于 0.05，则意味着这两种方法之间存在很大差异，反之亦然。

对于 DFS 和 L21FS，投影矩阵的维度设置为□ − *1*，就像传统的 LDA 一样。 对于五种算法的所有参数，通过十折交叉验证获得它们。使用 LIBSVM 对所选特征提供的样本进行分类，使用十折交叉验证。每种算法的平均精确度汇总在表 5-2 和表 5-3 中，加粗显示其中最好的精度。

表 5-2 选取 20 个特征的性能. (平均精度±方差, 时间: 秒, p-value)

Tab.5-2 The performances of the 20 selected features(Average ± STD, time: s, p-value)

	UMM	MCFS	LS	DFS	L21FS
USPS (2007x256, class:10)	73.94±2.52	87.84±1.58	53.31±0.98	**89.98±1.74**	89.63±1.67
	2.9441	0.1151	0.1284	**0.3622**	0.4324
	6.5402e-6	0.1574	2.8636e-10	**0.7805**	—
MADELON (2000x500,class:2)	61.00±1.65	60.25±1.30	61.10±1.57	60.60±0.93	**61.30±1.65**
	4.7316	0.0341	0.1872	1.9471	**2.3182**
	0.8038	0.3479	0.8654	0.4817	**—**

	UMM	MCFS	LS	DFS	L21FS
LUNG_DISCRETE (73x325,class:7)	67.04±8.1838	76.57±6.1677	57.52±8.78	71.23±2.43	**87.52±5.50**
	0.2948	0.0195	0.0027	0.6794	**0.8650**
	0.0032	0.0293	4.1081e-4	6.4027e-4	—
ISOLET5 (1559x617,class:26)	38.99±6.41	73.76±2.85	43.23±4.72	71.32±4.00	**76.77±2.06**
	4.6121	0.5380	0.1419	3.5905	**3.7723**
	3.5866e-6	0.1257	1.1493e-6	0.0419	—
ISOLET (1559x617,class:26)	33.07±3.22	73.58±3.11	54.93±4.92	68.33±5.89	**79.55±2.86**
	4.5595	0.5644	0.1404	3.6048	**4.1939**
	2.2656e-8	0.0227	2.4978e-5	0.0091	—
COIL20 (1440x1024,class:20)	67.08±1.93	87.56±3.70	61.66±4.95	**94.93±1.72**	91.66±2.48
	11.2796	0.7159	0.1890	**15.2525**	13.6754
	2.8113e-7	0.1034	4.6702e-6	**0.00629**	—
COLON (62x2000,class:2)	70.76±15.60	80.76±10.81	60.89±15.90	64.23±13.00	**85.38±6.32**
	47.2229	0.0616	0.0067	34.4736	**43.1703**
	0.1207	0.4822	0.0211	0.0191	—

表 5-3 选取 40 个特征的性能. (平均精度± 方差, 时间: 秒, p-value)

Tab.5-3 The performances of the 40 selected features(Average ± STD, time: s, p-value)

	UMM	MCFS	LS	DFS	L21FS
USPS (2007x256, class:10)	83.10±1.10	90.63±1.77	66.11±4.81	**91.62±1.23**	91.18±1.18
	2.9376	0.3461	0.1557	**0.3892**	0.7723
	8.7859e-6	0.6210	7.8923e-6	**0.6150**	—
MADELON (2000x500,class:2)	60.80±2.01	58.65±2.32	60.35±1.72	59.95±1.74	**60.85±1.55**
	4.8755	0.0590	0.1843	1.6079	**2.3469**
	0.9696	0.1545	0.6785	0.4628	—
LUNG_DISCRETE (73x325,class:7)	71.14±8.41	79.23±8.18	64.28±9.20	71.14±5.48	**84.85±3.16**
	0.3084	0.0482	0.0034	0.6491	**0.9327**
	0.0158	0.2362	0.0029	0.0025	—
ISOLET5 (1559x617,class:26)	49.90±4.00	86.14±1.36	63.05±2.94	82.36±3.11	**86.40±2.34**
	4.6838	1.1678	0.1440	3.6057	**5.3912**
	2.6778e-7	0.8540	1.6578e-6	0.0718	—
ISOLET (1559x617,class:26)	45.25±2.06	86.02±1.81	63.58±2.54	80.89±3.25	**89.03±3.11**
	4.7121	1.2450	0.1366	3.6701	**4.4487**
	1.1557e-8	0.1330	1.4194e-6	0.0068	—
COIL20 (1440x1024,class:20)	72.63±2.62	95.55±1.21	68.61±0.47	**97.22±1.07**	96.73±0.99
	11.4341	1.5625	0.1853	**15.4507**	16.8142
	1.3667e-7	0.1706	2.4107e-11	**0.5260**	—

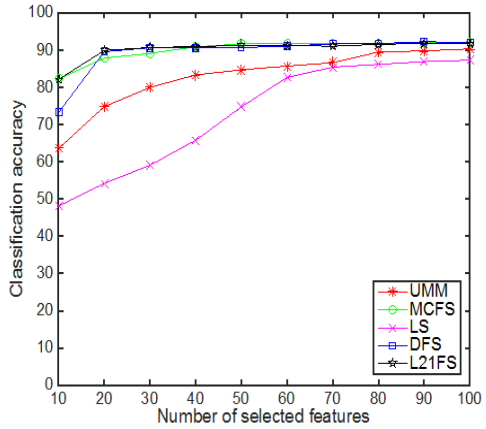
	UMM	MCFS	LS	DFS	L21FS
COLON (62x2000,class:2)	72.30±18.73	83.97±11.53	64.23±15.89	64.23±13.00	**85.38±6.32**
	49.3434	0.0670	0.0066	40.8704	**46.3781**
	0.2225	0.8356	0.0385	0.0191	—

