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)

## 2.4 本章小结

1. 基于L2p范数距离度量的TWSVM

## 3.1 范数定义

## 3.2 L2p范数应用

## 3.3 L2p-TWSVM模型推导

## 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