电子鼻信号分析关键算法研究

（申请清华大学工学博士学位论文）

培 养 单 位： 电子工程系

学　　科： 信息与通信工程

研　究　生： 闫　　轲

指 导 教 师： 张 大 鹏　教　授

二〇一六年五月

电子鼻信号分析关键算法研究

闫　轲

Research on Key Signal Analysis Algorithms for Electronic Noses

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Information and Communication Engineering

by

Yan Ke

Dissertation Supervisor： Professor David Zhang

May, 2016

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

电子鼻是一种利用气敏传感器阵列识别简单或复杂气味的设备，应用场景丰富，因此正在受到越来越多的重视。本文在搭建电子鼻系统的基础上，针对电子鼻数据的特点研究信号分析过程中的关键算法，从而提高电子鼻的准确性和实用性。本文主要从三个方面进行研究：基于电子鼻的呼气分析系统设计，针对相关特征的迭代特征删除算法，以及设备差异和时变漂移的补偿算法。

呼气分析是电子鼻的应用之一。通过检测呼出气体中与特定疾病相关的标志物，电子鼻可以辅助疾病筛查和监控。该方法具有无创、易操作等优点。我们针对糖尿病的呼气标志物筛选了气敏传感器阵列，设计并优化了呼气分析系统的硬件和软件，之后采集了包含 500 余个健康人和糖尿病人呼气样本的数据集。实验证明了该系统在糖尿病筛查和血糖预测问题上的有效性。

从电子鼻的信号曲线上可以提取多种特征，对候选特征进行排序和选择有利于提高预测准确率。本文发现，当部分候选特征高度相关时，经典的支持向量机迭代特征删除算法存在相关偏差问题，因此提出了一种改进策略。通过分析特征的排序结果，我们得到了不同气敏传感器和特征提取算法的有效性排名。

漂移是电子鼻实用中的难点问题，也是本文的研究重点。具体来说，同一型号的传感器敏感特性并不完全一致；同时，由于老化等原因，传感器的响应会随着时间变化。这些因素造成了电子鼻的设备差异和时变漂移。本文利用迁移学习的思想解决这两个问题，提出了四种漂移补偿算法。结合标准化误差的模型改进算法考察不同特征对漂移的鲁棒性，然后在已有预测模型中对不稳定的特征施加更高的惩罚权重。基于标定样本的多任务学习算法对每一个数据领域（即设备或时间段）学习一个模型，并针对两种漂移的不同先验分别采用并行和串行两种迁移策略；为了更精确地补偿时变漂移，我们还提出了一种动态模型策略。漂移补偿自编码器算法采用了深度学习框架，在学习更有鉴别能力的特征的同时，显式地建模和补偿特征中的设备差异和复杂时变漂移。上述算法都使用了标定样本，而最大独立领域适配算法能够在标定样本无法获取时利用漂移后的无标签样本进行补偿。该算法通过最大化子空间中样本与背景特征的独立性来学习领域不变子空间。这些算法拓展了多种迁移学习算法框架的适用范围。实验证明，它们能有效地提高对漂移后样本的预测精度。

关键词：电子鼻；特征选择；漂移补偿；迁移学习；呼气分析

Abstract

Electronic noses (e-noses) are devices that use gas sensor arrays to recognize simple or complex odors. They have a wide range of applications, thus are attracting more and more interest from researchers. In this thesis, we started from developing an e-nose system. Then, key algorithms for signal analysis were studied according to the characteristics of e-nose signals. The algorithms can enhance the accuracy and practical applicability of e-noses. Three topics were studied in this thesis: design of an e-nose-based breath analysis system, a recursive feature elimination algorithm for correlated features, and drift compensation algorithms for instrumental variation and time-varying drift.

As an application of e-noses, breath analysis is a non-invasive and convenient approach for disease screening and monitoring, which are achieved by detecting biomarkers in breath related to certain diseases. In this thesis, we selected a gas sensor array and designed the hardware and software of a breath analysis system for the purpose of diabetes screening and blood glucose prediction. A dataset with more than 500 breath samples was collected from healthy subjects and diabetics. Experimental results on the dataset confirmed the efficacy of the system.