表 5-2 和表 5-3 分别显示了使用七个数据集的前 20 和 40 个特征的分类性能的细节。如这两个表格所示，与其他四种特征选择方法相比，L21FS 表现最佳。在这七个数据集中，L1FS 在五个数据集上表现最好，DFS 在两个数据集中最好。这里应该注意一点，在四种情况下，其中 DFS 比本文所提出的新算法具有更好的平均准确度，但是几乎所有相应的 t 检验 p-Value 值都小于 0.05。这表明，在这些数据集中，两种方法的性能之间没有本质的区别，尽管数值显示出稍有不同。相反，在大多数情况下，p 值始终小于 0.05。p 值表明本文提出的算法在统计显着性上的其他四种算法明显不同。此外，L21FS 的标准偏差总是小于比较方法，这表明 L21FS 比其他方法更稳定，鲁棒性更好。

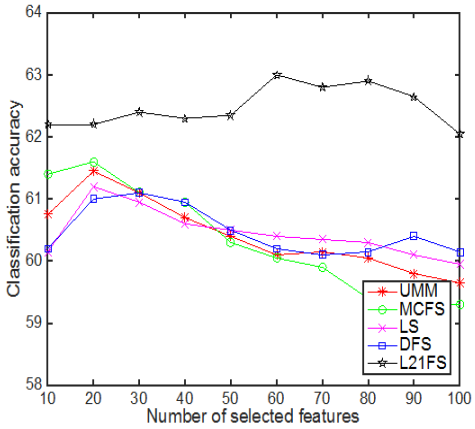
当关注计算消耗时，可以发现 MCFS 和 LS 比其他三种算法消耗更少的时间。这可以用时间复杂度来解释。对于 MCFS 来说，其时间复杂度大约是$O(n^2m)$。对于 LS，最耗时的步骤是计算瑞利商，相应的时间复杂度是$O(n^3)$。考虑到其他三种算法都是迭代方法，从这方面可以很好地解释时间差异。

此外，将本文的算法与其他四种算法进行比较不同数量特征情况下的分类性能。分类精度与所选特征数的变化如图 5-3 所示。

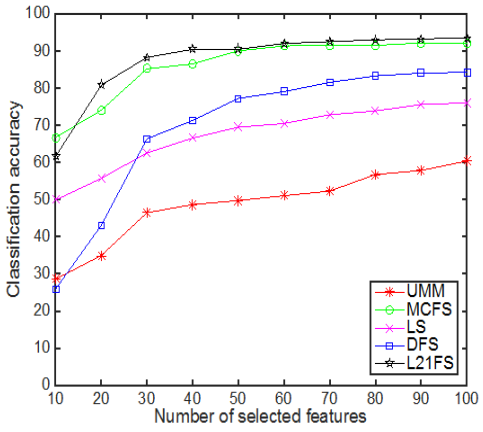
从图 5-3 可以看出，与其他特征选择方法相比，L21FS 在低维子空间中通常可以获得更高的分类准确率。这种现象在六个数据集上是一致的，尤其在 COLON，COIL20 和 MADELON 数据集中更为突出。图 5-3 所示的结果在视觉上表明，L21FS 确实比先前的算法具有更好的特征选择能力。



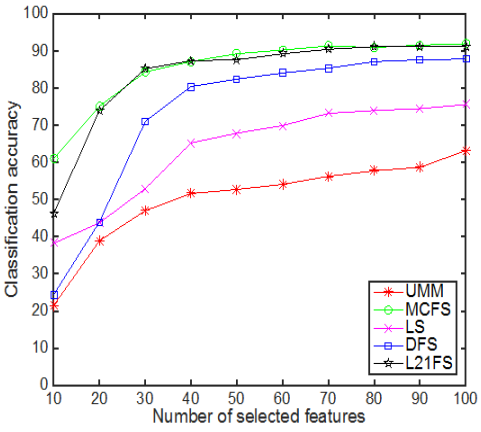
(a)USPS.



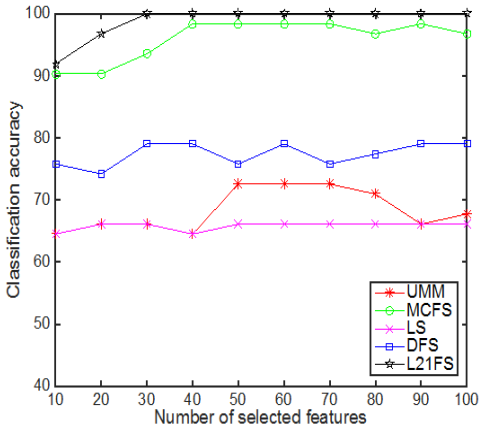
(b)MADELON.



