上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR’S THESIS

论文题目：射流火焰动态纹理分析

学生姓名：孟恒宇

学生学号：5134139017

专　　业：航空航天工程

指导教师：李元祥

学院(系)：航空航天学院

射流火焰动态纹理分析

摘要

射流火焰模型作为超燃冲压发动机的关键技术，具有很高的研究价值。但是射流火焰作为最复杂的湍流之一，高度非线性的特点导致传统建模手段难以描述，现有的射流火焰模型窘于分析手段，难以解决这一问题。本文基于射流火焰历史数据库进行了数据挖掘，利用 PCA 技术分析火焰数据的线性相关关系，利用 SVM、BP 神经网络和随机森林等机器学习方法进行火焰状态识别和特征选择。具体工作如下：

1.基于现有射流火焰模型，利用 PCA 分析了射流火焰各个成分的线性关系，得出射流火焰模型的主导成分，与计算流体力学领域的理论推导和实验共识吻合。并进行了特征标准化使得数据的每一个维度具有零均值和单位方差便于下一步机器学习。

2.火焰状态识别（不同喷射速度的一氧化碳火焰和不同射流速度的甲烷火焰，共四类），采用 SVM 对四类火焰状态进行识别；采用完全训练集训练的 dropout 神经网络和缺省测试集验证分类结果对各个特征的敏感性；采用随机森林同时完成特征重要性排序和状态识别并与 dropout 神经网络得到的敏感特征进行互相验证。

3.训练集和测试集均只保留 dropout 神经网络得到的敏感特征或随机森林得到的重要特征，再用 BP 神经网络进行分类，并将分类结果与完全数据比较。实验表明其分类正确率下降不明显，从而说明提取特征的重要性，有望大幅减轻射流火焰实验负担。

关键词：射流火焰，机器学习，特征选择，模式识别

DYNAMIC TEXTURE ANALYSIS FOR JET FLAME

ABSTRACT

As the key technology of scramjet, the Jet Flame model has great research value. However, as one of the most complex turbulence, the characteristic of high nonlinearity leads to the difficult description of traditional modeling method, and the existing jet flame model is embarrassed by the analysis method, it is difficult to solve this problem. In this paper, data mining is based on historical database, using PCA technique to analyze the linear correlation of flame data, the Flame State identification and feature selection were carried out by using SVM, BP neural network and stochastic forest. The specific work is as follows:

1. Based on the existing Jet flame model, the linear relationship between the components of the Jet flame is analyzed by PCA, and the dominant component of the Jet flame model is obtained, which is consistent with the theoretical deduction and experimental consensus in the field of computational fluid dynamics. The standardization of the feature makes the data every a dimensions have 0 mean and unit variance, which facilitates the next machine learning.

2. Flame state identification (carbon monoxide flame with different jet velocities and methane flame with different jet velocities, total four), using SVM to identify four kinds of flame states, using Dropout neural network and default test set trained by complete training set to verify the sensitivity of classification results to each characteristic, the random forest simultaneously completes the feature importance sorting and state recognition and verifies with each other the sensitive characteristics obtained by Dropout Neural network.

3. Both the training set and test set only preserving the sensitive features obtained by Dropout neural networks or the important characteristics of random forests, then classifies by BP neutral network and compares the classification results with the complete data. The experimental results show that the accuracy of classification doesn't decline obviously, which proves the importance of extracting feature, and it is expected to reduce the burden of Jet flame experiment greatly.

Key words: Jet Flame, data mining, feature extraction, pattern recognition

目 录

第一章 绪论

1.1 研究背景

随着信息技术的高速发展，人类社会进入信息化时代。各类传感器、输入输出设备和存储设备的普及，为迅猛增长的数据提供了产生和储存的条件，推动了大数据时代的产生和发展[1]。日常生活中我们时刻都在产生数据，仅仅通过互联网，每天产生的数据量已经十分庞大。因此在大数据的背景下，数据具有规模大、增长变化快、表达方式多、价值高等特点[2]。数据量的迅速增长带来了巨大的价值，但是同时也带来了很多问题[3][4]。为了解决这些问题，许多行之有效的算法被提出来，数据挖掘领域进步神速。与此同时，在一些学科领域，虽然也产生了相当多的数据，但是限于专业门槛，不能像生活信息一样结合数据挖掘技术，从海量的数据中得到新的知识以帮助理论的发展。本文试图为其中一个问题的解决提供新的思路。

1.1.1 超燃冲压发动机

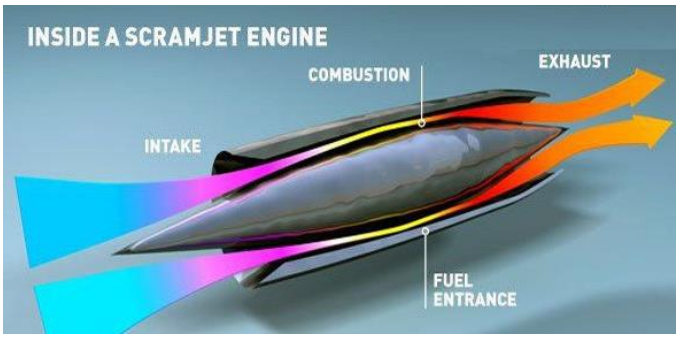


图 1 超燃冲压发动机原理示意图[5]

超燃冲压发动机一般是指运行马赫数大于 6，以超声速燃烧为核心技术的冲压发动机。一般应用于高超声速导弹、高超声速飞行器、空天飞机等未来的以火箭基组合循环为动力的单极低轨道运输系统。目前，高超声速导弹已经部分运用了超燃冲压发动机技术。

过去几十年，美俄两国均在超燃冲压发动机上投入巨资。俄罗斯（前苏联）中央航空发动机研究所进行了不间断的研究，在这一领域处于领先地位。1991 年俄罗斯成功进行了首次飞行试验，验证了超燃冲压发动机可行性。1998 年，首台实用的超燃冲压发动机投入使用，最大马赫数 6.5,在 Ma=3.5~6，5 区间内实现了燃烧模态转换。俄罗斯目前还在进行相应研究。

美国自 1964 年开始进行超燃冲压发动机研究，空军、海军、陆军、NASA 和 DARPA 合作研究。2004 年 3 月 27 日 X-43A 的在 3 万米高空，Ma=7 的情况下，启动超燃冲压发动机，工作时间 10s，飞行实验非常成功。该飞行器采用了升力体外形，接近实用飞行器水平，标志着美国在这一领域取得了世界领先水平。

我国在这一领域也开始了一些研究，部分单位实现了煤油超声速点火，有的实现了超燃点火，同时还开展了发动机冷却与热结构等研究。总而言之，国内在这一领域的研究还处于起步阶段。

1.1.2 射流火焰

高效的超声速燃烧室设计的关键在于有效控制燃烧模式及传播过程，包括火焰的点火，加减速，稳定性和熄灭的过程。而这一控制过程基于对同轴射流火焰的机理的理解。

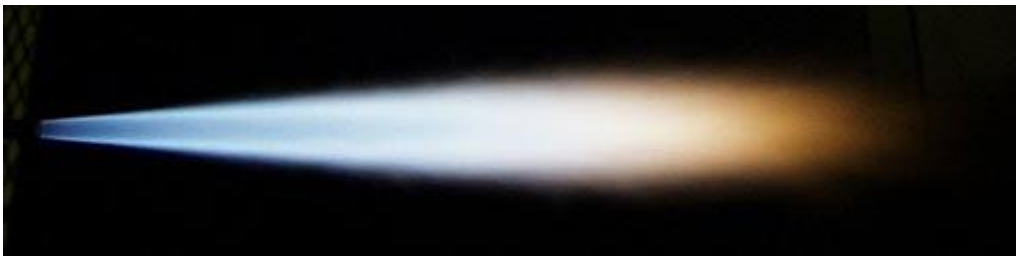


图 2 典型射流火焰

在近期的几项射流火焰研究[6][7][8]中，燃烧模式的描述模型并不完善，对于火焰的加减速，形状演化和稳定性的认识尚有很多空白。研究者需要火焰在所有发展阶段的量化分析，使得很有必要进行多工况多参数条件下的实验。

而与此同时，随着各类测量技术的发展，我们可以使用探针、激光、摄像机等来捕捉火焰燃烧现象的更多细节。比如本文基于的 Sandia 数据，建立于上世纪 90 年代，限于时代限制，只能使用探针对多个射流火焰流场的几十个测量点的当量比，温度，压力和十一种组分质量分数和摩尔分数作了数百数千次测量，从而形成流场在一段时间内各个测量点的详细参数分布。

但是如此大量的数据也带来一系列问题。首先，研究者不能从这一数据中获得直观的流体力学定性结论，也不能轻易由此得出更加完善的射流火焰模型。其次，在燃烧室设计中，需要一种简便的火焰状态判断方法来实现火焰状态的闭环控制，而传统的射流火焰实验庞大的数据量意味着很难使用传统条件逻辑判断方法。

1.2 研究意义

如此大量的数据更加详细地描述了事物，从而提供了更加丰富的信息，但是这同时必然会带来数据冗余的问题，并不一定需要如此高维的数据。在传统的火焰测量试验中，为了最大化利用每次昂贵实验的价值，测量结果都务求尽可能完善。这使得我们获取的大多数数据往往都是高度冗余的，冗余的数据使得其本质结构很难被发掘，分析特征、训练模型所需时间大大延长，同时高维样本容易引起“维度灾难”，模型也会更复杂，其推广能力下降。

因此，从冗余的数据中选取出数据的一个有用的特征子集、得到这些数据的本质结构，一方面能够减少特征个数，提高模型精确度，减少运行时间。另一方面，选取出真正相关的特征简化了模型，使研究人员易于理解数据产生的过程，有利于射流火焰领域的研究。这一过程被称之为特征选择（feature selection）,也被称为特征子集选择（feature subset selection. FSS）或属性选择（attribute selection）。

如果能通过特征选择，挑选出在火焰状态识别中最关键的一个或几个参数，有利于对射流火焰的深层次理解。传统实验过程中非常繁杂的每个点每个参数一一测量导致的庞大的流场参数分布表，将有望简化成一个或几个简单数据。目前射流火焰研究实验的一大阻碍就是高昂的实验成本。高效的特征选择不但可以大大减少测量次数，降低实验成本，还有利于对射流火焰模型的理解。

1.3 国内外研究现状

国内外对射流火焰的模式识别研究不多。J.E.Freech 使用火焰数字图像推导辛烷/空气射流火焰传播速度[9]。Ahsan.r.choudhuri 使用红外图像处理来得到不同组分的火焰参数[10]。s. David a. Rosenberg and james f. Driscoll 使用激光图像来测量部分预混火焰的参数并讨论了信噪比[11]。Jacob e. Temme 研究了涡轮发动机燃烧室燃烧图像的低频不稳定燃[12]。国内方面，孙玉洁使用火焰自由基图像来研究燃烧污染物，并且据此选择合适模型来预测火焰的传播

[13]。

然而这些都是基于传统的数字图像处理方法来研究火焰形态，没有在大数据基础上对火焰进行数据挖掘，进而得到超过传统理论架构的新知识。尤其在特征领域，采用了传统图像处理领域的一系列图像特征，对火焰模型认知帮助有限。

在相近领域，国内对于锅炉火焰的基于机器学习的状态识别已经有一定成果。尤其是南京航空航天大学吴一全教授实验室，先后提出了多阈值特征，交叉熵特征，KPCA 特征和 Krawtchouk 矩特特征等识别方案[14]。然而这些特征的选择多半是依赖人工经验，并没有从数据中学习得到特征，选择的特征在实验中表现也不尽如人意，不适用于本文研究的射流火焰。并且，现有研究采用的分类器多为简单的 SVM 分类器，对射流火焰的数据结构研究并不深入。

与此同时，机器学习的研究进步神速，大量算法和架构推陈出新，就分类器而言，传统的 SVM 分类器已经应用于诸多场景，深度学习（deep learning）包括深度神经网络（deep neutral network,）深度森林（deep forest）也在 ImageNet 等比赛中展现出潜力。机器学习不仅仅是分类器，更是研究数据内部本质结构如数据的稀疏性、凹凸性的有效手段[15]。对于射流火焰而言，机器学习可能成为新的模型描述手段。

总之，国内外研究现状中，尚无有效的射流火焰特征分析与选择手段，机器学习有望在射流火焰模型描述上取得突破。

1.4 论文的主要内容和章节安排

本文的主要工作是基于射流火焰经典数据 Sandia 数据开展研究。通过在数据上进行 PCA 分析，验证了射流火焰实验领域的普遍共识。本文在 Sandia 数据上使用了 KNN，SVM,BP 神经网络和随机森林这四种分类器成功进行了状态识别，并在识别过程中提出了特征选择问题，筛选出部分在状态识别中较为重要的特征。对仅有筛选特征的数据进行状态分类，实验证明筛选特征能较好的表达火焰状态。

根据上述内容，本文章节安排如下：

第一章为绪论，表述了本文的研究背景和意义，其次提出了本文核心——火焰数据特征选择，并对特征选择以降低试验费用的意义作了阐述。随后介绍了国内外研究现状，最后对本文主要内容和章节安排进行阐述。

第二章为数据预处理部分。首先从数据中采出机器学习必须的样本和真值，并将每个火焰流场对应测量点的每个参数测量值作为样本的维度，形成了机器学习概念上的数据集。其次对数据集进行 PCA 白化，分析了数据集的线性相关性，为下一步的机器学习作基础。

第三章为 KNN 和 SVM 分类。这一章使用了经典机器学习分类算法 KNN 和最常用的分类器支持向量机（Support Vector Machine, SVM）,探究了机器学习超参数对分类正确率的影响。

第四章为 BP 神经网络分类。这一章使用完备训练集训练设置 dropout 的神经网络，使用缺省测试集进行测试，以分类正确率下降多少为标准测试样本各个维度的重要性，并将重要维度单独组成数据集进行状态识别。

第五章为随机森林分类。利用随机森林分类与变量重要性同时完成的特点进行火焰状态分类，并筛选出重要变量。将重要变量单独抽出组成新的数据集进行分类测试得到识别正确率。

第六章为结束语，首先简单介绍了本文的主要工作以及创新点，然后分析研究过程的不完善之处，最后根据不完善之处得到下一步工作的研究方向。

第二章 数据预处理

本文采用公开数据作为试验对象。在公开数据中，Sandia 工作站是最为权威和详细的数据来源，故本章介绍 Sandia 工作站的基本情况和采用的 TNF 数据的详细信息，并将其制作为数据集。

2.1 数据来源

2.1.1 Sandia 工作站介绍

Sandia（http://www.ca.sandia.gov/TNF/abstract.html）是上世纪 90 年代由意大利那不勒斯大学建立的开放性的湍流燃烧国际联合工作站，之后又加入了荷兰代尔夫特大学、达姆施塔特大学、德国斯图加特大学、海德堡大学、札幌大学和芝加哥大学等研究机构。Sandia 专注于研究燃料气体湍流化学反应，整合了上述多所研究机构的实验数据，推出了在线数据库来帮助各国研究者验证模型和探究湍流火焰燃烧，提出了一系列框架来作测量结果与理论结果的比较并引导了未来实验和仿真的研究方向。