Many features can be extracted from the signals of e-noses. Prediction accuracies can be enhanced by ranking them and selecting the best feature subset. We used support vector machine recursive feature elimination (SVM-RFE) to rank e-nose features. However, when some candidate features are highly correlated, SVM-RFE will be influenced by the correlation bias problem. In this thesis, we proposed an improvement to tackle the problem. The importance of different gas sensors and features was assessed by analyzing the feature ranking results.

In practice, drift is a challenging issue for e-noses, which is also the focus of this thesis. Concretely, the sensitivity characteristics are often different for gas sensors of the same model. Moreover, owing to reasons such as the aging effect, the responses of gas sensors often change over time. These problems are known as instrumental variation and time-varying drift. To mitigate them, we proposed four drift compensation algorithms on the basis of transfer learning. The standardization-error-based model improvement (SEMI) algorithm investigates the robustness of different features against

drift, then imposes higher penalty weights to unstable ones. The transfer-sample-based multitask learning (TMTL) algorithm, on the other hand, defines each device or time period as a domain, then learns a prediction model for each domain. Two paradigms, parallel and serial transfer, were designed for TMTL because instrumental variation and time-varying drift correspond to two different types of prior knowledge. A dynamic model strategy was also proposed to refine the compensation of time-varying drift. Using the framework of deep learning, drift compensation autoencoder (DCAE) was developed to learn features with more discriminative power and less drift. It is able to explicitly model and compensate instrumental variation and complex time-varying drift. The algorithms introduced above rely on transfer samples, whereas the maximum independence domain adaptation (MIDA) algorithm can make use of unlabeled drifted samples to compensate for drift when transfer samples are not available. MIDA tries to learn a domain-invariant subspace in which the samples have the maximum independence with their domain features. These algorithms extended the scope of application of various transfer learning frameworks. Experimental results showed that they could improve the prediction accuracy on drifted samples.

Keywords: e-nose; feature selection; drift compensation; transfer learning; breath analysis

目　录

第 1 章　引言 1
1.1　课题背景 1
1.2　电子鼻工作原理简介 1
1.3　研究方向与意义 3
1.4　研究内容 5
1.4.1　基于电子鼻的呼气分析系统设计 5
1.4.2　针对相关特征的迭代特征删除 6
1.4.3　设备差异和时变漂移补偿 6
1.5　论文结构安排 8
第 2 章　基于电子鼻的呼气分析系统 10
2.1　相关工作 10
2.1.1　疾病与呼气标志物 10
2.1.2　基于电子鼻的呼气分析 11
2.2　系统设计与优化 12
2.2.1　气敏传感器阵列 12
2.2.2　电子鼻结构设计 13
2.2.3　采样流程 15
2.2.4　信号分析与特征增强 17
2.3　糖尿病呼气数据集 18
2.4　实验结果和讨论 19
2.4.1　丙酮浓度预测实验 19
2.4.2　糖尿病筛查 19
2.4.3　血糖预测 20
2.4.4　分析：个体训练样本数与血糖预测误差的关系 22
2.5　本章小结 23
第 3 章　针对相关特征的迭代特征删除 24
3.1　相关工作 24
3.1.1　特征选择概述 24
3.1.2　SVM-RFE 25