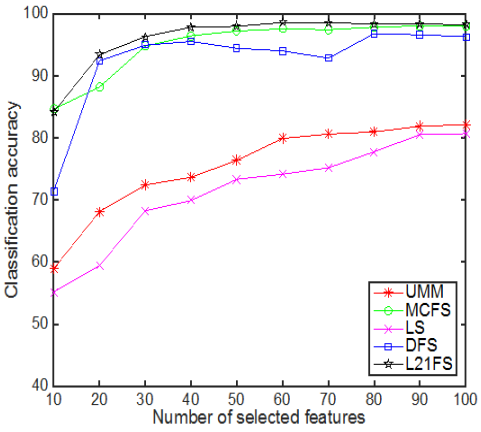
(c)ISOLET.



(d)ISOLET5.



(e)COLON.



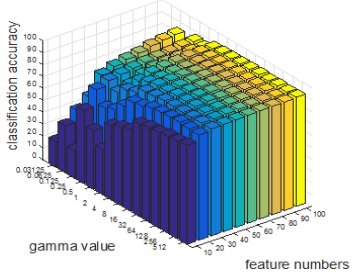
(f)COIL20.

图 5-3. 分类精度 VS 特征数目

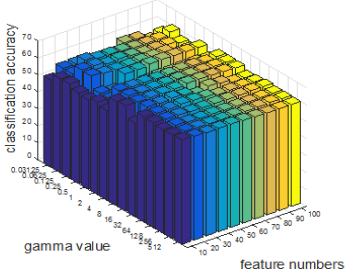
Fig. 5-3 Accuracy VS feature numbers

5.3.5 参数影响

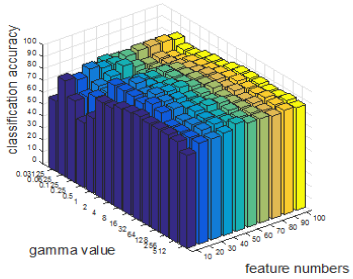
在这种新方法中，只有一个参数γ，可以平衡 L21FS 公式的稀疏性和凸性。γ值越大，L21FS 越稀疏，也就是说，投影矩阵的更多行被迫为零。在本小节中，主要关注γ对新方法性能的影响，改变γ从最小值到最大值的值，每个区间的值乘以 2。为了保持普遍性，在数据集COLON，ISOLET5和COIL20的所有实验中选择前[10,20,30,40,50,60,70,80,90,100]个特征。性能差异与所选特征的数量如图 5-4 所示。



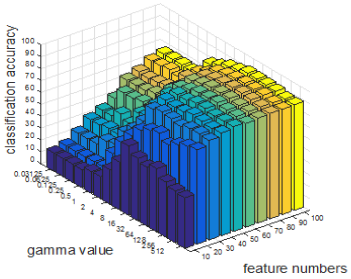
(a)USPS



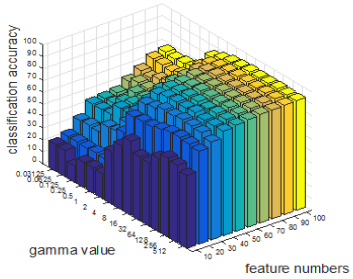
(b)MADELON



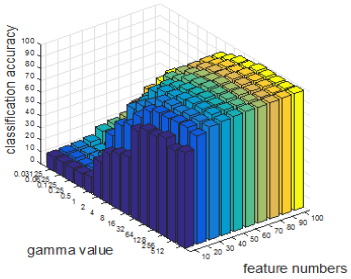
(c)LUNG_DISCRETE



(d)ISOLET5



(e)ISOLET



(f)COIL20

图 5-4. 算法性能变化与参数□的关系图.

Fig.5-4 Method performance w.r.t. paramter □

如图 5-4 所示，L21FS 在不同数据集上性能变化趋势都是相似的，但具有不同的最优参数γ。总体而言，具有相同数量的特征，γ值越大，准确度越高。值得注意的是，当γ数值达到 16 或者 32 时，精度已经达到了很高的水平，后续的增加对精度影响不大。此外，当选择的特征数量很少时，本文的方法的性能对该值更敏感。也就是说，由正则化参数γ引起的算法性能差异与所选特征的数量相关。

5.3.6 噪声数据上算法比较

由于本文提出的方法是一种鲁棒的特征选择方法，因此必须对噪声数据进行实验。为了仿真噪声数据样本，通过生成的噪声矩阵$\tilde{X} \in R^{m\times n}$来融入原始输入数据集$X = [X_1; \cdots; X_m] \in R^{m\times n}$，模拟整体噪声样本数据，噪声矩阵的元素是满足独立同分布的标准高斯变量。然后对模拟的样本数据$X + \theta\tilde{X}$进行与原始数据 X相同的实验，其中$\theta = nf\frac{\|X\|_F}{\|\tilde{X}\|_F}$并且 nf是一个给定的噪声因子。在本小节中，设定在所有的实验中$nf = 0.1$。使用与之前相同的实验设置将本文的方法与其他四种方法进行比较，并将结果汇总在表 5-4 和表 5-5 中。

表 5-4 噪声数据下选取 20 个特征的性能. (平均精度± 方差, 时间: 秒, p-value)

Tab.5-4 The performances of the 20 selected features with noise data(Average±STD, time: s, p-value)