该工作站目前依然在更新，注意力转移到火焰的碳氢化合物反应（甲烷，天然气和甲醇），包括这一过程的建模挑战、部分熄火、重新点火、分离或提升反应区、自动点火、流动回流区和漩涡。未来计划增加更复杂的燃料的燃烧模型，希望建立通用燃烧模型，包括非预混、部分预混、分层和预混火焰。未来 Sandia 研究火焰目标的选择将基于组织者和积极贡献着的科学共同体的需求。

2.1.2 TNF 数据介绍

湍流非预混火焰（Turbulent Non-premixed flames, TNF）数据是 Sandia 的核心，包含了多个进行化学过程的火焰湍流流场的尺度和速度数据，因此被称为是最合理完整、准确、适宜的湍流火焰的实验数据集,供研究者们验证各自的燃烧模型[16]。数据集包括几何形状相对简单、燃料组分相对单一、边界条件相对明确的射流火焰的详细的标量测量 (温度，主要成分，和一些次要成分) 和速度测量。

以苏黎世理工学院提供的 CO/H2/N2 Jet Flames 数据为例，共计包含两组实验数据 A 和 B，实验条件见下表：

表 1 CO/H2/N2 Jet Flames 主要参数

燃料组成		40%CO, 30%H2, 30%N2
计量混合分数		0.295
计量火焰长度		47 倍直径
雷诺数		16700
协流流速		0.75m/s
喉道内径	火焰 A	4.58mm
	火焰 B	7.72mm
扩张管外径	火焰 A	6.34mm
	火焰 B	9.46mm
喷射流速	火焰 A	76m/s
	火焰 B	45m/s

在每个火焰流场中取了 X/D=20,30,40,50,60 五个截面，每个截面取了 R=-1.5,1,3.5,6……mm 等若干个半径（各个截面半径不同）上的测量点，每个测量点测量当地当量比、温度、氧气质量分数、氮气质量分数、氢气质量分数、水质量分数、一氧化碳质量分数、二氧化碳质量分数、氢氧根离子质量分数、一氧化氮质量分数、一氧化氮离子质量分数、TDNR 这十二个参数，每个测量点测量 800-1000 次。

达姆施塔克大学和斯图加特大学提供的 CH4/H2/N2 Jet Flames,也包含两组实验数据 A 和 B，实验条件见下表：

表 2 CH4/H2/N2 Jet Flames 主要参数

喉道内径		8.0mm
燃料组分		22.1%CH4,33.2%H2,44.7%N2
协流速度		0.3m/s
协流温度		292K
协流 H2O 摩尔质量分		0.8%
射流速度	火焰 A	42.2m/s
	火焰 B	63.2m/s
雷诺数	火焰 A	15200
	火焰 B	22800

在每个火焰流场中取了 X/D=5,10,20,40,60,80 六个截面，每个截面取了半径等于 5,0,5,10……mm 等若干个半径（各个截面半径不同）上的测量点，每个测量点测量当地当量比、温度、氧气质量分数、氮气质量分数、氢气质量分数、水质量分数、甲烷质量分数、一氧化碳质量分数、二氧化碳质量分数、氢氧根离子质量分数、一氧化氮质量分数、一氧化碳离子质量分数、TDNR 这十三个参数，每个测量点测量 800-1000 次。

2.2 数据集制作过程

由上文可知苏黎世理工学院提供的 CO/H2/N2 Jet Flame 数据和达姆施塔克大学和斯图加特大学提供的 CH4/H2/N2 Jet Flames 数据，尽管所研究火焰类似，但是选取的测量点并不一一对应，每个测量点的测量参数也并不完全相同。为了保证每个样本的维度一致，应当选取不同火焰对应位置的测量点数据，每个测量点的测量量也应当取两类火焰测量量的公共子集。这样每个样本的变量来源如下表：

表 3 测量点分布表

测量点序号	来源截面	半径
1	20	-1.5
2	20	6
3	20	13.5
4	20	21
5	40	0
6	40	5
7	40	10
8	40	15
9	40	20

接表 3

10	40	25
11	40	30
12	60	0
13	60	10
14	60	20
15	60	30
16	60	40

每个测量点测量了 11 个量：当地当量比、温度、氧气质量分数、氮气质量分数、氢气质量分数、水质量分数、一氧化碳质量分数、二氧化碳质量分数、氢氧根离子质量分数、一氧化氮质量分数、TDNR。每个样本一共 176 个变量，或者说 176 个维度。

原有的苏黎世理工学院提供的 CO/H2/N2 Jet Flame 数据和达姆施塔克大学和斯图加特大学提供的 CH4/H2/N2 Jet Flames 数据中每个测量点测量次数并不均衡，少者只有 500 次，多则上千次。上文挑选的测量点，来源于小截面小半径的测量点测量次数均只有 550 次左右，而其他测量点测量次数多在 750 次至 900 次。为了充分利用数据同时平衡各类样本数目，决定提取出 750 个样本。测量次数不足 750 的测量点，采用插值法补全。

综上所述，从 Sandia 数据中提取出的机器学习数据集如下：真值一共 4 类，分别为不同喷射流速的一氧化碳火焰 A 和 B 和不同射流速度的甲烷火焰 A 和 B。每一类各有 750 个样本，合计 3000 个样本。每个样本包含 176 个维度，维度来源于 16 个测量点每个测量点 11 个测量量。机器学习的目标就是利用这一数据集制作火焰 4 类状态分类器并从每个样本 176 个维度中选择出最重要的维度。

随后进行特征标准化操作。特征标准化指的是独立地标准化数据的每一个维度使之具有零均值和单位方差。这是归一化中最常见的方法（例如，在使用支持向量机（SVM）时，特征标准化一般会作为第一步）。在实际应用中，特征标准化的具体做法是：在完整数据集上计算每一个维度上数据的均值，之后在每一个维度上都减去该均值。下一步便是在数据的每一维度上除以该维度上数据的标准差。特征标准化之后的数据将作为接下研究的基础。

2.3 PCA

主成分分析（Principal Components Analysis, PCA）是一种非常常用的数据降维算法[17]，属于无监督学习。其基本思想是构造一个超平面对正交属性空间内所有样本进行表达，该超平面应当具有两个性质：一，最近重构性：样本点到这个超平面的距离都足够近；二，最大可分性：样本点到这个超平面的投影尽可能分开（保留尽可能大的方差）。

算法描述如下：

表 4 PCA 算法流程

输入：样本集D $= \{X, X_2, X_3, \ldots, X_m\}$;
目标超平面维数 d;
过程：

1：对所有样本进行中心化：$x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$;

2.计算所有样本的协方差矩阵XX^T;
3.对协方差矩阵作特征值分解（2，3 两步也可代之以对 X 作奇异值分解）;
4.取最大的 d 个特征值所对应的特征向量$w_1, w_2, \ldots, w_d$;
输出：投影矩阵W $= (w_1, w_2, \ldots, w_d)$

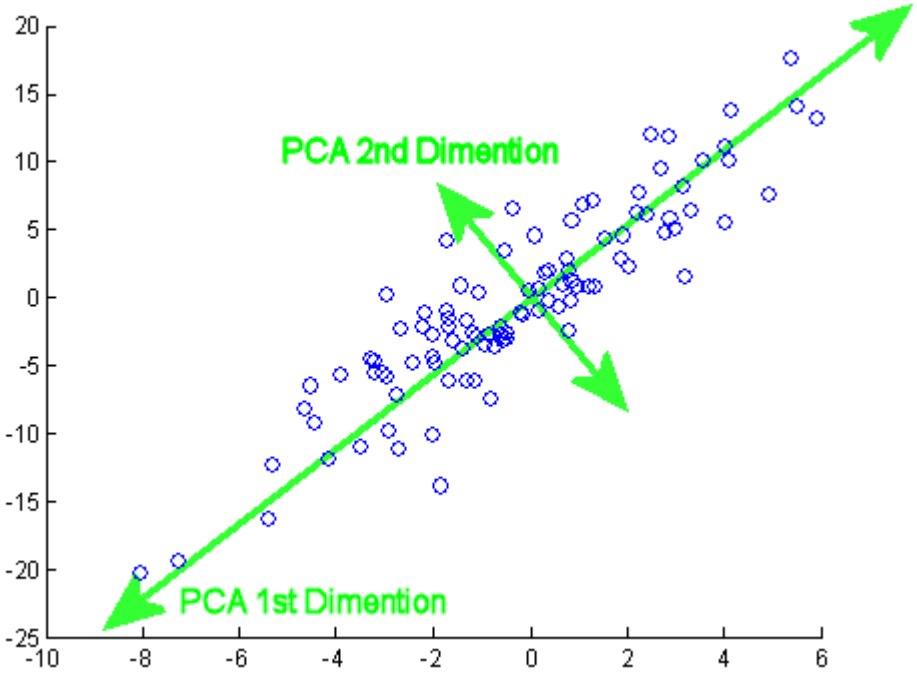


图 3 PCA 原理示意图

如上图所示，PCA 也可理解为旋转坐标轴。通过旋转各个特征维度坐标轴，使得样本的方差集中投影到前几个旋转平面上，之后的平面上样本投影方差极小可以忽略不计，如此达到了降维的效果。这样做的好处是突出了主导维度，同时当数据受到噪声影响时，小的特征值对应的特征向量往往含有较大的噪声分量，舍弃它们一定程度上可以去噪。

对达姆施塔克大学和斯图加特大学提供的 CH4/H2/N2 Jet Flames 数据作 PCA 操作结果如下：

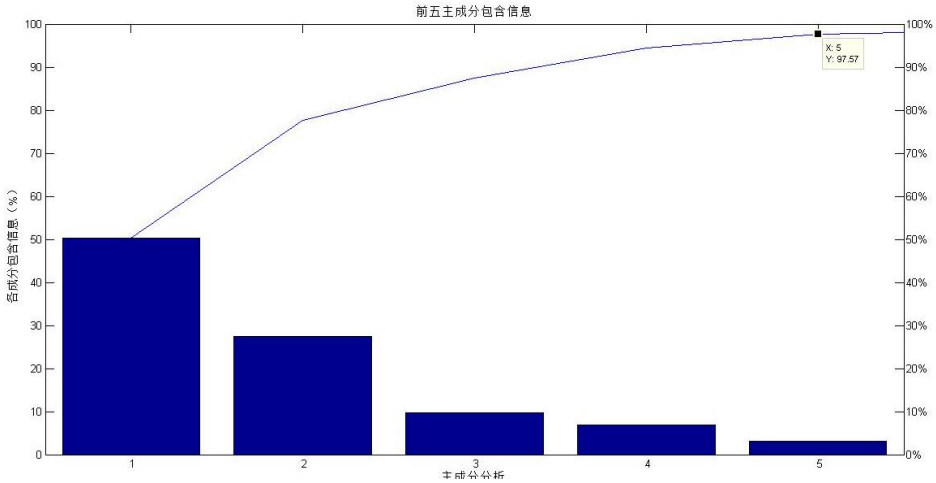


图 4 前五主成分包含信息

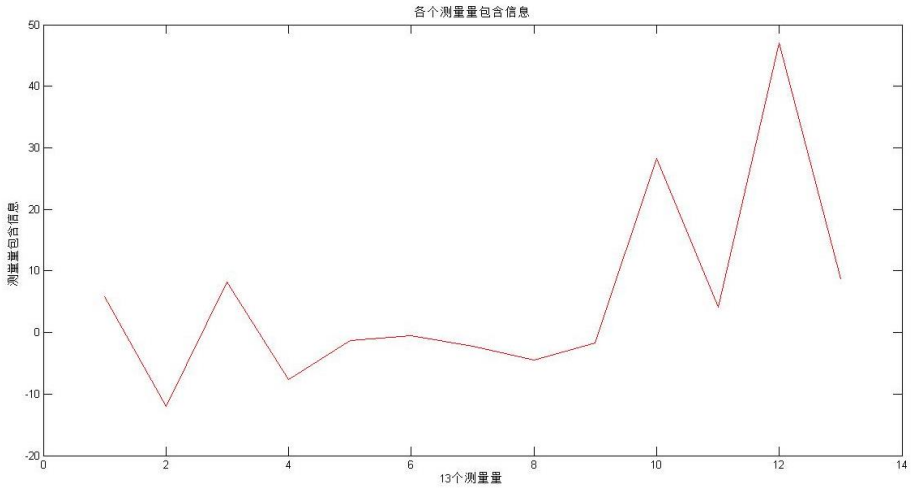


图 5 各个测量量包含信息

可以看到前五个主成分包含了 97.57%的信息，这证明该实验的 13 个测量量之间高度线性相关，这符合理论结果。而第 12 个量——一氧化碳质量分数在 13 个测量量之中也占极大主导地位。这与学界尤其是相关实验界对甲烷火焰的长久定性结论相符合。

对苏黎世理工学院提供的 CO/H2/N2 Jet Flame 数据作 PCA 操作结果如下图：

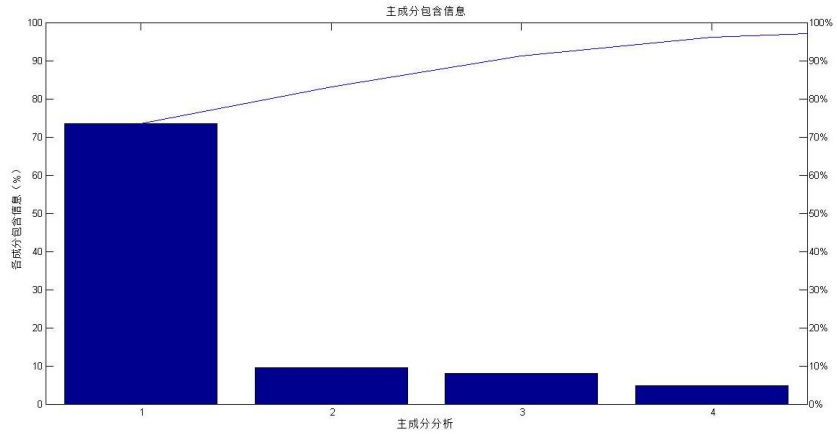


图 6 前四主成分包含信息

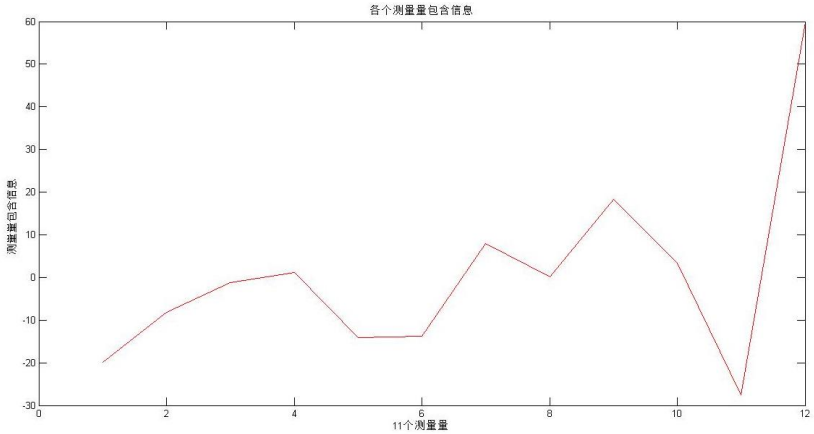


图 7 各个测量量包含信息

可以看到前四个主成分包含了 98.29%的信息，这证明该实验的 12 个测量量之间高度线性相关，甚至比甲烷实验的 13 个测量量更加线性相关，这符合理论结果。而第 12 个量——TDNR 在 12 个测量量之中的主导地位更为强势。这与甲烷火焰的性质有所不同。