3.2　改进 SVM-RFE：相关偏差缩减 26
3.2.1　相关偏差 26
3.2.2　相关偏差缩减 28
3.2.3　特征选择的稳定性和集成策略 29
3.3　候选特征集 30
3.4　实验结果和讨论 31
3.4.1　人工合成数据集 31
3.4.2　糖尿病呼气数据集 32
3.4.3　分析：稳定性和集成策略 35
3.4.4　分析：特征选择结果 36
3.5　本章小结 37
第 4 章　基于加权正则化和多任务学习的漂移补偿 38
4.1　相关工作 38
4.1.1　设备差异补偿 38
4.1.2　时变漂移补偿 40
4.1.3　标定样本选择 42
4.2　算法总结与面临的挑战 43
4.3　漂移数据集 44
4.4　基于加权正则化的漂移补偿 48
4.5　基于多任务学习的漂移补偿 49
4.5.1　双任务学习 50
4.5.2　多任务学习 53
4.5.3　动态模型策略 55
4.6　实验结果和讨论 56
4.6.1　阵列漂移数据集 56
4.6.2　Corn 数据集 60
4.6.3　呼气分析数据集 61
4.7　本章小结 62
第 5 章　漂移补偿自编码器 64
5.1　相关工作 65
5.1.1　自编码器 65
5.1.2　基于自编码器的迁移学习 66

5.2 背景特征 66
5.3 漂移补偿自编码器 67
5.3.1 网络结构与目标函数 68
5.3.2 处理复杂时变漂移 70
5.3.3 训练流程 71
5.4 实验结果和讨论 72
5.4.1 阵列漂移数据集 72
5.4.2 Corn 数据集 76
5.4.3 呼气分析数据集 77
5.4.4 分析：欠迁移与过迁移 78
5.4.5 分析：训练流程 79
5.5 本章小结 80
第 6 章 最大独立领域适配 81
6.1 相关工作 81
6.1.1 无监督领域适配 81
6.1.2 Hilbert-Schmidt 独立性准则 82
6.2 最大独立领域适配 83
6.2.1 无监督情况 83
6.2.2 半监督情况 85
6.2.3 与其他算法的联系 85
6.3 特征增强 86
6.4 实验结果和讨论 87
6.4.1 人工合成数据集 87
6.4.2 阵列漂移数据集 91
6.4.3 Corn 数据集 93
6.4.4 呼气分析数据集 94
6.5 本章小结 95
第 7 章 总结和展望 96
7.1 总结 96
7.2 创新点 97
7.3 展望 99
参考文献 101

致　谢 110

声　明 111

个人简历、在学期间发表的学术论文与研究成果 112

第1章　引言

1.1　课题背景

用传感器和信号分析的手段来扩展增强人类的“五感”一直是科技工作者的研究目标。在视觉和听觉领域，传感器相对成熟，分析算法也一直是热门研究课题。相比起来，人们对嗅觉的关注似乎没有那么多。事实上，气味也是一种重要的信息媒介。通过对其进行感知和分析，我们可以了解到被测物体或周边环境中与气体、气味相关的属性，从而做出判断和决策，给生产和生活带来便利。

电子鼻（electronic nose，e-nose），顾名思义，是一种利用电子手段来感知和分析气味的装置。最早对电子鼻的研究可以追溯到 20 世纪 60 年代[1]。近年来，随着各种应用的涌现，电子鼻正受到越来越多的关注[2]。人们最熟悉的应用当属酒驾的检测，也就是测试呼气中酒精的浓度。在工业领域，电子鼻常被用来检测空气中有毒、易燃、易爆气体的含量，以保证生产的安全和工人的健康[3]。空气质量是近几年社会关注的重点话题，电子鼻可以监控室内和室外的空气污染物浓度[4-6]。通过分析气味，电子鼻还可以判断农产品的成熟程度和质量[7,8]以及饮料的种类[9,10]。

医疗健康是电子鼻的另一个重要应用领域[11]。医学研究表明，多种疾病会导致呼出气体中相关标志物的浓度发生变化，例如丙酮与糖尿病和血糖值相关，氨气与肾衰竭相关，氢气与肠道疾病相关等[11-15]。呼气分析就是通过分析呼出气体中的标志物的组成来进行疾病的辅助诊断和监控。该方法具有无创、样本易收集的优点。传统的化学分析设备（如气相色谱仪、质谱仪[2]等）能够精确地确定混合气体中每一组份的浓度，因此可以用来对疾病和标志物的关系进行前期调研。比起气相色谱仪、质谱仪，电子鼻的优势在于易操作、便携、快速、成本低[15]，因此更适合推广到社区医疗中心和病人家庭中辅助疾病筛查和监控。除此之外，它还可以用于评价用户日常的身体状况，例如初创公司 Breathometer 开发了与手机连接的小型电子鼻，可以分析用户呼气中的酒精含量以及是否口臭①。

