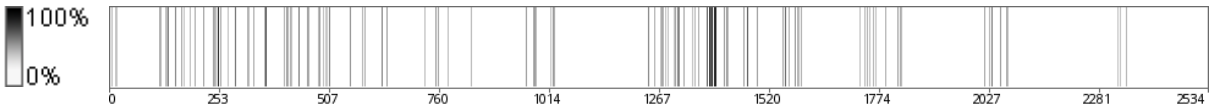
PaperPass旗舰版检测报告
简明打印版

比对结果(相似度):

总　体：12% (总体相似度是指本地库、互联网的综合对比结果)
本地库：12% (本地库相似度是指论文与学术期刊、学位论文、会议论文、图书数据库的对比结果)
期刊库： 7% (期刊库相似度是指论文与学术期刊库的比对结果)
学位库： 9% (学位库相似度是指论文与学位论文库的比对结果)
会议库： 2% (会议库相似度是指论文与会议论文库的比对结果)
图书库： 4% (图书库相似度是指论文与图书库的比对结果)
互联网： 1% (互联网相似度是指论文与互联网资源的比对结果)

编　号：5ADEF35DC61B5470D
版　本：旗舰版
标　题：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
长　度：49374字符(不计空格)
句子数：2534 句
时　间：2018-4-24 17:05:33
比对库：学术期刊、学位论文、会议论文、书籍数据、互联网资源
查真伪：http://www.paperpass.com/check

句子相似度分布图:



本地库相似资源列表(学术期刊、学位论文、会议论文、书籍数据):

1.相似度： 1% 篇名：《支持向量机算法及其在入侵检测中的应用》
来源：学位论文 河北大学 2015

互联网相似资源列表：

1.相似度： 2% 标题：《机器学习笔记（八）——决策树模型的特征选择 - ...》
http://www.lai18.com/content/9771010.html
2.相似度： 1% 标题：《经典分类算法——决策树 - Lai18.com ...》
http://www.lai18.com/content/7388803.html

全文简明报告:

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范数距离往往会放大噪声数据距离在整体数据距离中占比，导致算法的不鲁棒性。　基

于平方L2范数距离的不鲁棒缺陷，本文在分类问题和特征选择问题上分别通过L2p范数距离和L21范数距离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的分类器。 它的目标函数基于平方L2范数距离。 {47%：由于平方L2范数距离容易受到异常值的影响，因此TWSVM需要一个更加鲁棒的距离度量。} {43%：因为 L2 p范数距离比 L1范数距离或平方 L2范数距离能够更好地抑制异常值的影响，} {44%：因此在本文中，我们提出了一种基于 L2 p范数距离新的鲁棒的孪生支持向量机。} 然而，由于新的目标函数是不光滑并且非凸的，这导致目标问题难以解决。 作为本文一项重要工作，我们系统地推导出一种有效的迭代算法，来解决最小化L2p范数距离的问题。 理论研究表明这个迭代算法对于通过L2p范数取代平方L2范数距离来改进TWSVM是有效的。 {42%：大量的实验表明，L2p范数距离孪生支持向量机（pTWSVM）可以有效处理噪声数据，并且具有更好的精度。}

{57%：特征选择和特征抽取是降维的两种不同方法。} 但是，它们总是被分开讨论研究。{48%： 特征抽取旨在寻找新的特征子空间，而特征选择致力于选择原始特征集的子集。} {42%：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。} {51%： 新的目标函数能够更好提供算法稳健性。} 然而求解这个目标非常具有挑战性，因为它需要同时最小化和最大化非光滑的L21范数项。 {40%：针对这个问题，我们提出了一种迭代算法来解决L21范数距离的优化问题。} {59%：一系列理论证明了该算法的收敛性和计算效率。} {54%：各种数据集的实验结果证明了我们新方法的有效性。}

关键词： {57%： 鲁棒性， 孪生支持向量机， 特征选择， 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 a square L2 norm. 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, this paper improves the robustness of the algorithm by using the L2 p norm distance and the L21 norm distance respectively on the classification problem and feature selection problem .

Twin Support Vector Machine is an efficient classifier that is particularly suitable for XOR data. Its objective function is based on the square L2 norm distance. 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 L2 p norm distance, because the L 2 p 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 not smooth and non-convex, this causes the objective problem to be difficult to solve. 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 support shows that this iterative algorithm is effective in improving TWSVM by replacing the squared L2 norm distance by the L2p norm. 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.

Feature selection and feature extraction are two different methods of dimension reduction. 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 goal 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 to solve the optimization problem of the L21 norm distance. A series of theories proves the convergence and computational efficiency of the algorithm. The experimental results on various datasets prove the effectiveness of our new method.

Keywords: Robust, TWSVM, feature selection, L2p-norm, LDA

目录

第一章 前言5

1.1研究背景及意义5

1.2国内外研究现状6

1.3传统算法的缺陷7

1.4本文主要研究工作7

1.5本文内容安排7

第二章 支持向量机概述8

2.1 传统支持向量机8

2.2 广义特征值支持向量机9

2.3 孪生支持向量机10

2.4 本章小结12

第三章 基于L2p范数距离度量的TWSVM13

3.1 范数定义13

3.2 相关工作14

3.3 L2p-TWSVM模型推导15

3.3.1 模型推导15

3.3.2 迭代算法17

3.3.3 收敛性证明17

3.3.4 核函数L2p-TWSVM18

3.4 L2p-TWSVM算法实验20

3.4.1二进制数据20

3.4.2精度比较21

3.4.3参数p值研究23

3.4.4算法收敛性分析25

3.4.5.噪声数据实验25

3.5 算法总结28

第四章 特征选择概述29

4.1 特征选择与特征提取29

4.2特征选择分类29

4.3 纬度约减算法30

4.3.1 主成分分析法30

4.3.2 线性判别分析法31

4.3.3决策树32

4.4 本章小结32

第五章 基于L21范数距离度量的优化特征选择34

5.1 相关工作34

5.2 L21FS模型推导36

5.2.1 模型推导36

5.2.2 迭代算法39

5.2.3 收敛性证明40

5.2.4 时间复杂度分析41

5.2.5 评价标准41

5.3 L21FS算法实验41

5.3.1数据集描述42

5.3.2 ORL人脸数据集小实验42

5.3.3 Iris 鸢尾花小实验43

5.3.4算法比较44

5.3.5 参数影响47

5.3.6噪声数据上算法比较48

5.3.7噪声特征实验49

5.3.8收敛性分析50

5.4 算法总结51

第六章结束语52

6.1本文主要完成工作52

6.2未来工作展望52

{69%：攻读硕士学位期间的研究成果和发表的论文53}

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。

{52%：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。} {45%：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。} {49%：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。}

{40%：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]。} {46%：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。} {46%：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，支持向量机可以完美的通过内积的形式避开对高维数据的计算。 {44%：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。}

{47%：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。 {56%：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。} {44%：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。} {43%：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。} {63%：因此，学者们对特征选择进行了深入广泛的研究。}

{41%：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。} {43%：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。} {65%：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。} {44%：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。}

{43%：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 L2 p范数距离来改进算法[13]， {41%：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]。 {41%：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。 {60%：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。{54%：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。}

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。 {57%：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。} {49%：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。 {50%：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。}

{47%：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]。} {41%：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， {43%：作者提出了一个 L2 p范数的鲁棒的 LPP（ rLPP）算法来最小化 12范数的 p阶距离，} {41%：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。} 但是，解决这个目标函数非常具有挑战性，因为它不仅非光滑而且非凸。 {43%：因此作者系统地推导出一种有效的迭代算法来解决一般的p阶L2范数最小化问题。}