对两类火焰的合数据集作 PCA 结果如下图：

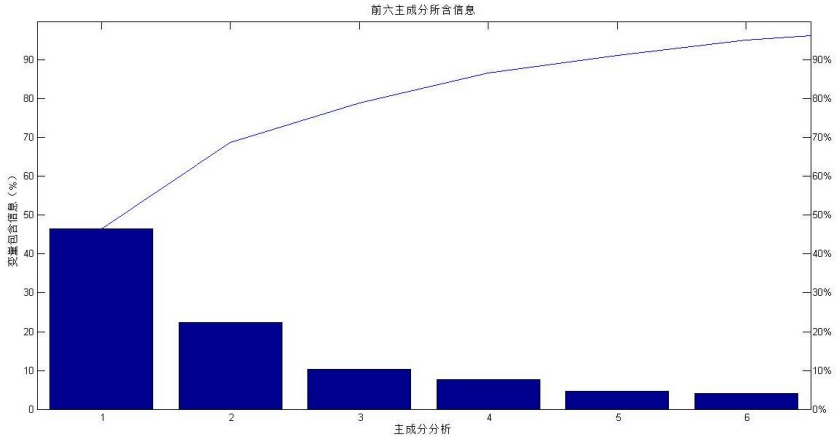


图 8 主成分包含信息

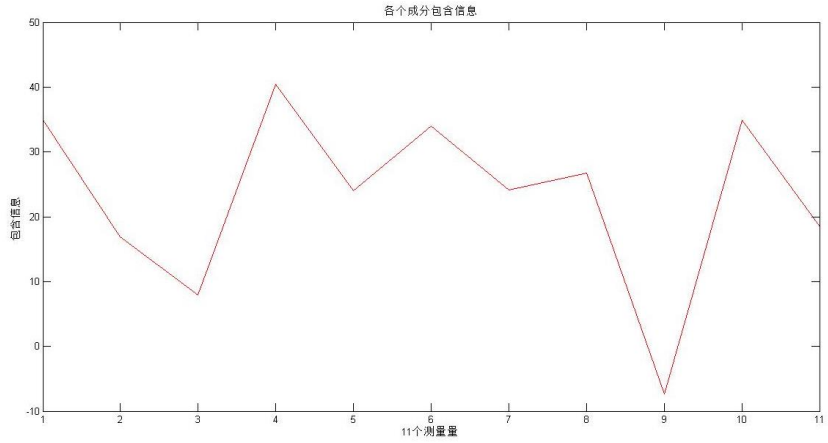


图 9 各个测量量包含信息

可以看到提取的前六个主成分包含了 95.28%的信息，这说明在删除了两类火焰一些非公共变量再将两类火焰合并之后，线性相关程度明显降低。在合并数据集中，各个测量量对整体的影响也有所降低，不再有明显的主导测量量。这有利于下一步机器学习。

2.4 本章小结

这一章介绍了作为本文基础的火焰数据集。首先介绍了数据来源——Sandia 工作站的主要情况，随后介绍了本文主要基于的 TNF 数据中苏黎世理工学院提供的 CO/H2/N2 Jet Flame 数据和达姆施塔克大学和斯图加特大学提供的 CH4/H2/N2 Jet Flames 数据，介绍了实验条件和数据由来。之后从原始数据中提取出了符合机器学习要求的数据集，并介绍了这一数据集的生成方式，数据分布和数量等属性。在选择特征标准化将数据集归一化之后，分别对 CO 火焰数据、甲烷火焰数据和合并数据集进行 PCA 操作，指出：对于甲烷火焰数据而言一氧化碳离子占主导地位；对于 CO 火焰数据而言，TDNR 占主导地位；而将两类火焰数据忽视非公有测量量而合并成的数据集而言，没有任一测量量占主导地位，各个测量量线性相关程度降低，有利于进一步机器学习。

第三章 基于 KNN 与 SVM 的火焰状态识别

K-最近邻（K-nearest Neighbor, KNN）与支持向量机（Support Vector Machine, SVM）是两种最经典的分类器，都具有实现简单、使用灵活的特点。由于本文基于的数据集样本容量人为设定非常平衡，因此两种分类器均适用于这个问题，因而首先尝试这两种分类器。

3.1 KNN 介绍

KNN 是数据挖掘中最简单的算法之一，所谓 K 近邻，就是最近的前 K 个邻居，也就是每个样本由其最近的 K 个邻居投票决定。KNN 算法的核心思想如下：如果一个样本在特征空间中的最相邻的样本中的前 K 个的大多数属于某一类别，则该样本也属于这个类别，并具有相应类别样本的性质。

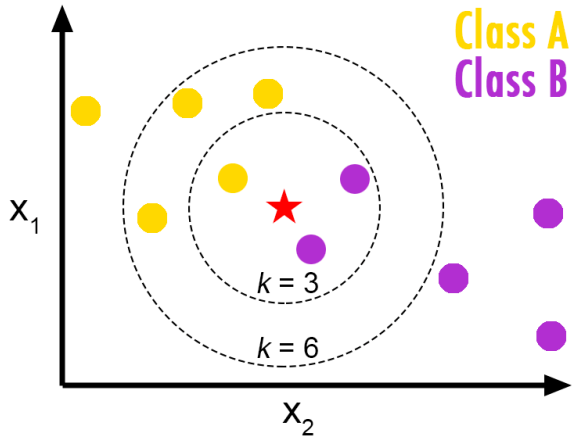


图 10 KNN 算法原理

KNN 算法原理如图所示。对于目标点（图中黑色叉号）而言，如果取 3 个最近邻点，则距离该点最近的 3 个点属性为两红一绿，该点则分类到红色一类。如果取 7 个最近邻点，该点最近的 7 个点属性为四红三绿，该点则分类到绿色一类。具体算法框架如下：

表 5 KNN 算法流程

1.设定唯一初始参数 K
2.遍历已有样本，求目前验证样本到所有样本的距离，并取出前 K 个距离最小点。
3.将 K 个最近邻点中占多数的性质赋给验证样本
4.对所有验证样本，重复 2~3
5.计算正确率，取不同的 K 值重复验证，取正确率最大的 K 值。

KNN 具有三个主要优点：1）算法架构简单，易实现，且没有训练过程。2）只有一个超参数（hyper parameter），容易取到最优解。3）适合于多分类问题（multi-model，或称多标签问题）。但也有两个缺点：1）样本数量不平衡大大影响分类结果，倾向于将新样本分类给样本数量较多的一类。2）计算量大，需要求出新样本到所有样本的距离，算法复杂度

O(n2)。因此渐渐淡出主流。

但在火焰数据分离中，由于数据集为人工选定，各类样本数量一致（均为 750），且样本数量较少（仅有 3000 个），维度相对也不高（仅有 176 维）;因此理论上这两个缺点影响不大，所以使用这个分类器尝试。

3.2 SVM 介绍

3.2.1 线性可分

对于分类问题中广泛存在的线性不可分问题，SVM 是一种常用的分类方法。SVM 基于结构风险最小化原理和 VC 维理论，属于有监督学习，适合解决小样本、非线性、高维识别等问题，可用于分类和回归等问题，1995 年由 Corinna Cortes 和 Vapnik 等首先提出。

SVM 的核心思想是在样本特征空间或样本特征映射而成的高维空间构造一间隔超平面，使得不同类的样本分布在间隔超平面的两边并使样本点到间隔超平面的距离最大。所谓支持向量，是指两类样本在空间内交界处的样本点，构造两个互相平行的超平面使得支持向量分别落在这两个超平面上，间隔超平面则与之平行并通过优化平面方向使得超平面间距离最大[18]。

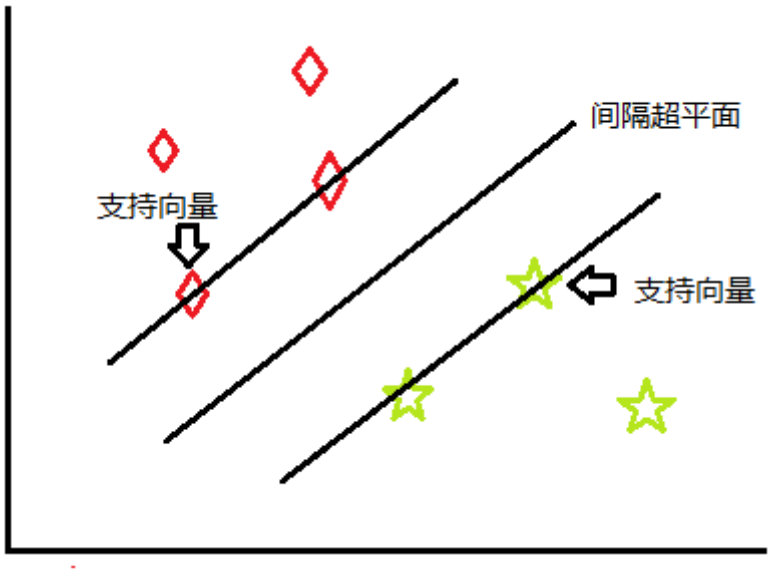


图 11 SVM 算法原理

为了量化这一间隔超平面的的优劣，引入几何间隔概念。对于高维空间内的超平面：

$$f(X) = w^T X + B \tag{1}$$

那么点到直线的距离可以写成

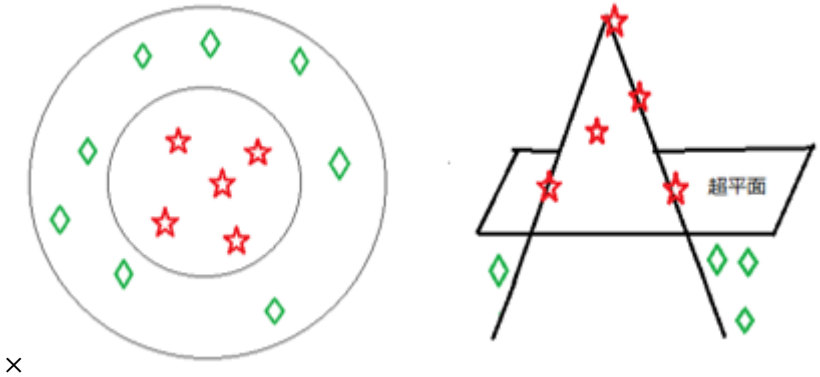
$$d = y \times \frac{w^T x + b}{\|w\|} (y\text{仅表示正负号}) \tag{2}$$

如果令$y(w^T x + b) = 1$，

则几何间隔最大值问题为

$$\max\left(\frac{1}{\|w\|}\right), s.t., y_i(w^T x_i + b) \geq 1, i = 1 \ldots\ldots n \tag{3}$$

转化为一个有约束优化问题。



×

图 12 核函数原理

3.2.2 核函数介绍

然而在大多数情况下不同类样本并不线性可分，如图 A，并不存在一个超平面可以完美分开两类样本。在这种情况下，我们引入映射概念，通过将数据映射到高维空间实现线性可分（如图 B）。简单来说，先在低维空间进行计算，再通过某种映射方式映射到高维空间，再在高维空间内构造最佳间隔超平面来实现线性分类。图 A,B 之间变换的映射如下：

图 A 中共有两个维度$\langle x_1, x_2\rangle$。

该平面内任意二次函数（图 A 中的圆周是一种特殊的二次函数）可以表达为：

$$a_1 \times x_1^2 + a_2 \times x_2^2 + a_3 \times x_1 + a_4 \times x_2 + a_5 \times x_1 \times x_2 + a_6 = 0 \tag{4}$$

如果我们令

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 \times x_2 \tag{5}$$

则上式转化为

$$\sum a_i \times z_i + z_6 = 0 \tag{6}$$

而这正是一个五维空间的超平面方程。换句话说，我们将二维空间内一个非线性的间隔曲面转化为五维空间内一个间隔超平面，这样就可以应用线性分类中的相关知识，将二维数据映射到五维空间内，构造支持向量最终得到间隔超平面。

但是这也存在一个问题，当原数据为二维时，映射之后数据为五维；当原数据为三维时，映射之后数据为十九维：映射之后的维度随原维度增长而指数增长直到无法运算。这时我们引入核函数的观点。核函数定义为满足以下关系式（以两个变量为例）：

$$\mathrm{k}(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j)\rangle = \varphi(x_i)^T \varphi(x_i) \tag{7}$$

即两个变量在特征空间的映射等于它们在原始低维空间中经过该核函数计算的结果，这样就可以在低维空间内计算而不必写出高维特征空间内的内积结果，这个映射又可称为隐式的定义了一个称之为“复生核希尔伯特空间”的特征空间，这样的映射就被称为核函数。

常见的核函数如下表所列：

表 6 常见核函数列表

名称	表达式	参数
多项式核	$\kappa(x, y) = (\langle x, y\rangle + c)^d$	多项式次数 d
高斯核	$\mathrm{k}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	高斯核带宽σ
拉普拉斯核	$\mathrm{k}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right)$	拉普拉斯核带宽σ

接表 6

Sigmoid 核	$k(x_i, x_j) = tanh(\beta x_i^T x_j + \theta)$	$\beta > 0, \theta > 0$

3.2.3 松弛变量介绍

但是，在真实数据中，必然存在噪点问题或者离群点问题。如下图所示，一个新样本的加入使得原本线性可分的问题又成为线性不可分。假若这个间隔超平面能完美划分现有的成千上万的点，则没有必要因为离群点而更改整个超平面，而应该忽略这个瑕疵点。因此引入“软间隔”的概念，即允许一些点不满足分类结果，而原本的要求所有点满足分隔则称为“硬间隔”[19]。

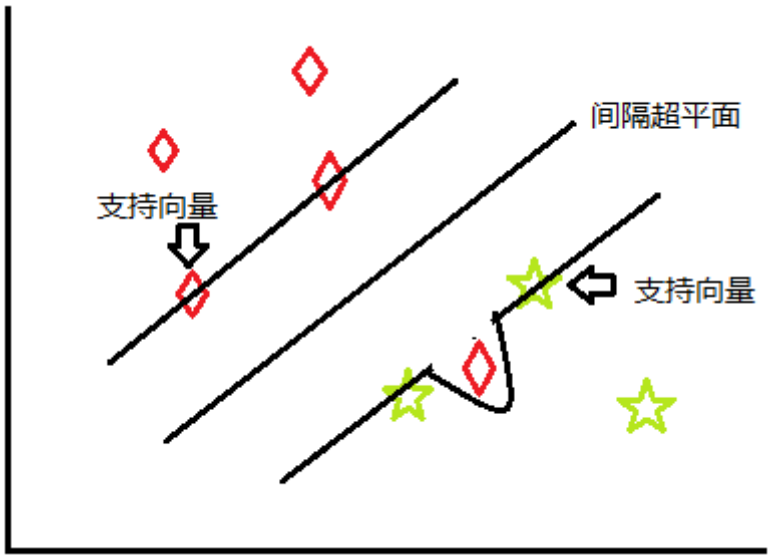


图 13 带松弛变量的 SVM

软间隔的数学表达式如下：

$$y_i(w^T x_i + b) \geq 1 - \delta_i \tag{8}$$

那么 SVM 的优化函数变为

$$\min\|w_i\| + C\sum_{i=1}^{n}\delta_i, \text{s.t.}, y_i(w^T x_i + b) \geq 1 - \delta_i, i = 1 \ldots\ldots n \tag{9}$$