	UMM	MCFS	LS	DFS	L21FS
USPS (2007x256, class:10)	73.64±2.33	87.94±0.89	53.71±1.55	**89.33±1.20**	89.2±1.75
	3.0750	0.1313	0.1311	**0.1498**	0.1754
	5.1935e-6	0.2255	1.5484e-9	**0.9284**	—
MADELON (2000x500,class:2)	61.05±1.60	60.20±1.20	60.95±1.52	56.90±1.67	**61.40±1.92**
	5.1963	0.0373	0.1935	0.5634	**0.9652**
	0.7870	0.3222	0.7239	0.0078	—
LUNG_DISCRETE (73x325,class:7)	67.14±7.66	76.57±6.16	57.52±12.17	67.04±8.18	**84.95±2.51**
	0.3043	0.0241	0.0049	0.2602	**0.8871**
	0.0022	0.0360	0.0022	0.0031	—
ISOLET5 (1559x617,class:26)	38.99±4.58	71.13±3.22	42.07±3.73	53.36±1.76	**74.66±1.52**
	4.7873	0.5766	0.1471	1.2686	**2.3271**
	4.3673e-7	0.0834	2.1851e-7	8.4249e-8	—
ISOLET (1559x617,class:26)	31.02±2.58	73.91±5.82	54.55±4.78	50.32±3.59	**78.58±3.37**
	4.7768	0.6071	0.1456	1.4804	**2.1248**
	1.6676e-8	0.2019	3.6108e-5	3.0244e-6	—
COIL20 (1440x1024,class:20)	67.63±1.85	89.86±2.94	58.19±3.17	89.93±0.93	**92.08±2.38**
	11.6559	0.7300	0.1829	5.7540	**5.7619**
	1.0762e-8	0.1455	3.2531e-8	0.0066	—
COLON (62x2000,class:2)	70.64±16.42	73.97±13.59	61.02±13.74	65.64±18.06	**77.43±8.06**
	46.4025	0.1317	0.0067	37.8078	**40.3341**
	0.4789	0.6730	0.0734	0.2672	—

从表 5-4 和表 5-5 中，可以发现下列的一些现象。首先，本文的方法在 7 个实验数据集中的大多数情况下表现最好，这表明 L21FS 方法比其他比较方法更稳健，并且更有可能在噪声数据上仍能学习到最具判别力的特征。其次，尽管本文的算法在原始数据集上的性能仅略好于其他算法，但本文的算法在噪声数据情况下分类精度下降最小。而且，当仔细观察标准差时，L21FS 的标准差变化总是比竞争方法的标准变化小得多。这也充分证明了本文算法的鲁棒性。

表 5-5 噪声数据下选取 40 个特征的性能. (平均精度± 方差, 时间: 秒, p-value)

Tab.5-5 The performances of the 40 selected features with noise data(Average±STD, time: s, p-value)

	UMM	MCFS	LS	DFS	L21FS
USPS (2007x256, class:10)	83.20±1.82	89.93±1.71	70.20±3.10	**92.17±1.41**	90.78±1.09
	2.9615	0.3138	0.1310	**0.1548**	0.2387
	9.8808e-5	0.4292	1.5564e-6	**0.1571**	—
MADELON (2000x500,class:2)	60.60±1.98	59.00±2.23	60.50±1.70	58.70±1.65	**60.70±1.95**
	4.8088	0.0567	0.1826	0.5200	**0.5276**
	0.9446	0.2857	0.8813	0.1573	—

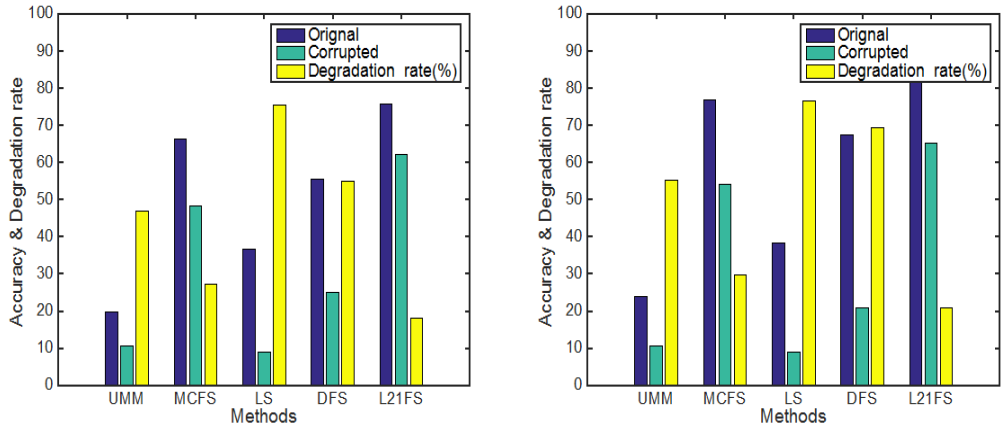
	UMM	MCFS	LS	DFS	L21FS
LUNG_DISCRETE (73x325,class:7)	69.80±8.49	76.57±7.47	64.28±9.20	72.47±9.15	**83.42±5.78**
	0.3220	0.0436	0.0024	0.2646	**0.4309**
	0.0293	0.1848	0.0078	0.0778	—
ISOLET5 (1559x617,class:26)	50.61±4.74	84.46±4.58	62.15±1.16	65.29±4.17	**85.8±2.21**
	4.7239	1.2656	0.1414	1.5309	**2.0919**
	8.7922e-7	0.5914	6.1962e-8	2.3321e-5	—
ISOLET (1559x617,class:26)	45.00±2.17	87.37±1.67	64.03±2.23	70.25±5.06	**88.58±2.17**
	4.7068	1.1739	0.1444	1.2809	**2.3248**
	2.5932e-9	0.4005	2.6152e-7	1.6017e-4	—
COIL20 (1440x1024,class:20)	72.98+2.69	95.69±0.71	67.22±3.23	93.05±1.45	**96.52±1.38**
	11.4180	1.4816	0.1872	5.7293	**7.7865**
	2.9313e-7	0.3171	1.7285e-7	0.0087	—
COLON (62x2000,class:2)	72.30±2.69	78.84±10.27	64.23±15.89	72.30±15.75	**82.17±8.29**
	11.4180	0.0787	0.0064	39.5869	**41.0819**
	0.3634	0.6272	0.0802	0.2996	—