1.3传统算法的缺陷

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，我们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。 由于传统的基于平方L2范数距离的算法不鲁棒，这导致在实际应用中往往不能得到令人满意的结果。

基于以上的缺陷，本文通过L2p范数距离，着重对分类算法中的TWSVM算法以及特征选择算法中的DFS算法进行研究改进。

1.4本文主要研究工作

在结合了上述前文的算法基础上，本文对分类算法TWSVM和特征选择算法DFS进行了下列的主要研究与改进：

{42%：（1）由于传统算法的学习函数都是基于平方 L2范数距离，然而传统的平方L2范数距离容易收到噪声数据以及离群数据的影响，} 使得算法不具有鲁棒性。 因此，本文从算法的公式推导中，通过L2p范数（包括L21范数）对算法目标函数进行了改进。

（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。 {43%：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，我们都对其收敛性以及时间复杂度在理论上进行了研究。

{44%：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，我们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：　{42%：首先介绍了本文的研究背景，然后介绍了模式识别中相关的一些算法研究现状，并依据介绍的内容，}　{52%：对介绍的相关算法进行了总结和概括，为后文的工作展开了铺垫。}

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。　{49%：通过对算法的分析，发现其中的不足之处，为后续对其进行改进做铺垫。}

第三章：　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优化目标。　{55%：接下来通过严格的理论证明，证明该算法的收敛性以及时间复杂度上的可行性。}　{42%：最后通过一系列的实验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。}

第四章：　介绍了特征选择的相关工作。　{53%：从降维工程的分类到特征选择和特征抽取的区别。}　并且介绍了一些降维工程中常用的算法，并且揭示了当前特征选择中的不足之处，为后续的改进提供了说明。

第五章：　详细介绍了我们的基于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。并且，由于目标函数的非凸性，本章节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。　理论显示算法严格收敛并且时间复杂度较低，实验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：　{58%：总结分析了全文的工作，并对后续工作进行了展望。}

第二章支持向量机概述

2.1　传统支持向量机

{43%：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，}　{52%：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。

假设有包含个点的数据集

x

1

，

x

2

，，

x

，该数据集可以记为X

，其中为样本个数，为样本维度。 如果第个点

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。 第个点

的标记可以表示为

+1，1

。

为第个样本的标签。 {43%：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

/(2-1)

其中表示平面。 {50%：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}

/ (2-2)

{47%：求公式(2-2)的最大值问题可以转化为如下的最小值问题}

/ (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。 {42%：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

/ (2-4)

其中

为第个样本的松弛变量，C为平衡系数。 {91%：将约束条件加入目标函数中，得到拉格朗日函数：}

/ (2-5)

{50%：将拉格朗日函数对，，，，分别求偏导，可以得到如下公式}

/ (2-6)

/ (2-7)

/ (2-8)

{50%：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}

/ (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。{48%：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。} {69%：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

/ (2-10)

其中

，

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]。 {56%：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。

{42%：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

1

对应的平面法向量为

1

，偏差为

1

，负类样本B

2

对应的平面法向量为

2

，偏差为

2

。 {67%：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；} {63%：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：

/ (2-11)

/(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，我们可以简写为

/ (2-13)

为了防止过拟合问题，我们给GEPSVM的目标加入一个L2正则项

/(2-14)

其中是一个非负的参数。 {48%：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，我们可以通过同样的方式获得。我们定义

/(2-15)

/(2-16)

{45%：其中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

/(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

/(2-18)

{76%：求解的目标即最小特征值对应的特征向量。} 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]。{47%：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。}{51%：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。{44%：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。}

假设有个数据点可以表示为一个矩阵X=

A

1

，

A

2

，

A

，A属于一个维的实值空间X

.同样,

+1，1

{49%：表示对应的第i个样本属于正类或事负类。} {40%：我们假设正类样本为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}

1

，

2

和偏差

1

，

2

：

/(2-19)

/(2-20)

其中

1

，

2

]0是非负的平衡参数,

1

,

2

是维度合适的单位列向量，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刀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。

{46%：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。} {52%：公式(2-19)对应的拉格朗日函数可以表达为如下形式：}

/(2-21)

{83%：其中，为拉格朗日乘子向量。} 通过KKT条件和对每一个变量求导，我们可以得到如下的公式

/ (2-22)

/ (2-23)

/(2-24)

/(2-25)

/(2-26)

/(2-27)

/(2-28)

结合(2-24)和(2-28)，我们可以得到

/(2-29)

接下来，通过将(2-22)与(2-23)相加可以得到

/(2-30)

我们定义

/(2-31)

通过这些定义，我们可以重写公式(2-30)为

/

/(2-32)

通过公式(2-32)可以发现，我们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

进行求逆运算，虽然

是一个半正定矩阵，但是仍有可能在某些情况下奇异。 因此我们给它添加一个正则项I，]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

/ (2-33)

但是在后续的工作中为了方便，我们仍然使用公式(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，我们可以得到第一个TWSVM平面的对偶形式如下：

/ (2-34)

通过 SMO算法[38-40]利用凸二次规划求解公式(34)，我们可以得到最优的值， 带入公式(32)或者公式(33)可以得到我们所求平面1的法向量

1

和偏差

1

。 同理，我们可以通过同样的方法来得到平面2 的法向量

2

和偏差

2

.

{49%：一旦两个平面确定，我们便可以确定一个新的点的类别。} 假设一个新的点

{46%：，那么我们可以通过这个点到两个平面的距离来判断新的点的类别：}

/(2-35)

如果点离平面1近离平面2远，那么点属于平面1对应的正类样本。 如果点离平面1远离平面2近，那么点属于平面2对应的负类样本。

2.4 本章小结

{53%：本章节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。} {54%：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，我们通过拉格朗日函数采用对偶形式来求解。 {47%：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。} {76%：通过求解一对广义特征值问题来获得两个非平行的分类平面。} {40%：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。 {69%：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。} {46%：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。}

但是我们可以注意到，上述的算法都是基于平方L2范数距离(欧式距离)进行求解。{46%：平方L2范数具有凸函数的性质，因此便于求解目标函数。} {42%：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算法的不鲁棒的问题，在下一章节中，本文将通过 L2 p范数距离来重新定义目标函数， {60%：提高算法的鲁棒性与算法的泛化能力。}

第三章基于L2p范数距离度量的TWSVM

3.1 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：

/ (3-1)

正齐次性:

/ (3-2)

三角不等式:

/ (3-3)

{49%：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。}

{65%：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：}

/(3-4)

{42%：向量的L2范数即传统的欧里几得距离(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：}

/(3-5)

{54%：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：}

/ (3-6)

{44%：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：

/ (3-7)

假设有矩阵X=

1

,

1

, ,

包含个数据点，并且每一个数据点

，那么矩阵X则是一个nd大小的矩阵。

,

{72%：代表矩阵X第i行第j列对应的元素。} {52%：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。}

{56%：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}

/(3-8)

{48%：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：}

/(3-9)

其中为

的最大特征值。 {44%：矩阵的无穷范数又称行和范数，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：}

/ (3-10)

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：

/ (3-11)

{41%：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 L2 p范数来优化改进一些模式识别算法， {57%：使得算法具有更好的泛化性和更好的效果。}

3.2 相关工作

{43%：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。 {44%：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。} {42%： 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。 {40%：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. {60%： Fungand等人提出了一种近似支持向量机算法(PSVM)。} {44%：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。 {54%： 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。 {42%： 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。} {41%： 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。}