1.2　电子鼻工作原理简介

Gardner 等人在 1994 年对电子鼻的综述中，认为电子鼻是“由具有部分选择

① https://www.breathometer.com/。

性的气敏传感器阵列和适当的模式识别方法组成的、具有识别单一或复杂气味能力的装置”[1]。气敏传感器是电子鼻的核心部件，其种类很多，包括电化学传感器、半导体传感器、光离子传感器、固体电解质传感器等[16]。其中，金属氧化物半导体（Metal Oxide Semiconductors，MOS）传感器[16]由于其成本低、寿命长、灵敏度高、响应速度快和电路简单等优点，获得了广泛的应用。典型 MOS 传感器的外观如图 1.1 所示。该类传感器中的金属氧化物材料表层在 200-500℃时会与特定敏感气体发生可逆的氧化还原反应，从而改变其电导率[16]。通过调整材料内部掺杂的成分，可以改变传感器的敏感特性，即敏感气体的种类和敏感度。对传感器的电阻进行测量，即可获得与被测敏感气体的种类和浓度相关的信息。

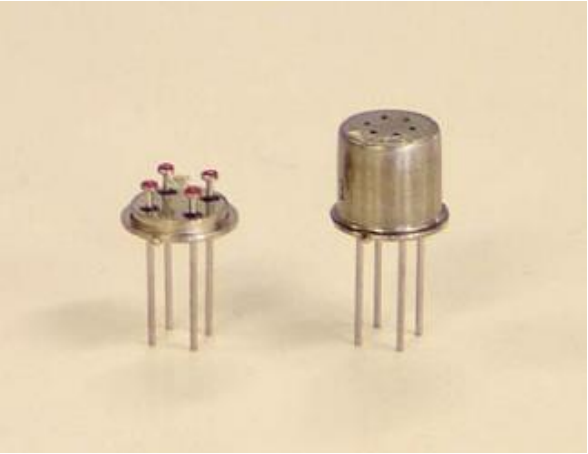


图1.1　MOS传感器TGS2600（日本费加罗公司）的外观和内部示意[17]

大多数 MOS 传感器具有交叉敏感性，对气体的选择性较差[16]。“交叉敏感”和“选择性较差”是指一个传感器会对多种气体敏感。TGS2600 的敏感特性曲线如图 1.2 所示，其横轴为气体浓度，单位是每百万分之一体积（parts per million，ppm）；纵轴是敏感气体存在情况下的传感器电阻 R_S 与纯净空气中电阻 R_O 的比值。可以发现传感器的电阻与气体浓度呈负相关，且传感器对多种气体敏感，但敏感的程度不同，对氢气最为敏感（电阻下降最快）。因此，如果被测气体成分不确定，只用一个传感器往往无法准确判断气体的种类。解决这个问题的一个办法是采用传感器阵列[18]。利用多个敏感特性各不相同的气敏传感器组成阵列，将它们的响应组合成多维特征向量输入模式识别算法，就可以较准确的判断被测气体的种类和浓度。事实上，这种策略正是电子鼻的核心思想之一[1,18]，利用适当选择的多个交叉敏感气敏传感器，既可以提高电子鼻预测的精度，也可以扩大其敏感气体的范围。