其中δ_i称为松弛变量，C 称为惩罚因子。惩罚因子的大小代表对离群点的重视，C 越小，越不重视离群点，C 越大离群点对分类器的影响越大。

在本文接下来的实验验证中，引入交叉验证的概念。交叉验证是统计学上一种常用的评估统计分析、一种泛化能力验证方法，其基本原理是将数据集切割为若干较小的子集，每次使用一个子集验证然后其他子集训练，分别称为测试集和训练集。交叉验证需要满足两个条件：一是采样尽量随机，训练集和测试集的分布均匀；二是训练集比例足够高，至少需要全部数据集的一半。一半有两种交叉验证方式：一是 K 折交叉验证（k-fold cross-validation），即将数据集随机切割为 k 个样本数量相同的数据子集，每次使用 k-1 个子集训练然后使用一个子集测试，如此重复 K 次，将 K 次测试的结果平均值作为最终正确率；二是留一交叉验证（least-one-out cross-validation，LOOCV)，即每次仅留一个样本测试，使用其余样本作训练,优点是几乎每一样本都参与每次验证，每次验证过程具有最强的泛华能力，且交叉验证过程不受随机因素影响，缺点是计算复杂度极高。限于计算能力，本节采用 K 折交叉验证。

3.3 分类实验

3.3.1 KNN 实验

对于机器学习而言，训练集与测试集样本数量之比对训练结果影响很大，因此本节比较了每类样本取 300、400、500、600 个（训练集共计 1200、1600、2000、2400 个样本），与之对应的测试集每类样本 450、350、250、150 个（测试集共计 1800、1400、1000、600 个样本），一共 4 个实验比较训练集与测试集数量对 KNN 算法识别正确率的影响。

另外对于 KNN 算法而言，占主导地位的超参数为最近邻个数 K，因此实验中比较了 K 取 1-20 时的算法识别正确率。另外最近邻的权重也有一定影响，一般有三种权重分配方式：最近邻等权重、最近邻权重与距离倒数成正比、最近邻权重与距离倒数平方成正比。因此本节也比较了三种权重分配方式对识别正确率的影响。4 个实验正确率折线图如下：

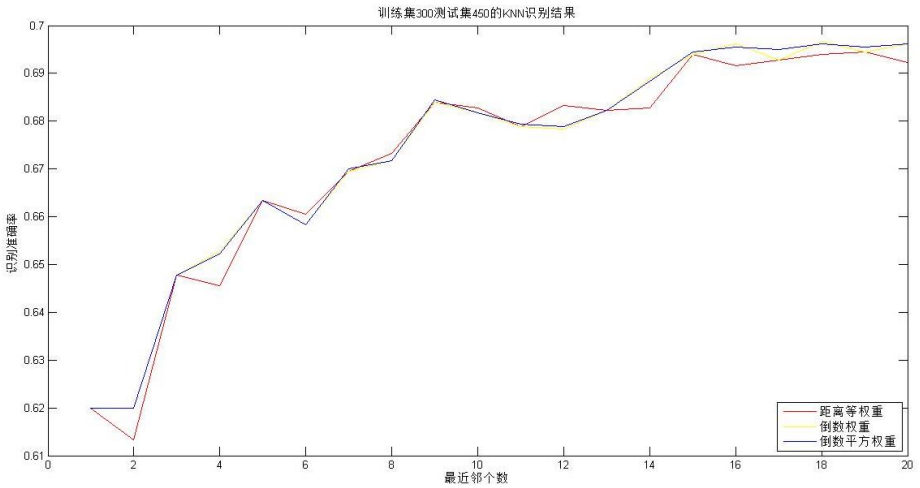


图 14 训练集 1200 个样本 KNN 识别正确率

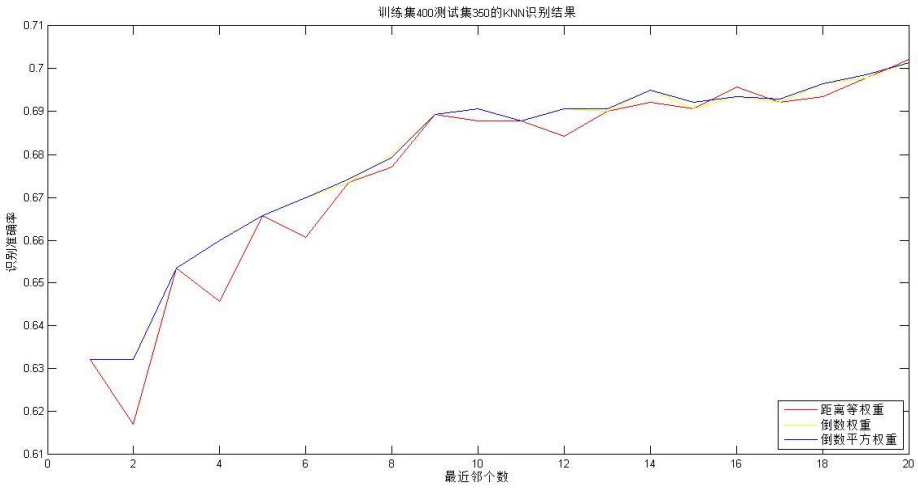


图 15 训练集 1600 个样本 KNN 识别正确率

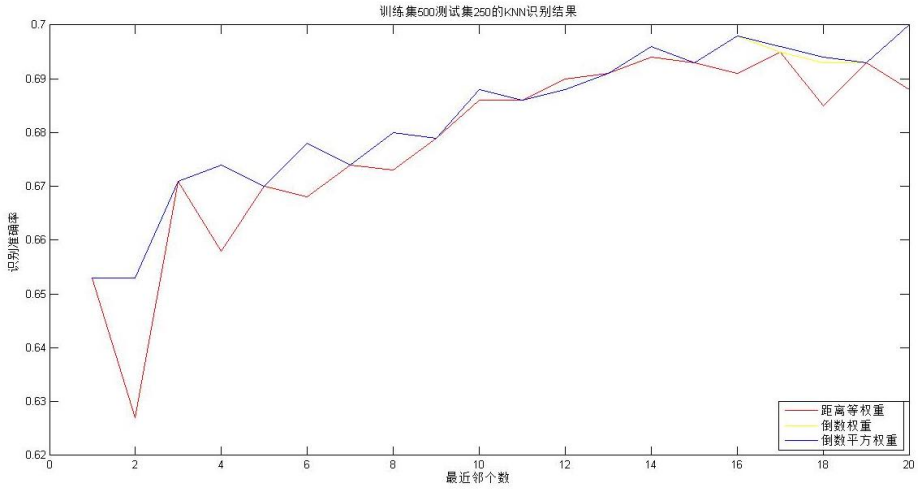


图 16 训练集 2000 个样本 KNN 识别正确率

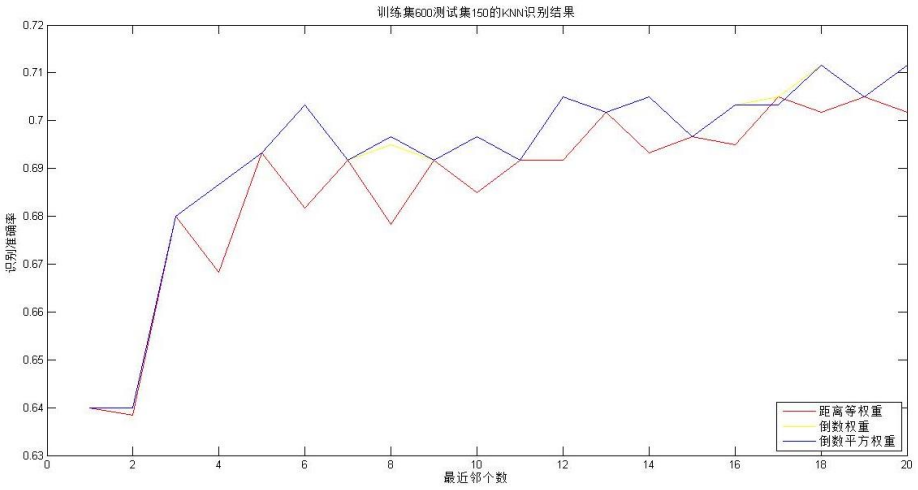


图 17 训练集 2400 个样本 KNN 识别正确率

首先可以得出结论，即 KNN 算法对本文的火焰数据集分类效果不佳，最高正确率仅仅略高于 70 %，一般在 60%到 70%之间波动。其次提高训练集占比并不能明显提高正确率，尤其在最后一次实验中训练集已经在整个训练集占比高达 80%，然而识别正确率依然在 64%到 71.2%之间。然后，最近邻权重分配的三种方式对识别正确率也没有太大影响，大多数情况下三种最近邻权重分配方式的正确率交替上升。最后，增加最近邻个数确实有利于提高正确率，可以看到在四张图中，随着最近邻个数从 1 升到 20，识别正确率大约提高了 8%到 10%。然而也可以从图中走势看出，增加最近邻个数并不能无限提高正确率，4 个实验的 12 根曲线都在最近邻个数取 20 左右进入平台区，基本稳定在 70%不再上升。

综上，可以认为 KNN 算法不适用于本文的火焰数据集识别，实验中的识别正确率不高并且没有预期也不能提高正确率。这可能是由 KNN 算法的本质决定的。可以发现火焰数据并不严格服从成簇的特点，大量样本分布于类别的分界面附近，或者说每一类样本的分布较为复杂，不能直接找到平滑的超曲面将各类分开。也有可能火焰数据样本分布在闵科夫斯基空间等非传统空间上符合一定有规律分布，但限于篇幅，本节不探讨这方面内容。

3.3.2 SVM 实验

a)有惩罚因子实验

如上文所言，原始数据收集于上世纪 90 年代，限于当时技术条件，这批数据并不精确，可能存在相当多的噪声点。因此本节采用了“软间隔”进行实验，来比较不同惩罚因子对分类正确率的影响。为了增加实验次数避免偶然性，选取了 Sigmod，将惩罚因子 C 与核函数

参数 gamma 结合起来制作成网格，在这一网格中寻找最优参数。

在本节实验中，设置 C=0.05:0.025:0.5,gamma=0.03:0.015:0.3,一共进行了一百组实验，正确率分布等高线如下;

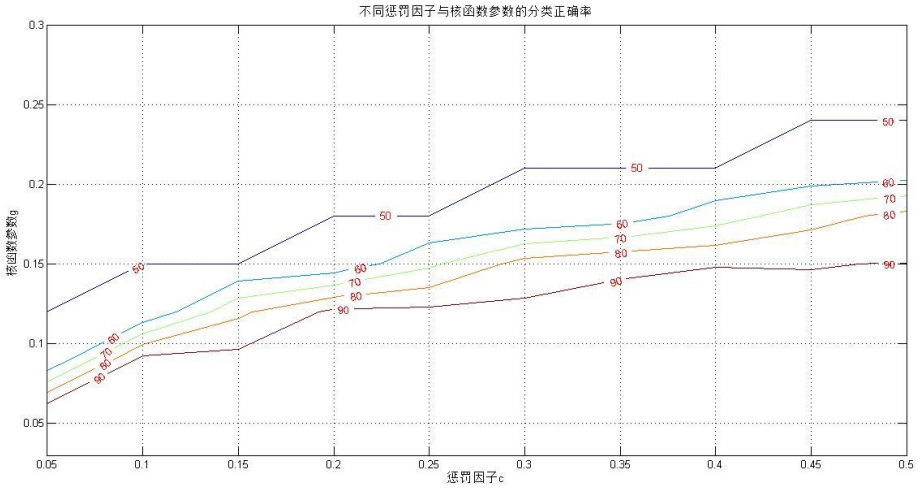


图 18 不同惩罚因子与核函数的分类正确率

可以看出，总体而言，惩罚因子越高、核函数参数越小，分类正确率越高。数据显示，分类正确率最高时，惩罚因子为 0.5，核函数参数为 0.03，最高分类正确率为 97.67%。由此可以得出，之前的猜测是错误的，原数据尽管当时测量技术老旧，但噪音并不大，完全可以使用硬间隔分类。

b)无惩罚因子实验

在排除了软间隔影响后，采用硬间隔分类，考察其他参数影响。对于 SVM 而言，识别正确率主要受训练集大小、核函数种类、惩罚选项、惩罚权重等等超参数。在本节实验中，依次采用了测试集样本数取数据集样本总数的 1/2，1/3，1/4，1/5，1/6，1/8，1/10 进行了七组实验。每一组尝试了七个核函数（二阶多项式核，三阶多项式核，四阶多项式核，高斯带宽为 0.5、1、2 的高斯核，以及 Sigmod 核、核函数参数选择 0.03）。每一次实验重复十次，每次随机取不同的训练集测试集，得到的实验正确率取十次平均，列表如下：

表 7 不同情况下 SVM 分类正确率

测试集占比	二阶多项式核	三阶多项式核	四阶多项式核	高斯核一	高斯核二	高斯核三	Sigmod 核
1/2	99.10	95.83	95.33	74.47	52.73	38.63	99.90
1/3	99.60	96.53	95.53	76.17	54.9	38.87	99.93
1/4	99.77	96.83	95.37	77.37	56.67	39.77	99.90
1/5	99.77	96.90	95.27	77.47	56.70	39.37	99.97
1/6	99.77	97.37	95.33	78.07	56.73	39.87	99.97
1/8	99.80	97.40	95.47	77.90	56.63	39.43	99.97
1/10	99.83	97.53	95.20	77.70	56.67	39.57	99.97

从上表可以明显看出不同测试集大小和不同核函数对火焰状态分类正确率的影响。首先，对于所有核函数，训练集占比越大测试集占比越小，总体上而言分类正确率越高，这符合机器学习基本常识。其次，更重要的是，不同核函数对分类正确率影响极大。三个高斯核函数的分类效果最差，最高分类正确率是高斯核带宽取 0.5 时的 78.07%，这仅仅比 KNN 算法的最好结果略好，可以认为高斯核函数最不适合本文的火焰数据分类。其次是三个多项式核，尤其是二阶多项式核，识别正确率随训练集占比增大而稳步提升，最高达到 99.83%，最低也有 99.10%，全部在 99%以上。对比三阶多项式核和四阶多项式核，阶数越高反而正确率越低，每提高一阶正确率下降 2 个百分点。由此可以得出，火焰的数据结构可能是一个低阶分布，因此低阶的多项式核反而可以更好将原数据映射到一个线性更可分的空间。最关键的是 Sigmod 核的结果。一般而言，作 SVM 分类时，很少用到 SIgmod 核，然而在本节实验中，Sigmod 核的分类效果最好，正确率随训练集占比提高而从 99.90%提高到 99.97%，几乎接近 100%的正确率。可以认为，Sigmod 核函数应当是最适合作火焰分类的核函数。