{46%：大量研究表明，使用 L1范数最小化和非平方 L2范数(L2 p范数，} 0[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。 {40%： 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。 {40%： 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。 {43%：行向量将通过列向量经由一个上标转置符号来定义。} {42%： 我们假设A表示正类的样本矩阵并且B表示负类的样本矩阵。} {60%：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m1n和m2n。 对于一个矩阵A，

A

i

表示矩阵的第行在实值空间

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：

/(3-12)

平方L2范数的公式表达可以推广至p序L2范数（L2p范数）：

/(3-13)

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。{47%：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。

3.3 L2p-TWSVM模型推导

3.3.1 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。 我们通过TWSVM获得的分类结果可能会被异常值所明显地影响。 {45%： 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。} 如果我们能找到合适的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的百分比。

TWSVM的改进可以通过解决以下问题来表示：

/(3-14)

/(3-15)

公式(3-14)的拉格朗日函数为

/(3-16)

其中，为拉格朗日乘子。

我们注意到公式(3-16)涉及到L2p范式，因此这个函数很难直接求解。 {42%：针对这样的问题，我们将含有L2p范数的项拆分为平方L2范数和(2)次方的L2范数的乘积：}

/ (3-17)

我们定义

/(3-18)

{56%：那么拉格朗日函数(51)可以重写为以下形式}

/(3-19)

我们对每一个参数进行求导计算，加上KKT条件，我们可以得到下列的公式

/ (3-20)

/ (3-21)

/ (3-22)

/ (3-23)

通过公式(3-22)和（3-23）我们可以得到

/ (3-24)

为了简化公式，我们定义

/ (3-25)

因此，公式(3-18)可以表达为

/ (3-26)

将公式(3-20)和公式(3-21)相加，我们可以得到

/ (3-27)

这个可以简化表达为

/ (3-28)

由公式（3-28）我们可以得到u的解析解为

/(3-29)

尽管

H

T

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，我们给公式(3-29)添加一个正则项，如下所示：

/(3-30)

其中]0并且I是一个合适维度的对角矩阵。

{46%：通过拉格朗日函数和KKT条件我们可以得到原L2p范数TWSVM的对偶问题的最小化形式，即}

/(3-31)

{45%：通过公式(3-31)，我们可以求解一个凸二次规划问题来得到最优的，} {43%：然后带入公式(3-29)即可得到第一个分类平面的法向量和偏差的值。} 同理，我们可以通过同样的方式来求解第二个平面。

3.3.2 迭代算法

但是我们注意到，是关于u的未知变量，因此我们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和在每次迭代中自动变化直至迭代收敛。

{48%：算法1 一个迭代算法解决L2p范数距离TWSVM问题}

Algorithm.1 An iterative algorithm for L2p norm distance TWSVM

输入： 训练数据集A

R

m1n

， B

R

m2n

，参数p，c1，c2。

步骤一： 计算H

R

m1

n+1

G

R

m2

n+1

I

R

n+1

n+1

步骤二： 初始化向量u

R

n+1

1

。

循环至收敛

步骤三： 计算=

H

u

T

2

p2

。

步骤四： {62%：通过公式(66)计算拉格朗日乘子。}

步骤五： 更新u，如果需要添加正则项。

结束循环

输出： u

R

n+1

1

同样，另一个平面向量可以通过同样的程序求得。

3.3.3 收敛性证明

{53%：为了证明这个算法的收敛性，我们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[2，下列不等式成立：

/(3-32)

证明： 假设函数f

x

=2

x

p

2

px+p2，我们对此函数进行求导，可以得到

/(3-33)

/(3-34)

很明显，当x]0并且0[2

f

x

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

p

x

2

+p20。 因此，可以得到以下公式：

/(3-35)

/

理论1： 该算法可以在每次迭代中单调地减小问题（3-14）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(3-14)用G，H表达，可以改写为

/(3-36)

公式(3-14)和公式（3-36）等价，这里用J来表示这个目标函数的值。 假设

u

是下一次迭代的的u的值，那么

/(3-37)

结合公式(3-35)我们可以得到

/

/

/(3-38)

根据定理1我们有

/(3-39)

结合公式(3-38)和公式(3-39)，我们可以得到

/

/

/(3-40)

{49%：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。

3.3.4 核函数L2p-TWSVM

{42%：为了将L2p范数距离TWSVM推广至非线性分类，我们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

/(3-41)

其中

C

T

=

A

T

;

B

T

，K

表示任意选择的核函数。 如果K

是一个线性核函数如K

x

T

，

C

T

=

x

T

{41%：C，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。}

我们构建如下最优化的核函数L2p范数距离TWSVM的目标函数：

/(3-42)

/(3-43)

{55%：公式(3-42)对应的拉格朗日函数可以表达为：}

/(3-44)

{43%：为了便于求解这个拉格朗日函数，我们将包含L2p范数距离的项拆分为如下形式：}

/(3-45)

{43%：公式(3-45)中，我们可以将乘积的前一项用来表示，那么新的拉格朗日函数可以表达为：}

/ (3-46)

我们可以通过求导以及KKT条件得到下列条件：

/(3-47)

/(3-48)

/(3-49)

/(3-50)

将公式(3-47)和公式(3-48)相结合，可以得到

/(3-51)

为了简化公式，我们定义

/(3-52)

并且用向量u=

1

,

1

来表示此分类平面。 因此，公式(86)可以改写为：

/(3-53)

由公式(3-51)我们可以得到关于超平面向量u的解析解：

/(3-54)

这样核函数的L2p范数距离TWSVM的最小化对偶形式为

/(3-55)

通过同样的方法，我们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，那么矩核技巧可以用来降低L2p范数距离TWSVM的维数。 {54%： 在线性情况下，正则化项往往能提高算法的性能。}

3.4 L2p-TWSVM算法实验

3.4.1二进制数据

为了直接比较TWSVM和L2p范数距离TWSVM之间的差异，我们对人造数据集进行了一个小实验。 我们构建了一个数据集，它包含两类数据，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。 {50%： 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。 {47%： 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3-1： 异或数据实验分类平面

Fig.3-1 classfication surfaces on XOR data

图3-1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3-1中分类平面几乎相同，结果符合我们的预期猜想。

为了引入野值，我们模拟了一些数据点，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。 {54%： 图3-2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3-2： 存在野值的异或数据实验分类平面

Fig.3-2 classfication surfaces on XOR data with noise data

从图3-2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。 {41%： 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。}

3.4.2精度比较

{43%：在本节中，我们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3-1数据集描述

Tab.3-1 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

{52%：为了公平起见，每个比较的算法都使用线性核。} 我们将我们提出的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 。 我们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。 {42%：我们在表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，我们进行了配对T检验，将这些方法与我们的新方法进行比较。 当配对 T检验中的 p- value[0.05时，我们认为该算法与我们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。

{44%：表格3-2算法分类精度比较(平均精度 标准差，时间：} 秒，p-value值)

