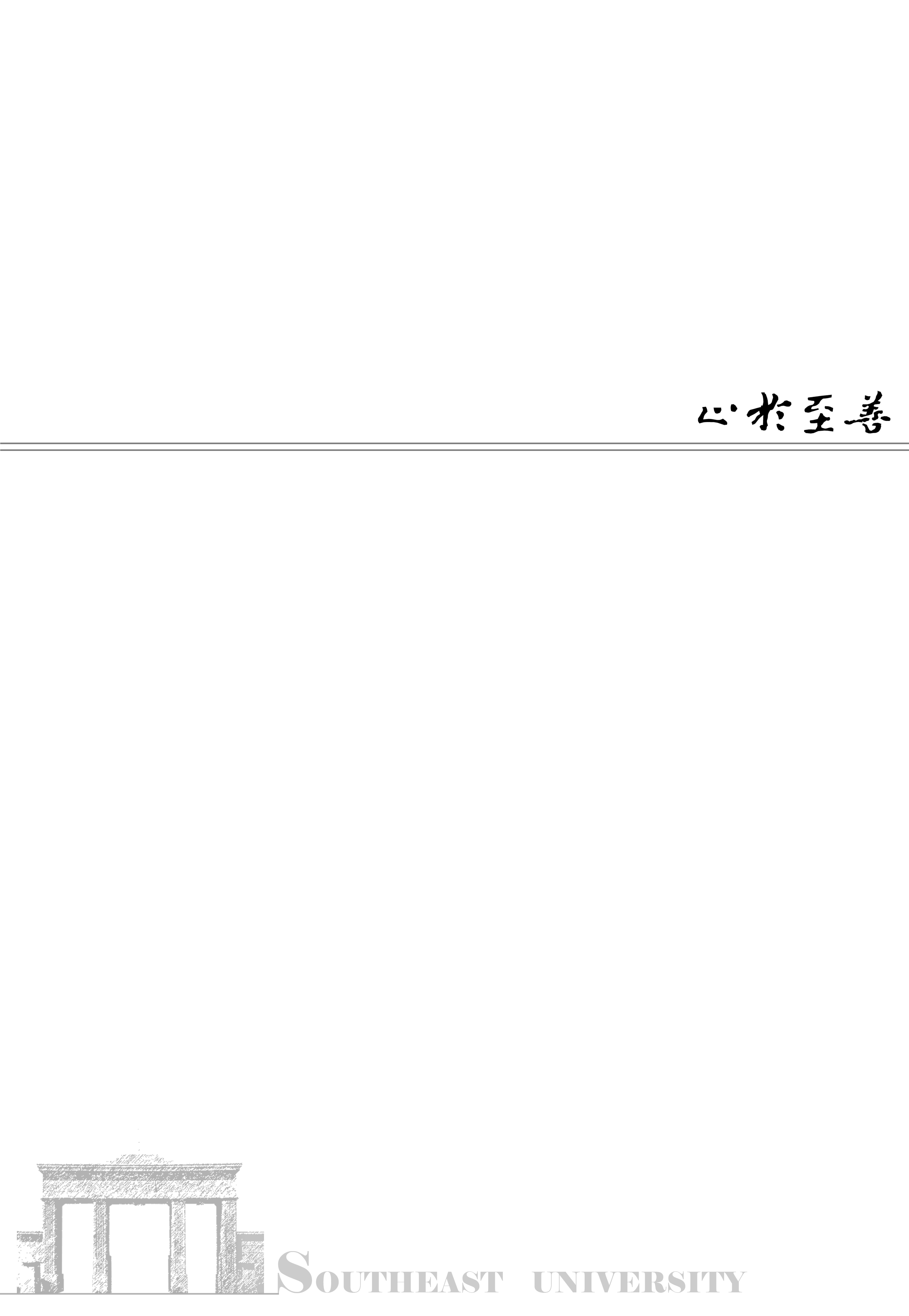
学校代码： 10289

分类号：

密 级：

论文题目 姓 名 江苏科技大学

学 号：

**江苏科技大学**

**工程硕士学位论文**

基于优化核函数支持向量机

的船舶柴油机故障诊断系统研究

|  |  |
| --- | --- |
| 研究生姓名 汪小杰 | 导师姓名  **杨兴林** |
| 申请学位类别 工程硕士 | 学位授予单位  **江 苏 科 技 大 学** |
| 专业领域 船舶与海洋工程 | 论文提交日期 **2017 年 04月 30 日** |
| 研究方向**动力装置性能与优化** | 论文答辩日期 **2017 年 06 月 12 日** |
| 答辩委员会主席 | 评 阅 人 |
|  |  |

2017 年 04月 22 日

分类号：

密 级：

学 号：

基于优化核函数支持向量机

的船舶柴油机故障诊断系统研究

|  |  |
| --- | --- |
| 学生姓名 | 汪小杰 |
| 指导教师 | 杨兴林教授 |

江苏科技大学

二O一七 年 五 月

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Master of Engineering

Research on Fault Diagnosis system of Marine Diesel Engine based on improved kernel function Support Vector Machine

Submitted by

Name Wang Xiaojie

Supervised by

Professor Name Yang Xinglin

Jiangsu University of Science and Technology

May, 2017

**江苏科技大学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，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。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标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

年 月 日

**江苏科技大学学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权江苏科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于：

(1)保密□，在 年解密后适用本授权书。

(2)不保密□。

学位论文作者签名： 指导教师签名：

年 月 日 年 月 日

# 摘 要

近年来，柴油机故障诊断技术已经成为国内外船舶行业比较热门的研究方向之一。它通过分析和处理船舶柴油机的实时状态和历史数据，来预测出柴油机未来可能会发生的故障状态。船舶柴油机具有往复式机械的典型特征，其故障具有复杂性和多样性的特点，因此传统的故障诊断方法很难达到高效的结果。

由于船舶柴油机故障数据样本的局限，因此本文引入支持向量机来进行预测柴油机故障类型。支持向量机作为机器学习算法，它可以在小样本情况下表现出非常好的性能。近些年来，支持向量机在故障诊断领域发展迅速。本文以基于Python语言的Scikit-Learn模块，编写了针对船舶柴油机燃油系统故障数据的代码来实现支持向量机，并根据学习过程建立了故障识别的模型，再对未知的故障进行预测。

本文采用了四个典型的核函数以及本文提出的基于线性核函数优化的支持向量机来实现预测柴油机故障类型，都取得了很好的效果。线性核函数的准确率保持在86.42%；多项式核函数的准确率能达到88.89%；径向基核函数的准确率保持在86.42%；而基于线性核函数优化的支持向量机预测效果最好，其准确率可以达到90.12%。

**关键词** 机器学习；支持向量机；核函数；船舶柴油机；燃油系统；故障诊断

# Abstract

In recent years, the fault diagnosis technology of diesel engine has become one of the most popular research directions. By analyzing and dealing with the real-time and historical data of the marine diesel engine, the fault state of the diesel engine in the future can be predicted. The diesel engine has the typical characteristics of reciprocating machinery, its fault has the characteristics of complexity and diversity, so the traditional fault diagnosis method is difficult to achieve high efficiency. Feature extraction is an important part of machine learning, which is of great significance for fault diagnosis.

Due to the limitation of fault data samples of marine diesel engine, this paper introduces the support vector machine. Support vector machine as an algorithm for machine learning, it can show very good performance in small sample case. In recent years, SVM has developed rapidly in the field of fault diagnosis. This paper is based on the Scikit-Learn module of the Python language based on the marine diesel engine fault data learning, and established a fault recognition model based on the learning process, and to predict the unknown fault, and achieved good result.

In this paper, four typical kernel functions and the support vector machine based on linear kernel function are used to predict the fault types of diesel engine. The accuracy of linear kernel function in 86.42%; the accuracy of polynomial kernel function can reach 88.89%; the accuracy of RBF in 86.42%; and the support vector machine prediction results of linear kernel function optimization based on the best, the accuracy rate can reach 90.12%.

**Key words** machine learning; support vector machine; kernel function; marine diesel engine; fuel system; fault diagnosis

**目录**

[摘 要 I](#_Toc476312366)

[Abstract II](#_Toc476312367)

[第1章 引言 1](#_Toc476312368)

[1.1 研究背景 1](#_Toc476312369)

[1.2 船舶柴油机故障诊断技术研究现状 2](#_Toc476312370)

[1.2.1 热力参数分析法 2](#_Toc476312371)

[1.2.2 油液分析法 2](#_Toc476312372)

[1.2.3 振动分析法 3](#_Toc476312373)

[1.2.4 瞬时转速法 3](#_Toc476312374)

[1.2.5 基于专家系统的智能化诊断方法 3](#_Toc476312375)

[1.2.6 基于神经网络的诊断方法 4](#_Toc476312376)

[1.2.7 基于灰色系统理论的诊断方法 4](#_Toc476312377)

[1.2.8 基于信号处理技术的诊断方法 5](#_Toc476312378)

[1.2.9 基于混合系统的智能诊断方法 5](#_Toc476312379)

[1.2.10 基于支持向量机的诊断方法 6](#_Toc476312380)

[1.3 本文研究的意义及主要内容 6](#_Toc476312381)

[1.3.1 本文研究的意义 6](#_Toc476312382)

[1.3.2 本文的主要内容 7](#_Toc476312383)

[第2章 柴油机故障诊断理论 8](#_Toc476312384)

[2.1柴油机系统结构 8](#_Toc476312385)

[2.2 柴油机常见故障与原因 9](#_Toc476312386)

[2.3 柴油机燃油系统及常见故障 13](#_Toc476312387)

[2.3.1 柴油机燃油系统简介 13](#_Toc476312388)

[2.3.2 燃油压力波及测量 15](#_Toc476312389)

[2.3.3 燃油系统常见故障分析 15](#_Toc476312390)

[第3章 支持向量机 17](#_Toc476312391)

[3.1 机器学习 17](#_Toc476312392)

[3.1.1 机器学习的定义 17](#_Toc476312393)

[3.1.2 机器学习的发展史 17](#_Toc476312394)

[3.1.3 机器学习分类 18](#_Toc476312395)

[3.2 支持向量机 19](#_Toc476312396)

[3.2.1 VC维 19](#_Toc476312397)

[3.2.2 结构风险最小化 19](#_Toc476312398)

[3.2.3 支持向量机的基本原理 20](#_Toc476312399)

[3.2.4 核函数 22](#_Toc476312400)

[3.3 SVM实现与Python介绍 25](#_Toc476312401)

[3.3.1 Python语言简介 25](#_Toc476312402)

[3.3.2 Scikit-learn 25](#_Toc476312403)

[第4章 SVM船舶柴油机故障诊断技术 28](#_Toc476312404)

[4.1 燃油系统常见故障分析 28](#_Toc476312405)

[4.2 支持向量机的学习条件 31](#_Toc476312406)

[4.2.1 交叉验证 31](#_Toc476312407)

[4.2.2 惩罚项参数和松弛参数 35](#_Toc476312408)

[4.2.3 多分类学习 35](#_Toc476312409)

[4.3 SVM故障诊断技术实现 37](#_Toc476312410)

[4.3.1 线性核函数分类学习 38](#_Toc476312411)

[4.3.2 多项式核函数分类学习 38](#_Toc476312412)

[4.3.3 径向基核函数多分类学习 39](#_Toc476312413)

[4.3.4 Sigmoid核函数分类学习 39](#_Toc476312414)

[4.3.5 优化线性核函数分类学习 40](#_Toc476312415)

[4.4 多种核函数算法评估对比 41](#_Toc476312416)

[4.4.1 准确率 43](#_Toc476312417)

[4.4.2 混淆矩阵、查准率和查全率 43](#_Toc476312418)

[4.4.3 F1度量 47](#_Toc476312419)

[第5章 总结与展望 48](#_Toc476312420)

[5.1 本文工作总结 48](#_Toc476312421)

[5.2 进一步研究的展望 48](#_Toc476312422)

[参考文献 50](#_Toc476312423)

[致谢 53](#_Toc476312424)

# 第1章 引言

## 1.1 研究背景

柴油发动机自1892年问世以来已有百年历史，由于当时航运事业的飞速发展，柴油发动机开始被引入船舶领域。

柴油发动机是一种压缩发火的往复式内燃机，它使用的是挥发性很差的柴油或者劣质燃料油作为燃料，然后采用将燃料和空气在其气缸内部形成可燃混合气，并采用空气压缩形成的高温而自行发火的方法来做功。这种工作特点使柴油机在热机领域内具有较高的热效率，一般可达到55%左右，而且适合作为船用发动机使用[1]。

20世纪60年代，船用低速柴油发动机开始进入了黄金时代，其在船舶动力装置发展历程中取得了鲜明的具有压倒性的优势。由此，柴油发动机技术也开始趋于完善，这个时期的船用低速柴油发动机的性能参数大致范围为：

表 1.1 船用低速柴油发动机的性能参数

Table 1.1 Marine Low Speed Diesel Engine Performance Parameters

|  |  |
| --- | --- |
| 缸径 |  |
| 行程 |  |
| 单缸有效功率 |  |
| 油耗率 |  |
| 有效热效率 |  |

此期间内船用低速柴油机发展的特点，按顺序大致为增大机组效率，提高可靠性，提高经济性。

柴油机系统具有零部件多、运动复杂、互相关联和工作环境恶劣等特点，并且柴油机的目的是实现高速高功率，其负载的程度越来越高，这就使得柴油发动机发生故障的可能性增大。根据对在日本某油田野外作业的柴油机的故障统计分析，这个小型小型柴油机的故障发生率达到大约每小时0.4×10-3次[2]，同样根据M. Hernqvist从海难事故的统计中，发现机械故障发生率在众多海难事故中位居第一，高达22%。其中，主机故障在机械故障发生率占了45%，由此可见船舶柴油机的故障发生率早已严重影响着航运的经营效益。随着内燃机故障引起的大型海难、公路交通事故、空难、工业事故等大型事故的愈演愈烈，这就使得人们开始重新认识和探索柴油发动机的实际运行状况从而来降低柴油发动机的故障发生率以及故障程度。自此，故障诊断技术开始受到人们的广泛关注[2]。

19世纪60年代的时候开始出现机械故障诊断学，并在最近十几年来取得迅速发展的一门学科。机械故障诊断学的研究内容是机组或机器的运行状态变化在诊断信息中的反映，是机械运行状态的学科[3]。机械故障诊断学不断吸收近现代科学技术发展的最新成就，诊断和应用技术都有了长足的进步，目前已发展成为一门集数学、物理、计算机与微电子技术、力学、化学、信息处理技术、人工智能等多种专业学科于一体的新兴交叉学科。