5.3.7 噪声特征实验

如前所述，本文的方法不仅在正则化项中替换了距离度量方法，而且还替换了了学习目标的距离度量方法。*L21*范数距离特征选择方法对于异常值和噪声特征都是鲁棒的。因此，在本小节中，进行实验来测试本文提出的方法在人脸图像数据集（ORL）上的特征鲁棒性。为了评估特征的鲁棒性，将大小为8×8的黑色方块随机放置在图像上以模拟噪声的特征。图 5-5 显示了被黑块遮挡的图像。分别选择了前 20 个和前 40 个特征进行分类，并且将精确度和精度下降率汇总在图 5-6 中。

图 5-5 六个随机选择的人像 VS 对应被随机遮罩的人像

Fig.5-5 Six randomly selected portraits VS corresponding to randomly masked portraits



(a)前 20 个特征.　　(b) 前 40 个特征.

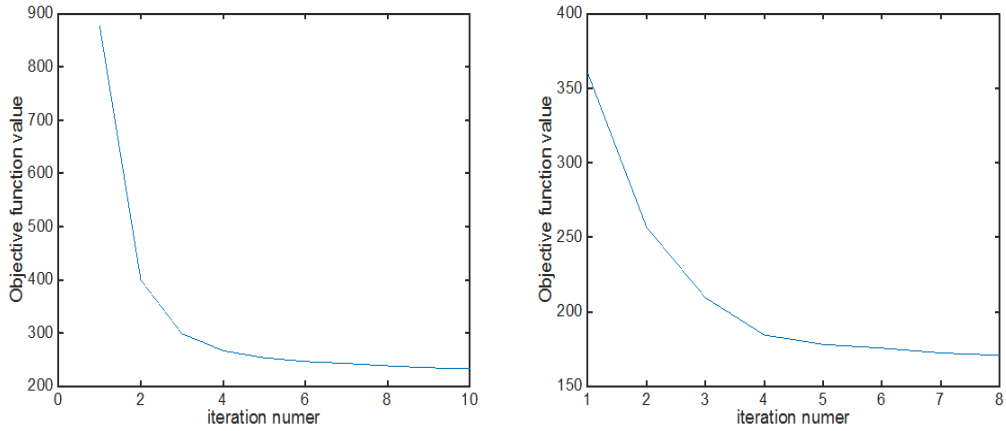
图 5-6 20 特征和 40 特征下的精确度和精确度下降率

Fig.5-6 Accuracy and Degradation Rates for 20 Features and 40 Features

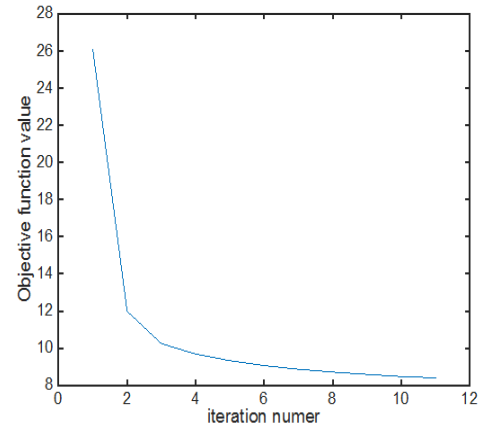
如图 5-6 所示，本文的方法无论是在原始图像上还是特征噪声图像上都比其他算法表现出了更好的分类精度。而且，当关注精度下降率时，本文提出的方法的性能下降率很小，这提供了更具体的证据来支持 L21FS 的鲁棒性。

5.3.8 收敛性分析

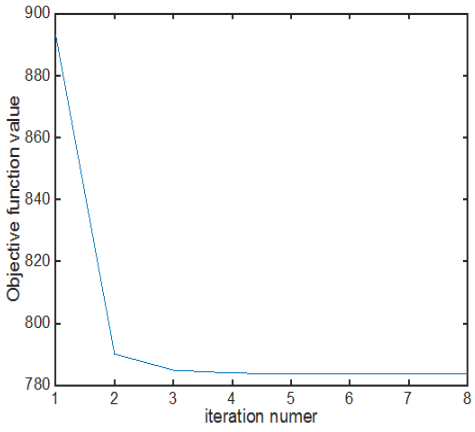
最后，通过在实验数据集上目标函数值的变化曲线来评估本文提出的方法的计算效率。如前所述，目标函数值最终将收敛到局部最优值。 因此，迭代次数是方法的效率中最重要的部分之一，它决定了算法收敛的速度。图 5-7 绘制了 6 个数据集的目标函数值收敛曲线。



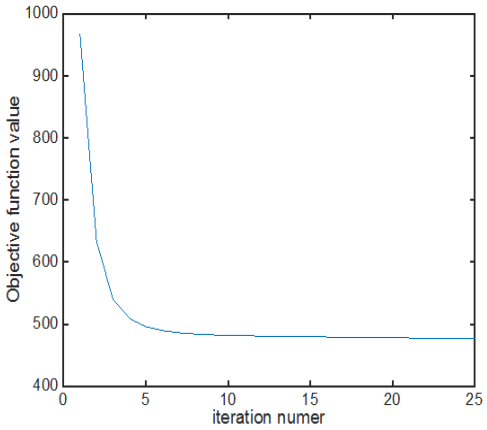
(a)USPS.　　(b)MADELON.