Tab.3-2 Methods Comparision(AverageSTD， 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

heart

0.842.77 0.1623

0.785.56 0.0132 0.0720

0.823.95 0.0071 0.5675

0.823.00 0.9383 0.4698

0.794.38 0.7859 0.1113

0.672.48 0.0427 5.97e-5

australian

0.842.52 1.2176

0.674.96 0.0211 8.62e-6

0.844.00 0.1180 0.8613

0.851.65 8.1210 0.6857

0.664.66 1.0614 1.65e-6

0.573.06 0.8684 2.55e-7

pima

0.763.82 1.1706

0.754.05 0.0137 0.5455

0.752.30 0.0412 0.5268

0.753.43 1.8497 0.5713

0.744.24

0.9329 0.3572

0.744.26 0.9378 0.3558

monk1

0.707.07 0.3543

0.793.98 0.0125

5.06e-7

0.703.18 0.0934

0.7047

0.559.29 0.1614

5.10e-7

0.762.29 0.8432

0.0515

0.664.55 0.0777 0.1641

sonar

0.6810.04 0.3965

0.714.88 0.0158

0.0816

0.685.55 0.0079

0.8062

0.743.57 1.5948 0.0293

0.729.52 4.2953 0.0184

0.726.15 0.0257 0.2800

spect

0.791.50 0.1442

0.584.83 0.0187 2.02e-6

0.795.49 0.0062 0.9740

0.714.40 1.5253 0.0041

0.785.09 2.7397 0.6591

0.795.49 0.0253 0.9951

cancer

0.961.28 1.4262

0.917.14 0.0159 0.0033

0.961.63 0.0925 0.9934

0.971.16 0.2452 0.6237

0.952.26 1.0705 0.4251

0.951.80 0.3123 0.3608

ionodata

0.901.90 0.2017

0.824.49 0.0140 3.19e-4

0.855.65 0.0094 0.0121

0.863.17 1.4361 0.0204

0.794.40 2.1234 6.43e-4

0.865.68 0.3446 0.2272

haberman

0.6319.51 0.1335

0.754.78 0.0123 1.80e-4

0.735.17 0.0079 0.0014

0.6421.10 0.2823 0.6761

0.745.02 0.7074 6.57e-4

0.735.28 0.0204 0.0105

monk3

0.825.97 0.6786

0.872.17 0.0142 0.0908

0.782.78 0.0361 0.0240

0.483.51 0.1020 8.39e-9

0.793.63 0.8342 0.0619

0.773.37 0.5351 0.0323

wpbc

0.785.84 0.1236

0.727.38 0.0131 0.0042

0.767.06 0.0060 0.1583

0.736.79 1.8354 0.0298

0.766.56 1.4888 0.1706

0.767.59 0.0634 0.5226

bupa

0.693.35 0.2546

0.545.15 0.0118 3.99e-5

0.674.72 0.0091 0.1921

0.666.25 0.8766 0.0232

0.53914.03 0.7477

4.03e-6

0.627.25 0.0975 0.1008

checkdata

0.534.87

1.3981

0.575.84 0.0197 0.0134

0.504.76 0.0785 0.0703

0.514.74 0.6881 0.1096

0.525.78 0.9978 0.4931

0.513.90 0.5537 0.4245

从表3-2我们可以发现，与其他几种算法相比，L2p范数距离TWSVM在绝大多数数据集上表

现最好。 单独比较 pTWSVM和 TWSVM，我们可以发现 pTWSVM总是比 TWSVM分类更准确，虽然它不是在每一个数据集上都好， 但是在不好情况中，配对 T检验表明算法的精度只有小于0.1%的差异，这是可以被忽略的。 这种情况可以解释为TWSVM是L2p范数距离TWSVM的特例。 {46%：当L2p范数距离TWSVM的参数p固定为2时，L2p范数距离TWSVM将退化为TWSVM。} 我们小实验的结果表明，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。 {42%：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。} 这意味着我们提出的新方法具有更好的鲁棒性，并且我们的算法具有更高的稳定性，这符合我们的预期。

{41%：显然，表二中许多 p- value值小于0.05，即在大多数数据集上} {53%：我们提出的 L2 p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 L2 p范数距离 TWSVM和 NLPTSVM的 p- value值分别为0.0121和0.0240， 我们因此得出结论 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，我们观察到在一些数据集中，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 L2 p范数距离 TWSVM与它们的 T检验的 p- value值分别为0.6857和0.6237，这让我们得出结论，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。

当我们关注表3-2所示的运算时间时，我们注意到NLPTSVM总是比L2p范数距离TWSVM更快。这可以从他们的公式来解释。 {40%：虽然它们都是基于传统 TWSVM的迭代算法，但是 L2 p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）。

{42%：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。}

3.4.3参数p值研究

新提出的方法有一个关于如何确定p值的问题。 {46%：考虑到目标函数，我们认为p值受异常值的影响。} {50%：为了获得更高的精确度，噪音的比例越大，p值越小，反之亦然。} {42%：公式(49)很明显的表明p值直接影响目标函数的结果。} 我们将公式分成两部分：{53%：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。因此我们认为参数p值可以直接影响实验精度。

{45%：我们以上述数据集中几个基准数据集为例进行实验。} 为了测量精度的影响，我们将其余参数设置为特定值1=2=1。 然后我们记录不同p值下算法的正确率。 为了研究其对分类性能的影响，我们将p值的目标范围固定在0.1到2之间变化。 {56%：通过实验数据，我们模拟了相应的正确率曲线。} 所有的记录如图3-3所示。

////

//

图3-3： {61%：不同p值下算法正确率折线图}

Fig.3-3: Accracy line with different p

{45%：图3-3显示，参数p值的确定与特定数据集密切相关。} 我们可以在图三中找到两

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，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17，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。 {42%：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。

3.4.4算法收敛性分析

由于该算法是一种迭代算法，因此我们新方法的收敛性是一个重要的问题。 在前文中，我们从理论上严格证明了它的收敛性，现在我们从实验中研究它的收敛性。 我们试验了几个数据集，并且固定p值。 {42%：我们将算法在每次迭代中的目标值绘制在图四中。}

//

//

图3-4 迭代次数 vs. 目标值差异

Fig. 3-4 Iteration numbers vs. objective value difference

{41%：图3-4显示我们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。 根据这些实验结果，我们在实验中设定了一个停止阈值为

10

5

{54%：，这足以在收敛性方面取得令人满意的结果。}

3.4.5.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，因此我们将重点关注在以下异常值处理数据集的实验中。 {47%：为了模拟异常值数据样本，给定输入数据集X=[}

x

1

，，

x

n

]

R

mn

，我们给它加入一个噪声矩阵

X

R

dn

{51%：，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，我们在X+

X

{50%：数据集上执行与原始数据相同的计算程序，其中=nf}

X

F

X

F

并且 nf是一个给定的噪音因子。 在我们所有的实验中，我们设定nf=0.1。 我们将我们的新方法与以前的其他方法进行比较，并将结果总结在表格三中。

表格3-3加入20%噪声的分类效果 (平均精度 标准差，p-value值)

Tab.3-3 Methods Comparision with 20% noise(AverageSTD， 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