近年来，机器学习伴随着人工智能的飞速发展，已广泛应用于各个领域。其中基于统计学理论的支持向量机（Support Vector Machine）在解决小样本学习中有很大的优势，在很多领域都得到了广泛的应用，比如：图像识别、文本分类等领域。故障诊断的瓶颈之一是故障样本的缺乏,鉴于SVM能在训练样本很少的情况下很好地达到分类，推广的目的,本文尝试使用SVM对船舶柴油机进行故障诊断[4]。

## 1.2 船舶柴油机故障诊断技术研究现状

在检测和诊断船舶柴油机故障时有多种方法，传统的方法主要包括瞬时转速法、油液分析法、热力参数分析法、振动分析法等。现代的方法主要有基于专家系统的智能化诊断方法、基于神经网络的诊断方法、基于灰色系统理论的诊断方法、基于信号处理的诊断方法、基于混合系统的智能化诊断方法、基于支持向量机的诊断方法等。

### 1.2.1 热力参数分析法

热力参数分析法是利用船舶柴油发动机工作时性能参数来判断其工作状态的，这些参数主要包括气缸压力示功图、转速、滑油温度、排气温度、冷却水进出口温度及排放等。热力参数分析法主要对柴油机性能指标做出判断，在这些性能参数中，示功图所包含的信息量最多，示功图可以计算出压力升高率、指示功、压缩压力，从而可以判断各缸功率是否平衡和燃烧质量的好坏，但测试用压力传感器的寿命及可靠性是影响示功图诊断柴油机性能的主要问题。

### 1.2.2 油液分析法

油液分析法是综合利用铁谱分析、油品化验、通过对润滑油中磨粒浓度、磨粒形状以及大小的变化、含铁量检查、含铁量的变化进行整体磨损状态分析和油质的变化，从而来判断柴油发动机的故障状态和磨损状态。油液分析法主要分为两种方法，第一种方法是分析油液中不洁物质的技术，其中包括光谱分析、铁谱分析和颗粒计数法等；第二种方法是油液本身化学物理性能的分析。在具体实施过程中，油液分析提取信息法实施起来比较方便，可以避免声振技术的频谱干扰等缺陷；而光谱技术诊断发的代价十分昂贵；相较于光谱技术而言，铁谱技术手动操作过程较复杂，而且速度慢，识别的要求和判断分析的要求较高，需要经验丰富的技术人员，使得它的使用受到一定的限制，并且标准图谱积累需时较长。

### 1.2.3 振动分析法

振动分析法是一种利用船舶柴油发动机在其工作时各组成部分所产生的振动信号，经测试、处理和分析，然后对其内部零件状态进行诊断的方法。该振动分析法具有准确率高、诊断速度快和在线诊断等优势。目前国内外振动分析法的主要应用方向有：可以用瞬时的转速推算气缸内压力的变化；可以通过机体表面的振动信号去识别柴油打动机气缸内的压力示功图；利用小波分析和时频分析等信号分析处理的方法来处理柴油发动机的表面振动信号。振动监测诊断柴油机故障的方法在国内已经取得了许多的研究成果。比如在气缸头部安装振动传感器，然后通过分析振动信号来诊断缸内的故障，利用振动信号来诊断柴油机的主轴承故障，利用润滑油管内的压力信号来诊断柴油机轴承的故障等等。其中在利用机器表面振动信号气缸磨损、诊断活塞、主轴承状态和气阀漏气等方面，武汉理工大学已经有了大量的实验研究，并研制出了柴油机智能诊断仪，可实现在不解体的前提下诊断船舶柴油机活塞—缸套磨损和气阀漏气等故障。

### 1.2.4 瞬时转速法

以多缸柴油机为例，多缸柴油机的曲轴瞬时转速的信号能实时反映出柴油机各缸的工作状态，所以通过对瞬时转速的波动信号的分析，可以得到柴油机的运行状态和其相关的故障信息。在通常情况下，各缸的动力性能应该基本一致，柴油发动机的运转也很平稳，柴油机各缸的瞬时转速波动虽然有所差异，但总在一个浮动不大的范围内，并始终呈现出某种规律性；但当柴油机某气缸发生故障，开始不正常工作，其动力的一致性会遭到一定的破坏，柴油机的运转平稳性会变差，瞬时转速的波动信号就会产生畸形，所以可以由此来判断柴油机缸内工作过程的稳定。然而瞬时转速法还存在着这些不足：利用瞬时转速的波动信号虽然能识别出非正常工作的缸位，但不能确定造成故障的具体原因：假如缸内压力降低造成曲轴瞬时的转速变化，也可能是燃油系统故障造成燃烧不充分所致，可能是活塞环或者缸套磨损引起气密性变差所导致的等等各种原因；当缸数较多的时候，特别是缸数达到十六缸以上的时候，利用瞬时转速可以准确地诊断柴油发动机是否有故障，定位故障缸都十分困难，主要原因是单缸对转速的影响在瞬时转速得波形信号中所占曲柄的转角较小，其相邻缸之间作功的重复曲柄的转角较大，所以提取单缸作功信息的能力会很弱。

### 1.2.5 基于专家系统的智能化诊断方法

专家系统的定义是利用该研究领域专家的专业知识去进行推理从而解决专业性的高难度的实际问题的智能系统。故障诊断专家系统，作为专家系统的一个重要分支，它是人们根据大量的故障诊断知识和长期的实践经验，而设计出的一种计算机程序智能系统，用来解决难以用数学模型来精确描述的故障诊断问题。

专家系统的核心思想主要包括四个部分：知识库、推理机、知识获取部分和解释部分。在知识表达方面，大多数的专家系统都是以框架进行知识和产生式规则来表达实现的。产生式规则得以进行知识表达，一方面得益于现存的人工智能语言；另一方面受益于其表达合乎人的心理逻辑，便于进行知识的获取，比较容易让人接受。在诊断预测方面，主要在对推理逻辑和推理模型的研究方面。

值得关注的是，近期有学者提出了基于模型的知识库理论，这也就使推理机制发生了根本改变，如定性物理模型、神经网络模型和可视觉模型等，这表明了人工智能领域已经注入了新的活力。

### 1.2.6 基于神经网络的诊断方法

基于神经网络的机械故障诊断技术，顾名思义就是通过对诊断经验和故障实例的训练学习，用分布在网络内部的权值来表达所学习的故障诊断知识的技术，它具有对故障模式的模式匹配、联想记忆和相似归纳能力，从而去实现故障与特征之间复杂的映射关系。目前，神经网络在柴油机故障诊断中的运用主要有：

1．神经网络直接用于故障诊断。挑选有效参数作为输入，故障的参数作为输出，通过大量典型样本的学习所得的权值进行识别。

2．自适应式神经网络模式识别。它通过神经网络分布式信息存储和并行处理，巧妙避开了模式识别中最复杂的模型建立和特征提取，然后剔除特征提取不当所带来的影响，使得故障易于进行识别。

3．神经网络信号处理。神经网络用于信号处理，主要是因为其最优化算法以及其智能化识别的特点。

4．模糊神经网络。具有准确的非线性拟合以及学习能力。

5．专家系统与神经网络的结合识别。神经网络只要和专家系统巧妙结合起来，可以实现互补长短，从而克服神经网络的缺乏经验、无推理性，也可以解决专家系统的知识“瓶颈问题”等缺陷。

### 1.2.7 基于灰色系统理论的诊断方法

1982 年由华中理工大学学者邓聚龙教授创立了灰色系统理论。灰色系统理论的广泛的适用性和新颖的思路在工程界得到了广泛的注意，并在社会、经济及工程等许多领域获得了广泛的应用。

将灰色系统理论用于船舶柴油机故障诊断，其原理是将柴油发动机系统认为是一个复杂的灰色系统，将已知信息整理然后去推测含有故障模式的不可知信息的状态、特征以及发展的趋势，从而对未来的发展进行预测。整个过程就是一个灰色系统的白化过程。

灰色理论在机械故障诊断技术的应用中主要包括了灰色系统建模、关联度分析、灰色模型预测这三个步骤。灰色系统理论可以实现故障的预测，具有准确率高，计算量小，易于微机实现等优势。灰色系统理论的学习过程中人工干预较少，只需收集到一定数量典型状态的样本经过简单运算，就可以构成典型状态模式向量。与人工神经网络方法相比，其学习方法不仅编程简单，便于调试，而且容易实现再学习。实践证明了基于灰色系统理论的诊断方法对大型复杂机械设备的故障诊断是十分有效的，该方法是一种很有潜力的船舶柴油机故障诊断方法。

### 1.2.8 基于信号处理技术的诊断方法

作为20 世纪80 年代后期发展起来的应用数学分支，小波变换具有许多优良的特性。利用小波变换的诊断方法是基于信号处理技术的诊断方法的中心思想。

小波变换作为一种信号预处理方法，可以用于故障特征信号的提取并进行信号去噪。适当的选取小波的尺度，然后用这些尺度的小波对信号进行重构，去掉高频、工频噪声频段内的小波尺度，可以得到重构信号的系统运行信息及故障信息。由于基于小波变换的故障诊断方法无需数学模型，可以进行在线实时故障检测，对输入信号的要求较低，同时灵敏度高，计算量较小，抗干扰能力强等优势，可以克服前面几种诊断方法的缺点，因此基于小波变换的故障诊断方法是一种很有发展潜力的故障诊断方法。

### 基于混合系统的智能诊断方法

基于混合系统的智能诊断方法本质上是神经网络方法与专家系统方法的结合。传统的专家系统在开发研制过程中存在着知识维护难、实时性差、推理能力弱以及知识获取的“瓶颈”问题，而利用神经网络响应快、较好的容错性、自组织和自适应等优点可以弥补这些不足。神经网络与专家系统的结构有两种方法：

1. 使用神经网络来构造专家系统，即把传统专家系统的符号推理变成基于数值运算的推理，从而提高专家系统的执行效率，并解决专家系统的自学习的问题；

2. 将神经网络看作是知识源的表达与处理方式，这种模式与其它知识表达方式，如框架、规则等共同来表达该领域专家的知识，从而实现面向不同推理的机制。

人工神经网络和模糊推理的结合在两者的知识存储、知识的表示、推理速度等方面发挥了很大的作用。神经网络具有模仿人脑神经元的功能，数据的直接处理能力和具有强大的自学能力等优点；模糊推理则模仿人脑的逻辑思维，具有较强的结构性知识表达能力等特点。通过比较两种方法的优缺点，可以将神经网络与模糊推理结合起来，继承两者的优点，从而实现故障诊断系统对模糊信息的处理，同时还使得基于规则的结构性知识可以进行学习和调整。神经网络方法与模糊逻辑方法结合的方式也有两种：

1. 一种是将传统神经网络方法模糊化，这种方式保留神经网络原有的结构，而将神经元进行模糊化处理，使其具有可以模糊信息的能力。

2. 一种是基于联结主义的神经的模糊推理协作系统，它根据模糊规则或模糊分类算法构造相应的网络结构[6]。

### 1.2.10 基于支持向量机的诊断方法

支持向量机作为机器学习领域的研究热点，在很多研究领域已经取得了很广泛的应用。支持向量机是Vapnik在上世纪90年代提出的一种新型的机器学习方法。支持向量机具有完备的统计学习理论基础和出色的学习性能这两大特点。

支持向量机本身是二分类器，而柴油机故障种类繁多，所以将支持向量机应用在柴油机故障诊断上就需要构造多类分类器，主要思想就是用二分类的支持向量机去构造多值的分类器。

目前多类分类器的研究过程主要包括以下四个步骤：

1. 归一化预处理。在得到数据后的第一步处理就是数据的初始化，其目的是为了减少故障样本中不同参数的数值大小对支持向量机诊断方法的影响。

2. 第二步是对样本进行学习，通过对故障样本的学习，从而得到不同故障类型的最优超平面来训练分类器。

3. 然后是对故障进行识别，把新的样本输入到分类器中，从而对新样本进行故障识别，得到样本的故障类型。

4. 最后一步是新故障处理，将支持向量机方法中新的故障类型输入到分类器中，通过迭代学习获得新的最优超平面，从而提高故障诊断的识别能力，本文就是使用这种方法[5]。

## 1.3 本文研究的意义及主要内容

### 1.3.1 本文研究的意义

本文以船舶柴油发动机作为研究对象，使用支持向量机方法进行故障诊断，其主要的研究意义包括以下几点：

1. 本文为支持向量机在柴油发动机故障诊断研究的发展做一些基础的学习和研究，为支持向量机方法作为柴油机故障诊断技术的一种做了理论论证，并致力于支持向量机在柴油机故障诊断领域或其他领域得到推广。

2. 通过故障诊断预测的故障类型，快速精确地分析出柴油发动机故障的原因，从而实现减少维修费用，缩短维修时间，降低因柴油机停机而造成的衍生损失等等。同时为该诊断系统进行实时故障记录的诊断结论，为该柴油发动机和同类柴油机诊断和维修提供有效有理论支撑的参考。

3. 为当前支持向量机中参数选择、数据缺少等问题，进行了编程优化，并进一步地进行了支持向量机在多值分类问题中的学习和研究，最终讨论了支持向量机应用于柴油机故障诊断领域的可行性和优势。

### 1.3.2 本文的主要内容

论文主要分为5章：

第1章是引言部分，主要介绍了国内外柴油机故障诊断技术研究的现状与发展，简单的介绍了诸如机械振动、油液分析、热力分析、瞬时转速、神经网络、专家系统，支持向量机等方法，最后还介绍了本文的主要内容。

第2章主要是介绍柴油机故障诊断理论基础，包括柴油机系统结构、柴油机常见故障及其产生原因和柴油机燃油系统及常见故障。

第3章主要是机器学习的简介，主要包括机器学习的定义、发展和分类。由此引出支持向量机，然后介绍统计学习理论和支持向量机的理论基础，主要有VC维、经验风险最小化原则、支持向量机的基本原理等理论，还介绍了在支持向量机中最优超平面的构造、核函数统计学习的有关理论和支持向量机的基本理论。最后介绍了用于实现支持向量机算法的编程语言Python，主要介绍了Python语言的特点和主要用于机器学习的程序包Scikit-learn。

第4章主要说明了燃油系统常见故障类别及其数据特点、支持向量机算法实现的前提、多种不同核函数支持向量机预测结果与分析。主要讲述了四个典型的核函数支持向量机以及一个自定义和函数支持向量机，并基于多种分类学习算法评估方法来评价五个算法的优劣。