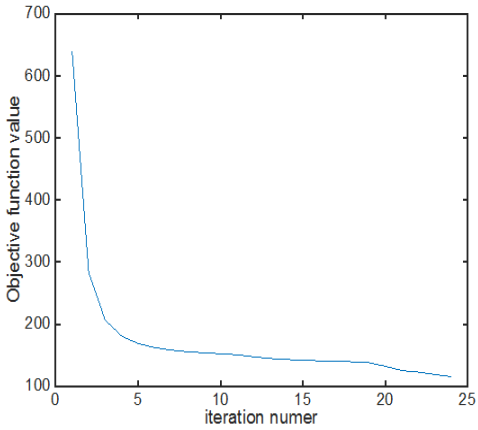
(c)LUNG_DISCRETE.



(d)ISOLET5.



(e)ISOLET.



(f)COIL20.

图 5-7 目标函数值 VS 迭代次数

Fig.5-7 Objective value vs iteration numbers

如图 5-7 所示，本文提出的方法在所有数据集上的目标函数值随着迭代过程而不断减小，这与理论分析完全一致。此外，可以发现该算法通常在约 7 次迭代内收敛到局部最优。这种少量的迭代数确保了本文提出的算法的效率和可行性。因此，由于收敛速度快，本文提出的 L21FS 方法在实践中表现良好。

5.4 本章总结

本章提出了一种结合传统特征抽取方法 LDA 和 L21 范数距离的特征选择方法。与以前的带 L21 范数正则化项的特征选择方法不同，在学习函数中也使用了 L21 范数距离。这种新颖的目标函数导致了一个非光滑的非凸优化问题。为了最大化类内散度值的 L21 范数距离并同时最小化类间散度值的 L21 范数距离，本文引入了一种有效的迭代算法。严格的理论收敛性证明和时间复杂度分析表明 L21FS 具有高效，快速的收敛性。广泛的实验结果表明，本文提出的方法相比相关最先进的方法更加高效。此外，对各种类型数据集的实验表明，本文提出的方法对噪声数据和噪声特征都是鲁棒的。

第六章　结束语

在模式识别应用里，距离度量是各种算法中一项必不可少的工作。传统的机器学习算法往往因为平方 L2 范数距离度量所具有的凸性以及便于求解的特点，而采用平方 L2 范数距离作为其距离度量标准。但是随着科学技术的不断发展，人们对算法的泛化能力要求越来越高，由于现实世界的数据往往伴随着噪声或者野值的存在， 使得传统的基于平方 L2 范数距离度量的算法不能满足现实环境的需求。因此，寻找一种合适的距离度量标准非常必要。本文对传统的分类算法中的孪生支持向量机以及特征选择算法中的判别特征选择通过 L2P 范数距离加以改进，使得算法具有较高的鲁棒性与稀疏性，从而大大提高了算法的表现性能。

6.1 本文主要完成工作

本文主要对 TWSVM 以及 DFS 等算法通过 L2P 距离度量进行改进。不同于之前流行的在正则项中应用范数距离度量，在整个学习函数中使用了 L2P 范数距离度量，这使得本文的分类算法 L2P-TWSVM 具有更好的鲁棒性以及本文的特征选择算法 L21FS 具有更好的稀疏性。

针对基于 L2P 范数距离度量的 TWSVM，通过采用在学习函数中使用 L2P 范数距离，重新规划了点到分类平面的距离度量，有效抑制了噪声数据点到平面距离过大带来的影响。本文设计了一个简单有效的针对 L2P 范数 TWSVM 的迭代算法，使得目标函数值能够收敛到一个局部最优解。理论证明了该算法的可行性，实验结果表明了算法的有效性。

针对原先的 DFS 特征选择算法，理论分析证实了其学习函数仍然不够鲁棒，因此本文采用 L21 范数距离重新度量了其类间散度以及类内散度的计算方式。通过固定其中一项，实现了同时最大最小化 L21 范数项的目标。并且针对 L21FS 设计的迭代算法能够使得在极少次数的迭代后找到一个局部最优解。改进的 L21FS 不仅对噪声数据鲁棒，而且对噪声特征鲁棒，这使得算法性能相比传统的特征选择算法大大提高。由于是基于 L21 范数，本文的 L21FS 具有更好的稀疏性，因此能够寻找到最具有代表性的特征。

6.2 未来工作展望

1、本文在处理奇异性问题时，都是通过添加正则项来避免奇异性。但是通过这种方法往往会导致算法的精确度下降，如何寻找一个新的方法来解决奇异性问题仍是今后值得思考的问题。

2、在改进 DFS 特征选择算法时，本文直接使用的是 L21 范数距离。众所众知，L21 范数是 L2P 范数的一个特例，如果将 L21FS 推广至 L2P 范数的特征选择将是今后需要继续关注的问题。

3、在决定 L2P 范数的 p 值时，通过不同的 p 值下的算法表现来决定 p 值的大小。但是这仍然不够具有代表性。因此，p 值对算法性能的影响以及 p 值的确定仍是今后需要研究的一个重点问题。

攻读硕士学位期间的研究成果和发表的论文

已发表论文：

[1] 马旭, 刘应安, 业宁,等. 基于核PCA与SVM算法的木材缺陷识别[J]. 常州大学学报(自然科学版), 2017, 29(3):60-68.