0.708.84 —

0.676.68

0.2522

0.681.17

0.4363

0.704.53

0.9993

0.655.92

0.0820

0.662.15

0.0922

australian

0.682.97

—

0.627.25

0.0037

0.654.77

0.1955

0.593.15

3.39e-4

0.6105.52

0.0068

0.573.19

7.08e-4

pima

0.753.20

—

0.723.18

0.2118

0.745.07

0.7443

0.743.31

0.5015

0.722.73

0.2411

0.715.53

0.0258

monk1

0.684.16

—

0.804.21

0.0010

0.652.93

0.1227

0.546.60

1.47e-5

0.802.84

3.41e-5

0.664.55

0.2543

sonar

0.758.01

—

0.709.19

0.0232

0.688.60

0.0127

0.746.90

0.9782

0.732.17

0.3558

0.726.53

0.1043

spect

0.764.35

—

0.555.25

1.68e-6

0.795.01

0.2039

0.725.43

0.0331

0.773.67

0.7666

0.795.49

0.1662

cancer

0.961.93

—

0.950.59

0.9173

0.961.55

0.7347

0.960.99

0.5161

0.951.50

0.5975

0.951.95

0.7156

ionodata

0.902.80

—

0.814.42

1.26e-4

0.864.67

0.0593

0.872.33

0.0948

0.813.75

3.03e-5

0.875.19

0.0874

haberman

0.744.69

—

0.744.06

0.8491

0.725.06

0.3094

0.742.63

0.8794

0.754.66

0.5593

0.725.16

0.3124

monk3

0.865.00

—

0.843.63

0.0813

0.791.93

0.0010

0.7014.79

1.75e-6

0.792.94

2.86e-4

0.773.37

2.29e-4

wpbc

0.797.17

—

0.687.52

1.74e-5

0.745.26

0.0181

0.603.72

5.20e-7

0.765.71

0.0560

0.767.59

0.0166

bupa

0.685.10

—

0.613.73

0.0065

0.6410.58

0.1574

0.644.35

0.1724

0.514.88

1.24e-5

0.636.83

0.0771

checkdata

0.534.51

—

0.574.69

0.0920

0.512.72

0.2867

0.501.99

0.1932

0.531.94

0.8108

0.513.84

0.4727

如表3-3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3.41 e-5， 0.1662，0.5161，0.5593，0.0920。{44%： 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，我们在实验中采取了不同的值。 {46%： 以下图片总结了不同算法在不同值的基准数据集上的性能。}

//

{60%：图3-5不同噪声程度下算法分类精度}

Fig.3-5 accuracy with different noise factor

从图3-5中，我们可以得到以下几点：

首先，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。

其次，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， L2 p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， 但在我们的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。 所以我们提出的方法相比传统TWSVM方法分类精度提高了12.78%=(0.74810.6633)/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%=(0.86670.8296)/0.8296。 这一现象耶普片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。

最后，图3-5同时也显示 L2 p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 L2 p范数距离 TWSVM方法比原始 TWSVM方法更快且更容易趋于稳定。 {47%：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。} 3.5 算法总结

我们提出了一个基于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。 {48%： 与传统的TWSVM相比，新方法具有更多挑战性的优化问题。} {44%： 为了解决这个问题，我们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。 首先，处理奇异性的问题。 在我们的上文的研究中，这是通过正规化项解决的。 其次，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。 最后，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。

第四章特征选择概述

4.1 特征选择与特征提取

{43%：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， {49%：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。} {51%：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。}

{45%：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。} {51%：这使得样本数据的特征工程任务复杂化。} {46%：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。 {48%：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。 {41%：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。 {43%：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。

{42%：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。 {61%：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。 {61%：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。

4.2特征选择分类

{47%：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，{40%：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。} {58%：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。

{50%：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。}

/

{53%：通过特征选择的方式，我们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。 {43%：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。} {62%：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。}

{56%：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吧最终的学习器的分类精度当作特征子集好坏的评价标准。} {60%：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。

{67%：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。

4.3 纬度约减算法

{53%：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。

4.3.1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，{50%：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

，那么样本点

在超平面W上的投影点即

{51%：，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}

最可能的分散。 于是优化目标可以写为：

/(4-1)

对于公式(4-1)，我们将投影向量提出，可以得到

/(4-2)

我们定义求和项为全局散度矩阵

，那么公式(4-2)可以简写为

/(4-3)

{56%：对公式(4-3)使用拉格朗日橙子法可得}

/(4-4)

于是，只需要对散度矩阵

{47%：进行特征值分解，将所有的特征值按照降序排序，再取前个特征值对应的特征向量，组合成}

/(4-5)

即主成分分析的解。

{41%：传统的降维方法降维后的维数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。

/(4-6)

其中

是第i个特征值。

4.3.2线性判别分析法

{45%：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，}
{62%：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=

，

=1

，令

，

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。

{48%：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；} {53%：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：

/(4-7)

我们定义类内散度矩阵为

，内间散度矩阵为

，则公式(4-7)可以简化为

/(4-8)

{45%：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(4-8)，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，我们将公式(4-8)转换成以下形式：

/(4-9)

{75%：对公式(4-9)使用拉格朗日乘子法可得：}

/(4-10)

{54%：我们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。} {41%：我们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。}

4.3.3决策树

{62%：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。} {40%：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。} {73%：对于决策树而言，特征的选择是决定用哪个特征来划分特征空间。} {85%： 特征选择是要选取出对悬链数据集具有分类能力的特征，这样可以提高决策树的学习效率。} {100%：如果利用某一个特征进行分类与随机分类的结果没有很大的差别，则称这个特征是没有分类能力的。} 这样的特征可以丢弃。 {100%：常用的特征选择的准则是信息增益和信息增益比。}

信息增益是熵的一种增益变化情况。 {100%：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。} {61%：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：}

/(4-11)

{65%：从公式(4-11)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。} {100%：熵越大，随机变量的不确定性就越大。}

{97%：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。} {91%：

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}
{100%：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：}

/(4-12)

{65%：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。} {95%：根据信息增益准则进行特征选择的方法是：} {90%：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。}

{94%：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。 {95%：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:

/(4-13)

/(4-14)

其中n 是特征A取值的个数。

4.4 本章小结

上述文章介绍了降维工程中相关的一些原理以及常见的算法。 {41%：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。} 但是我们发现传统算法中，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。{41%：基于此问题，我们 在下文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。}

第五章 基于L21范数距离度量的优化特征选择

5.1 相关工作

{45%：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。 {47%：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。 {53%：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。 {42%：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。} {41%：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。}

{57%：Fisher线性判别分析（LDA）是最受欢迎的监督特征抽取方法之一。} {57%：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。} {42%：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。} {44%：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。} {56%：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。} {41%：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。 {44%：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。 {41%：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。

在本文中，我们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。{40%： 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，据我们所知，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。 {50%： 受到DFS的启发，我们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。

{50%：在本节中，我们将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。 假设我们有个属于类的数据点

1

，

2

。 为了方便起见，数据集可以用矩阵表示。 此外，

表示第个数据点并且

表示第 个特征。 {52%： 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，我们引入了基于平方L2范数距离的散度矩阵：

/(5-1)

/(5-2)

/(5-3)

其中，

，

，

{73%：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

类别数

数据点数

总体样本平均值

第个类别

类间散度矩阵

类内散度矩阵

总散度矩阵

投影矩阵

很明显，

是

和

的总和，可以写成：

/(5-4)

LDA的目标函数可以写成如下形式：

/(5-5)

其中就是我们所寻找的投影矩阵。

{51%：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。} {52%： 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：

/(5-6)

{50%：对于投影矩阵，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

，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：

/(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强制投影矩阵独立于

。 此外，L21范数正则项用来代替

，1

范数正则项。 新公式可以写成如下形式：

/(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

，1

范数和

2，1

范数都是

1

范数的拓展，DFS同样也利用了

2，

范数正则项来代替

，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。 {42%： 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：

/

/

/(5-9)

如公式（5-9）所示，很明显，目标函数涉及平方

2

范数项。 众所周知，平方

2

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

在本节中，我们将首先使用21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21范数项并且目标函数不是一个凸函数问题。然后，我们引入一个能够有效解决问题的迭代算法。 接下来我们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21范数，但DFS的学习函数仍是基于平方2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。 在我们的新方法中，21范数距离不仅用于正则化项中，还用于学习函数中。 因此我们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，我们首先给出了一些21范数相关的概念和定义，然后提出我们新方法的目标函数。 {50%： 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=[

]