第5章是总结与展望，对全文工作进行总结，提出对进一步工作的展望。

# 第2章 柴油机故障诊断理论

## 2.1柴油机系统结构

表2.1柴油发动机系统结构

Table 2.1 Diesel Engine System Structure

|  |  |  |
| --- | --- | --- |
| 柴油发动机 | 燃油系统 | 燃油箱 |
| 燃油泵 |
| 低压油管 |
| 过滤器 |
| 喷油泵 |
| 高压油管 |
| 燃烧室 |
| 喷油室 |
| 润滑油系统 | 润滑油想 |
| 过滤器 |
| 润滑油泵 |
| 限压阀 |
| 冷却器 |
| 润滑部件 |
| 进排气系统 | 增压器 |
| 进气阀 |
| 排气总管 |
| 排气阀 |
| 尾气净化 |
| 滤清器 |
| 气门机构 |
| 冷却系统 | 气缸盖 |
| 气缸体 |
| 散热片 |
| 风扇 |
| 风压室 |

船舶柴油发动机是一个复杂的机电系统，其主要结构有：机组体、配气机构、曲柄连杆机构、冷却系统、供给系统、起动系统和润滑系统。柴油发动机系统按照划分常规以及柴油机故障特征，将其分为燃油系统、进排气系统、润滑油系统和冷却系统等四大部分。

表2.1说明了柴油机的结构分级：第一级是柴油发动机；第二级是组成柴油机的四个子系统；第三级是组成四个子系统的各个部件。三级结构的划分可以有效地为组织不同的诊断知识提供了条件，通过对这些结构的描述，可以将诊断系统条目化、清晰化，从而可以高效的实现诊断知识的获取和诊断策略的控制。

## 2.2 柴油机常见故障与原因

柴油机系统结构复杂，出于简化问题的出发点，效仿柴油机的结构，现将柴油机故障结构也分为四个子系统：润滑油系统故障、燃油系统故障、冷却系统故障和进排气系统故障。系统故障和各子系统故障的逻辑关系图详见表2.2。

表2.2 柴油机故障分类

Table 2.2 Diesel Engine Fault Classification

|  |  |  |  |
| --- | --- | --- | --- |
| 柴油机系统故障 | | | |
| 燃油系统故障 | 润滑油系统故障 | 进排气系统故障 | 冷却系统故障 |

燃油系统常见故障与产生原因（如表2.3所示）

表2.3 燃油系统常见故障及其产生原因

Table 2.3 Common Faults of Fuel System and it`s Causes

|  |  |  |
| --- | --- | --- |
| 故障现象 | 产生原因 | |
| 启动困难 | 柴油温度低 | 环境温度低，加热设备不起作用 |
| 柴油液位过低 | 柴油消耗将尽，供油不连续 |
| 供油量不足 | 燃油泵故障 |
| 柱塞副磨损严重 |
| 柴油燃油中水分过多 |
| 燃油过滤器堵塞或漏气 |
| 高压油泵故障 |
| 气缸内混合气体压力太低 | 活塞、缸套、活塞环等零件磨损严重 |
| 喷油器故障 | 喷油压力过低，造成喷油提前，雾化不良 |
| 发动机运转时  声音异常 | 喷油过早 | |
| 各缸供油不均 | |
| 喷油器针阀卡死 | |
| 发动机转速  不稳定 | 喷油泵调速器反应滞后 | 各缸喷油量、雾化效果不一致 |
| 各缸供油不均 | 个别进、排气阀出现故障 |
| 个别气缸活塞环出现串气、断裂、磨损过度 |
| 调速弹簧失效，喷油器压力过低 | |
| 气缸压缩压力不足 | 汽缸垫损坏 | |
| 进气门和排气门的气门间隙调整不当 | |
| 进气门和排气门与座密封不严 | |
| 活塞环安装不正确，没有按原厂规定安装 | |
| 活塞环与汽缸磨损严重，或活塞环卡死在活塞环槽内 | |
| 柴油机突然停车 | 燃油耗尽 | |
| 燃油系统进入空气或油管破裂、接头松动 | |
| 燃油中有水 | |
| 燃油滤清器堵塞 | |
| 喷油嘴柱塞卡死 | |
| 喷油嘴柱塞弹簧断裂 | |
| 进气管或空气过滤器堵塞 | |
| 气门弹簧断裂或者气门卡死 | |
| 润滑油压力过低 | |
| 过度磨损 | 拉缸 |
| 主轴承或连杆大端轴承烧瓦 |
| 最高爆发压力下降 | 喷油泵或喷油器经长期工作后，喷油质量变差，产生滴漏，雾化不良 | |
| 供油定时不对，喷油提前角过小，燃烧太迟 | |
| 工作粗暴，个别气缸最高燃烧压力太高 | 油量调节机构失灵或喷油器针阀卡死在全开的位置，造成过多的燃油 | |
| 喷油过早 | |
| 轴承间隙过大或运动部件连接螺栓松动 | |
| 扫气压力下降 | 涡轮增压器损坏，或空气滤器和空气冷却器污堵，阻力太大 | |
| 涡轮背压大高，增压器转速降低 | |
| 增压器气封环间隙过大或安装不正确造成漏气 | |
| 换气系统中的进气阀阀片断裂或扫气泵活塞环磨损 | |
| 排气回转阀损坏或间隙太大 | |
| 扫气压力升高 | 在直流扫气中，由于排气阀关闭不严，造成废气能量大，使涡轮转速 | |
| 燃烧不良、后燃严重，排气温度升高 | |
| 发生了活塞环咬住、损坏或轴承油膜破坏等故障而未能及时发现 | |

润滑油系统常见故障与产生原因（如表2.4所示）

表2.4 润滑油系统常见故障与产生原因

Table 2.4 Lubricating Oil System Common Faults and Causes

|  |  |  |
| --- | --- | --- |
| 故障现象 | 产生原因 | |
| 润滑油消耗偏多 | 润滑油油位偏高、偏低 | |
| 活塞与气缸间隙过大 | |
| 活塞环特别是油环弹性差 | |
| 缸套、活塞环过度磨损，活塞环被粘住、对口，或扭曲环装反 | |
| 活塞环与环槽边隙和侧隙过大，或活塞上油环回油孔被积炭阻塞 | |
| 增压柴油机涡轮增压器弹力密封装置失效 | |
| 气门杆与导管配合间隙过大或油封失效 | |
| 润滑油的压力过低 | 油底壳的油位过低，油泵不能泵到油 | |
| 油路里有空气 | |
| 润滑油被柴油稀释，粘度下降 | |
| 润滑油泵故障造成压力过低 | |
| 润滑油滤清器调压弹簧过软 | |
| 润滑油过滤器堵塞 | |
| 油道泄漏 | |
| 润滑油的压力过高 | 润滑油滤清器调压弹簧过硬 | |
| 调压阀的压力调整不当或堵塞、卡死 | |
| 油温过低造成润滑油粘度过高 | |
| 细滤器脏，旁通油量过小 | |
| 主油道或油管堵塞 | |
| 润滑油温度过高 | 冷却不良 | 机油冷却器效率不高，散热片脏堵 |
| 气缸盖和气缸体散热片表面脏堵严重 |
| 冷却风短路或冷却风量不够 |
| 润滑油容量不足，油底壳的液位太低 |
| 柴油机工况不良 | 喷油器雾化不良 |
| 柴油机长期处于超负荷状态 |
| 空气过滤器脏、堵使柴油燃烧不充分 |
| 过度磨损 | 活塞和缸套过度磨损为拉缸 |
| 主轴承、曲柄销轴承磨损过度 |
| 溢流阀压力低，机油从溢流阀溢出过多，通过冷却器的机油量减少 | |

进排气系统常见故障与产生原因（如表2.5所示）

表2.5 进排气系统常见故障与产生原因

Table 2.5 Common Failures and Causes in Intake and Exhaust System

|  |  |  |
| --- | --- | --- |
| 故障现象 | 产生原因 | |
| 尾气冒黑烟 | 排气温度整体升高 | 鼓风机风扇传动皮带过松，使风扇转速降低 |
| 空气滤清器堵塞 |
| 涡轮增压器失效 |
| 喷油器开启阀压力太低，喷油器漏油，喷孔部分堵塞 |
| 扫气压力和压缩压力过低 |
| 排气阀漏气或气口严重结炭 |
| 超负荷运行或由于负荷分配不均而浩成某必缸韶负荷 |
| 燃油品质差 |
| 冒白烟 | 水漏入气缸 | |
| 尾气冒蓝烟 | 油底壳的机油液面太高 | |
| 润滑油进入燃烧室 | |
| 活塞环相互对口，造成串气 | |
| 单杠不做功 | 气阀故障 | 进气阀常开或者常闭 |
| 排气阀常开或者常闭 |
| 燃油系统故障 | 单缸喷油器不喷油，或没有雾化 |

冷却系统故障与产生原因（如表2.6所示）

表2.6 冷却系统故障与产生原因

Table 2.6 Cooling system failures and causes

|  |  |  |
| --- | --- | --- |
| 故障现象 | 产生原因 | |
| 柴油机机体温度过高，冷却效果不佳 | 鼓风机风扇转速低 | 液力偶合器的热力膨胀阀不正常 |
| 散热片脏、堵 | 尘埃，油污结成污垢 |
| 风压室密封不严 | 风压室盖板变形 |
| 机油温度过高 | 机油冷却器内部脏、堵严重 | 机油脏 |
| 机油冷却器外部脏堵 | 尘埃，油污结成污垢 |

通过对柴油机的这些故障机理分析，得出柴油发动机各子系统主要的故障。

表2.7 柴油发动机各子系统主要的故障

Table 2.7 Diesel Engine Main Subsystems of Failure

|  |  |  |  |
| --- | --- | --- | --- |
| 燃油系统 | 润滑油系统 | 进排气系统 | 冷却系统 |
| 低压油路供油不畅  输油泵活塞弹簧折断  输油泵止回阀密封不严  高压油管堵塞  喷油嘴偶件磨损和卡死 | 机油泵压力不足  机油泵压力过高  机油润滑不力 | 气门间隙异常  气门漏气  空气滤清器堵塞  排烟不正常 | 冷却系统结构简单，出故障机会较少，这里不考虑 |

通过上述故障类型及其产生机理的分析，以上四大类的故障机理分析对系统检测柴油发动机的功能及性能参数，建立专家系统的知识库具有十分重要的意义[9]。

## 2.3 柴油机燃油系统及常见故障

船舶柴油发动机的燃油系统作为柴油机的重要组成部分之一，在其整个作业过程中，燃油质量的优劣直接影响到柴油机的工作性能。燃油系统的功能是根据柴油发动机的工作流程和负载的大小，定量、定时、定压地将雾化质量良好的高质量柴油按相应的喷油喷入气缸，与压缩空气混合，从而保证气缸内燃油正常燃烧[7]。

柴油发动机燃油系统的状态信息主要体现在高压油管的压力波形中，在某处发生故障时，必然会使原有的供油状态发生变化，燃油流动的流速和压力等工作参数也会产生相应的改变，这就会反映在压力信号的波形上定会导致波形的参数值和波形的状态的变化。然后对燃油压力的波形信号进行分析，并提取相应的特征信息，为方便判断系统实时的工作状态，最终达到诊断和预测系统故障类型的目的[8]。

### 2.3.1 柴油机燃油系统简介

柴油机的燃油喷射系统由柴油滤清器、柴油箱、输油泵、喷油器、高压油泵、高压油管、低压油管和回油管组成。它的主要组成部分的性能和作用如下：

1. 喷油泵：作用是定量、定时地向喷油器注入高压燃油。尤其在多缸柴油机中喷油泵应该保证以下几点：

（1） 每个缸的供油顺序应符合所要求的发动机的工作顺序。

（2） 每个缸的供油量分布均匀，不均匀度在标定工况下不大于3％到4％。因为额定供油率与发动机的额定转速和额定功率相对应的。为了实现转速平稳，就要求各缸的供油率确保均匀。

（3） 供油的提前角需要保持一致，相差不得大于0.5度。喷油泵的供油提前角在一方面要求与发动机的曲轴保持同步（也就是第一缸的喷油起始时间要对应上发动机的曲轴转角零位的标记），在另一方面还要求每个缸供油的时间间隔必须保持一致。供油提前角的准确性就是正常喷油的保证，从而防止喷油的时间过长所造成的燃烧不良。为减少喷油滴漏的现象，喷油泵还需要保证供油停止的迅速。

（4） 燃油雾化良好。由于柴油的挥发性比较差，为了使其在很短的时间内形成混合气，要求喷出的燃油束要颗粒细小均匀、雾化良好、方向和形状与燃烧室相适应。

2. 喷油器：喷油器能够将喷油泵的高压燃油雾化成更加细的颗粒，并以一定的角度往发动机高速运转需要，从喷油器喷向气缸中的燃油一定要尽快燃烧，并在合适的时刻迅速燃烧完毕，从而将燃油的化学能最大限度的转化为推动发动机运行的机械能。这就需要以下要求：

（1） 喷油器要具有一定的喷注贯穿距离和喷油压力。气缸中的空气在压缩后，压强和温度都会大大地增加，如果喷油器的喷油压力不超过高压就无法实现燃油喷射。

（2） 有良好的雾化性能。为保证喷射的燃油与空气迅速混合而获得充分燃烧。

（3） 当喷油停止的时候，应该能够迅速完全的切断燃油供给，以免发生异常喷油现象。当气体温度和气缸压力因活塞下行而迅速下降后，异常喷油引起的燃烧会造成积炭增多、燃烧不良、油耗增加和排气异常[10]。

柴油发动机的每一个循环作业都会分别向每个缸喷射一次燃油。根据四冲程柴油机的工作原理，其每个循环都要经过进气、压缩、作功、排气这四个进程。柴油发动机在进气进程时进的是空气，在压缩进程时，柴油经喷油泵将油压提高至10MPa以上，然后通过喷油器来喷入气缸。当柴油发动机每转动两圈，就喷油一次，喷油泵的供油时由外廓曲线来控制，而曲轴和凸轮轴之间就会通过传动比为1: 2的齿轮来实现减速传动，也就是说，曲轴每转两圈，凸轮转一圈[10][11]。

在从柴油箱到喷油泵入口的油路中，油压是由输油泵建立的，输油泵的出油压力一般在0.15~0.3MPa，因此这段油路称为低压油路，只用于向喷油泵供给滤清的柴油。而在从喷油泵到喷油器这段油路中，喷油泵建立的油压，通常在10MPa以上，因此称这段油路为高压油路。在燃油喷射的过程中，高压油管内的最高压力一般为60MPa~70MPa或更高，而喷射结束后油管内的剩余压力远不足它的十分之一。所以，燃油喷射过程是高压系统内压力剧变的过程[10][12][13][14]。

### 2.3.2 燃油压力波及测量

柴油机的燃油系统实际上是燃油压力波的传播系统，在燃油供给直至最后喷射到气缸内整个过程中，在高压油管中燃油都是以压力波的形式传播。高压油管内各处的压力随着时间和位置的变化而变化，并带有一定的随机性。但是，基本的变化规律(或称压力波形的模式)是由喷油泵总成、喷油器总成、高压油管、及供油角联动装置来决定的，反映燃油系统状态信息最多的就是高压油管中的压力信号[10][11]。当某处发生故障时，必然会使原有的供油状态发生变化，燃油流动的压力和流速等参数也会有相应的改变，反映在压力波形上将导致波形形态和波形参数值的变化。如果我们抓住波形的局部形态畸变特征，并结合柴油机专业领域知识，对异常波形的状态进行诊断和分析[14]，就可以将燃油系统的故障诊断任务转化为对燃油压力波形的模式识别问题。

柴油机中，对每一缸而言，喷油阶段仅对应曲轴转30～40度范围。所以喷油压力信号的主频较低，大约在100Hz内。喷油压力将会使高压油管管壁产生径向膨胀。由于系统的特点，所提取的特征必须是反应整机各部位配合工作下的运行状态。发动机在运行的过程中，许多部件都是运动的，它们的状态无法直接测量，只能间接测量。而且要求对系统实时检测，就不允许在机器上打孔、钻眼来加装测试器件，但是状态信号提取的正确与否会直接影响分析结果，所以恰当的选择测试部位，合理的配置测试器件就显得十分关键[15][16][17][18]。

测试部位选取的总原则是：

（1） 要保证提取信号的真实性和完整性。如果测量某振动源的信号，测试应选择在尽量靠近振源、信号传递路径最短、传递界面最少的位置.

