1. 绪论
   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，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，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。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，支持向量机可以完美的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。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，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. 国内外研究现状
  2. 传统算法的缺陷
  3. 本文主要研究工作
  4. 本文内容安排

1. 支持向量机概述

## 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。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，该数据集可以记为，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，如果其为负类，那么标记该点为。第个点的标记可以表示为。为第个样本的标签。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

(1)

其中表示平面。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

(2)

求公式(2)的最大值问题可以转化为如下的最小值问题

(3)

然而公式(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

(4)

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

 (5)

将拉格朗日函数对分别求偏导，可以得到如下公式

 (6)

 (7)

 (8)

将公式(6)(7)(8)带入拉格朗日函数，可以得到整个对偶目标函数

 (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

 (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)。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。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。

假设正类样本对应的平面法向量为，偏差为，负类样本对应的平面法向量为，偏差为。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：

 (11)

 (12)

公式(11)(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

 (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

 (14)

其中是一个非负的参数。公式(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

 (15)

 (16)

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

 (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

(18)

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，但是寻找这两个分类平面的方法完全不同。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。

假设有个数据点可以表示为一个矩阵，属于一个维的实值空间.同样，表示对应的第个样本属于正类或事负类。我们假设正类样本为,负类样本为，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和偏差:

 (19)

 (20)

其中是非负的平衡参数，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(19)和(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(19)和(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(19)对应的拉格朗日函数可以表达为如下形式：

 (21)

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

 (22)

 (23)

 (24)

 (25)

 (26)

 (27)

 (28)

结合(24)和(28)，我们可以得到

 (29)

接下来，通过将(22)与(23)相加可以得到

 (30)

我们定义

 (31)

通过这些定义，我们可以重写公式(30)为



 (32)

通过公式(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，其中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

 (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：

 (34)

通过SMO算法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的法向量和偏差。同理，我们可以通过同样的方法来得到平面2 的法向量和偏差.

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,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：

 (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

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。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。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。通过求解一对广义特征值问题来获得两个非平行的分类平面。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。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算法的不鲁棒的问题，在下一章节中，本文将通过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。

1. 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。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：

 (36)

正齐次性:

 (37)

三角不等式:

 (38)

对于任意向量，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：

 (39)

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：

 (40)

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：

 (41)

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：

 (42)

假设有矩阵包含个数据点，并且每一个数据点,那么矩阵则是一个大小的矩阵。代表矩阵第行第列对应的元素。对于矩阵，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

 (43)

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：

 (44)

其中为的最大特征值。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：

 (45)

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：

 (46)

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。

## 3.2 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，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，)最小化可以为目标函数提供鲁棒性，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。 我们假设表示正类的样本矩阵并且表示负类的样本矩阵。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的实值空间。因此，所有矩阵和矩阵的大小分别为和。 对于一个矩阵，表示矩阵的第行在实值空间中的第行。的平方L2范数可以表示为。 因此，矩阵的平方L2范数定义如下：

 (47)

平方L2范数的公式表达可以推广至p序L2范数（L2p范数）：

 (48)

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

## 3.3 L2p-TWSVM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。 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，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：

 (49)

 (50)

公式(49)的拉格朗日函数为

 (51)

其中为拉格朗日乘子。

我们注意到公式(51)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，我们将含有L2p范数的项拆分为平方L2范数和次方的L2范数的乘积：

 (52)

我们定义

 (53)

那么拉格朗日函数(51)可以重写为以下形式

(54)

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

 (55)

 (56)

 (57)

 (58)

通过公式(58)和（57）我们可以得到

 (59)

为了简化公式，我们定义

 (60)

因此，公式(53)可以表达为

 (61)

将公式(55)和公式(56)相加，我们可以得到

 (62)

这个可以简化表达为

 (63)

由公式（63）我们可以得到的解析解为

(64)

## 3.4 L2p-TWSVM算法实验

## 3.5 算法总结

1. 特征选择概述

## 4.1 特征选择与特征提取

## 4.2特征选择分类

## 4.3 特征选择算法

## 4.4 本章小结

1. 基于L21范数距离度量的优化特征选择

## 5.1 L21范数应用

## 5.2 L21FS模型推导

## 5.3 L21FS算法实验

## 5.4 算法总结

1. 结束语