，我们定义的第行为

，

{66%：表示样本矩阵中的第个数据点。} 同样，我我们用

来表示矩阵的第列，即矩阵的第个特征。 因此，

{52%：表示样本中第个数据点的第个特征。}

对于矩阵，传统的平方L2范数距离定义如下：

/(5-10)

相应的21范数距离定义如下：

/(5-11)

{48%：与传统的线性判别分析算法类似，L21FS也需要通过21范数距离来定义类间散度矩阵和类内散度矩阵。} {41%： 假设投影空间为，那么新空间中类间散度的距离可表示为：}

/(5-12)

其中

是矩阵 的平均值，是类别个数，

是第类样本的平均值，

是第类样本的样本个数。 那么类间数据点矩阵

可以定义为

/(5-13)

{47%：同样，投影后的21范数距离类内散度矩阵值可以表示为}

/(5-14)

其中类内数据点矩阵为

/。

{56%：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。} {40%： 利用这个思想，我们可以通过以下目标函数实现最优21范数投影矩阵：}

/(5-15)

/

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。 {47%： 到目前为止，我们可以通过求解这个最小优化问题来获得21范数距离的最优投影矩阵。} 对于矩阵

，它的每一行都对应一个特征。 {44%： 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21范数距离LDA转换成特征选择方法，我们必须强制更多的行为零。 因此，有必要引入21范数距离正则项：

/(5-16)

/

其中]0是可以调节投影矩阵的行稀疏程度的参数。 {47%： 越大的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

/(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，下面我们给出一个迭代算法来解决这个同时解决最小最大化问题。

在推导出我们的新方法之前，我们将首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，我们有以下等式：

/(5-18)

/

根据引理1，类间散度值可以被重写为

/ (5-19)

其中

是一个对角矩阵，定义为

/. (5-20)

那么类间散度矩阵可以表示为

/. (5-21)

同样，类间散度值可以写为

/ (5-22)

其中，

/(5-23)

那么类内散度矩阵可以表示为

/. (5-24)

根据引理1，我们可以得到

/ (5-25)

其中，

/. (5-26)

回顾上述L21FS的公式，可以用传统21范数距离公式来解决：

/

/. (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

是对应于最小特征值的特征向量：

/. (5-28)

5.2.2 迭代算法

不过需要注意的是,

,

和都依赖于投影矩阵，因此它们也是未知的变量。 {41%： 我们在这里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, 参数.

初始化列正交矩阵, 参数.

计算类间数据点矩阵

和类内数据点矩阵

.

while不收敛

计算

,

和 .

通过

,

构建散度矩阵

,

.

通过特征值问题求解公式(5-28).

通过获得的特征向量更新.

end while

备注1： 由于

的秩等于类别数减1(1)，即

是不满秩的，所以对

{48%：进行求逆运算会存在奇异性问题。} 为了处理这个问题，我们在

的主对角元素上增加一个小的值。

备注2： 请注意，在实际问题中，的某些行将是零，它会导致

，

和的一些元素不存在。 同样，我们替换

2

为

2

+

。

5.2.3 收敛性证明

在这一小节中，我们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，我们将引入以下引理。

引理2： 对于函数

，

=

2

2

+

2

2

，给定任意非零值，，，0

，并且0，下面的不等式成立：

/. (5-29)

因此，对于任意非零向量， ，

，

，我们有

/. (5-30)

定理1： 在固定(

)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，我们通过

表示更新的W。 在每次迭代中，我们都有

/ (5-31)

/，

这表明

/ (5-32)

/.

为了方便，我们定义矩阵W的第行为

，那么

/

/

/

/

/. (5-33)

根据公式(5-30)，我们有

/ (5-34)

/.

结合公式(5-33)和(5-34)，我们可以得到

/. (5-35)

公式(5-35)可以重写为

/(5-36)

因此，在

{41%：=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。

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

1

+

=的特征值问题。 该操作的时间复杂度近似于(

3

)。 {43%： 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。} {42%： 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

一旦我们找到了最优投影矩阵

{58%：，接下来就是确定特征重要性的评估原则。} {41%： 在本算法中，我们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

2

{63%：值越大，则相应的特征就越重要。} 有了这个原则，我们可以按照我们的期望得到排名靠前的特征。

5.3 L21FS算法实验

{45%：在本节中，我们进行了大量实验来评估我们新方法的性能。} 所有的代码都写在MATLAB_R2014b中。 实验环境： 2.7 GHz Intel Core i5 CPU，8 GB 1867 MHz DDR3内存。 我们采用LIBSVM算法对数据点进行分类。 为了更加精确，我们的方法的测试精度是使用传统的十折交叉验证来计算的。 {46%： 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，我们进行两个小实验来展示我们的新算法能够找出最具判别特征的能力。 {45%：然后我们将我们的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，我们通过实验来显示参数对L21FS的性能的影响。 为了进一步研究我们的新方法和其他方法之间的分类精度的差别，我们还采用了配对T检验方法。 {43%： 然后，我们分别对具有噪声数据和噪声特征的数据集进行实验。} {43%： 最后，我们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

在我们的实验中，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：

{65%：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3232。 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1616手写数字图像，其中包括7291个训练图像和2007个测试图像。

{44%：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。} {50%： 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。 {42%： 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

为了直观地展示我们的新方法能够选择最显着特征的能力，我们在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。 在这里我们随机选择两个人的照片数据作为训练数据。 为了绘制图片，我们选择排名靠前的

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中我们可以发现，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，我们注意到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，这与我们的期望一致。

5.3.3 Iris 鸢尾花小实验

{47%：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。} {77%： 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，我们对其进行实验来直观地表达我们提出的方法的效果。

在这个实验中，我们从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。 我们遍历了这四个特征的所有可能组合并将其画出。 此外，我们还绘制了通过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。 当我们仔细观察图5-2（2）时，我们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

为了显示我们新方法的性能，我们在公开数据集上进行了试验，并将我们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]。 {40%： 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。

为了描述我们新方法的效果，我们采用了以下几个指标：

平均精度： {52%：我们采用十折交叉验证来评估每种方法的性能。} {43%：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。} {44%：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： {48%：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： 我们执行配对T检验来比较L1FS和其他方法[31]。 {43%： T检验的p值表示两个分类准确度值之间差别的概率。} {45%： 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，我们通过十折交叉验证获得它们。 {46%： 我们使用LIBSVM对所选特征提供的样本进行分类，使用十折交叉验证。} {56%： 每种算法的平均精确度汇总在表5-2和表5-3中。} 我们加粗显示其中最好的精度。

表格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(AverageSTD, time: s, p-value)

UMM

MCFS

LS

DFS

L21FS

USPS (2007x256, class: 10)

73.942.52 2.9441 6.5402e-6

87.841.58 0.1151

0.1574

53.310.98 0.1284

2.8636e-10

89.981.74 0.3622

0.7805

89.631.67 0.4324

—

MADELON (2000x500, class: 2)

61.001.65 4.7316 0.8038

60.251.30 0.0341

0.3479

61.101.57 0.1872

0.8654

60.600.93 1.9471

0.4817

61.301.65 2.3182

—

LUNG_DISCRETE (73x325, class: 7)

67.048.1838 0.2948 0.0032

76.576.1677 0.0195

0.0293

57.528.78 0.0027

4.1081e-4

71.232.43 0.6794

6.4027e-4

87.525.50 0.8650

—

ISOLET5 (1559x617, class: 26)

38.996.41 4.6121 3.5866e-6

73.762.85 0.5380

0.1257

43.234.72 0.1419

1.1493e-6

71.324.00 3.5905

0.0419

76.772.06 3.7723

—

ISOLET (1559x617, class: 26)