（2） 要保证传感器的安装、调试是在不影响机器的正常运行的情况下。测试点应该在机器的外表，为了不影响测试精度和机器的正常工作，测试部件的质量与传感器的质量之比要在两个数量级以上。

（3） 要保证机器的安全性。在重要部位和故障多发的部位应多设检测忍占[12][18]。

本系统利用夹持式压力传感器检测油压波形，只需要方便地将其夹持在高压油管的外端，就可实现压力波形的不解体在线检测。系统测得的油压波形是通过测量高压油管的膨胀变形量来间接得到，测取的波形清晰、峰值突出、特征点明显，并有效地抑制了机体振动对波形的影响。

### 2.3.3 燃油系统常见故障分析

表2.8是柴油机工程师和用户协会出版的报告。从表2.8中可以看出：造成柴油机停机故障的各种原因中，燃油系统故障占27％，所占比例最大，是柴油机的一个重要的故障源。所以，如何有效地诊断供油系统的故障，便成为当前柴油机维修领域的紧迫课题[19]。

表2.8 柴油机故障分类与故障出现率

Table 2.8 Diesel Engine Fault Classification and Fault Rate

|  |  |
| --- | --- |
| 故障的分类 | 故障出现率/% |
| 喷油系统及燃油系统故障 | 27.0 |
| 漏水故障 | 17.3 |
| 阀门及阀座故障 | 11.9 |
| 轴承故障 | 7.0 |
| 活塞组件故障 | 6.6 |
| 漏油及润滑系统故障 | 5.2 |
| 涡轮增压系统故障 | 4.4 |
| 齿轮及驱动装置故障 | 3.9 |
| 调速器齿轮故障 | 3.9 |
| 燃油泄漏 | 3.5 |
| 漏气 | 3.2 |
| 除下列专门标题外的其他破坏及破裂 | 2.5 |
| 其他故障 | 2.5 |
| 基座故障 | 0.9 |
| 曲轴故障 | 0.2 |

本系统的故障模拟实验是在某油泵试验台上进行的，分别在喷端和泵端采集不同油泵转速、不同油量和不同故障工况的实验样本共400多例。其中，油泵转速为750r/min（发动机的转速为1500 r/min），发火顺序为1-5-3-6-2-4。实验台上模拟了燃油系统常见的9种故障（其中正常状态也作为故障的一种情况），正常工作情况（100％供油），供油量不足（包括75％供油、50％供油、25％供油、怠速供油），针阀卡死至油孔堵塞（包括小油量和标定油量下），针阀泄漏，出油阀失效。

# 第3章 支持向量机

## 3.1 机器学习

### 3.1.1 机器学习的定义

机器学习（Machine Learning）是一门人工智能的学科，该领域的主要研究对象是人工智能，热别是如何在经验学习中改善具体算法的性能。机器学习主要应用于专家系统、规划和问题求解、网络信息服务、故障诊断、机器人、认知模拟、数据挖掘、图像识别、自然语言理解和博弈等领域。

### 3.1.2 机器学习的发展史

机器学习是人工智能研究较为年轻的分支，它的发展过程大体上可分为4个时期。第一阶段是在20世纪50年代中叶到60年代中叶，属于热烈时期。第二阶段是在20世纪60年代中叶至70年代中叶，被称为机器学习的冷静时期。第三阶段是从20世纪70年代中叶至80年代中叶，称为复兴时期。机器学习的最新阶段始于1986年。

机器学习进入新阶段的重要表现在下列诸方面：

(1) 机器学习已成为新的边缘学科并在高校形成一门课程。它综合[应用心理学](http://baike.baidu.com/view/46434.htm)、生物学和神经生理学以及数学、自动化和[计算机科学](http://baike.baidu.com/view/92404.htm)形成机器学习理论基础。

(2) 结合各种学习方法，取长补短的多种形式的集成学习[系统研究](http://baike.baidu.com/view/1446932.htm)正在兴起。特别是连接学习[符号学习](http://baike.baidu.com/view/6217848.htm)的耦合可以更好地解决连续性[信号处理](http://baike.baidu.com/view/642820.htm)中知识与技能的获取与求精问题而受到重视。

(3) 机器学习与人工智能各种基础问题的统一性观点正在形成。例如学习与问题求解结合进行、知识表达便于学习的观点产生了通用智能系统的组块学习。类比学习与问题求解结合的基于案例方法已成为[经验](http://baike.baidu.com/view/21717.htm)学习的重要方向。

(4) 各种学习方法的应用范围不断扩大，一部分已形成商品。归纳学习的知识获取工具已在诊断分类型专家系统中广泛使用。连接学习在声图文识别中占优势。分析学习已用于设计综合型专家系统。[遗传算法](http://baike.baidu.com/view/45853.htm)与强化学习在工程控制中有较好的应用前景。与[符号系统](http://baike.baidu.com/view/6552866.htm)耦合的[神经网络](http://baike.baidu.com/view/5348.htm)连接学习将在企业的智能管理与智能机器人运动规划中发挥作用。

(5) 与机器学习有关的学术活动空前活跃。国际上除每年一次的机器学习研讨会外，还有计算机学习理论会议以及遗传算法会议。

### 3.1.3 机器学习分类

机器学习算法有很多，其分类方式也有多种。主要的分类方式有基于学习策略的分类、基于所获取知识的表示形式分类、按应用领域分类、综合分类、学习形式分类等。其中最为普遍接受的是按学习形式分类。

机器学习按学习形式分类有监督学习（Supervised Learning）、非监督学习（Unsupervised Learning）和强化学习等几种类型。实践中应用较多的也就是监督学习和非监督学习。

监督学习就是数据样本会告诉计算机在该情形下的正确输出结果，希望计算机能够在面对没有见过的输入样本时也给出正确的输出结果，从而达到预测未知的目的。就像一个学生通过做多套高考模拟卷并订正答案的方式来提高高考成绩。在这种情形下，数据就像是监督计算机进行学习的教师，故而得名。

根据输出结果是离散还是连续，监督学习可以分为分类问题和回归问题两大类。他们在文字、语音、图像识别、垃圾邮件分类与拦截、网页检索、股票预测等方面有着广泛应用。

非监督学习是指数据样本中没有给出正确的输出结果信息。就像做了很多套没有答案的模拟卷，最后还要去参加高考。但是其实做没有答案的试卷也是能学到许多有价值的信息的，比如说哪些题出现频率高之类的等等。无监督学习希望从数据中挖掘的正是这一类信息，常见的例子有聚类、关联规则挖掘、离群点检测等等。

强化学习是智能系统从环境到行为映射的学习，以使奖励信号(强化信号)函数值最大，强化学习不同于连接主义学习中的监督学习，主要表现在教师信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价（通常为标量信号），而不是告诉强化学习系统RLS（Reinforcement Learning System）如何去产生正确的动作。由于外部环境提供的信息很少，RLS必须靠自身的经历进行学习。通过这种方式，RLS在行动评价的环境中获得知识，改进行动方案以适应环境。

## 3.2 支持向量机

Cortes和Vapnik于1995年首先提出支持向量机(Support Vector Machine)，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他[机器学习](http://lib.csdn.net/base/machinelearning)问题中。

支持向量机是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以期获得最好的泛化能力。

### 3.2.1 VC维

VC维（Vapnik-Chervonenkis Dimension）的概念是为了研究学习过程一致收敛的速度和推广性，由统计学习理论定义的有关函数集学习性能的一个重要指标。

VC维的定义是：对一个指标函数集，如果存在N个样本能够被函数集中的函数按所有可能的2的K次方种形式分开，则称函数集能够把N个样本打散；函数集的VC维就是它能打散的最大样本数目H。若对任意数目的样本都有函数能将它们打散，则函数集的VC维是无穷大，有界实函数的VC维可以通过用一定的阀值将它转化成指示函数来定义。

VC维反映了函数集的学习能力，VC维越大则学习机器越复杂（容量越大），遗憾的是，目前尚没有通用的关于任意函数集VC维计算的理论，只对一些特殊的函数集知道其VC维。例如在N维空间中线形分类器和线形实函数的VC维是N+1。

### 3.2.2 结构风险最小化

结构风险最小化（Structural Risk Minimization）是指把函数集构为一个函数子集序列，使各个子集按照VC维的大小排列，在每个子集中寻找最小经验风险，在子集间折中考虑经验风险和置信区间，取得实际风险的最小化。也就是SRM准则。

这个假设与问题真实解之间的误差，就叫做风险。我们选择了一个假设之后，真实误差无从得知，但我们可以用某些可以掌握的量来逼近它。最直观的想法就是使用分类器在样本数据上的分类的结果与真实结果之间的差值来表示。这个差值叫做经验风险Remp(w)。

统计学习因此而引入了泛化误差界的概念，就是指真实风险应该由两部分内容刻画，一是经验风险，代表了分类器在给定样本上的误差；二是置信风险，代表了我们在多大程度上可以信任分类器在未知文本上分类的结果。很显然，第二部分是没有办法精确计算的，因此只能给出一个估计的区间，也使得整个误差只能计算上界，而无法计算准确的值。

置信风险与两个量有关，一是样本数量，显然给定的样本数量越大，我们的学习结果越有可能正确，此时置信风险越小；二是分类函数的VC维，显然VC维越高，推广能力越差，置信风险会变大。泛化误差界的公式为：

公式中R(w)就是真实风险，Remp(w)就是经验风险，Ф(n/h)就是置信风险。统计学习的目标从经验风险最小化变为了寻求经验风险与置信风险的和最小，即结构风险最小。

### 3.2.3 支持向量机的基本原理

#### 3.2.3.1 间隔与支持向量



图3.1 存在多个划分超平面将两类训练样本分开

Fig 3.1 Multiple Classification Hyperplane deparated the two kinds of Training Sample

给定如图上训练样本集，其中方点和圆点分别代表两个类别，分类学习最基本的思路就是基于训练集在样本空间中找到一个划分超平面，将不同类别的样本分开，但能将样本分开的划分超平面可能有很多，如图就有无数条直线可以将样本划分开。直观的说，我们应该寻找位于两类样本的最中间的划分超平面，就是图中的红线。虽然图中五条直线都能将训练集的样本划分的很好，但是对于未知的测试集而言，数据的分布有可能很接近中间的分割线，这就会使得很多的划分超平面不能有效的划分，而红色的超平面受到的影响最小。这就是划分超平面的泛化能力和鲁棒性的体现。

划分超平面的表示如下：

其中为法向量；为位移项，决定了超平面与原点之间的距离。

超平面由决定，记为。样本空间的任意一点到超平面的距离：

假设超平面能将训练集样本正确分类，即对于，若，则有；若，则有。令

如图所示，距离超平面最近的几个训练样本点使式（3.4）的等号成立，这些点就是“支持向量”（support vector），两个异类支持向量到超平面的距离之和就是

这就是间隔（margin）。



图3.2 支持向量与间隔

Fig 3.2 Support vector and margin

只有确保r最大，也就是间隔最大，才能保证对于测试集的样本具有最准确的判断。想要找到具有“最大间隔”的划分超平面，也就是要找到能满足式（3.4）中约束的参数，使得r最大，即

显然，为了最大化间隔，仅需最大化,这等价于最小化。于是，式（3.6）可以改成为

这就是支持向量机的基本型。

#### 3.2.3.2 对偶问题

根据式（3.6）得到大间隔划分超平面所对应的模型

其中和是模型参数。注意到式（3.6）是一个凸二次规划（Convex Quadratic Programming）问题，能直接用现成的优化计算包求解，但我们可以有更高效的办法。

对式（3.6）使用拉格朗日乘子法可得到其“对偶问题”（Dual Problem）。具体来说，对式（3.6）的每条约束添加拉格朗日乘子，则该问题的拉格朗日函数可写为

其中。令对和的偏导数为零可得

将式（3.10）代入式（3.9），即可将中的和消去，在考虑式（3.11）的约束，就得到式（3.7）的对偶问题

解出后，求出和即可得到模型

### 3.2.4 核函数

在前面的内容中，我们假设训练样本是线性可分的，即存在一个划分超平面能将训练样本正确分类。但是往往在实际应用中，样本往往不是线性可分的，这就使得我们需要寻求一个新的方式使得线性不可分的样本变得线性可分。



图 3.3 线性不可分问题

Fig 3.3 Linear inseparable problem

图3.3的坐标系中横轴上的两点a和b之间的红色部分定义为一类，两边的黑色部分定义为另一类。现在要寻找一个超平面来将两类分割开来，这个超平面很明显不可能是一条直线，但是我们可以找到一条曲线，例如下面这一条：



图3.4 线性不可分问题变为线性可分问题

Fig 3.4 Linear Inseparable Problem into a Linear Separable Problem

如图3.4可知，在这条曲线的上方就是红色部分的一类，曲线下方就是黑色部分的一类。这条曲线是一条二次曲线。

线性回归的线性不是在看自变量是否线性相关，而是看回归系数是不是线性。这样的函数又是线性函数了。因此我们将一些非线性问题通过向高维转化，将其转化成线性问题来求解。这一过程最关键的部分就在于向高维转化的方法。