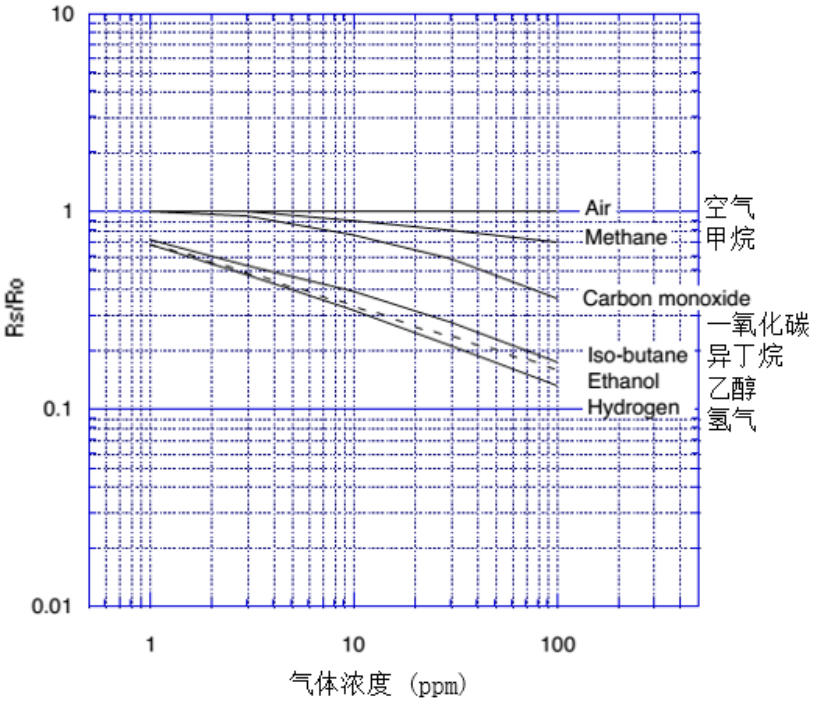


图1.2　MOS传感器TGS2600的敏感特性曲线[17]

分析传感器阵列响应的常用步骤包括信号预处理、特征提取、分类和回归等[19]。首先需要对响应曲线进行预处理，例如基线补偿、幅值归一化[19]等，以去除简单的基线漂移、提高特征区分度。接下来需要在信号曲线上提取有利于分类或回归的特征。研究者们已经发明了多种特征，例如稳态响应、暂态特征（曲线在各个时间点的幅值、导数等）、主成分特征、相位特征、小波特征、曲线拟合特征等[18,19]。在提取多种特征后，一个可选的步骤是利用特征选择算法选择最优的特征子集。最后，如果分析的目的是判断气味的种类或被测物质的某种离散属性值（如疾病/健康），那么可以对提取的特征进行分类，电子鼻文献中常用的分类器包括 K 近邻、神经网络、支持向量机、逻辑回归等；如果分析的目的是判断气味的浓度或被测物质的某种连续属性值（如血糖值），那么可以采用回归算法，电子鼻文献中常用的算法包括最小二乘拟合、偏最小二乘回归、神经网络、支持向量回归等。

1.3　研究方向与意义

基于电子鼻的疾病筛查与监控具有无创、易操作等优点，近年来正逐渐受到重视。目前已有的商用和实验用电子鼻在传感器阵列的针对性和设备的实用性上还有不足。因此，我们将首先在电子鼻系统搭建方面进行研究，设计一套呼气分析系统。该系统可以用于辅助疾病筛查和监控（例如糖尿病筛查和血糖预测），具

有较高的实用价值。另外，电子鼻信号的研究尚不如图像、语音等领域那样成熟，公共数据集较少。我们将利用该系统采集呼气数据集，为后续的算法研究提供测试平台。

从气敏传感器的响应曲线上可以提取多种特征。用于预测时，这些特征的有效性各不相同。通过对特征进行选择，可以在降低预测模型的时空复杂度的同时提高其准确性。更重要的是，它可以帮助我们更好地理解数据。气敏传感器信号曲线上提取的特征具有较高相关性。我们将研究适用于高度相关特征的特征选择算法，并用它来对呼气数据集进行特征选择。通过分析特征选择的结果，我们可以得到关于系统软硬件的有用结论，为后人研究提供参考。