33.073.22 4.5595 2.2656e-8

73.583.11 0.5644

0.0227

54.934.92 0.1404

2.4978e-5

68.335.89 3.6048

0.0091

79.552.86 4.1939

—

COIL20 (1440x1024, class: 20)

67.081.93 11.2796 2.8113e-7

87.563.70 0.7159

0.1034

61.664.95 0.1890 4.6702e-6

94.931.72 15.2525

0.00629

91.662.48 13.6754

—

COLON

(62x2000, class: 2)

70.7615.60 47.2229

0.1207

80.7610.81 0.0616

0.4822

60.8915.90 0.0067

0.0211

64.2313.00 34.4736

0.0191

85.386.32 43.1703

—

表格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(AverageSTD, time: s, p-value)

UMM

MCFS

LS

DFS

L21FS

USPS (2007x256, class: 10)

83.101.10 2.9376 8.7859e-6

90.631.77 0.3461

0.6210

66.114.81 0.1557

7.8923e-6

91.621.23 0.3892

0.6150

91.181.18 0.7723

—

MADELON (2000x500, class: 2)

60.802.01 4.8755

0.9696

58.652.32 0.0590

0.1545

60.351.72 0.1843

0.6785

59.951.74 1.6079

0.4628

60.851.55 2.3469

—

LUNG_DISCRETE (73x325, class: 7)

71.148.41 0.3084

0.0158

79.238.18 0.0482

0.2362

64.289.20 0.0034

0.0029

71.145.48 0.6491

0.0025

84.853.16 0.9327

—

ISOLET5 (1559x617, class: 26)

49.904.00 4.6838

2.6778e-7

86.141.36 1.1678

0.8540

63.052.94 0.1440

1.6578e-6

82.363.11 3.6057

0.0718

86.402.34 5.3912

—

ISOLET (1559x617, class: 26)

45.252.06 4.7121

1.1557e-8

86.021.81 1.2450

0.1330

63.582.54 0.1366

1.4194e-6

80.893.25 3.6701

0.0068

89.033.11 4.4487

—

COIL20 (1440x1024, class: 20)

72.632.62 11.4341

1.3667e-7

95.551.21 1.5625

0.1706

68.610.47 0.1853

2.4107e-11

97.221.07 15.4507

0.5260

96.730.99 16.8142

—

COLON

(62x2000, class: 2)

72.3018.73 49.3434

0.2225

83.9711.53 0.0670

0.8356

64.2315.89 0.0066

0.0385

64.2313.00 40.8704

0.0191

85.386.32 46.3781

—

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。 {50%：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，我们发现了总共四个情况下，其中 DFS比我们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值表明我们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。

{41%：当我们关注表5-2和表5-3中所示的计算消耗时，我们发现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(

2

)。 {45%： 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(}

3

)。 考虑到其他三种算法都是迭代方法，从这方面我们可以很好地解释时间差异。

{40%：此外，我们将我们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。} {65%： 分类精度与所选特征数的变化如图3所示。}

{43%：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，可以平衡L21FS公式的稀疏性和凸性。 {46%： 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，我们主要关注对新方法性能的影响。 {53%：我们改变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，我们在数据集 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。{50%： 总体而言，具有相同数量的特征，值越大，准确度越高。} 值得注意的是，当数值达到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，我们的方法的性能对该值更敏感。 {43%： 也就是说，由正则化参数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

由于我们提出的方法是一种鲁棒的特征选择方法，因此我们必须对噪声数据进行实验。为了仿真噪声数据样本，我们通过生成的噪声矩阵

来融入原始输入数据集=[

1

; ;

]

{42%：, 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。}然后我们对模拟的样本数据+

进行与原始数据 相同的实验, 其中=

并且 是一个给定的噪声因子。 在本小节中, 我们设定在所有的实验中=0.1。 我们使用与之前相同的实验设置将我们的方法与其他四种方法进行比较, 并将结果报告在表5-4和表5-5中。

表格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(AverageSTD, time: s, p-value)

UMM

MCFS

LS

DFS

L21FS

USPS (2007x256, class: 10)

73.642.33 3.0750

5.1935e-6

87.940.89 0.1313

0.2255

53.711.55 0.1311

1.5484e-9

89.331.20 0.1498

0.9284

89.21.75 0.1754

—

MADELON (2000x500, class: 2)

61.051.60 5.1963

0.7870

60.201.20 0.0373

0.3222

60.951.52 0.1935

0.7239

56.901.67 0.5634

0.0078

61.401.92 0.9652

—

LUNG_DISCRETE (73x325, class: 7)

67.147.66 0.3043

0.0022

76.576.16 0.0241

0.0360

57.5212.17 0.0049

0.0022

67.048.18 0.2602

0.0031

84.952.51 0.8871

—

ISOLET5 (1559x617, class: 26)

38.994.58 4.7873

4.3673e-7

71.133.22 0.5766

0.0834

42.073.73 0.1471

2.1851e-7

53.361.76 1.2686

8.4249e-8

74.661.52

2.3271

—

ISOLET (1559x617, class: 26)

31.022.58 4.7768

1.6676e-8

73.915.82 0.6071

0.2019

54.554.78 0.1456

3.6108e-5

50.323.59 1.4804

3.0244e-6

78.583.37 2.1248

—

COIL20 (1440x1024, class: 20)

67.631.85 11.6559

1.0762e-8

89.862.94 0.7300

0.1455

58.193.17 0.1829

3.2531e-8

89.930.93 5.7540

0.0066

92.082.38 5.7619

—

COLON

(62x2000, class: 2)

70.6416.42 46.4025

0.4789

73.9713.59 0.1317

0.6730

61.0213.74 0.0067

0.0734

65.6418.06 37.8078

0.2672

77.438.06 40.3341

—

表格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(AverageSTD, time: s, p-value)

UMM

MCFS

LS

DFS

L21FS

USPS (2007x256, class: 10)

83.201.82 2.9615

9.8808e-5

89.931.71 0.3138

0.4292

70.203.10 0.1310

1.5564e-6

92.171.41 0.1548

0.1571

90.781.09 0.2387

—

MADELON (2000x500, class: 2)

60.601.98 4.8088

0.9446

59.002.23 0.0567

0.2857

60.501.70 0.1826

0.8813

58.701.65 0.5200

0.1573

60.701.95

0.5276

—

LUNG_DISCRETE (73x325, class: 7)

69.808.49 0.3220

0.0293

76.577.47 0.0436

0.1848

64.289.20 0.0024

0.0078

72.479.15 0.2646

0.0778

83.425.78

0.4309

—

ISOLET5 (1559x617, class: 26)

50.614.74 4.7239

8.7922e-7

84.464.58 1.2656

0.5914

62.151.16 0.1414

6.1962e-8

65.294.17 1.5309

2.3321e-5

85.82.21 2.0919

—

ISOLET (1559x617, class: 26)

45.002.17 4.7068

2.5932e-9

87.371.67 1.1739

0.4005

64.032.23 0.1444

2.6152e-7

70.255.06 1.2809

1.6017e-4

88.582.17 2.3248

—

COIL20 (1440x1024, class: 20)

72.98+2.69 11.4180

2.9313e-7

95.690.71 1.4816

0.3171

67.223.23 0.1872

1.7285e-7

93.051.45 5.7293

0.0087

96.521.38 7.7865

—

COLON

(62x2000，class: 2)

72.302.69 11.4180

0.3634

78.8410.27 0.0787

0.6272

64.2315.89 0.0064

0.0802

72.3015.75 39.5869

0.2996

82.178.29 41.0819

—

从表5-4和表5-5中，我们发现下列的一些现象。 首先，我们的方法在7个实验数据集中的大多数情况下表现最好，这表明 L21 FS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。 {41%： 其次，尽管我们的算法在原始数据集上的性能仅略好于其他算法，但我们的算法在噪声数据情况下分类精度下降最小。} {42%：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。} 这也充分证明了我们算法的鲁棒性。

5.3.7噪声特征实验

如前所述，我们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。 21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，我们进行实验来测试我们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，我们将大小为88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。 我们分别选择了前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，我们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，我们注意到我们提出的方法的性能下降率很小，这为我们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，我们通过在实验数据集上目标函数值的变化曲线来评估我们提出的方法的计算效率。{45%： 如前所述，目标函数值最终将收敛到局部最优值。} 因此，迭代次数是我们提出的方法的效率中最重要的部分之一，它决定了我们算法收敛的速度。 {49%： 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.