令表示将映射后的特征向量，于是，在特征空间中划分超平面所对应的模型可表示为

其中和是模型参数。类似式（3.6），有

其对偶问题是

求解式（3.16）涉及到计算，这是样本映射到特征空间之后的内积。由于特征空间位数可能很高，甚至可能是无穷维，因此直接计算通常是困难的。为了避开这个障碍，可以设想这样一个函数：

即在特征空间的内积等于他们在原始样本空间中通过函数计算的结果。有了这样的函数，我们就不用直接去计算高维甚至无穷维特征空间中的内积，于是式（3.16）可以写为

求解后可得到

这里的函数就是“核函数”（kernel function）。式（3.19）显示出模型最优解可通过训练样本的核函数展开，这一展开式就是“支持向量展式”（support vector expansion）。

核函数必须满足下面的条件：当且仅当对于任意数据,核矩阵K总是半正定的。

也就是说只要一个对称函数所对应的核矩阵半正定，它就能作为核函数使用。

## 3.3 SVM实现与Python介绍

### 3.3.1 Python语言简介

自从20世纪90年代初Python语言诞生至今，它已被逐渐广泛应用于系统管理任务的处理和[Web](http://baike.baidu.com/subview/3912/15992867.htm)编程。

Python是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言。它具有多种特点：易于学习、易于阅读、易于维护、广泛的标准库、交互式、可移植、可扩展性、数据库GUI编程、可嵌入等。Python凭借着自身的诸多优势，广泛应用于系统编程、图形处理、数学处理、文本处理、数据库编程、网络编程、Web编程、多媒体应用、黑客编程和数据挖掘等领域或方向。

Python具有丰富和强大的库，并能够把用其他语言制作的各种模块联结在一起。常见的一种应用情形是，使用Python快速生成程序的原型，然后对其中有特别要求的部分，用更合适的语言改写，比如[3D游戏](http://baike.baidu.com/view/96860.htm)中的图形渲染模块，性能要求特别高，就可以用C/C++重写，而后封装为Python可以调用的扩展类库。需要注意的是在您使用扩展类库时可能需要考虑平台问题，某些可能不提供[跨平台](http://baike.baidu.com/view/469855.htm)的实现。

众多开源的科学计算软件包都提供了Python的调用[接口](http://baike.baidu.com/view/159864.htm)，例如著名的计算机视觉库[OpenCV](http://baike.baidu.com/view/1343775.htm)、三维可视化库VTK、医学图像处理库ITK。而Python专用的科学计算扩展库就更多了，例如经典的科学计算扩展库：NumPy、SciPy和Matplotlib，它们分别为Python提供了快速数组处理、数值运算以及绘图功能。而本论文在使用这三个标准库的基础上，还使用面向机器学习的Scikit-learn标准模块库。

### 3.3.2 Scikit-learn

Scikit-learn是一个开源的基于Python的科学计算工具包。基于SciPy，目前开发者们针对不同的应用领域已经发展出了为数众多的分支版本，它们被统一称为Scikits，即SciPy工具包的意思。而在这些分支版本中，最有名，也是专门面向机器学习的一个就是Scikit-learn。

Scikit-learn项目最早由数据科学家David Cournapeau在2007年发起，需要NumPy和SciPy等其他包的支持，是Python语言中专门针对机器学习应用而发展起来的一款开源框架。

和其他众多的开源项目一样，Scikit-learn目前主要由社区成员自发进行维护。可能是由于维护成本的限制，Scikit-learn相比其他项目要显得更为保守。这主要体现在两个方面：一是Scikit-learn从来不做除机器学习领域之外的其他扩展，二是Scikit-learn从来不采用未经广泛验证的算法。

Scikit-learn的基本功能主要被分为六大部分：分类，回归，聚类，数据降维，模型选择和数据预处理。

分类是指识别给定对象的所属类别，属于监督学习的范畴，最常见的应用场景包括垃圾邮件检测和图像识别等。目前Scikit-learn已经实现的算法包括：支持向量机（SVM），最近邻，逻辑回归，随机森林，决策树以及多层感知器（MLP）[神经网络](http://www.leiphone.com/news/201505/t3T1XQy2g3spCUdd.html)等等。需要指出的是，由于Scikit-learn本身不支持[深度学习](http://www.leiphone.com/news/201701/LqwiP7VUJO9DgBPi.html)，也不支持GPU加速，因此这里对于MLP的实现并不适合于处理大规模问题。有相关需求的读者可以查看同样对Python有良好支持的Keras和Theano等框架。

回归是指预测与给定对象相关联的连续值属性，最常见的应用场景包括预测药物反应和预测股票价格等。目前Scikit-learn已经实现的算法包括：支持向量回归（S[VR](http://www.leiphone.com/category/arvr)），脊回归，Lasso回归，弹性网络（Elastic Net），最小角回归（LARS ），贝叶斯回归，以及各种不同的鲁棒回归算法等。可以看到，这里实现的回归算法几乎涵盖了所有开发者的需求范围，而且更重要的是，Scikit-learn还针对每种算法都提供了简单明了的用例参考。

聚类是指自动识别具有相似属性的给定对象，并将其分组为集合，属于无监督学习的范畴，最常见的应用场景包括顾客细分和试验结果分组。目前Scikit-learn已经实现的算法包括：K-均值聚类，谱聚类，均值偏移，分层聚类，DBSCAN聚类等。

数据降维是指使用主成分分析（PCA）、非负矩阵分解（NMF）或特征选择等降维技术来减少要考虑的随机变量的个数，其主要应用场景包括可视化处理和效率提升。

模型选择是指对于给定参数和模型的比较、验证和选择，其主要目的是通过参数调整来提升精度。目前Scikit-learn实现的模块包括：格点搜索，交叉验证和各种针对预测误差评估的度量函数。

数据预处理是指数据的特征提取和归一化，是机器学习过程中的第一个也是最重要的一个环节。这里归一化是指将输入数据转换为具有零均值和单位权方差的新变量，但因为大多数时候都做不到精确等于零，因此会设置一个可接受的范围，一般都要求落在0-1之间。而特征提取是指将文本或图像数据转换为可用于机器学习的数字变量。

总结来说，Scikit-learn实现了一整套用于数据降维，模型选择，特征提取和归一化的完整算法/模块，虽然缺少按步骤操作的参考教程，但Scikit-learn针对每个算法和模块都提供了丰富的参考样例和详细的说明文档。

总体上来说，作为专门面向机器学习的Python开源框架，Scikit-learn可以在一定范围内为开发者提供非常好的帮助。它内部实现了各种各样成熟的算法，容易安装和使用，样例丰富，而且教程和文档也非常详细。

应该明确的一点是：虽然概括地说Scikit-learn并不适合深度学习问题，但对于某些特殊场景而言，使用Scikit-learn仍然是明智的选择。例如要创建连接不同对象的预测函数时，或者在未标记的数据集中为了训练模型对不同的对象进行分类时，面对这些场景Scikit-learn只通过普通的旧机器学习模型就能很好地解决，而并不需要建立数十层的复杂神经网络。

# 第4章 SVM船舶柴油机故障诊断技术

## 4.1 燃油系统常见故障分析

燃油系统常见的故障有供油量不足，针阀卡死至油孔堵塞，针阀泄漏，出油阀失效等几种故障。本系统诊断的故障也是这九种故障，依次为：100%供油量（正常情况T1），75%供油量（T2），50%供油量（T3），25%供油量（T4），怠速油量（T5），针阀卡死1（小油量 T6），针阀卡死2（标定油量T7），针阀泄漏（T8），出油阀失效（T9）。

表4.1 正常喷油（T1）状态下的9组数据

Table 4.1 Normal Injection (T1) Condition of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.7557 | 0.2281 | 0.15586 | 0.15471 | 0.1547 | 0.16109 | 0.12739 | 0.14335 |
| 2 | 0.74378 | 0.27833 | 0.16303 | 0.14216 | 0.13223 | 0.17908 | 0.12458 | 0.13799 |
| 3 | 0.7409 | 0.25091 | 0.16728 | 0.15815 | 0.14868 | 0.16707 | 0.11976 | 0.13723 |
| 4 | 0.73224 | 0.23097 | 0.16936 | 0.14613 | 0.14371 | 0.16887 | 0.11523 | 0.1317 |
| 5 | 0.73144 | 0.25232 | 0.16874 | 0.15266 | 0.1392 | 0.17303 | 0.11881 | 0.13966 |
| 6 | 0.72911 | 0.23732 | 0.15872 | 0.15528 | 0.13195 | 0.17902 | 0.12094 | 0.13352 |
| 7 | 0.72524 | 0.26616 | 0.17069 | 0.13656 | 0.15055 | 0.16867 | 0.1299 | 0.14379 |
| 8 | 0.72113 | 0.27973 | 0.1643 | 0.14411 | 0.13204 | 0.16895 | 0.12301 | 0.13987 |
| 9 | 0.7199 | 0.26779 | 0.16372 | 0.13804 | 0.14056 | 0.16546 | 0.12122 | 0.13104 |

表4.2 75%油量（T2）状态下的9组数据

Table 4.2 75% oil (T2) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.63866 | 0.24583 | 0.16032 | 0.13619 | 0.14454 | 0.16335 | 0.12175 | 0.13013 |
| 2 | 0.63945 | 0.22378 | 0.15052 | 0.13072 | 0.1395 | 0.16156 | 0.11221 | 0.14157 |
| 3 | 0.65915 | 0.22362 | 0.17094 | 0.1206 | 0.13678 | 0.17068 | 0.10899 | 0.13969 |
| 4 | 0.65476 | 0.20755 | 0.15371 | 0.13923 | 0.13552 | 0.16192 | 0.10796 | 0.13522 |
| 5 | 0.65435 | 0.23175 | 0.1607 | 0.13400 | 0.13761 | 0.16500 | 0.11131 | 0.13305 |
| 6 | 0.65383 | 0.21006 | 0.15928 | 0.12955 | 0.13437 | 0.16080 | 0.11594 | 0.13722 |
| 7 | 0.649 | 0.24696 | 0.15175 | 0.14132 | 0.13986 | 0.15836 | 0.11953 | 0.13676 |
| 8 | 0.64525 | 0.2176 | 0.14548 | 0.13141 | 0.14362 | 0.16199 | 0.11669 | 0.12979 |
| 9 | 0.63985 | 0.24984 | 0.13923 | 0.14583 | 0.13332 | 0.16815 | 0.11611 | 0.13562 |

表4.3 50%油量（T3）状态下的9组数据

Table 4.3 50% oil (T3) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.63601 | 0.20301 | 0.13059 | 0.13624 | 0.13365 | 0.17147 | 0.11153 | 0.1296 |
| 2 | 0.63341 | 0.21267 | 0.12507 | 0.15017 | 0.13234 | 0.17239 | 0.11243 | 0.13088 |
| 3 | 0.63969 | 0.19848 | 0.13967 | 0.1501 | 0.13726 | 0.15942 | 0.11114 | 0.12769 |
| 4 | 0.6258 | 0.20312 | 0.14007 | 0.12774 | 0.12774 | 0.16223 | 0.12119 | 0.13590 |
| 5 | 0.63303 | 0.2169 | 0.15614 | 0.13074 | 0.14562 | 0.16944 | 0.11076 | 0.12668 |
| 6 | 0.63081 | 0.2534 | 0.15453 | 0.13511 | 0.12827 | 0.16339 | 0.10925 | 0.12936 |
| 7 | 0.63026 | 0.22838 | 0.14968 | 0.12263 | 0.13745 | 0.16913 | 0.11182 | 0.1314 |
| 8 | 0.63012 | 0.18489 | 0.15069 | 0.13565 | 0.13967 | 0.15574 | 0.11694 | 0.12756 |
| 9 | 0.62716 | 0.24847 | 0.14083 | 0.14644 | 0.12596 | 0.16922 | 0.12227 | 0.13057 |

表4.4 25%油量（T4）状态下的9组数据

Table 4.4 25% oil (T4) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.59585 | 0.16801 | 0.12161 | 0.15593 | 0.13886 | 0.17724 | 0.11401 | 0.13236 |
| 2 | 0.59107 | 0.1768 | 0.12974 | 0.15775 | 0.12655 | 0.17657 | 0.11766 | 0.13552 |
| 3 | 0.59106 | 0.16276 | 0.12493 | 0.16816 | 0.1261 | 0.15913 | 0.11242 | 0.1225 |
| 4 | 0.58964 | 0.16963 | 0.13721 | 0.17359 | 0.13188 | 0.16418 | 0.11257 | 0.13114 |
| 5 | 0.58921 | 0.19515 | 0.13816 | 0.16046 | 0.13049 | 0.17993 | 0.11011 | 0.13759 |
| 6 | 0.58549 | 0.21066 | 0.12431 | 0.14306 | 0.11699 | 0.16357 | 0.12468 | 0.13031 |
| 7 | 0.58386 | 0.24012 | 0.14612 | 0.12278 | 0.14166 | 0.17095 | 0.11248 | 0.13388 |
| 8 | 0.59651 | 0.15270 | 0.12573 | 0.16935 | 0.13109 | 0.16254 | 0.11023 | 0.12884 |
| 9 | 0.58425 | 0.17945 | 0.11960 | 0.15620 | 0.13700 | 0.16996 | 0.11467 | 0.13457 |

表4.5 怠速供油（T5）状态下的9组数据

Table 4.5 the idle oil (T5) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.4976 | 0.15437 | 0.12598 | 0.11538 | 0.12413 | 0.16705 | 0.10656 | 0.14112 |
| 2 | 0.49516 | 0.14475 | 0.10971 | 0.13055 | 0.12924 | 0.16583 | 0.11556 | 0.12718 |
| 3 | 0.48842 | 0.15474 | 0.11329 | 0.13325 | 0.12826 | 0.15826 | 0.12064 | 0.12546 |
| 4 | 0.48163 | 0.16928 | 0.11871 | 0.12261 | 0.12645 | 0.16429 | 0.11859 | 0.12671 |
| 5 | 0.48668 | 0.15295 | 0.11100 | 0.11897 | 0.12589 | 0.16169 | 0.11263 | 0.12821 |
| 6 | 0.49771 | 0.17716 | 0.11540 | 0.10921 | 0.12592 | 0.1614 | 0.11492 | 0.13006 |
| 7 | 0.48644 | 0.17303 | 0.12527 | 0.12977 | 0.12542 | 0.15877 | 0.12227 | 0.13528 |
| 8 | 0.48501 | 0.15383 | 0.11210 | 0.12587 | 0.13185 | 0.16194 | 0.12476 | 0.13107 |
| 9 | 0.48422 | 0.17133 | 0.11847 | 0.12382 | 0.13013 | 0.15795 | 0.11015 | 0.12773 |