3.4 本章小结

这一章使用了两种经典的机器学习分类算法 KNN 和 SVM 对火焰状态进行分类。首先介绍了两种算法的原理，对两种机器学习算法中重要的超参数（KNN 算法的近邻数 K，SVM 的核函数种类）的概念和意义作了重点说明。对于 KNN 算法，在不同的训练集/测试集占比的火焰数据集上进行了实验，得出不同计算距离方式对分类结果无影响，增大近邻个数可以提高正确率但提高也有限，总之 KNN 算法不适合于本文的火焰数据集。对于 SVM 算法，首先讨论了软硬间隔对分类效果的影响，得出原数据中并不存在明显的离群点。随后在硬间隔基础上讨论不同核函数对分类效果的影响，得出使用 Sigmod 核函数的 SVM 算法对火焰数据集的分类效果最好，分类正确率高达 99.9%以上。本章节提出了有效的火焰状态分类方法，这也为下一步的火焰特征选择提供了参考。

第四章 基于 BP 神经网络的状态识别与特征选择

BP 神经网络（back propagation neutral network, BP NN）是目前较为流行的分类器，且带 Dropout 的神经网络可以用来比较各个维度的重要性，因此本章采用 BP 神经网络对火焰状态进行分类，使用带 Dropout 的神经网络作特征选择。

4.1 神经网络算法

4.1.1 神经网络原理

神经网络（neutral network, NN）由 1943 年心理学家 W.S.McCulloch 和数理逻辑学家 W.Pitts 首先建立了数学模型，称为 MP 模型。1986 年，Rumelhart, Hinton, Williams 提出了革命性的反向传播（back propagation, BP）算法,构建了 BP 神经网络[20]。BP 神经网络是一种误差前向传播的多层前馈网络，能够从大量的输入-输出映射自动学习，而无需输入-输出映射的数学模型。BP 神经网络采用梯度下降法，通过反向传播来不断调整网络中各个神经元的阈值和神经元链接的权重，使网络输出与训练集预期输出的误差平方最小。1989 年 Michael Nielsen 证明，BP 神经网络是一个万能逼近函数，即闭区间内的任何连续函数，都可以用含有一个隐含层的 BP 网络来无限逼近，这是 BP 神经网络能够拟合任意非线性映射的数学基础。一个典型的 BP 神经网络模型包括输入层（input）、隐层(hidden layer)和输出层(output layer)，其结构如图所示：

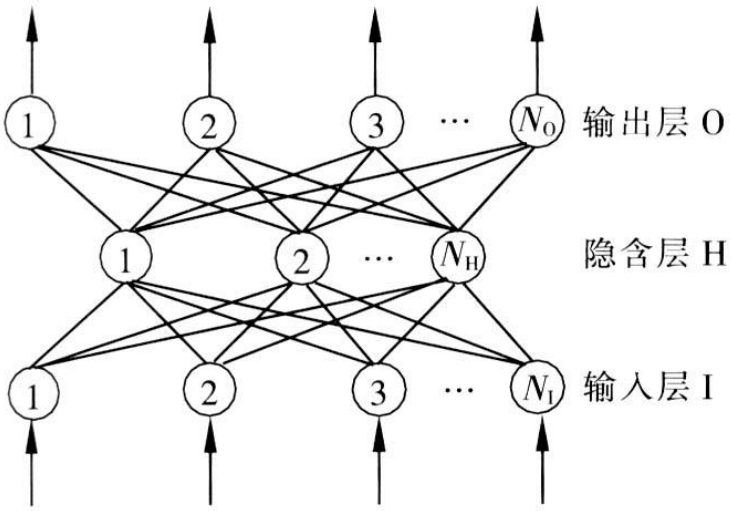


图 19 BP 神经网络原理

BP 神经网络的运算过程很简单，分为正向传播过程和反向传播过程。正向传播过程中现在设节点 i 和节点 j 之间的权值为w_{ij}，节点的阀值为b_i，每个节点的输出值为x_j，而每个节点的输出值是根据上层 m 个节点的输出值乘以当前节点与上一层 m 个节点的权值加上当前节点的阀值然后代入激活函数f来实现的。当前层第 j 个节点的输出x_j如下

$$S_j = \sum_{i=0}^{m-1} w_{ij} x_i + b_i \, , x_j = f(S_j) \tag{1}$$

正向传递的过程就是按照顺序计算上式。在 BP 神经网络中，默认阈值为 0。

反向传播过程是神经网络的核心。假设输出层第 j 个节点的输出值为d_j，则误差函数为

$$E(w,b)=\frac{1}{2}\times\sum_{j=0}^{n-1}(d_i-y_i)^2 \tag{2}$$

求这一误差函数对权重的偏导数 ε

$$\frac{\partial E(w,b)}{\partial w_{ij}}=\frac{1}{\partial w_{ij}}\times\frac{1}{2}\times\sum_{j=0}^{n-1}(d_i-y_i)^2 \tag{3}$$

$$\delta_{ij}=(d_i-y_i)\times f'(x), f'(x)\text{是激活函数导数} \tag{4}$$

$$\frac{\partial E(w,b)}{\partial w_{ij}}=x_i\times\delta_{ij} \tag{5}$$

在得到导数之后，根据梯度下降法，可以得到从后一层第 j 个节点到前一层第 i 个节点的神链接权重调整公式：

$$w_{ij}=w_{ij}-\varepsilon_{ij}\times\delta_{ij}\times x_i\text{，其中}\varepsilon_{ij}\text{是学习速率} \tag{6}$$

这样就可以通过多次迭代权重来使得神经网络无限逼近目标的非线性函数。

4.1.2 传递函数

决定 BP 神经网络性能的主要参数包括隐含层层数、激活函数类型、非线性优化方法和网络链接方式。隐含层层数对分类效果影响的量化分析目前学界还没有共识，一般而言，隐含层层数越多分类效果越好，但在有些场景中，隐含层层数越多分类效果反而下降。在下一节实验中，选取了 1-10 层隐含层来比较分类正确率。激活函数则一般包括以下四种函数[21]：

表 8 常见激活函数列表

函数名称	函数解析式
S 形函数 Sigmod	$\sigma(x)=\frac{1}{1+e^{-x}}(0<\sigma(x)<1)$
双 S 形函数 Tanh	$\sigma(x)=\frac{2}{1+e^{-x}}-1(-1<\sigma(x)<1)$
ReLU 函数	Max(x,0)
Leaky ReLU 函数	$\sigma(x)=\begin{cases}kx(x<0)\\x(x>0)\end{cases}(k\text{是一个很小的正数，如 }0.001)$

S 形函数 Sigmod 是历史最悠久、使用最为广泛的激活函数，特点是将任意输入值映射到 0~1 之间，尤其是无限小的负数趋近于 0，无限大的正数趋近于 1，对于二类分类而言可以将连续的输入值映射到离散的两类（0 和 1）并且衡量该输入值划分到两类的概率，而在神经网络中神经元的激活与未激活状态是一个典型的二类分类问题，因此 Sigmod 是最常见的激活函数类型。

双 S 形函数 Tanh 则与 Sigmod 类似，不同的是将连续的输入值映射到-1 和 1，这样输出的均值即为 0，这有利于下一步的梯度下降实现反向传播。

ReLU 函数是近期投入使用且迅速流行的激活函数。当输入值大于 0 时，输出即为输入本身。当输入值小于 0 时，输出为 0，相当于处于不激活状态。在激活状态时，ReLU 是线性函数，一方面可以使收敛速度加快，另一方面反向传播时线性函数的求导过程远远比非线性的 Sigmod 和 Tanh 简单。缺点是容易坏死，大多数神经元一旦关闭，导数永远为 0，不会再次投入网络运行。

Leaky ReLU 函数则部分改善了 ReLU 函数的缺点，减少了神经元坏死。当输入值小于 0 时，激活函数称为一个斜率较小的函数，保留了一些负输入的输出，有利于下一步运算。

4.1.3 优化方法

一般而言，神经网络采用梯度下降法。梯度下降法的直观解释就是沿着当前位置的绝对值最大的负梯度方向前进，犹如下山一样，始终沿着最陡峭的方向下坡。然而在这一过程中，可以找到全局极值点，也有可能停在局部极值点而出不去。对于神经网络而言，如果目标函数严格满足凸函数，则优化总能找到全局极值点。

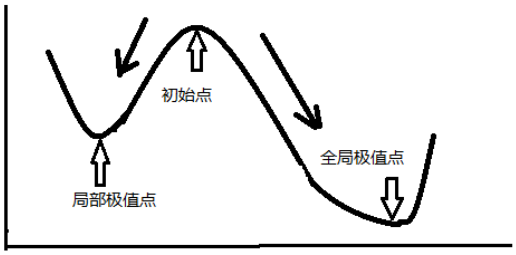


图 20 梯度下降法原理

梯度下降法的算法流程如下：

表 9 梯度下降法算法流程

1.设置目标函数初始位置θ_0，每次梯度下降的步长 a。

2.计算初始位置的梯度$\frac{\partial}{\partial\theta}J(\theta)$

3.步长乘以梯度得到下降距离$\mathrm{a}\times\frac{\partial}{\partial\theta}J(\theta)$

4.计算每一个维度的下降距离ε是否小于目标下降距离ε_0。如果小于，寻优完成跳出。如果不小于，继续步骤 5。

5.计算新的位置$\theta=\theta-\mathrm{a}\times\frac{\partial}{\partial\theta}J(\theta)$。重复步骤 2，3，4。

但是，如果目标函数不满足严格凸函数定义，梯度下降法非常容易陷入局部最小值点。为了减小陷入局部极值点的几率，引入 Levenberg-Marquardt 算法(下称 LM 算法)。LM 算法是介于牛顿法与梯度下降法之间的一种优化方法，最大优点在于对冗余参数不敏感，使代价函数不容易陷进局部极值点，这使得 LM 算法广泛应用于机器学习领域[22]。LM 算法的流程如下：

表 10 LM 算法

1.设置初始值，包括初始点P_0，终止变量ε，阻尼因子μ

2.计算 Jacobi 矩阵J_k,计算$\overline{N_k}=J_k^TJ_k+\mu_kI$,构造增量正规方程$\overline{N_k}\cdot\delta_k=J_k^T\varepsilon_k$。

3.求解增量正规方程得到δ_k

（1）如果$\|x-f(p_k+\delta_k)\|<\varepsilon_k$,令$p_{k+1}=p_k+\delta_k$。若$\|\delta_k\|<\varepsilon$,跳出循环，输出结果；否则令$\mu_{k+1}=\mu_k/\gamma$，转到步骤 2。

（2）如果$\|x-f(p_k+\delta_k)\|\geq\varepsilon_k$,令$\mu_{k+1}=\mu_k\cdot\gamma$，重新解增量正规方程得到$\delta_k$，返回步骤 1。

在下一节的实验中，将比较两种优化方法的优劣。

4.2 基于 BP 神经网络的火焰状态识别

本节的实验主要是尝试利用前馈神经网络进行火焰状态分类，在不同训练集/测试集占比上比较不同隐含层层数和不同优化方式对神经网络性能的影响。在本节中，先后选取了样本数目为 400，800，1200，1600，2000，2400 的训练集，隐含层层数选取了 1-10 层，优化方式则比较了梯度下降法和 Levenberg-Marquardt 算法。下表第一列为训练集样本个数，第一行为隐含层层数，表格内为分类正确率（%）。

表 11 基于梯度下降法的状态识别

训练集	1 隐含层	2 隐含层	3 隐含层	4 隐含层	5 隐含层	6 隐含层	7 隐含层	8 隐含层	9 隐含层	10 隐含层
400	52.23	96.19	99.35	99.08	99.35	99.50	98.35	99.88	99.46	99.23
800	51.82	99.64	74.18	99.86	99.82	99.91	99.95	100.0	100.0	99.95
1200	50.00	99.22	74.83	99.89	99.89	99.94	100.0	100.0	100.0	99.94
1600	50.00	99.21	75.00	99.79	99.93	99.86	100.0	100.0	100.0	100.0
2000	49.80	99.40	99.90	99.80	99.80	99.90	100.0	100.0	99.90	99.90
2400	91.83	99.83	100.0	100.0	99.83	99.50	100.0	100.0	99.83	99.67

表 12 基于 LM 算法的状态识别

训练集	1 隐含层	2 隐含层	3 隐含层	4 隐含层	5 隐含层	6 隐含层	7 隐含层	8 隐含层	9 隐含层	10 隐含层
400	74.04	97.77	99.50	91.54	99.27	99.77	99.08	99.85	99.54	98.81
800	99.50	99.50	99.95	99.95	99.95	94.55	99.95	100.0	100.0	99.68
1200	99.78	99.78	99.94	99.83	99.89	99.94	99.44	99.72	100.0	100.0
1600	99.93	99.86	99.93	100.0	98.36	100.0	100.0	100.0	100.0	100.0
2000	99.90	99.90	100.0	100.0	99.90	100.0	99.70	100.0	99.90	100.0
2400	100.0	99.83	99.83	100.0	99.83	99.67	99.67	100.0	100.0	100.0

从上表可以看出，大多数情况下神经网络的分类效果还是很好的，分类正确率轻松到达 90%,,甚至在一些实验中能够达到 100%。从趋势上来说，毫无疑问，训练集占比越大，分类准确率越高。仅取 400 个样本作为训练集时，最低准确率只有 52.23%和 74.4%；取 2400 个样本作为训练集时，分类准确率都在 99%以上，这一趋势符合机器学习的基本常识。从隐含层层数的影响而言，也有很明显趋势，即隐含层层数越多分类准确率越高。但在少数情况下，也有隐含层层数增加而性能降低，这可能数据样本数太少有关。相对更重要的是，发现两种优化方式存在明显差异。在所有训练集所有神经网络上，LM 算法的效果始终好于梯度下降法，能够在较苛刻的条件下实现 100%分类准确率。因此，可以认为 LM 算法更加适合于火焰数据集分类问题。

4.3 基于 Dropout 神经网络的特征选择

从第三章和第四章第二节可以看出，火焰数据的状态分类是较为简单的，经过简单参数寻优之后，SVM 和 BP 神经网络都可以实现接近甚至达到 100%的识别正确率。于是引出了一个新问题：在样本数据的 176 个维度中，哪些维度才是有用的。显然，状态识别正确率极高说明这批数据的维度是高度冗余的，因而有极大必要从中筛选出真正重要的维度，这些维度真正决定了火焰的状态。

4.3.2 Dropout 原理介绍

利用神经网络筛选出重要的决定火焰状态的维度，很自然的想法是保留训练集的维度不变，利用完备维度的训练集训练神经网络。使用这一神经网络在缺省某些维度的测试集来做测试，再在完备维度的测试集上做测试，比较两者分类准确率的下降，以分类准确率的下降多少反映这些缺省维度在分类的重要性；或是使用这一神经网络在只保留某些维度的测试集上作测试，将准确率与在完备维度的测试集的实验准确率对比，以准确率高低反映保留维度的重要性。