在实用中，信号漂移是电子鼻面临的一个重大挑战。气敏传感器大多存在鲁棒性问题[18,19]，其敏感特性会受到多种干扰因素的影响，信号的可重复性较差。具体来说，同一个型号、甚至同一个批次的传感器，其敏感特性并不会完全相同，因此导致了电子鼻的设备差异（instrumental variation）。随着时间的推移，传感器会发生老化，采样环境、甚至样本本身的构成也可能会随时间变化，这些因素共同导致了电子鼻的时变漂移（time-varying drift）。其他干扰因素还包括环境的温湿度等。漂移问题已经成为了制约电子鼻推广和实用化的一大障碍[19]，是电子鼻的“短板”，因此也正在被更多研究者关注。在谷歌学术搜索关键词“e-nose drift”，可以发现相关文献数目逐年增多，如图 1.3 所示。本文将对漂移补偿算法进行重点研究，利用迁移学习[20]的思想，针对不同的应用场合提出多种漂移补偿算法，缓解电子鼻实用中的难点问题：设备差异和时变漂移。

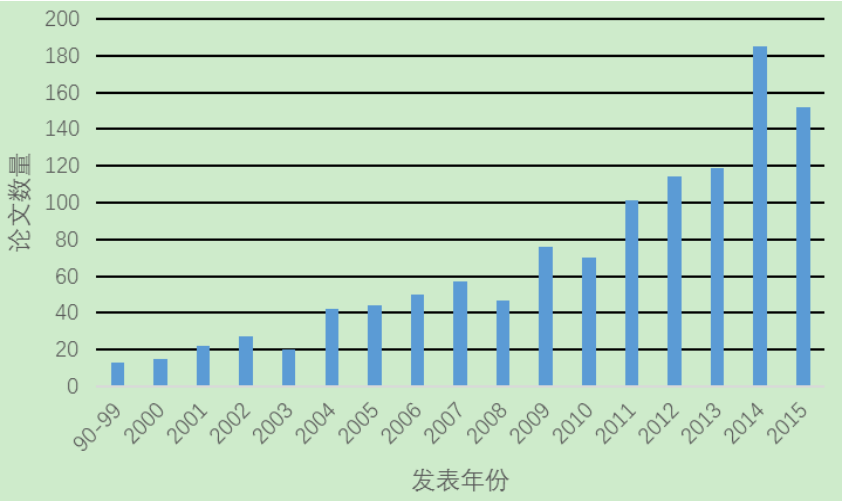


图1.3　在谷歌学术搜索“e-nose drift”得到的每年发表文献数

总体来说，本文从系统搭建出发，之后针对电子鼻的数据特点（特征高度相关和存在漂移）研究了适合的特征选择和漂移补偿算法，提高了电子鼻的准确性

和实用性。

1.4　研究内容

本文的主要围绕三个方面进行研究：基于电子鼻的呼气分析系统设计和样本采集，针对相关特征的迭代特征删除算法，以及设备差异和时变漂移的补偿。这三个方面不是孤立的，它们都是电子鼻系统的关键组成部分，如图 1.4 所示。系统设计和样本采集为后续的算法研究提供了测试平台；特征选择可以提高系统精度，同时帮助我们更好地理解数据；而漂移补偿则是提高系统鲁棒性和实用性的重要步骤。本文提出的特征选择和漂移补偿算法可以用于本文系统及其他电子鼻系统。