表4.6 针阀卡死1（T6）状态下的9组数据

Table 4.6 needle valve stuck 1 (T6) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.67008 | 0.26384 | 0.17326 | 0.16429 | 0.10947 | 0.16584 | 0.11185 | 0.12091 |
| 2 | 0.62966 | 0.31009 | 0.18241 | 0.13428 | 0.1375 | 0.15234 | 0.10000 | 0.12354 |
| 3 | 0.67854 | 0.22976 | 0.15464 | 0.17939 | 0.13158 | 0.14428 | 0.10510 | 0.12154 |
| 4 | 0.65998 | 0.20437 | 0.13257 | 0.16685 | 0.12162 | 0.1724 | 0.11943 | 0.12557 |
| 5 | 0.63853 | 0.27977 | 0.16641 | 0.14078 | 0.14675 | 0.15808 | 0.11176 | 0.12886 |
| 6 | 0.64923 | 0.22569 | 0.14889 | 0.18055 | 0.13315 | 0.14265 | 0.10678 | 0.11557 |
| 7 | 0.64851 | 0.22854 | 0.14177 | 0.15824 | 0.13375 | 0.17444 | 0.11711 | 0.12459 |
| 8 | 0.64103 | 0.32548 | 0.18744 | 0.13053 | 0.14048 | 0.14895 | 0.10206 | 0.12834 |
| 9 | 0.64038 | 0.32152 | 0.18154 | 0.13718 | 0.15908 | 0.14838 | 0.10717 | 0.13622 |

表4.7 针阀卡死2（T7）状态下的9组数据

Table 4.7 needle valve stuck 2 (T7) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.86874 | 0.29701 | 0.18691 | 0.16495 | 0.14632 | 0.14877 | 0.11690 | 0.12636 |
| 2 | 0.86561 | 0.30018 | 0.19140 | 0.1633 | 0.14303 | 0.16268 | 0.12125 | 0.12507 |
| 3 | 0.86156 | 0.32864 | 0.20148 | 0.18497 | 0.13756 | 0.16203 | 0.12158 | 0.13075 |
| 4 | 0.86021 | 0.30849 | 0.19929 | 0.17966 | 0.13714 | 0.16386 | 0.12408 | 0.13268 |
| 5 | 0.85837 | 0.38187 | 0.18893 | 0.16082 | 0.15434 | 0.14493 | 0.10045 | 0.13701 |
| 6 | 0.90000 | 0.19040 | 0.16108 | 0.14358 | 0.15992 | 0.12018 | 0.12700 | 0.29788 |
| 7 | 0.84418 | 0.35251 | 0.1986 | 0.17765 | 0.15517 | 0.1577 | 0.11762 | 0.13701 |
| 8 | 0.85781 | 0.31088 | 0.20372 | 0.1874 | 0.14614 | 0.1502 | 0.11325 | 0.13490 |
| 9 | 0.85069 | 0.34529 | 0.19014 | 0.17468 | 0.14273 | 0.15499 | 0.11341 | 0.13317 |

表4.8 针阀泄漏（T8）状态下的9组数据

Table 4.8 needle valve leakage (T8) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.69644 | 0.22654 | 0.15001 | 0.15022 | 0.13759 | 0.16145 | 0.11456 | 0.12451 |
| 2 | 0.69128 | 0.26862 | 0.1506 | 0.13995 | 0.12654 | 0.15927 | 0.10652 | 0.12452 |
| 3 | 0.6884 | 0.22367 | 0.15434 | 0.15514 | 0.14198 | 0.15405 | 0.11635 | 0.12645 |
| 4 | 0.68592 | 0.23286 | 0.12923 | 0.15764 | 0.12569 | 0.15864 | 0.11518 | 0.13066 |
| 5 | 0.68496 | 0.24594 | 0.16993 | 0.14384 | 0.12628 | 0.15579 | 0.11548 | 0.13391 |
| 6 | 0.6845 | 0.22787 | 0.15609 | 0.13999 | 0.13234 | 0.1522 | 0.11710 | 0.12979 |
| 7 | 0.68202 | 0.2496 | 0.15431 | 0.13532 | 0.13087 | 0.15811 | 0.10670 | 0.12397 |
| 8 | 0.67888 | 0.21911 | 0.14028 | 0.16365 | 0.13674 | 0.16136 | 0.10575 | 0.11996 |
| 9 | 0.70786 | 0.23190 | 0.13993 | 0.14184 | 0.13266 | 0.14887 | 0.11745 | 0.13671 |

表4.9 出油阀泄漏（T9）状态下的9组数据

Table 4.9 delivery valve leakage (T9) of 9 sets of data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 1 | 0.76061 | 0.25103 | 0.15455 | 0.16977 | 0.15036 | 0.16779 | 0.12840 | 0.14224 |
| 2 | 0.76042 | 0.22351 | 0.14963 | 0.17073 | 0.13551 | 0.17105 | 0.11886 | 0.13910 |
| 3 | 0.75545 | 0.2457 | 0.14289 | 0.16287 | 0.12928 | 0.16842 | 0.12769 | 0.13402 |
| 4 | 0.75269 | 0.24863 | 0.1751 | 0.14807 | 0.13969 | 0.16781 | 0.11634 | 0.13952 |
| 5 | 0.74851 | 0.29003 | 0.17448 | 0.13540 | 0.13377 | 0.17046 | 0.12993 | 0.13592 |
| 6 | 0.75103 | 0.23735 | 0.1676 | 0.1615 | 0.14311 | 0.16944 | 0.12232 | 0.13836 |
| 7 | 0.75035 | 0.24078 | 0.16833 | 0.15853 | 0.14357 | 0.16213 | 0.12113 | 0.14087 |
| 8 | 0.74986 | 0.24547 | 0.16595 | 0.15154 | 0.14302 | 0.16982 | 0.12104 | 0.14655 |
| 9 | 0.76067 | 0.21935 | 0.15168 | 0.16554 | 0.14553 | 0.15967 | 0.11545 | 0.13909 |

这组数据来自于[19]，其中每一行都是一组数据，F0、F1、F2、F3、F4、F5、F6、F7和F8表示油压波形信号的八个频段，并将其作为RBF网络的输入向量，然后利用数据的归一化最后得到以上81组数据。

## 4.2 支持向量机的学习条件

### 4.2.1 交叉验证

交叉验证（Cross Validation）是用来验证分类器性能的一种统计分析方法。基本思想是把在某种意义下的原始数据（Dataset）进行分组，一部分作为训练集（Training set），另一部分作为验证集（Testing set）。首先用训练集对分类器进行训练，再利用验证集来测试训练得到的模型（Model）。以此作为评价分类器的性能指标。常见的交叉验证方法有：留出法（Hold-out Method）、K折交叉验证（K-fold Cross Validation）、留一法（Leave-one-out Cross Validation）。

#### 4.2.1.1 留出法

留出法直接将数据集D划分为两个互斥的集合，其中一个集合作为训练集S，另一个作为测试集T，且两者没有交集。在S上训练出模型后，用T来评估其测试误差，从而作为对泛化误差的估计。

通常而言，一般选用数据集样本数的70%作为训练集，另外30%作为测试集。以测试集的分类错误作为其错误率，以测试集的分类正确率作为其精度。

需要注意的是，训练集和测试集的划分要尽可能保持数据的一致性，避免因数据划分过程引入额外的偏差而对最终结果产生影响。例如在分类任务中至少要保持样本的类别比例相似。如果从采样的角度来看待数据集的划分过程，这保留类别比例的采样方式通常被称为“分层采样”。例如通过对D进行分层采样而获得含有70%样本的训练集S和含有30%样本的测试集T。若D包含500个正例和500个反例，这分层采样得到的S应该包含350个正例和350个反例，而T也包含150个正例和150个反例。若S、T中样本类别比例差别很大，则误差估计将由于训练、测试数据分布的差异而产生偏差。

#### 4.2.1.2 K折交叉验证

K折交叉验证先将数据集D划分为k个大小相似的互斥子集，即每个子集之间都没有交集。每个子集都尽可能保持数据分布的一致性，即从D中通过分层采样得到。然后每次k-1个子集的并集作为训练集，余下的那个子集作为测试集，这样就可以获得k组训练集和测试集，从而可进行k次训练和测试，最后返回的是这k个测试结果的均值。

显然，k折交叉验证评估结果的稳定性和保真性在很大程度上取决于k的取值，为强调这一点，所以这一方法的名字称为“K折交叉验证”。K最常用的取值是10，此时成为10这交叉验证；其他常用的k值有5、20等。图给出了10这交叉验证的示意图。

与留出法相似，将数据集D划分为k个子集同样存在多种划分方式。为减少因样本划分不同而引入的差别，k折交叉验证通常要随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的有“10次10折交叉验证”。



图4.1 十折交叉验证示意图

Fig 4.1 10 Fold Cross-Validation

#### 4.2.1.3 留一法

假定数据集D中包含m个样本，若令k=m，这得到了交叉验证法的一个特例：留一法。显然留一法不受随机样本划分方式的影响，因为m个样本只有唯一的方式划分为m个子集，每个自己包含一个样本。留一法使用的训练集与初始数据集相比只少了一个样本，这就使得绝大多数情况下，留一法被实际评估的模型与期望评估的用D训练出的模型很相似。因此，留一法的评估结果往往被认为比较准确。然而，留一法也有其缺陷：在数据集样本数较大时，训练m个模型的计算开销可能是演绎忍受的，比如数据集包含一百万个样本，则需训练一百万个模型。而这还是未考虑算法调参数的情况下。另外，留一法的估计结果也未必永远比其他评估方法准确。

在本文中，我们使用K折交叉验证，其中K值取9。这就表示我们将81组数据集分成9大组数据集，每一次学习依次选择一组作为测试集，另外八组作为训练集来训练支持向量机的模型。其中本次学习的交叉验证代码如下：

num=-1

kfold1 = KFold(9, n\_folds=9,shuffle=True)

data1 = np.loadtxt("d1", dtype = np.float)

X1 = data1[:,:num]

y1 = data1[:,num:]

data2 = np.loadtxt("d2", dtype = np.float)

X2 = data2[:,:num]

y2 = data2[:,num:]

data3 = np.loadtxt("d3", dtype = np.float)

X3 = data3[:,:num]

y3 = data3[:,num:]

data4 = np.loadtxt("d4", dtype = np.float)

X4 = data4[:,:num]

y4 = data4[:,num:]

data5 = np.loadtxt("d5", dtype = np.float)

X5 = data5[:,:num]

y5 = data5[:,num:]

data6 = np.loadtxt("d6", dtype = np.float)

X6 = data6[:,:num]

y6 = data6[:,num:]

data7 = np.loadtxt("d7", dtype = np.float)

X7 = data7[:,:num]

y7 = data7[:,num:]

data8 = np.loadtxt("d8", dtype = np.float)

X8 = data8[:,:num]

y8 = data8[:,num:]

data9 = np.loadtxt("d9", dtype = np.float)

X9 = data9[:,:num]

y9 = data9[:,num:]

for train, test in kfold1:

X1\_train, X1\_test = X1[train], X1[test]

y1\_train, y1\_test = y1[train], y1[test]

X2\_train, X2\_test = X2[train], X2[test]

y2\_train, y2\_test = y2[train], y2[test]

X3\_train, X3\_test = X3[train], X3[test]

y3\_train, y3\_test = y3[train], y3[test]

X4\_train, X4\_test = X4[train], X4[test]

y4\_train, y4\_test = y4[train], y4[test]

X5\_train, X5\_test = X5[train], X5[test]

y5\_train, y5\_test = y5[train], y5[test]

X6\_train, X6\_test = X6[train], X6[test]

y6\_train, y6\_test = y6[train], y6[test]

X7\_train, X7\_test = X7[train], X7[test]

y7\_train, y7\_test = y7[train], y7[test]

X8\_train, X8\_test = X8[train], X8[test]

y8\_train, y8\_test = y8[train], y8[test]

X9\_train, X9\_test = X9[train], X9[test]

y9\_train, y9\_test = y9[train], y9[test]

X\_train=np.concatenate((X1\_train,X2\_train,X3\_train,X4\_train,X5\_train,X6\_train,X7\_train,X8\_train,X9\_train),axis=0)

y\_train=np.concatenate((y1\_train,y2\_train,y3\_train,y4\_train,y5\_train,y6\_train,y7\_train,y8\_train,y9\_train),axis=0)

X\_test =np.concatenate((X1\_test,X2\_test,X3\_test,X4\_test,X5\_test,X6\_test,X7\_test,X8\_test,X9\_test),axis=0)

y\_test =np.concatenate((y1\_test,y2\_test,y3\_test,y4\_test,y5\_test,y6\_test,y7\_test,y8\_test,y9\_test),axis=0)

### 4.2.2 惩罚项参数和松弛参数

大多数学习算法都有有些参数（Parameter）需要设定，参数配置不同，学得模型的性能往往有显著差别。因此，在进行模型评估与选择时，除了要对适用学习算法进行选择，还需对算法参数进行设定，这就是通常所说的“参数调节”或简称“调参”（Parameter Tuning）。

对每种参数配置都悬链出模型，然后把对应最好模型的参数作为结果。这样的考虑基本是正确的，当学习算法的很多参数是在实数范围内取值，因此，对每种参数配置都训练出模型来是不可行的。现实中常用的做法，是对每个参数选定一个范围和变化步长。

显然这样选定的参数往往不是“最佳”值，但这是在计算开销的性能估计之间进行折中的结果，通过这个折中，学习过程才变得可行。事实上，即便在进行这样的折中之后，调参往往仍很困难。

在本次学习中我们选定惩罚项参数C为0.001-1000，Gamma为0.001-1000。步长为一个次方。

### 4.2.3 多分类学习

现实中常遇到多分类学习，有些二分类学习方法可直接推广到多分类，但在更多情形下，我们是基于一些基本策略，利用二分类学习器来解决多分类问题。

一般来说，考虑N个类别，多分类学习的基本思路是拆解法，即将多分类任务拆为若干个而分类任务求解。具体来说，先对问题进行拆分，然后为拆除的每个二分类任务训练一个分类器；在测试时，对这些分类器的预测结果进行集成以获得最终的多分类结果。这里的关键是如何对多分类任务进行拆分，以及如何对多个分类器进行集成。

经典的拆分策略有三个：“一对一”（One vs. One，简称OvO）、“一对其余”（One vs. Rest，简称OvR）和“多对多”（Many vs. Many，简称MvM）。