[2] Xu Ma, Yingan Liu, Qiaolin Ye. P-Order L2-norm distance Twin Support Vector Machine[M]// The 4th Asian Conference on Pattern Recognition,2017 (ACPR 2017). (EI)

[3] Xu Ma, Qiaolin Ye, He Yan. L2P-Norm Distance Twin Support Vector Machine[J]. IEEE Access, 2017, 5: 23473-23483.

在审论文：

[1] Xu Ma, Qiaolin Ye, Yingan Liu, He Yan, Robust Feature Selection via L21-Norm Minimization and Maximization,Neurocomputing. (Under review)

参考文献

[1]杨健. 线性投影分析的理论与算法及其在特征抽取中的应用研究[D]. 南京理工大学, 2002.

[2]张小洵, 贾云得. 基于互补子空间线性判别分析的人脸识别[J]. 北京理工大学学报, 2006, 26(3):206-210.

[3]刘勇进, 赵敬红. 基于稀疏恢复的l_1范数凸包分类器在人脸识别中的应用[J]. 沈阳航空航天大学学报, 2016, 33(1):42-46.

[4]丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1):2-10.

[5]杜树新, 吴铁军. 模式识别中的支持向量机方法[J]. 浙江大学学报(工学版), 2003, 37(5):521-527.

[6]杨绪兵, 潘志松, 陈松灿. 半监督型广义特征值最接近支持向量机[J]. 模式识别与人工智能, 2009, 22(3):349-353.

[7]高斌斌, 刘霞, 李秋林. 改进孪生支持向量机的一种快速分类算法[J]. 重庆理工大学学报, 2012, 26(11):98-103.

[8]丁世飞, 张健, 张谢锴,等. 多分类孪生支持向量机研究进展[J]. 软件学报, 2018(1):89-108.

[9]王娟, 慈林林, 姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27(12):68-71.

[10]张静远, 张冰, 蒋兴舟. 基于小波变换的特征提取方法分析 [J]. 信号处理, 2000, 16(2): 156-62.

[11]Alalga A, Benabdeslem K, Taleb N. Soft-constrained Laplacian score for semi-supervised multi-label feature selection[J]. Knowledge & Information Systems, 2016, 47(1):75-98.

[12]Guyon I, Elisseeff A. An Introduction to Variable Feature Selection[J]. Journal of Machine Learning Research, 2003, 3:1157-1182.

[13]胡正平, 王玲丽. 基于 L1 范数凸包数据描述的多观测样本分类算法[J]. 电子与信息学报, 2012, 34(1):194-199.

[14]Huang H, Feng H, Peng C. Complete local Fisher discriminant analysis with Laplacian score ranking for face recognition[J]. Neurocomputing, 2012, 89(10):64-77.

[15]Yu L, Zhang M, Ding C. An efficient algorithm for L1-norm principal component analysis[C]，IEEE International Conference on Acoustics, Speech and Signal Processing, 2012:1377-1380.

[16]Yang M S, Tsai H S. An Alternative Fuzzy Compactness and Separation Clustering Algorithm[M]，Advanced Concepts for Intelligent Vision Systems. Springer Berlin Heidelberg, 2005:146-153.

[17]叶天语. 基于范数与范数均值比较的印刷防伪水印算法[J]. 光电工程, 2011, 38(6):126-133.

[18]Yan H, Ye Q, Zhang T, et al. L1-Norm GEPSVM Classifier Based on an Effective Iterative Algorithm for Classification[J]. Neural Processing Letters, 2017(4):1-26.

[19]刘建伟，李双成，付捷,等. L1 范数正则化 SVM 聚类算法[J]. 计算机工程，2012, 38(12):185-187.
[20]Wang R, Nie F, Yang X, et al. Robust 2DPCA With Non-greedy ℓ_1 -Norm Maximization for Image Analysis.[J]. IEEE Transactions on Cybernetics, 2017, 45(5):1108-1112.
[21]Yi S, Lai Z, He Z, et al. Joint Sparse Principal Component Analysis[J]. Pattern Recognition, 2017, 61:524-536.
[22]谭龙，何改云，潘静,等. 基于近似零范数的稀疏核主成成分算法[J]. 电子测量技术, 2013, 36(9):27-30.
[23]Wang H, Nie F, Huang H. Learning robust locality preserving projection via p-order minimization[C]，Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015:3059-3065.
[24]马丽，董唯光，梁金平,等. 基于随机投影的正交判别流形学习算法[J]. 郑州大学学报(理学版), 2016, 48(1):102-109.
[25]陈兵飞，江兵兵，周熙人,等. 基于稀疏贝叶斯的流形学习[J]. 电子学报，2018, 46(1):98-103.
[26]Lin T, Zha H. Riemannian manifold learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 30(5):796-809.
[27]黄石青. 基于流行学习 LPP 算法与 Dijkstra 算法结合的交通路径控制研究[J]. 科技创新与应用, 2013(31):290-291.
[28]张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1):32-42.
[29]Shao Y H, Deng N Y, Yang Z M. Least squares recursive projection twin support vector machine for classification[J]. International Journal of Machine Learning & Cybernetics, 2016, 7(3):411-426.
[30]高斌斌，王建军. 多分类最大间隔孪生支持向量机[J]. 西南师范大学学报(自然科学版), 2013, 38(10):130-135.
[31]徐金宝，业巧林，业宁. 基于简单特征值问题的修正 GEPSVM[J]. 计算机工程，2009, 35(21):183-185.
[32]MANGASARIAN O L, WILD E W. Multisurface proximal support vector machine classification via generalized eigenvalues [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(1): 69-74.
[33]Li C N, Shao Y H, Deng N Y. Robust L1-norm non-parallel proximal support vector machine[J]. Optimization, 2016, 65(1):169-183.
[34]杨绪兵，陈松灿，杨益民. 局部化的广义特征值最接近支持向量机[J]. 计算机学报，2007, 30(8):1227-1234.
[35]JAYADEVA, KHEMCHANDANI R, CHANDRA S. Twin Support Vector Machines for pattern classification [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(5): 905-910.
[36]Chen X, Yang J, Liang J, et al. Robust and Sparse Twin Support Vector Regression via Linear Programming[J]. Soft Computing, 2014, 18(12):2335-2348.