但是一般而言，在完备维度上训练的普通参数的神经网络，只能对完备维度起到分类效果。简单而言，假如目标测试集由 n 维样本组成，却用 m(m>n)维的样本组成的训练集训练神经网络，那么可以认为，这一神经网络一定是过拟合的，这一过拟合的网络不能实现对 n 维样本的成功分类。

过拟合是机器学习中常见的概念（overfitting，或称 over-trained 过度训练），是指在训练一个机器学习模型时，采纳了过多冗余信息。已知数据集信息总量不变的情况下，训练出的模型只要过分复杂，总能完美满足所有的输入——输出映射。过拟合违反奥卡姆剃刀原则（如无必要，勿增实体，或称简单有效原则）。过拟合会大大减弱模型的泛化能力，在不同的测试集上表现会大大降低。具体来说，机器学习的模型是有训练集的输入——输出映射总结得到，即可以预期学习模型的输出结果。然而，学习模型应该有一定的泛化能力，应该在更广泛的，未能预知输出结果的输入上作出合适映射。然而，由于只能预期训练集的输入—输出映射并偏向于完美总结出这种映射，学习模型却会去适应训练集中离群性的随机性的特征，特别是训练过程太长或者训练样本数太少时。如果模型陷入了过拟合，随着泛化（generation）能力下降，模型特化性上升，当训练集测试结果表现更好时，测试集上的表现一定更差。

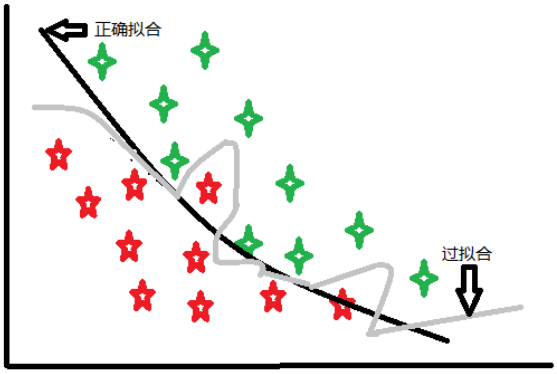


图 21 正确拟合与过拟合

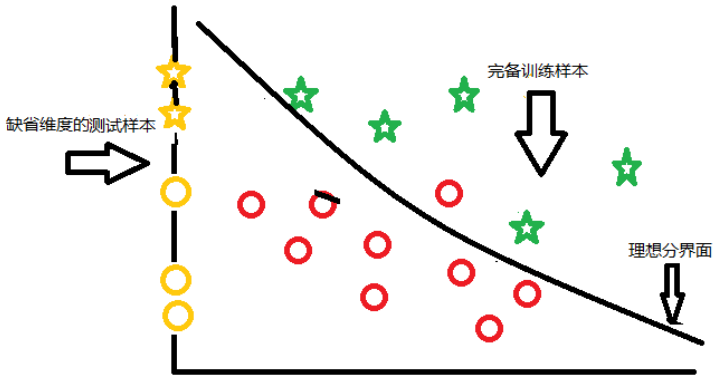


图 22 二维训练集训练的分隔面

上文提到的完备训练集与缺省测试集，也可以认为是一个过拟合问题。如上图所示，使用二维的训练集，训练出的分界面更倾向于完美划分二维数据，但在在缺失一个维度的一维测试集上不能做出划分。理想的分界面应该像下图所示，既能对二维数据作不错的分类效果，又具有很强的泛化能力，对一个维度缺失的数据也能做到分类（图中黄色的测试集数据，可以看到被分界面分成两半）。因此，在这一节的神经网络筛选特征实验中，首先要搭建的就是一个能够较完美解决过拟合问题的神经网络。

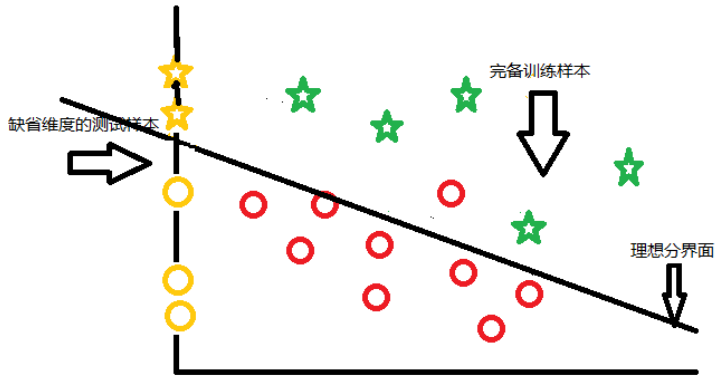


图 23 理想分隔面

在机器学习中，为了避免过拟合现象，研究者有很多成熟但未必一定适用的方法（如交叉验证、early stopping、贝斯信息量准则（英语：Bayesian information criterion）、赤池信息量准则或 model comparison），以指出何时会有更多训练而没有导致更好的一般化。在本节中采用 dropout 来减小这种过拟合的影响。

Dropout 是由 Nitish Srivastava, Geoffrey E Hinton 于 2014 年提出的[23]，通过阻止特征检测器的共同作用来提高神经网络的性能，是一种简单的防止过拟合的方法。神经网络的训练过程可以认为是神经元链接的权重和神经元输入阈值不断更新以减小网络输出和预期输出的差别，hinton 提出，在这一过程中，可以暂时的随机令一些权重置 0，在随机梯度下降中，如果随机丢弃一些权重，相当于用很多 mini-batch 训练不同的网络。

4.3.2 Dropout 原理推导

对于 dropout 为何有效，学界暂时没有达成共识，一般而言有两派观点。一方观点由 Hinton 本人提出。Hinton 认为在深度学习的训练过程中，计算消耗时间过长和模型过拟合是相伴而生的两大缺点。为了解决这两个问题，一般而言会采用整体训练方法，即训练多个模型作组合，举图中的例子，即训练出不同的分界面，再对这些分界面作一定程度的组合，也

许能组合出所需的泛化能力更强的分界面。但是这一训练方法的问题在于不但训练单一模型过程变成训练多个模型，大大增加了训练时间，而且对于组合方法也没有理论研究或者经验公式，测试组合方法本事大大增加了训练难度。Hilton 提出的 dropout 解决了这个问题，由于权重置零是完全随机的，相当于从一个网络神经中每次找出一个更瘦的网络来更新权重。假定一个有 n 个权重的神经网络，每个权重被随机丢弃的概率为二分之一，则这一个神经网络相当于2^n个神经网络的组合（当然实践中未必这么理想，随机丢弃过程不能保证遍历每一个点），且这个组合的组合方式遵从神经网络训练的过程自动实现最优组合，但需要训练的参数还是 n 个，没有增加计算复杂度，部分解决了费时的问题。

以上的解释是 Hilton 最初的解释，认为 dropout 是整体训练的近似，但实际上，dropout 是在一个神经网络上训练出一套权重和阈值参数，这唯一的一套权重和阈值是应当从动机上分析有效性。Hilton 举了一个生物学的例子：生物繁殖的方式主要有两种，有性繁殖和无性繁殖，前者指后代基因从父母双方各继承一般。无性繁殖的优点是可以保留亲代大段的优秀基因，有性繁殖则将大段基因随机拆开重组，似乎破坏了优良基因的完整性。但是大自然选择了有性繁殖，大多数高等动植物采用有性繁殖，因为基因联合的能力强于单个基因（从概率上讲，总有一组随机组合的效果不弱于于固定组合）。有性繁殖不但能讲优秀基因传递下来，还可以提高基因的泛化能力，使得大段即优良基因的联合适应性变成一系列小段基因的联合适应性。Dropout 也能达到类似的效果，将呆板的神经元链接打散，强迫随机挑选出的神经元链接共同工作，这样可以降低整个神经网络的联合适应性，增强泛化能力。

有了 dropout 的神经网络示意图和理论推导如下：

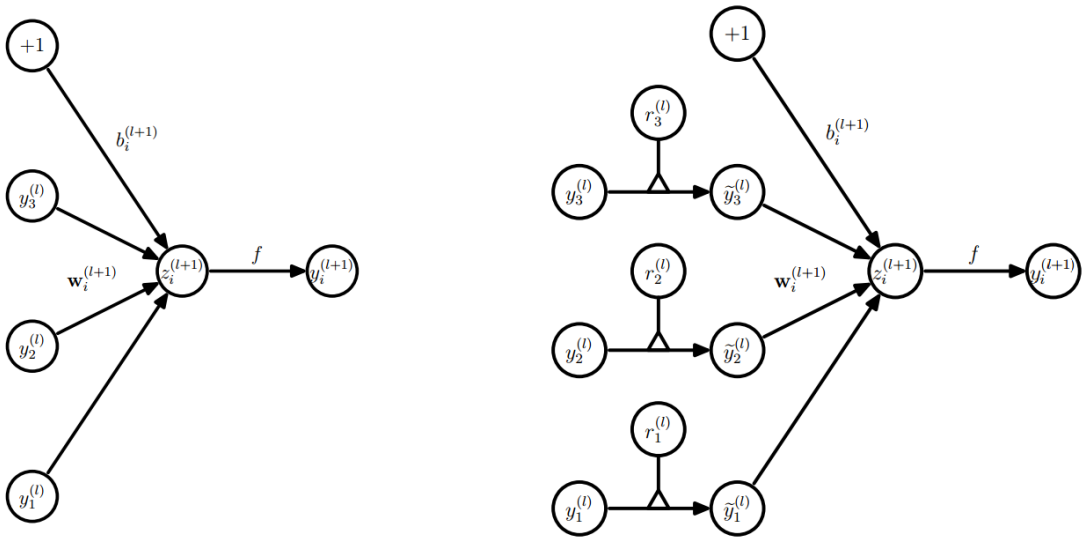


图 24 有无 dropout 神经网络对比

对于没有 dropout 的神经网络：

$$z_i^{l+1} = w_i^{l+1} y^l + b_i^{l+1} \tag{7}$$

$$y_i^{l+1} = f(z_i^{l+1}) \tag{8}$$

有 dropout 的神经网络：

$$r_j^l \sim Bernoulli(p) \tag{9}$$

$$\check{y}^l = r^l \times y^l \tag{10}$$

$$z_i^{l+1} = w_i^{l+1} \check{y}^l + b_i^{l+1} \tag{11}$$

$$y_i^{l+1} = f(z_i^{l+1}) \tag{12}$$

交叉验证和理论推导都可以证明 dropout 率取 0.5 时效果最好，因为此时生成的子网络结构最多。

测试时每一个预测单元预乘以 P

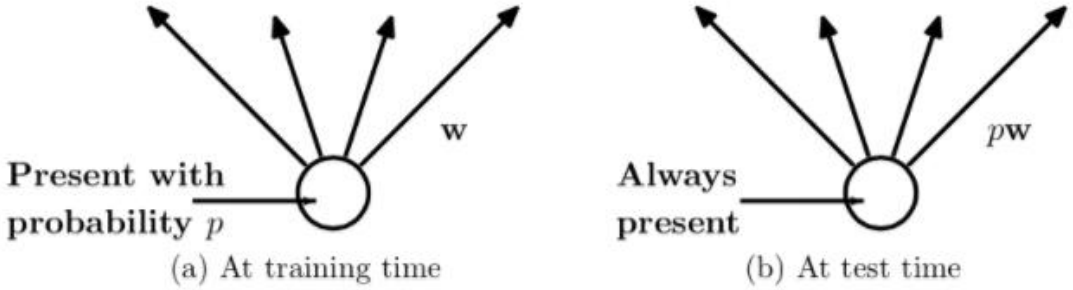


图 25 有无 dropout 神经元对比

另一派观点认为 dropout 网络事实上是在训练时对训练集作数据集扩张，总可以找到一个样本，使得在有无 dropout 设置的网络上达到一样的效果[24]。比如，对于某一样本，训练中某一层，某次更新之后的链接权重为（0，1，2，3，5），那么该样本一定对应着一个训练中该层该次更新的链接权重为（1，1，2，3，5）只要第一个 1 是被 dropout 的权值。同理（2，1，2，3，5），（3，1，2，3，5）甚至（1.5，1，2，3，5）。这样，每一 dropout 等价于增加了很多样本。这里作者引入了两个观点。一是部分空间学习。当属于某个特定类的数据点在输出空间内沿线性流形或子空间分布时，学习少量特征就足够描述整个流形的特性。但是，当数据服从高度非线性的不连续流形分布时,合理描述这种分布的学习应当能够通过描述输入空间的局部子空间的特征，有效地 "平铺" 空间来定义非线性决策边界。二是局部簇。在作非线性分类时，如果数据量足够，即便数据间的重叠度很高，不必担心过拟合行为。但当数据量不够时，应当通过增加数据的稀疏性，来增加特征的区分度。在使用了 dropout 情况下，相当于得到了更多的局部簇，在同等数据量的情况下，簇变多意味着稀疏性变大，更加可分。

在此，定义了 dropconnect 公式：

Dropout:

$$h_n = \vec{w}_n^T(\vec{r} \cdot \vec{x}) + b_n \tag{13}$$

DropConnect

$$h_n = (\vec{r} \cdot \vec{x})^T \vec{x} + b_n \tag{14}$$

4.4 基于 Dropout 神经网络的特征选择实现

如上文所言，本文的特征选择并不是纯统计意义的特征选择，而是有一定物理背景的需要符合物理理论的特征选择，特征选择的目的，不仅仅是提高分类准确率，更是找出射流火焰的主导因素以帮助简化昂贵的射流火焰实验测量。因此，本节将特征维度划有两个类属性：测量点类属性和测量量类属性，176 个维度中，1-11 个维度为第一个测量点（截面 20，半径 -1.5），12-22 个维度为第二个测量点（截面 20，半径 6）依次类推。第 1、12、23、34、45、56、67、78、89、100、111、122、133、144、155、166 个维度都是属于第一个测量量当地当量比，第 2、13、24、35、46、57、68、79、90、101、112、123、134、145、156、167 个维度都属于第二个测量量温度，其他测量量包括氧气质量分数、氮气质量分数、氢气质量分数、水质量分数、一氧化碳质量分数、二氧化碳质量分数、氢氧根离子质量分数、一氧化氮质量分数、TDNR 以此类推。由此可以将实验分为两类，即测量点重要性排序和测量量重要性排序。

4.4.1 验证 Dropout 有效性

首先验证 dropout 的有效性，采用带有 dropout 和不带有 dropout 的神经网络对缺失任一测量量维度的测试集进行状态分类，比较正确率差异得到不同测量量正确率。在本节实验中，

首先使用完备的训练集训练不带 dropout 神经网络，测试时每次将测试集的一个测量量维度删去，如测试当地当量比的重要性，将测试集的第 1、12、23、34、45、56、67、78、89、100、111、122、133、144、155、166 个维度置 0，得到缺省当地当量比维度的测试集的分类正确率。测试温度的分类重要性，将测试集的第一个测量量当地当量比，第 2、13、24、35、46、57、68、79、90、101、112、123、134、145、156、167 个维度删去，得到缺省温度维度的测试集的分类正确率。采用了训练集样本数 800 个，测试集样本数 2200 个，神经网络采用 2 层隐含层，寻优方法采用 LM 算法。实验结果如下表：