#### 4.2.3.1 OvO

给定数据集C，OvO将这N个类别两两配对，从而产生N(N-1)/2个二分类任务，例如OvO将为区分类别Ci和Cj训练一个分类器，该分类器把Ci类样例作为正例，Cj类样例作为反例。在测试阶段，新样本将同时提交给所有分类器，于是将得到个分类结果，最终结果可通过投票产生：即把被预测的最多的类别作为最终分类结果，如图4.2。



图4.2 OvO示意图

Fig 4.2 OvO Schematic Diagram

#### 4.2.3.2 OvR

OvR是每次将一个类的样例作为正例，所有其他类的样例作为反例来训练N个分类器。在测试时若仅有一个分类器预测为正类，则对应的类别标记作为最终分类结果，如图4.3。若有多个分类器预测为正类，则通常考虑各分类器的预测置信度，选择置信度最大的类别标记作为分类结果。

不难看出，OvR只需训练N个分类器，而OvO需训练N(N-1)/2个分类器，因此，OvO的存储开销和测试时间开销通常比OvR更大。但在训练时，OvR的每个分类器均使用全部训练样例，而OvO的每个分类器仅用到两个样例。因此，在类别很多的时候，OvO的训练时间开销通常比OvR更小。至于预测性能，则取决于具体的数据分布，在多数情况下两者差不多。



图4.3 OvR示意图

Fig 4.3 OvR Schematic Diagram

#### 4.2.3.3 MvM

MvM是每次将若干个类作为正类，若干个其他类作为反类。显然，OvO和OvR是MvM的特例。MvM的正反类构造必须有特殊的设计，不能随意取。通常使用的MvM技术主要是纠错输出码（Error Correcting Output Codes）。

## 4.3 SVM故障诊断技术实现

本课题共进行了5组基于不同核函数的分类学习，其中核函数分别是线性核函数、多项式核函数、径向基核函数、sigmoid核函数和自定义的优化线性核函数。5组分类学习都采用九折交叉验证来对数据集处理产生训练集和测试集，采用OVR进行多分类处理，其中惩罚项参数从10-3取到103，步长为一个次方，松弛参数从10-5取到105，步长也是一个次方。

### 4.3.1 线性核函数分类学习

线性核函数的表达式如（4.1）。

主要用于线性可分的情形，参数少，速度快，对于一般数据，尤其是线性可分的数据，分类效果已经很理想了。这实际上就是原始空间中的内积。这个核存在的主要目的是使得“ 映射后空间中的问题”和“映射前空间中的问题”两者在形式上统一起来了。

表4.10 线性核函数分类学习在不同参数情况下的测试集预测准确率

Table 4.10 Linear Kernel Function Classification Learning under the condition of different parameters of Testing Set Prediction Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 0.001 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 0.01 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 0.1 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 1 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 10 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 100 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |

### 4.3.2 多项式核函数分类学习

多项式核函数是典型的全局核函数，多项式核函数的表达式如（4.2）。

其中，参数d为多项式的阶数。由上式可以看出两者之间的实际距离的多少，训练集X中每个点都能对测试集Y的核函数值产生影响。相对于径向基核函数而言，多项式核函数虽然插值能力相对较弱，但比较善于提取样本的全局特性。

表4.11 多项式核函数分类学习在不同参数情况下的测试集预测准确率

Table 4.11 Polynomial Kernel Function Classification Learning under the condition of different parameters of Testing Set Prediction Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 0.001 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.888889 |
| 0.01 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.901235 |
| 0.1 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.888889 | 0.901235 |
| 1 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.901235 | 0.901235 |
| 10 | 0.876543 | 0.876543 | 0.876543 | 0.876543 | 0.888889 | 0.901235 |
| 100 | 0.876543 | 0.876543 | 0.876543 | 0.888889 | 0.901235 | 0.901235 |

### 4.3.3 径向基核函数多分类学习

径向基核函数的表达式如下:

径向基核函数主要处理非线性的情况，这个核函数就是最开始提到过的会将原始空间映射为无穷维空间的那个家伙。不过，如果选得很大的话，高次特征上的权重实际上衰减得非常快，所以实际上（数值上近似一下）相当于一个低维的子空间；反过来，如果选得很小，则可以将任意的数据映射为线性可分。当然，这并不一定是好事，因为随之而来的可能是非常严重的过拟合问题。不过，总的来说，通过调控参数，高斯核实际上具有相当高的灵活性，也是使用最广泛的核函数之一。

表4.12 径向基核函数分类学习在不同参数情况下的测试集预测准确率

Table 4.12 RBF Kernel Function Classification Learning under the condition of different parameters of Testing Set Prediction Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 0.001 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 0.01 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 0.1 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 |
| 1 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.888889 | 0.864198 |
| 10 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.864198 | 0.876543 |
| 100 | 0.864198 | 0.864198 | 0.864198 | 0.888889 | 0.901235 | 0.901235 |

### 4.3.4 Sigmoid核函数分类学习

Sigmoid核函数来自于神经网络，在实际应用中受到限制，主要是因为只有满足某些特定条件（参数和满足一定的条件），Sigmoid才能满足对称、半正定的核函数条件。其表达式如式（4.4）。

当参数和分别取不同值时，Sigmoid核函数有着不同性质。当时，是输入数据的一个幅度调节参数，是一个控制映射阈值的位移参数。当时，输入数据的内积不仅被缩放而且方向也变了。不同参数组合下的sigmoid核函数的性质见表，可以看出：对于Sigmoid核函数来说，第一种情况，即当时，是最合适的选择。

表4.13 不同参数组合下的Sigmoid核函数性质

Table 4.13 Sigmoid Kernel Function Properties under different parameter combinations

|  |  |  |
| --- | --- | --- |
|  |  | 结论 |
|  |  | 当很小时，K是CPD的，同时当趋于0时，与RBF核相似 |
|  |  | 一般来说，不如第一种情况好 |
|  |  | 当足够大时，目标函数趋于负无穷 |
|  |  | 目标函数很容易趋于负无穷 |

具体情形可参考文献[49].

1998年，Platt提出了专门处理大规模问题的新的支持向量机的训练算法，就是序列最小最优化（SMO）算法[60]。SMO算法可以看成是分解算法的一个特例，它将之问题的规模减少到了最小（两个样本）。由于子问题的优化只涉及两个变量，而且其中一个变脸可用另一个变量线性表出，所以迭代过程中每一步的子问题的最优解可以直接用解析的方法求出来，而不用在算法的迭代中求解二次规划问题。

表4.14 径向基核函数分类学习在不同参数情况下的测试集预测准确率

Table 4.14 RBF Kernel Function Classification Learning under the condition of different parameters of Testing Set Prediction Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 0.001 | 0.864198 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |
| 0.01 | 0.864198 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |
| 0.1 | 0.679012 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |
| 1 | 0.679012 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |
| 10 | 0.864198 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |
| 100 | 0.864198 | 0.864198 | 0.864198 | 0.679012 | 0.679012 | 0.679012 |

### 4.3.5 优化线性核函数分类学习

优化线性核函数是本文基于线性核函数提出的一项改进后的核函数。其表达式如下：

其中，A作为调整矩阵，根据数据集的分布做过很多次调整，确保每组数据的每一维数据在各个类别中能占到特征选择的较大比重，这个方法可以明确的调整核函数，使之有利于实现分类的准确性。

其中A的矩阵表示如下：

表4.15 优化线性核函数分类学习在不同参数情况下的测试集预测准确率

Table 4.15 Optimization of Linear Kernel Function Classification Learning under the condition of different parameters of Testing Set Prediction Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 0.001 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |
| 0.01 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |
| 0.1 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |
| 1 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |
| 10 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |
| 100 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 | 0.901235 |

## 4.4 多种核函数算法评估对比

通常，对于分类学习算法评估一般有这几种评估参数：准确率（accuracy）、查准率（Precision）、查全率（Recall）、混淆矩阵（Confusion Matrix）、P-R曲线和F1 score。

以二分类问题为例，可将样例根据其真实类别与学习器预测类别的组合划分为四类：真正例（True Positive，TP）、假正例（False Positive，FP）、真反例（True Negative，TN）和假反例（False Negative，FN）。令TP、FP、TN、FN分别表示其对应的样例数，则显然有TP+FP+TN+FN=样本总数。分类结果的混淆矩阵如表所示。

表4.16 二分类结果混淆矩阵

Table 4.16 Binary Classification Results Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 预测结果 | |
| 正例 | 反例 |
| 真实情况 | 正例 | TP（真正例） | FN（假反例） |
| 反例 | FP（假正例） | TN（真反例） |

其中，查准率P与查全率R分别定义为：

查准率和查全率一般是一对互相矛盾的度量。一般来说，查准率高时，查全率往往偏低，而查全率高的时候，查准率往往偏低。在很多情形下，我们可根据学习器的预测结果对样例进行排序，排在前面的是学习器认为“最可能”是正例的样本，排在最后的则是学习器认为“最不可能”是正例的样本。按此顺序逐个吧样本作为正例进行预测，着每次可以计算出当前的查全率、查准率。以查准率为纵轴、查全率为横轴作图，就得到了查准率-查全率，简称“P-R曲线”，显示该曲线的图称为“P-R图”。

P-R图直观地显示出学习器在样本总体上的查全率、查准率。在进行比较时，若一个学习器的P-R曲线被另一个学习器完全“包住”，则可断言后者的性能优于前者。但是假如两个算法的P-R曲线发生了交叉，则很难的断言两个算法的比较。但是在很多情况下，往往还是要将两个算法比较出高低。着一个比较合理的判据是比较P-R曲线下面积的大小，他在一定程度上表征了学习器在查准率和查全率上取得相对“双高”的比例。但这个值不太容易估算。因此，人们设计了一些综合考虑查准率和查全率的性能度量。

在一些应用中，对查准率和查全率的重视程度不同。例如在柴油机燃油故障预测中，预测出结果第一时间去修理，不能出现仔细检查之后发现故障并不是预测的结果。所以这个时候查准率更重要。我们通常使用F1度量算法。F1度量的一般形式是F，能让我们根据需求决定对查准率和查全率的不同偏好。他定义为

其中度量了查全率对查准率的相对重要性。时退化为标准的F1；时查准率有更大影响；查全率有更大影响。

图4.4 5个核函数分类学习在不同参数下的预测准确率

Fig 4.4 5 Kernel Function Classification Learning under different parameters Prediction Accuracy

### 4.4.1 准确率

从图中可以看出，五个核函数测试集的分类准确度随着惩罚项参数和松弛参数的变化而变化。其中，sigmoid核函数的分类学习测试集准确率在0.679012至0.864198之间；线性核函数的分类学习测试集准确率保持在0.864198；多项式核函数的准确率保持在0.876543到0.888889之间；径向基核函数的准确率保持在0.864198；而基于线性核函数优化的分类学习测试集准确率可以达到0.901235。并且在不同的惩罚项系数和松弛参数的情况下都能保持在0.901235不变。仅从分类学习准确率我们可以看出，基于线性优化的核函数不但具有较高的预测精度，同时还具有稳定性很高的预测效果。

### 4.4.2 混淆矩阵、查准率和查全率

下表描述了五个核函数分类学习的混淆矩阵，其中左侧序列表示测试集的真实柴油机故障类别，上侧序列表示支持向量机预测的柴油机故障类别。每一列的总数表示该类别的数据实例的数目。每一列中的数值表示真实数据被预测为该类的数目。比如表4.17的第一行第一列的276表示有276个实际属于第一类的实例被预测为第一类，同理，第一行第九列的48表示归属于第一类的实例被错误的预测为第九类。

表4.17、表4.18、表4.19、表4.20和表4.21分别为线性核函数、多项式核函数、径向基核函数、sigmoid核函数和自定义的优化线性核函数的支持向量机分类学习的混淆矩阵。

表4.17 优化线性核函数分类学习在不同参数情况下的混淆矩阵

Table 4.17 Optimization of Linear Kernel Function Classification Learning under the condition of different parameters of Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 |  |
| T1 | 276 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 0.8518 |
| T2 | 0 | 324 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| T3 | 0 | 36 | 288 | 0 | 0 | 0 | 0 | 0 | 0 | 0.8888 |
| T4 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 0 | 1 |
| T5 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 1 |
| T6 | 0 | 36 | 66 | 0 | 0 | 156 | 0 | 66 | 0 | 0.4814 |
| T7 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 1 |
| T8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 1 |
| T9 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 288 | 0.8888 |
|  | 0.8846 | 0.8181 | 0.8135 | 1 | 1 | 1 | 1 | 0.8307 | 0.8571 |  |

从图中可以看出，数据分布主要集中在对角线上说明算法预测效果越好，分布在非对角线的数值越小也说明算法预测效果越好。从下面五张表中，我们可以明显发现图一的数据比其他数据更加集中在对角线，并且非对角线的数据值也很小。所以可以判定基于线性核函数优化的支持向量机算法预测的效果更好。

表4.18线性核函数分类学习在不同参数情况下的混淆矩阵

Table 4.18 Linear Kernel Function Classification Learning under the condition of different parameters of Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |
| 1 | 252 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 72 | 0.7777 |
| 2 | 0 | 324 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 72 | 252 | 0 | 0 | 0 | 0 | 0 | 0 | 0.7777 |
| 4 | 0 | 0 | 36 | 288 | 0 | 0 | 0 | 0 | 0 | 0.8888 |
| 5 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 36 | 72 | 0 | 0 | 144 | 0 | 66 | 0 | 0.4444 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 1 |
| 9 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 288 | 0.8888 |
|  | 0.875 | 0.75 | 0.7 | 1 | 1 | 1 | 1 | 0.8307 | 0.8 |  |

表4.19多项式核函数分类学习在不同参数情况下的混淆矩阵

Table 4.19 Polynomial Kernel Function Classification Learning under the condition of different parameters of Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |
| 1 | 223 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 101 | 0.6882 |
| 2 | 0 | 315 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0.9722 |
| 3 | 0 | 27 | 297 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9166 |
| 4 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 83 | 0 | 0 | 0 | 178 | 0 | 66 | 0 | 0.5493 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 315 | 0 | 0.9722 |
| 9 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 281 | 0.8672 |
|  | 0.8383 | 0.7411 | 1 | 1 | 1 | 0.908 | 1 | 0.8267 | 0.7356 |  |