[37]李凯，李娜，卢霄霞．一种模糊加权的孪生支持向量机算法[J]．计算机工程与应用，2013, 49(4):162-165.
[38]郑奇，段会川，孙海涛．间隔值辅助的 SMO 算法改进研究[J]．计算机工程与应用, 2017, 53(4):64-69.
[39]张召，黄国兴，鲍钰．一种改进的 SMO 算法[J]．计算机科学, 2003, 30(8):128-129.
[40]周晓剑，马义中，朱嘉钢. SMO 算法的简化及其在非正定核条件下的应用[J]．计算机研究与发展, 2010, 47(11):1962-1969.
[41]Tao H, Hou C, Nie F, et al. Effective Discriminative Feature Selection With Nontrivial Solution[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(4):796-808.
[42]Sun S, Xie X, Dong C. Multiview Learning With Generalized Eigenvalue Proximal Support Vector Machines[J]. IEEE Transactions on Cybernetics, 2018, PP(99):1-10.
[43]Zhao C, Changqin W U, Hua G E. Robust L1-norm non-parallel proximal support vector machine via efficient iterative algorithm[J]. Journal of Computer Applications, 2017.
[44]Wang H, Lu X, Hu Z, et al. Fisher Discriminant Analysis With L1-Norm[J]. IEEE Transactions on Cybernetics, 2017, 44(6):828-842.
[45]Kong D, Ding C, Huang H. Robust nonnegative matrix factorization using L21-norm[C]，ACM International Conference on Information and Knowledge Management. ACM, 2011:673-682.
[46]王晓慧．线性判别分析与主成分分析及其相关研究评述[J]．中山大学研究生学刊(自然科学.医学版), 2007(4):50-61.
[47]马旭，刘应安，业宁,等．基于核 PCA 与 SVM 算法的木材缺陷识别[J]．常州大学学报(自然科学版), 2017, 29(3):60-68.
[48]吴晓婷，闫德勤．数据降维方法分析与研究[J]．计算机应用研究, 2009, 26(8):2832-2835.
[49]王和勇，郑杰，姚正安,等．基于聚类和改进距离的 LLE 方法在数据降维中的应用[J]．计算机研究与发展, 2006, 43(8):1485-1490.
[50]宋枫溪，高秀梅，刘树海,等．统计模式识别中的维数削减与低损降维[J]．计算机学报, 2005, 28(11):1915-1922.
[51]Xue B, Zhang M, Browne W N, et al. A Survey on Evolutionary Computation Approaches to Feature Selection[J]. IEEE Transactions on Evolutionary Computation, 2016, 20(4):606-626.
[52]Yao C, Han J, Nie F, et al. Local Regression and Global Information-Embedded Dimension Reduction[J]. IEEE Transactions on Neural Networks & Learning Systems, 2018, PP(99):1-12.
[53]He X, Cai D, Niyogi P. Laplacian Score for Feature Selection.[C]， International Conference on Neural Information Processing Systems. MIT Press, 2005:507-514.
[54]Kwak N. Principal Component Analysis Based on L1-Norm Maximization[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 30(9):1672-1680.
[55]周大可，杨新，彭宁嵩．改进的线性判别分析算法及其在人脸识别中的应用[J]．上海交通大学学报, 2005, 39(4):527-530.
[56]Nemirko A P. Transformation of feature space based on Fisher’s linear discriminant[J]. Pattern Recognition & Image Analysis, 2016, 26(2):257-261.

[57]Tanha J, Someren M V, Afsarmanesh H. Semi-supervised self-training for decision tree classifiers[J]. International Journal of Machine Learning & Cybernetics, 2017, 8(1):355-370.
[58]路翀，徐辉，杨永春. 基于决策树分类算法的研究与应用[J]. 电子设计工程，2016, 24(18):1-3.
[59]黄春华, 陈忠伟, 李石君. 贝叶斯决策树方法在招生数据挖掘中的应用[J]. 计算机技术与发展, 2016, 26(4):114-118.
[60]王飒，郑链. 基于 Fisher 准则和特征聚类的特征选择[J]. 计算机应用，2007, 27(11):2812-2813.
[61]Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data[C]， ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010:333-342.