{65%：图5-7 目标函数值 VS 迭代次数}

Fig.5-7 Objective value vs iteration numbers

很明显，如图5-7所示，我们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， 这与我们之前的理论分析完全一致。 此外，我们可以发现该算法通常在约7次迭代内收敛到局部最优。 {45%： 这种少量的迭代数确保了我们提出的算法的效率和可行性。} 因此，由于收敛速度快，我们提出的L21FS方法在实践中表现良好。

5.4 算法总结

本大章节提出了一种结合传统特征抽取方法LDA和21范数距离的特征选择方法。 与以前的带21范数正则化项的特征选择方法不同，我们在学习函数中也使用了21范数距离。 {56%：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21范数距离并同时最小化类间散度值的21范数距离，我们引入了一种有效的迭代算法。 {41%：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。} {53%：广泛的实验结果表明，我们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，我们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。 {41%：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。} {41%：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。 {66%：因此，寻找一种合适的距离度量标准非常必要。} 在本文中，我们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 L2 p范数距离加以改进， {47%：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

在本文中，我们主要对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，我们在整个学习函数中使用了 L2 p范数距离度量，这 使得我们的分类算法 L2 p- TWSVM具有更好的鲁棒性以及我们的特征选择算法 L21 FS具有更好的稀疏性。

{45%：针对基于 L2 p范数距离度量的 TWSVM，我们通过采用在学习函数中使用 L2 p范数距离，} {42%：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。} 并且，我们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。 {50%：并且第三章节从理论到实验证明了该算法的有效可行性。}

针对原先的 DFS特征选择算法，我们理论分析证实了其学习函数仍然不够鲁棒，因此我们 {44%：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，我们实现了同时最大最小化L21范数项的目标。 {41%：并且针对L21FS设计的迭代算法能够使得我们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。并且，由于是基于L21范数，我们的L21FS能够得到更好的稀疏性，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本文值得思考的一点。

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。 我们知道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我们的后续工作。

3、在决定L2p范数的p值时，我们通过不同的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。 {42%：因此，p值对算法性能的影响以及p值的确定仍是本文需要

研究的一个重点问题。}

{76%：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)

在审论文：

[1] Xu Ma， Qiaolin Ye， Yingan Liu， He Yan， Robust Feature Selection via L21-Norm Minimization and Maximization， IEEE Access. (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): 206-10.

[3]刘勇进，赵敬红.基于稀疏恢复的 1_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)： 42-6.

[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): 521-7.

[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): 349-53.

[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， et al. 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]DASH M， LIU H. Feature Selection for Classification [M]. IOS Press， 1997.

[12]GUYON, ISABELLE, ELISSEEFF, et al. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3(6): 1157-82.

[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): 194-9.

[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]TSAGKARAKIS N, MARKOPOULOS P P, PADOS D A. L1-norm Principal-Component Analysis of Complex Data [J]. 2017,

[16]YANG M S, HUNG W L, CHUNG T I. Alternative Fuzzy Clustering Algorithms with L 1-Norm and Covariance Matrix; proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, F, 2006 [C].

[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): 126-33.

[18]YAN A H, YE B Q, LIU C Y A, et al. The GEPSVM Classifier Based on L1-Norm Distance Metric [M]. Springer Singapore, 2016.

[19]刘建伟, 李双成, 付捷, et al. L1范数正则化SVM聚类算法 [J]. 计算机工程, 2012, 38(12): 185-7.

[20]NIE F, HUANG H. Non-Greedy L21-Norm Maximization for Principal Component Analysis [J]. 2016,

[21]DU L, ZHOU P, SHI L, et al. Robust Multiple Kernel K-means Using L21-Norm [J]. 2015,

[22]谭龙, 何改云, 潘静, et al. 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 [J]. 2015,

[24]李富杰. 基于流行学习的人脸表情识别研究 [D]; 杭州电子科技大学, 2013.

[25]许熳锋. 无监督流行学习算法的若干探讨 [D]; 浙江大学, 2010.

[26]陈晋音, 保星彤, 陈心怡, et al. 一种基于自适应密度聚类的非线性流行学习降维方法 [M]. 2017.

[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): 290-1.

[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]王国胜, 钟义信. 支持向量机的若干新进展 [J]. 电子学报, 2001, 29(10): 1397-400.

[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): 130-5.

[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): 183-5.

[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]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.

[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): 001227-1234.

[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.

[36]TANVEER M. Robust and Sparse Linear Programming Twin Support Vector Machines [J]. Cognitive Computation, 2015, 7(1): 137-49.

[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): 162-5.

[38]BACH F R, LANCKRIET G R G, JORDAN M I. Multiple kernel learning, conic duality, and the SMO algorithm; proceedings of the Proc International Conference on Machine Learning, F, 2004 [C].

[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): 128-9.

[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): 1962-9.

[41]HONG T, 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]FUNG G, MANGASARIAN O L. Proximal support vector machine classifiers; proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, F, 2001 [C].

[43]DODGE Y. on Statistical data analysis based on the L1-norm and related methods [J]. Computational Statistics Data Analysis, 2002, 6(4): R3-R.

[44]WANG H, LU X, HU Z, et al. Fisher Discriminant Analysis With L1-Norm [J]. IEEE Transactions on Cybernetics, 2013, 44(6): 828-42.

[45]KONG D, HUANG H, HUANG H. Robust nonnegative matrix factorization using L21-norm; proceedings of the ACM International Conference on Information and Knowledge Management, F, 2011 [C].

[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, et al. 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): 60-8.

[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): 2832-5.

[49]王和勇, 郑杰, 姚正安, et al. 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): 1485-90.

[50]宋枫溪, 高秀梅, 刘树海, et al. 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): 1915-22.

[51]KWAK N, CHOI C H. Input feature selection for classification problems [J]. IEEE Transactions on Neural Networks, 2002, 13(1): 143.

[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; proceedings of the International Conference on Neural Information Processing Systems, F, 2005 [C].

[54]KWAK N. Principal component analysis based on l1-norm maximization [J]. IEEE Tpami, 2008, 30(9): 1672-80.

[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): 527-30.

[56]WELLING M. Fisher Linear Discriminant Analysis [J]. Department of Computer Science, 2009, 16(94): 237-80.

[57]刘小虎, 李生. 决策树的优化算法 [J]. 软件学报, 1998, 9(10): 797-800.

[58]唐华松, 姚耀文. 数据挖掘中决策树算法的探讨 [J]. 计算机应用研究, 2001, 18(8): 18-9.

[59]杨学兵, 张俊. 决策树算法及其核心技术 [J]. 计算机技术与发展, 2007, 17(1): 43-5.

[60]HONG T, 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.

[61]CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data; proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, F, 2010 [C].