表 13 缺省测试集在无 dropout 神经网络上测试

缺省维度	当地当量比	一氧化氮	水质量分数	氮气	氢气	氧气
正确率%	45.29	39.05	33.3871	48.93	34.50	40.18
缺省维度	一氧化碳	二氧化碳	氢氧根离子	温度	TDNR	
正确率%	42.30	51.66	25.00	42.27	35.12	

表 14 缺省测试集在有 dropout 神经网络上测试

缺省维度	当地当量比	一氧化氮	水质量分数	氮气	氢气	氧气
正确率%	66.50	35.05	48.7727	72.13	94.55	74.91
缺省维度	一氧化碳	二氧化碳	氢氧根离子	温度	TDNR	
正确率%	72.32	92.86	76.36	61.82	97.05	

从上表可以明显看出有无 dropout 的神经网络对缺省维度测试的影响。在无 dropout 情况下，神经网络对缺省数据集的样本分类效果极差，分类准确率大多分布在 30%到 50%之间，而根据先验知识可知，11 个测量量中必然有对分类影响不大的测量量，在缺省这些测量量时，分类准确率应当基本不变。同时 11 个测量量中必然有对分类占主导低维的若干测量量，在缺失这些测量量时，分类准确率应当大幅下降。由此可知，无 dropout 的神经网络不适合用来作特征选择。相反有 dropout 的神经网络符合本节的需求，正确率最高为缺失 TDNR 时的 97.0455%，说明 TDNR 特征在状态分类中毫无重要性；正确率最低为一氧化氮质量分数 35.0455%，说明一氧化氮质量分数在状态分类占主导地位。

4.4.2 火焰测量量重要性实验

本节实验比较各个测量量和测量点在分类准确率的重要性，首先使用完备维度的训练集训练一个 dropout 等于 0.5、隐含层层数为 2、优化方法采用 LM 算法的神经网络，再依次置零测试集样本各个测量量所属维度和各个测量点所属维度，根据正确率下降多少得到测量量和测量点的重要性排序，再将数据集中重要性排序靠前的测量量与测量点提出构建训练集和测试集进行测试，得到重要的测量量和测量点对火焰状态的代表性。

表 15 缺失单一测量量时的火焰状态识别正确率

缺省维度	当地当量比	一氧化氮	水质量分数	氮气	氢气	氧气
正确率%	66.50	35.05	48.7727	72.13	94.55	74.91
缺省维度	一氧化碳	二氧化碳	氢氧根离子	温度	TDNR	
正确率%	72.32	92.86	76.36	61.82	97.05	

由上表可以看出，测量量重要性排序为一氧化氮、水、温度、当地当量比、氮气、一氧化碳、氧气、氢氧根离子、二氧化碳、氢气、TDNR。为了验证代表性，只保留前 n 重要的测量量再作分类测试，结果如下表：

表 16 保留前 n 测量量的识别正确率

保留前 n 维	1	2	3	4	5	6
正确率%	95.45	99.86	99.14	99.18	99.05	99.59
保留前 n 维	7	8	9	10	11	
正确率%	99.77	100.00	98.36	97.36	99.50	

由上表可以得出，仅需要保留前两个测量量一氧化氮质量分数和水质量分数就可以使得分类正确率达到 99%以上。

4.4.3 火焰测量点重要性实验

表 17 缺失单一测量点时的火焰状态识别正确率

测量点序号	1	2	3	4	5	6	7	8
分类正确率	93.00	83.18	96.50	90.09	95.45	94.86	92.73	87.82
测量点序号	9	10	11	12	13	14	15	16
分类正确率	90.00	95.73	89.77	96.73	92.73	96.50	82.77	85.05

表 18 缺失 n 个测量点时，最低识别正确率

测量点个数	1	2	3	4	5
测量点序号	15	2，5	1，2，5	1，2，5，12	1，2，5，7，12
分类准确率	82.77	70.09	59.23	67.19	79.50

可以看出表一与表二是互相矛盾的，如果只去掉一个测量点，可以看出 15 号测量点是影响最大的，导致分类正确率下降到 82.77%。但是如果去掉多个测量点，则前 n 重要的测量点变为 2，5，1，12，7 号点。因此，只保留上表 n 个测量点组合测试识别正确率：

表 19 只保留 n 个测量点组合测试识别正确率

测量点个数	1	2	3	4	5
测量点序号	15	2，5	1，2，5	1，2，5，12	1，2，5，7，12
分类准确率	79.68	90.09	99.23	97.19	99.50

由上表可以看出，只需要保留 1，2，5 号测量点，火焰状态识别正确率达到 99%以上。如果只保留 15 号测量点，正确率仅为 79.68%。由此可以看出应当保留 1，2，5 号测量点。

4.5 本章小结

本章首先介绍了 BP 神经网络的算法原理并将其用于火焰状态识别，比较了不同训练集/测试集样本数之比、不同隐含层层数、不同优化方式的 BP 神经网络分类效果，提出一个使用 Levenberg-Marquardt 优化算法的 4 层隐含层 BP 神经网络可以完美完成火焰状态分类任务。接着提出通过比较缺失不同维度的测试集的识别正确率来进行特征选择任务，介绍了 dropout 概念，并比较了有无 dropout 的神经网络在上述缺省测试集的分类效果。然后在 dropout 神经网络上比较各个维度的重要性，通过完备训练集-缺省数据集和提取训练集-提取测试集两种角度对各个测量量和测量点进行了重要性排序。最终得到 1、2、5 号三个测量点和一氧化氮质量分数和水质量分数两个测量量在火焰状态中占主导地位。

第五章 基于随机森林的状态识别和特征选择

由上文可知，我们采用带 Dropout 的神经网络作了状态分类和特征选择，但在随机森林中，这两步可以同时完成。因此本章采用随机森林方法同时进行状态识别和特征选择。

5.1 随机森林算法

5.1.1 分类树介绍

分类树（Classification Tree）属于基于 ID3 算法的决策树的一种，是一种简单高效的有监督学习分类器。与其他机器学习算法一样，通过训练集学习模型，对测试集作出分类测试。分类树的原理如图所示，是一个树结构，每一个分叉节点代表样本的一个维度的线性分类器，每一个分支代表该维度在这个线性分类器的分类结果，每一个叶子（树枝末端）代表一个类别。使用时，测试集样本从根部开始，依次经过每一个分叉节点测试每一个维度的特征，输出到下一个分支，直到该样本到到达最终的叶子，获得该样本的分类结果。分类树的分类结果简单易懂，被广泛应用于各个领域。

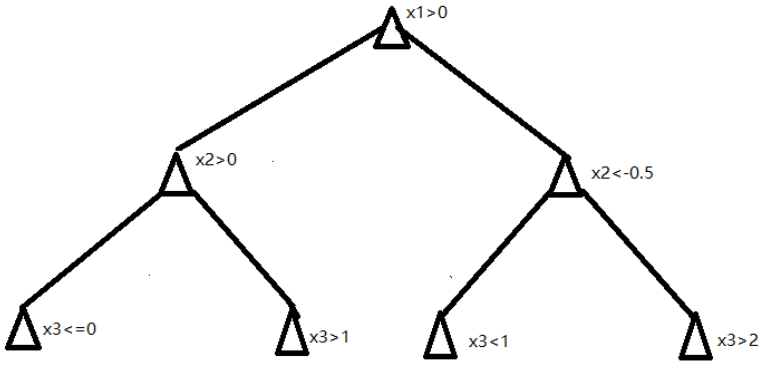


图 26 分类树简图

分类树的训练过程可以认为是更新各个分叉节点的线性分类阈值，直到训练集中每一个样本被分类到对应的类别中，或者说每一个叶子内的样本都属于同一个类别（如果样本维度用完了叶子内的样本还不属于同一个类别，则通过多数表决投票来决定叶子的类别）。引入信息熵的概念来更新分叉节点。假设事件 A 的所有可能性为$A_1, A_2, A_3, \dots A_n$，发生这些可能的概率为$P_1, P_2, P_3, \dots P_n$，则信息熵的定义为：

$$\text{entropy(A)} = \text{entropy}(P_1, P_2, P_3, \dots P_n) = -\sum_{i=1}^{n} P_i log_2 P_i \tag{15}$$

假设将事件 A 按照A_1属性划分，则期望信息熵为

$$\text{entropy}(P_1) = \sum_{j=1}^{n} \frac{|A_j|}{|A|} entropy(P_J) \tag{16}$$

定义信息增益：

$$\text{gain(A)} = \text{entropy(A)} - \text{entropy}(P_1) \tag{17}$$

每个节点分裂时应当遵循使信息增益最大的原则。因此分类树的算法流程如下：

表 20 分类树算法流程

1 将所有样本视为节点 2.遍历所有维度的所有分隔方式，寻找信息增益最大的节点分化成两个分支 B1 和 B2。 3.对 B1,B2 分别执行 2，3 两步，直到每个叶子仅包含一个类别的样本，或所有维度的分割方式用完。 4,如果维度用完，叶子内的样本还不属于同一个类别，则通过多数表决投票来决定叶子的类别

5.1.2 随机森林介绍

随机森林是利用多棵上述的分类树组成的有监督学习分类器[25]。随机森林是目前较为流行的算法，拥有很多优点：1）在小范围数据集上，能够达到更高精度，且不容易过拟合。2）有效处理高维数据，不需要对样本维度预处理，且有很强的抗噪声能力。3）创建随机森林时，使用无偏估计，没有泛化误差。4）训练完成后，能够检测维度的相互影响，得到维度重要性排序。5）训练速度快，实现很简单，容易并行计算。

这五个优点都非常适合于火焰状态识别。目前的火焰状态数据集不过 3000 个样本，属于随机森林适用的小数据集；每个样本维度高，受源数据影响噪声大；无偏估计有利于结合应用于火焰理论研究；火焰维度重要性排序有助于简化火焰实验，节约实验费用；训练速度相对于 BP 神经网络非常快，节约测试时间：因此本章采用随机森林来作火焰状态识别和特征选择。

随机森林的算法流程如下：

表 21 随机森林算法流程

1.假设有 N 个样本，有放回的的选择 N 个样本（即样本可以重复选择），使用 N 个样本训练一棵决策树，作为树根节点的样本。 2.每个样本有 M 个维度，在决策树节点分裂时，随机从 M 个维度中不放回选取 m 个维度进行分裂（m<<M），每个节点选取 m 个维度中一个维度依据信息增益最大化进行分裂。 3.重复步骤 2，直到子节点使用的维度与父节点的属性重复，则该决策树生成，每个样本到达叶子节点完成分类。 4.重复步骤 1-3，建立大量决策树，构成随机森林。

决定每一棵决策树的样本和维度均为随机挑选，这；两个随机性是随机森林与其他监督学习模型的最大差异。

5.2 基于随机森林的状态分类和特征选择实验

决定随机森林表现的一般有两个因素，一是决策树的多少，而是维度采样 m 的大小。在本节中，作实验比较这两个因素对分类正确率的影响，同时得出特征重要性排序。

表 22 随机森林状态分类正确率

采样维度	100 棵树	200 棵树	300 棵树	400 棵树	500 棵树
10	89.08%	92.24%	92.48%	99.04%	99.57%
12	92.34%	95.78%	97.05%	99.73%	100%
14	94.70%	96.09%	99.15%	100%	100%

接表 22

16	88.92%	90.34%	92.89%	94.83%	98.85%
18	88.90%	89.75%	90.11%	97.55%	99.73%

由上表可以看出，随机森林算法可以用于火焰状态分类，分类正确率可以达到 100%。实验结果也与随机森林相关理论符合，即树的棵树越多越好，维度采样值 m 应当约等于 M 的平方根。

表 23 基于随机森林的测量量重要性排序

测量量	当地当量比	一氧化氮	水质量分数	氮气	氢气	氧气
出现次数	842	3201	1945	442	2	451
测量量	一氧化碳	二氧化碳	氢氧根离子	温度	TDNR	
出现次数	123	107	11	690	0	

由上表可以得到 12 个测量量的重要性排序：一氧化氮、水、当地当量比、温度、氧气、一氧化碳、氢气、二氧化碳、TDNR，这与 BP 神经网络的结果基本相同。

表 24 基于随机森林的测量点重要性排序

测量点序号	1	2	3	4	5	6	7	8
出现次数	875	1509	15	34	1337	192	543	99
测量点序号	9	10	11	12	13	14	15	16
出现次数	75	0	83	398	0	2	921	5

由上表可以得出 16 个测量点的重要性排序，明显 2，5，15，1，7，12 号测量点重要性靠前。BP 神经网络得出的结论是 1，2，5 号点重要性最高，但 15 号点重要性也居前但单独提出时就不能主导状态分类。而第四章基于 BP 神经网络的特征选择得出 1、2、5 号三个测量点和一氧化氮质量分数和水质量分数两个测量量在火焰状态中占主导地位。可见，两种方法得出结论互相验证。

5.3 本章小结

本章将随机森林算法应用于火焰状态分类和特征选择上。首先介绍了随机森林算法和其基础决策树算法，解释了为何随机森林算法非常适合于本文的问题解决。其次采样不同决策树棵树和维度采样数的随机森林对火焰状态进行分类，比较了分类效果异同，验证了随机森林算法的有效性。最好，从训练好的随机森林中统计不同测量点和测量量在分裂节点中的出现次数，得到了测量量和测量点的重要性排序，并与 BP 神经网络结果相验证，结论一致。

第六章 全文总结

在大数据的背景下，机器学习成为很多学科领域作数据挖掘的关键工具。射流火焰研究作为超燃冲压发动机的关键技术，具有极高的研究价值，但射流火焰作为最复杂的湍流的一种，具有极高的非线性属性，传统的数据分析手段不能轻易挖掘其数据分布特点并帮助描述射流火焰模型。本毕业设计原计划分析射流火焰的动态纹理，采用字典学习和深度学习网络完成动态纹理的建模和验证。但是由于实验室事故，产生射流火焰的冲压炉被烧毁，导致不能获得一手的射流火焰图像和数据，只能临时变更研究方向，采用公开数据进行研究。本文在射流火焰中引入机器学习手段，使用多种算法对射流火焰进行状态分类和特征选择，为数据挖掘技术大幅度应用于该领域研究提供了一些思路和参考。

6.1 本文的主要工作

本文的主要工作如下：

1.首先对射流火焰 Sandia 工作站的历史数据集 TNF 数据进行分析，介绍了 TNF 数据的来源，了解了 TNF 数据中各个变量的意义，尤其是测量量与测量点两个概念。从 TNF 数据中 4 个射流火焰流场，选取了 11 个测量量和 16 个测量点组成 176 维的样本，每个火焰流场从时序数据中选取了 750 个样本，总共 3000 个样本，这样构成了本文的机器学习数据集。然后对数据集作 PCA 处理，并与未筛选的射流火焰流场数据对比，证明本文数据集线性相关程度低，适于机器学习。