表4.20径向基核函数分类学习在不同参数情况下的混淆矩阵

Table 4.20 RBF Kernel Function Classification Learning under the condition of different parameters of Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |
| 1 | 250 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 74 | 0.7716 |
| 2 | 0 | 323 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9969 |
| 3 | 0 | 69 | 255 | 0 | 0 | 0 | 0 | 0 | 0 | 0.7870 |
| 4 | 0 | 0 | 33 | 291 | 0 | 0 | 0 | 0 | 0 | 0.8981 |
| 5 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 33 | 66 | 0 | 0 | 155 | 0 | 66 | 0 | 0.4783 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 324 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 322 | 0 | 0.9938 |
| 9 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 287 | 0.8858 |
|  | 0.8710 | 0.76 | 0.7183 | 1 | 1 | 0.9872 | 1 | 0.8298 | 0.7950 |  |

表4.18 Sigmoid核函数分类学习在不同参数情况下的混淆矩阵

Table 4.18 Sigmoid Kernel Function Classification Learning under the condition of different parameters of Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |
| 1 | 153 | 0 | 0 | 0 | 54 | 0 | 18 | 0 | 99 | 0.4722 |
| 2 | 9 | 216 | 0 | 0 | 54 | 0 | 0 | 0 | 45 | 0.6666 |
| 3 | 9 | 60 | 156 | 0 | 54 | 0 | 0 | 0 | 45 | 0.4814 |
| 4 | 9 | 0 | 24 | 192 | 54 | 0 | 0 | 0 | 45 | 0.5925 |
| 5 | 9 | 0 | 0 | 0 | 216 | 0 | 54 | 0 | 45 | 0.6666 |
| 6 | 9 | 18 | 48 | 0 | 54 | 102 | 0 | 48 | 45 | 0.3148 |
| 7 | 33 | 0 | 0 | 0 | 54 | 0 | 180 | 0 | 57 | 0.5555 |
| 8 | 15 | 0 | 0 | 0 | 54 | 0 | 12 | 198 | 45 | 0.6111 |
| 9 | 33 | 0 | 0 | 0 | 54 | 0 | 6 | 0 | 231 | 0.7129 |
|  | 0.5483 | 0.7346 | 0.6842 | 1 | 0.3333 | 1 | 0.6666 | 0.8048 | 0.3515 |  |

除了观察混淆矩阵，我们还需要考察五个不同核函数的支持向量机算法的查准率和查全率的分布，并将其绘制成P-R图，描绘P-R曲线。如图，我们可以发现曲线1明显可以包住其他曲线。这说明我们基于线性核函数优化的支持向量机预测效果更好。通过平衡点（Break Even Point）的取值可以发现基于线性优化的核函数的曲线的平衡点最大。所以基于BEP的比较，可认为基于线性优化的核函数的支持向量机的预测效果最好，其次是线性核函数、多项式核函数、径向基核函数与sigmoid核函数。

图4.5 5个核函数分类学习的P-R曲线

Fig 4.5 5 kernel function classification learning curves of P-R

### 4.4.3 F1度量

F1度量的参数，我们选择标准的F1度量，从下图中，我们可以看出从F1度量的角度来看仍然是基于线性优化的核函数的支持向量机预测的效果更好。其次依次为多项式核函数、径向基核函数、线性核函数和sigmoid核函数。

表4.22 5个核函数分类学习的F1度量

Table 4.22 5 kernel function classification learning F1 measure

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kernel | 线性 | 多项式 | 径向基 | Sigmoid | 线性优化 |
| F1 | 0.873974 | 0.889759 | 0.876215 | 0.616636 | 0.90638 |

综合以上三个小结的内容，通过对分类学习算法的评估方法：准确度、混淆矩阵、P-R曲线、F1度量四个方面，都可以直观的发现基于线性优化的核函数的支持向量机算法对于本例的数据预测的效果更准确、更稳定。

# 第5章 总结与展望

## 5.1 本文工作总结

支持向量机有着一套坚实的理论基础，作为一种新的基于统计学习理论的机器学习方法在理论方面有很突出的优势。但遗憾的是它的应用研究现在还比较滞后，目前大部分研究报道还都是仿真和对比实验，因此可以判断支持向量机的应用研究是一个大有作为的方向，是一个值得大力研究的领域，对它的研究将对机器学习等学科领域产生重要影响。

论文围绕支持向量机在柴油机故障诊断应用中的问题进行研究，现将本文研究的主要内容与成果总结如下：

1. 通过对国内外柴油机故障诊断技术现状和研究方向的综述，指出使用支持向量机做柴油机故障诊断的优势，并举例了一个用SVM进行柴油机燃油系统故障诊断的实例。

2．以柴油机中产生故障最频繁的柴油机燃油系统为研究对象，使用支持向量机的多类分类器对故障样本进行了分类试验，并对常规的SVM中的惩罚参数C和核函数参数gamma如何选择做了编程优化。

## 5.2 进一步研究的展望

在论文课题研究中，随着课题的开展和深入，作者深切感觉到还有很多问题需要深入研究。但由于本人时间、精力、能力有限，本文只是做了使用支持向量机进行柴油机故障诊断的前期研究，仍然有很多需要进一步研究和探讨的问题。需要在以下几个方面进行深入研究：

1．本文研究的重点是故障发生后故障类型的诊断，但在实际中如何避免故障发生和轻微故障发生后如何避免故障变大是国际社会对故障诊断这一课题研究的重点方向，预测成为故障诊断领域研究的新方向，因硬件条件和时间等因素限制，本文未涉及故障预测部分，这将是后期继续研究工作的重点

2．柴油机故障数据库的建立。支持向量机可以通过一定数量的柴油机故障实际数据整理出反故障与特征量之间的关系来进行真正意义上的故障诊断，因此柴油机故障数据库的建立十分有意义，这将极大的推动柴油机故障诊断技术在实际中的运用。

3．目前支持向量机主要是采用离线训练在线识别的方法进行学习，如何使它能够在线学习和辨识是下一步研究的问题。国内外支持向量机的应用研究现在主要还处于仿真及实验室阶段，应不断拓宽其应用领域，加深其应用深度，在柴油机故障诊断方面深入研究可以诊断出多种柴油机系统故障的基于SVM的柴油机系统故障诊断软件，推动SVM在柴油机故障诊断领域的发展。

# 参考文献

[1] 王福根. 船舶柴油机及安装[M]. 哈尔滨工程大学出版社, 2011.

[2] 张百慈. 船舶柴油机故障诊断现状及发展趋势[J]. 中国水运:理论版, 2006, 4(9).

[3] 徐立华. 柴油机故障诊断技术的现状及发展趋势[J]. 铁道机车与动车, 2011(5):35-37.

[4] 朱志宇, 张冰, 刘维亭. 模糊支持向量机在船舶柴油机故障诊断中的应用[J]. 中国造船, 2006, 47(3):64-69.

[5] 郭晶亮. 基于支持向量机的柴油机故障诊断系统研究[D]. 武汉理工大学, 2011.

[6] 郭江华, 侯馨光, 陈国钧,等. 船舶柴油机故障诊断技术研究[J]. 中国航海, 2005(4):75-78.

[7] 周龙保. 内燃机学.第3版[M]. 机械工业出版社, 2011.

[8] 曹龙汉. 柴油机智能化故障诊断技术[M]. 国防工业出版社, 2005.

[9] 刘鑫. 柴油发动机运行状态监测和故障诊断系统的设计与实现[D]. 电子科技大学, 2006.

[10] 张连方, 刘炽棠, 顾宏中. 柴油机原理[M]. 上海交通大学出版社, 1987.

[11] 李彦凤. 基于模糊模式识别的柴油机故障诊断系统[D]. 山东大学, 2004.

[12] Song Q, Grigoriadis K M. Diesel engine speed regulation using linear parameter varying control[C] American Control Conference, 2003. Proceedings of the. IEEE, 2003:779-784 vol.1.

[13] Donoho D L. De-noising by soft-thresholding[J]. IEEE Transactions on Information Theory, 1995, 41(3):613-627.

[14] 赵芳. 基于模糊模式识别的船用柴油机状态监测和故障诊断[D]. 山东大学, 2005.

[15] 赵文仓. 基于模式识别的柴油机故障诊断系统的研究[D]. 山东大学, 2002.

[16] 李玉峰. 基于神经网络的柴油机燃油系统故障诊断的研究和实现[D]. 山东大学, 2007.

[17] 曹龙汉. 柴油机智能化故障诊断技术[M]. 国防工业出版社, 2005.

[18] 虞和济, 陈长征. 基于神经网络的智能诊断[J]. 振动工程学报, 2000, 13(2):202-209.

[19] 李凤鸣. 基于小波神经网络的柴油机燃油系统故障诊断的设计与实现[D]. 山东大学, 2010.

[20] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[21] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2012, 12(10):2825-2830.

[22] Rossum G V, Drake F L. Python 3 Reference Manual[J]. Department of Computer Science [CS], 1995, 111(254):1–52.

[23] Rossum G V, Drake F L. Python 2.6 Reference Manual[M]. 1995.

[24] Bowles M. Machine Learning in Python[J]. 2015.

[25] Kramer O. Scikit-Learn[M] Machine Learning for Evolution Strategies. 2016.

[26] Garreta R, Moncecchi G. Learning scikit-learn[J]. 2013.

[27] Varoquaux G, Buitinck L, Louppe G, et al. Scikit-learn: Machine Learning Without Learning the Machinery[J]. Getmobile Mobile Computing & Communications, 2015, 19(1):29-33.

[28] Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python[M]. Packt Publishing, 2013.

[29] Hauck T. Scikit-learn Cookbook[J]. 2014.

[30] Cawley G C. Leave-One-Out Cross-Validation Based Model Selection Criteria for Weighted LS-SVMs[J]. 2014:1661-1668.

[31] Liang T, Davier A A V. Cross-Validation[J]. Applied Psychological Measurement, 2014, 38:281-295.

[32] Kernel Function[M] Encyclopedia of Microfluidics and Nanofluidics. Springer New York, 2015:1499-1499.

[33] Lottati I, Nissim E. Nonplanar, supersonic, three-dimensional, oscillatory, piecewise continuous-kernel function method[J]. Journal of Aircraft, 2015, 24(1):45-54.

[34] Kulkarni S, Harman G. VC Dimension[M] An Elementary Introduction to Statistical Learning Theory. John Wiley & Sons, Inc. 2011:125-136.

[35] Vapnik V, Levin E, Cun Y L. Measuring the VC-dimension of a learning machine[J]. Neural Computation, 1994, 6(5):851-876.

[36] Guyon I. Structural Risk Minimization for Character Recognition.[C] Advances in Neural Information Processing Systems. DBLP, 1991:471-479.

[37] Zhang X. Structural Risk Minimization[J]. 2016.

[38] 张浩然, 韩正之, 李昌刚. 支持向量机[J]. 计算机科学, 2002, 29(12):135-137.

[39] NelloCristianini, JohnShawe-Taylor. 支持向量机导论[M]. 电子工业出版社, 2004.

[40] 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):1397-1400.

[41] 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.

[42] 张学工. 关于统计学习理论与支持向量机[J]. Acta Automatica Sinica, 2000, 26(1):32-42.

[43] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程, 2004, 30(10):6-9.

[44] 刘志刚, 李德仁, 秦前清,等. 支持向量机在多类分类问题中的推广[J]. 计算机工程与应用, 2004, 40(7):10-13.

[45] 高学, 金连文, 尹俊勋,等. 一种基于支持向量机的手写汉字识别方法[J]. 电子学报, 2002, 30(5):651-654.

[46] 董春曦, 饶鲜, 杨绍全,等. 支持向量机参数选择方法研究[J]. 系统工程与电子技术, 2004, 26(8):1117-1120.

[47] 郑勇涛, 刘玉树. 支持向量机解决多分类问题研究[J]. 计算机工程与应用, 2005, 41(23):190-192.

[48] 李盼池, 许少华. 支持向量机在模式识别中的核函数特性分析[J]. 计算机工程与设计, 2005, 26(2):302-304.

[49] 刘明. 支持向量机中Sigmoid核函数的研究[D]. 西安电子科技大学, 2009.

[50] 奉国和. SVM分类核函数及参数选择比较[J]. 计算机工程与应用, 2011, 47(3):123-124.

[51] 高涛. SVM在船舶柴油机故障检测系统中的应用研究[J]. 机电设备, 2007, 24(7):26-30.

[52] 刘鑫. 柴油发动机运行状态监测和故障诊断系统的设计与实现[D]. 电子科技大学, 2006.

[53] 李晓伟. 船舶柴油机故障诊断方法的研究[J]. 计算机仿真, 2012, 29(5):215-218.

[54] 张百慈. 船舶柴油机故障诊断现状及发展趋势[J]. 中国水运:理论版, 2006, 4(9).

[55] 魏巍, 詹玉龙, 赵倍聪,等. 基于支持向量机的船舶柴油机层次故障诊断的研究[J]. 南通航运职业技术学院学报, 2009, 8(1):53-57.

[56] 朱志宇, 刘维亭. 基于支持向量机的船舶柴油机故障诊断[J]. 船舶工程, 2006, 28(5):31-33.

[57] 詹玉龙, 翟海龙, 曾广芳. 基于支持向量机的船舶柴油机故障诊断的研究[J]. 中国航海, 2007(2):89-92.

[58] 朱志宇, 张冰, 刘维亭. 模糊支持向量机在船舶柴油机故障诊断中的应用[J]. 中国造船, 2006, 47(3):64-69.

[59] 林新通, 詹玉龙, 周薛毅,等. 支持向量机在船舶柴油机废气涡轮增压器故障诊断中的应用[J]. 上海海事大学学报, 2012, 33(2):18-21.

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

# 致谢

在本论文即将完成之际，谨此向我的导师XXX老师致以衷心的感谢和崇高的敬意!本论文的工作是在XXX老师的悉心指导下完成的。他以敏锐的洞察力、渊博的知识、严谨的治学态度、精益求精的工作作风和对科学的献身精神给我留下了刻骨铭心的印象，这些使我受益匪浅，并将成为我终身献身科学和献身事业的动力。

在攻读硕士的这X年里，导师不仅为我创造了优越的科研和学习环境，使我得以在船舶领域中自由翱翔，同时在思想上、人生态度和意志品质方面给予了谆谆教诲，这些教益必将激励着我在今后的人生道路上奋勇向前。

真诚感谢师兄师姐师妹以及同届同学给予的帮助，他们不仅在学术上给我指引，而且在生活上予以帮助，从他们身上我学到很多知识。

衷心的感谢我的父母和其他亲朋好友对我的关心、支持和理解，没有他们对我的关心、鼓励和支持，我无法完成现在的硕士学业。

最后，感谢曾经教育和帮助过我的所有老师。衷心地感谢为评阅本论文而付出宝贵时间和辛勤劳动的专家和教授们!

忘不了学院所有的老师，忘不了班里的同学，谢谢你们在这人生最后的校园生活里的陪伴。接着，我们都将踏上人生崭新的旅程，愿我们都越来越好!