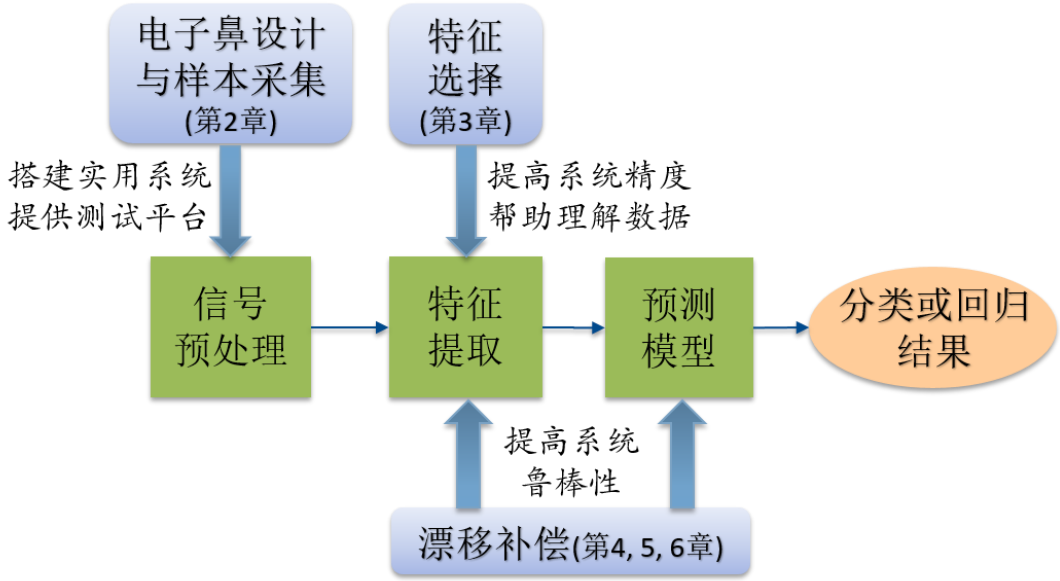


图1.4　电子鼻应用中的关键步骤，其中标出章节的部分为本文研究内容

1.4.1　基于电子鼻的呼气分析系统设计

根据现有医学文献中的呼气标志物研究结论，我们选择了针对糖尿病呼气标志物的气敏传感器阵列，并且对电子鼻中气室等结构进行了优化，设计并制作了一套呼气分析系统。利用这套系统，我们到医院中采集了 500 余个健康人和糖尿病人的呼气样本，进行了糖尿病筛查和血糖预测实验，取得了初步的成果。这些工作验证了呼气分析系统的有效性。我们还提出了一种特征增强策略，用以改善血糖预测中存在的个体差异问题。

1.4.2　针对相关特征的迭代特征删除

为了获取最优的特征组合，我们在上面提到的糖尿病呼气数据集上提取了 7 种瞬态特征，利用生物信息学领域应用广泛的支持向量机迭代特征删除算法（Support Vector Machine Recursive Feature Elimination，SVM-RFE）进行了特征排序和选择实验。然而，由于电子鼻信号的特点，候选特征之间存在较大相关性。这种情况下，SVM-RFE 的特征评价准则将会受到影响，相关的特征会得到过低的评价。我们研究了这种现象的原因，同时提出了一种改进方案——相关偏差缩减（Correlation Bias Reduction，CBR）。实验表明，CBR 能够提高传统 SVM-RFE 选择得到的特征的分类准确率。最后，通过分析特征的排名结果，我们得到了关于传感器、特征提取算法和温度调制电压[18]的有效性排名结论。

1.4.3　设备差异和时变漂移补偿

在影响电子鼻鲁棒性的因素中，我们主要研究影响较大也较复杂的两种因素：设备差异和时变漂移。事实上，除了电子鼻，一些其他的传感器或测量设备也面临相似的鲁棒性问题，例如光谱仪[21]。由于后者的信号是与气敏传感器类似的一维曲线，因此电子鼻和光谱学领域的漂移补偿算法往往可以通用。

从机器学习和迁移学习[20]的角度来看，设备差异和时变漂移本质上都是样本的边缘分布 $P(X)$ 和条件分布 $P(Y|X)$ 在训练样本和测试样本中存在差异，导致从训练样本得到的预测模型在测试样本上准确率下降。借用迁移学习的术语，我们可以假设无漂移的样本来自源领域 $\mathcal{D}_S$（source domain），有漂移的样本来自目标领域 $\mathcal{D}_T$（target domain），两者的边缘分布或/和条件分布不同。源领域中的有标签样本通常比较充足，而目标领域中的有标签样本较少甚至没有。因此，单独为目标领域训练一个预测模型是较困难的，需要将源领域的知识迁移到目标领域。对于设备差异来说，初始用来采集训练样本的设备对应着源领域，而每一个后来制作的新设备都对应着一个目标领域，不同领域间的分布变化是离散的；对于时变漂移来说，初始用来采集训练样本的时间段对应着源领域，而之后的时间段对应着目标领域，但领域之间的分布变化是连续的。在本文中，为了简洁起见，有时把设备差异和时变漂移统称为“漂移”，因为它们都对应了特征空间中样本分布的移动。但在一些文献中[19]，只有时变漂移补偿被称作漂移补偿（drift compensation），而设备差异补偿对应的术语是标定迁移（calibration transfer）。