2.使用 KNN、SVM、BP 神经网络、随机森林四种算法对火焰状态进行分类。首先介绍了这些算法的技术原理，给出了各自算法流程，探讨了可能影响分类效果的参数设置。然后在火焰数据集上进行了四种算法的分类试验，在不同训练集/测试集占比的试验中，比较了不同超参数对分类正确率的影响，得出 KNN 算法不适用于火焰状态分类，SIGMOD 核函数的 SVM 与 4 层隐含层、采用 Levenberg-Marquardt 寻优算法的 BP 神经网络可以完美实现火焰状态分类，而随着决策树棵数的增多，随机森林算法也可以很好实现状态分类效果。

3.在使用 SVM 和 BP 神经网络作状态分类过程中，轻易达到极高正确率说明样本的维度信息是高度冗余的，结合射流火焰实验费用高昂亟须简化测量次数的需求，本文提出在火焰数据集上进行特征选择。采用带 dropout 的神经网络对缺省测试集测试，比较缺省不同维度时正确率下降多少，再将重要维度从训练集与测试集中提出再次进行分类验证重要性，得出结论：一氧化氮质量分数和水质量分数在火焰状态测量量中占主导地位，1、2、5 号点在火焰状态测量点中占主导地位。采用随机森林算法，在作火焰状态分类的同时得到了各个测量量和测量点的重要性排序，结果也是一氧化氮和水的重要性居前，1、2、5 号测量点重要性居前这与 BP 神经网络结果一致。

6.2 本文的主要创新点

本文工作的主要创新点为：

1.将基于机器学习的数据挖掘技术应用于射流火焰研究中。通过挖掘射流火焰的历史数据，证明射流火焰数据的线性相关性不高，提出了实用的射流火焰状态分类方法，可以由数次简单的测量结果反馈目前的射流火焰状态，为未来实现超燃冲压发动机燃烧状态闭环控制提供了参考。

2.通过射流火焰状态识别的正确率高低，筛选出样本的射流火焰状态主导维度，原来 176 维的样本使用 6 维即可代替且不影响状态识别。由于射流火焰实验测量的昂贵性，本文研究方法可以大幅降低实验费用，并减轻未来的超燃冲压发动机燃烧闭环控制的反馈传感器负担。

6.3 后续研究工作

本文是基于机器学习的射流火焰数据挖掘，由于机器学习是射流火焰领域内较少采用的技术，本文所获参考资料有限，研究过程中存在着很多不足：

1.限于历史数据有限，本文仅采用了 2 类 4 个火焰流场作状态分类和特征选择。对于更多火焰组分和更广泛流场条件的火焰，不能说明本文的结论同样适用。未来，应当尝试获取更广泛的流场数据，在更大的数据集上验证本文结论。

2.使用 PCA 分析数据的线性相关关系，仅仅说明了各个测量量之间线性相关程度不高。尤其是各个测量点的线性相关与否，并没有找到行之有效的方法论证。未来，应当考虑使用聚类分析和无监督学习的方式来分析测量点之间的关系。

3.本文在作特征选择时，发现第 15 号测量点始终反常。在仅删去一个测量点作火焰分类和随机森林分类并作重要性排序时，15 号测量点的重要性均在前茅。然而在仅保留一个测量点的训练集和数据集上作状态分类，又发现 15 号测量点主导的分类效果不佳。15 号测量点位于火焰 X=60D 截面、R=30mm 处。这一现象是属于数据噪声的偶然现象还是存在更深层次机理，值得探讨。

在未来研究中，首先继续尝试获取火焰的动态纹理，完成本毕业设计。其次，在获取一手射流火焰数据后，在更广泛的数据上验证本文结论，补全以上不足。

参考文献

[1] 郭丽. 基于稀疏表征的降维算法研究[D]. 浙江师范大学, 2013.
[2] 王元卓,靳小龙,程学旗. 网络大数据:现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
[3] 张引,陈敏,廖小飞. 大数据应用的现状与展望[C]. 中国计算机学会大数据学术会, 2013.
[4] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.
[5] 贺武生. 超燃冲压发动机研究综述[J]. 火箭推进, 2005, 31(1): 29-32.
[6] Moshe Matalon. Intrinsic Flame Instabilities in Premixed and Non-premixed Combustion [J]. Annual Reviewa of Fluid Mechanics, 2007, 39(1): 63–91.
[7] Vladimir A. Sabelnikov, Andrei N. Lipatnikov. Recent Advances in Understanding of Thermal Expansion Effects in Premixed Turbulent Flames [J]. Annual Review of Fluid Mechanics, 2017, (49): 91–117.
[8] Jieyu Jiang,Xi Jiang,Min Zhu. A computational study of preferential diffusion and scalar transport in nonpremixed hydrogen-air flames [J]. International journal of hydrogen energy, 40(2015): 15079-15722.
[9] J.E.Freech,K Kumar,H.Huang and C.J.Sung. Laminar Flame Speeds of Preheated Iso-Octane/Air and n-Decane/Air Flames Using Digital Particle Image Velocimetry[C]. 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit.
[10] Ahsan r.Choudhuri and Mahesh Subramanya. Infrared Thermographic Image Processing For Flame Equivalence Ratio Measurements[C]. 1st International Energy Conversion Engineering Conference.
[11] David A. Rosenberg and James F. Driscoll. A Method to Image Flame Index in Partially Premixed Flames[C]. 50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition.
[12] Jacob E. Temme, Patton M. Allison and James F. Driscoll. Low Frequency Combustion Instabilities Imaged in a Gas Turbine Combustor Flame Tube[C]. 50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition.
[13] 孙玉洁.基于火焰自由基和软测量技术的燃烧过程污染物预测[D]. 华北电力大学, 2014.
[14] 吴一全. 基于 Krawtchouk 矩特征的锅炉火焰图像稳定性判断[J]. 自动化学报, 2015, 34(5): 734-740.
[15] 周志华. 机器学习[M]. 清华大学出版社, 2016.
[16] 葛冰. 加湿旋流扩散燃烧火焰的实验研究及大涡模拟[D]. 上海交通大学, 2010.
[17] 王彬. 聚类结构保持的降维技术及实现[D]. 南京师范大学, 2015.
[18] Nello Cristianini, John Shawe-Taylor. 支持向量机导论[M]. 电子工业出版社, 2008.
[19] 陈宝林. 最优化理论与算法(第 2 版)[M]. 清华大学出版社, 2010.
[20] Rumelhart, McClelland. Parallel distribution processing: explorations in the microstructure of cognition[M]. 清华大学出版社, 1986.
[21] Andrew Kirillov. Neural Networks on C# [J/OL]. http://www.codeproject.com/KB/recipes/aforge_neuro.aspx, 2006.
[22] 张鸿燕，狄征. Levenberg-Marquardt 算法的一种新解释[J]. 计算机工程与应用, 2009,

45(19): 5-8.

[23] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[24] Xavier Bouthillier, Kishore Konda, Pascal Vincent, Roland Memisevic. Dropout as data augmentation [J/OL]. http://arxiv.org/abs/1506.08700, 2015.

[25] Vladimir Svetnik, Andy Liaw. Random Forest: A classification and regression tool for compound classification and QSAR modeling[J]. Journal of Chemical Information and Computer Sciaences, 2003, 15(1): 755-780.

谢辞

首先，我要衷心感谢我的导师李元祥老师和陈方老师，两位老师分别从两个领域出发，指导完成了这个交叉学科课题。两位老师的学识人品，兢兢业业的态度，追赶前沿的努力深表敬佩。还有周拥军老师和丁拥科博士，两位老师在我的论文完成中帮助很大，感谢两位老师的教诲。

此外，我还要感谢在我毕业设计完成过程中，帮助我的诸位学长学姐，包括彭希帅博士、徐俊硕士、陆永帅硕士、黄文宣硕士、王彦植硕士、施雨舟硕士、刘村硕士、虞达礼硕士、李子佳硕士、陈辰硕士。以及与我同届的刘嘉玮、张宇轩两位同学，九月研究生入学，再一起奋斗！

最后，感谢我的家人，他们一直在身后默默支持我，是我心灵最温暖的港湾！

再次对所有关怀我的人表示最衷心的感谢！

英文大摘要

With the rapid development of information technology, human society enters the information age. The popularization of various kinds of sensors, input and output devices and storage equipment provides the conditions for the generation and storage of the rapidly growing data, and promotes the generation and development of the large data age. In everyday life we are generating data, just through the Internet, the amount of data produced every day is very large . Therefore, under the background of large data, the data has the characteristics of large scale, fast growth, many expressions and high value. The rapid growth of data has brought great value, but it also brings a lot of problems. In order to solve these problems, many effective algorithms have been proposed, the Data mining field progress rapidly. At the same time, in some areas of discipline, although there is a considerable amount of data, but limited to the professional threshold, not the same as the life information combined with data mining technology, from the vast number of data to get new knowledge to help the development of the theory. This paper tries to provide a new idea for solving one problem.

Sandia was an open international Joint station for turbulent combustion, established by the University of Naples in the 90, and then joined the research institutions of Delft University, Darmstadt University, Stuttgart University, Heidelberg University, Sapporo University and the University of Chicago. Sandia focuses on the study of the turbulent chemical reaction of fuel gases, integrates the experimental data of the above-mentioned institutions, launches an online database to help researchers verify the model and explores turbulent flame combustion, and presents a series of frameworks to compare the results of the measurements with the theoretical results, and to guide the research direction of future experiments and simulations

The workstation is still being updated, focusing on the hydrocarbon reactions of the flames (methane, natural gas and methanol), including the modeling challenge of the process, partial flameout, ignition, separation or elevation of the reaction zone, automatic ignition, flow back zone and whirlpool. Future plans to increase the combustion model of more complex fuels, hoping to build a general combustion model, including non premixed, partially premixed, layered and premixed flames. Future Sandia Research on the choice of flame targets will be based on the needs of organizers and actively contributing scientific communities.

Turbulent non premixed flames (Turbulent Non-Premixed Flames, TNF) data is the core of Sandia, which contains a number of scale and velocity data of the flame turbulent flow field in chemical process, so it is called the experimental data set of the most reasonable and complete, accurate and suitable turbulent flame, which is used by the researchers to verify their combustion models. Datasets include detailed scalar measurements (temperature, main components, and some minor components) and velocity measurements of a jet flame with relatively simple geometrical shapes, relatively homogeneous fuel components, and relatively clear boundary conditions.

The main content of this paper is based on the Sandia data of the classic data of Jet flame. The general consensus in the field of Jet flame experiment is verified by PCA analysis on data. In this paper, we use the KNN SVM, neural network and random forest in Sandia data to identify the four

kinds of classifier successfully, and put forward the feature selection problem in the process of recognition, and filter out some important features in state recognition. The experimental results show that the selected feature can be used to express the state of the flame.

In the background of large data ego, machine learning becomes a key tool for data mining in many disciplines. However, data mining is rarely used in the field of aerodynamics, which is the goal of this article. As the key technology of scramjet, research on the field of Jet Flame has great value, but the Jet flame, as one of the most complex turbulence, has very high non-linear property, traditional data analysis method can't easily excavate its data distribution characteristic and help to describe the jet flame model. In this paper, a method of machine learning is introduced in Jet flame, and a variety of algorithms are used to classify and select the Jet Flame, which provides some ideas and references for the application of data mining technology in this field. The main work of this paper is as follows:

1. Firstly, the TNF data of the jet Flame Sandia Workstation is investigated, the sources of TNF data are introduced, and the significance of each variable in the TNF data is understood, especially the two concepts of measurement and measuring points. From the 4-Jet flame flow field in the TNF data, 11 measurements and 16 measuring points were selected to form 176-dimensional samples, and each flame flow field selected 750 samples from the time series data, which made up the machine learning dataset in this paper. Then the data set is processed by PCA, and compared with the unfiltered flow field data, it is proved that the data set is of low linearity and suitable for machine learning.

2. Using KNN, SVM, neural network, random forest Four kinds of algorithms to classify the flame state. Firstly, the technical principle of these algorithms is introduced, the respective algorithm flow is given, and the parameter setting which may affect the effect of classification is discussed. Then the classification experiments of four kinds of algorithms are carried out on the flame dataset, and in the experiment of different training sets/test sets, the effect of different super parameters on classification accuracy is compared, the KNN algorithm is not suitable for flame state classification, the Sigmod kernel function SVM and 4 layer hidden layer, the neural network using Levenberg optimization algorithm can realize the Flame state classification perfectly, and with the increase of tree number of decision-making, The random forest algorithm can also realize the state classification effect well.

3. In the process of using SVM and neural network for State classification, it is easy to achieve extremely high correct rate indicating that the dimension information of sample is highly redundant, combined with the high cost of Jet flame experiment, it is necessary to simplify the demand of measurement times, this paper presents the feature selection on the flame data set. Using a neural network with Dropout to test the default test set, and comparing the accuracy of the default different dimensions, the importance of classifying and verifying the important dimensions from the training set and the test set is concluded: The quality fraction of nitric oxide and the water quality fraction dominate the flame state measurement, and the 1, 2 and 5th points dominate in the flame state measurement. A random forest algorithm was used to classify the flame state, and the importance of each measurement and measuring point was obtained, and the importance of nitric oxide and water was also measured before the importance of 1, 2 and 5th was in line with the neural network.

The main innovations in this work are:

1. The data mining technology based on machine learning is applied to the research of jet flame. By digging the historical data of Jet flame, it is proved that the linear correlation of jet flame data is

not high, and a practical method of Jet flame State classification is proposed, which can feedback the current Jet flame state by several simple measurement results, and provide a reference for realizing the closed-loop control of scramjet combustion state in the future.

2. Through the correct rate of the Jet flame state recognition, the dominant dimension of the Jet flame state of the sample is screened out, and the original 176-D sample is replaced by 6 dimensions and does not affect state identification. Due to the high cost of the experimental measurement of jet flame, this result can greatly reduce the costs of the experiment and reduce the burden of the feedback sensor on the combustion closed loop control of the scramjet.

Follow-up research work

This paper is based on machine learning jet flame data mining, because machine learning is a less used technology in the field of research, resources are seriously limited for this article, and there are many deficiencies in the research process:

1. Limited to historical data, this paper only uses 2 categories of 4 flame flow field for State classification and feature selection. For more flame components and more extensive flow field conditions, it is not clear that the conclusions of this article are equally applicable. In the future, you should try to obtain more extensive data on the Flow field and verify this conclusion on a larger dataset.

2. Using PCA to analyze the linear correlation of data only shows that the linear correlation between the measurements is not high. In particular, there is no effective method to prove the linear correlation of the measurement points. In the future, consideration should be given to the use of cluster analysis and unsupervised learning to analyze the relationship between measurement points.

3. In this paper, we find that the 15th measurement point is always abnormal when making feature selection. The importance of the No. 15th point is at the forefront when only one measurement point is deleted for flame classification and random forest classification and the importance is sorted. However, in order to classify the training set and data set of only one measuring point, it is found that the classification effect of No. 15th measurement point is poor. The No. 15th Measurement Point is located in the Flame Xu 60d section and R 30mm. This phenomenon is the accidental phenomenon of data noise or the existence of a deeper mechanism, it is worth discussing.