对于漂移，根本的解决办法之一是制造鲁棒性更好的传感器。然而在成本受限以及现有工艺无法取得突破的条件下，利用信号分析算法进行漂移补偿就是比较合适的选择了。很多漂移补偿算法用到了标定样本[18,19,21]。所谓标定样本就是在

每个领域（不同的设备或时间段）分别采集的一组辅助样本，一般是预先选定的几种标准气体。由领域之间一一对应的标定样本对可以推断出分布移动的相关信息，从而提高补偿精度。然而采集标定样本需要一定的专业知识和工作量，不太适用于非专业的电子鼻使用者。也有一些方法不需要标定样本，而是利用目标领域的有标签或无标签样本来得到漂移的信息并进行补偿。已有的漂移补偿算法在准确度和辅助样本的需求数量上还不能让人满意，尤其是在漂移比较复杂的时候。同时，已有算法大多只能处理设备差异和时变漂移之中的一种，而在现实中，如果电子鼻被批量制造和长期使用，这两种漂移会同时存在。

为了解决以上问题，本文提出了四种漂移补偿算法。其中前三种利用了标定样本，最后一种方法基于目标领域的无标签样本进行补偿。下面分别简要介绍：

- 结合标准化误差的模型改进（Standardization-Error-based Model Improvement，SEMI）。特征标准化是一种简单的特征层面的补偿方法，该方法依据标定样本将不同领域的对应特征进行匹配，但匹配后的特征仍然可能存在差异。SEMI 在已有的预测模型中加入一个加权的 Tikhonov 正则项，对匹配较差的特征施以较大的惩罚权重，从而使预测模型更加依赖那些匹配较好的特征。这种方法将特征标准化与预测模型训练有机结合，提高模型对漂移的鲁棒性。
- 基于标定样本的多任务学习（Transfer-sample-based MultiTask Learning，TMTL）。该方法直接在模型层面进行迁移，对每一个领域（设备或时间段）学习一个模型。该方法的基本假设是：不同设备的模型应当是相似的；同一个设备在时间段 t 的模型应当和时间段 t-1 的模型相似，也就是时变漂移应当具有平缓性。另外，不同领域的标定样本在各自的模型投影下应当对齐。另外，本文还提出了一种动态模型策略用来生成任意时间点的模型。TMTL 中应用了上面介绍的 SEMI 策略，以进一步提高漂移补偿精度。
- 漂移补偿自编码器（Drift Compensation AutoEncoder，DCAE）。机器学习算法中，为了获得较好的准确率，提高特征的鉴别能力十分重要。近年来，基于深度网络的特征学习算法被成功地应用在图像、语音等信号上，然而在电子鼻信号上的应用还比较少。本文对自编码器算法进行改进，在学习更有鉴别能力的特征的同时，显式地建模和补偿特征中的设备差异和时变漂移。为了达到这个目的，首先定义了“背景特征”来描述一个样本的采集设备和时间等信息，接着在自编码器的编码和解码的阶段分别显式加入和减除一组补偿项，从而保证网络隐层中学习到的特征都是去除了漂移的。为了更好地补偿非线性时变漂移，我们还设计了非线性补偿层。DCAE 的目标函数使其可以同时最小化样本的重构误差、有标签样本的预测误差和标定样本的对齐误差。

- 最大独立领域适配（Maximum Independence Domain Adaptation，MIDA）。该方法的目标是在标定样本无法获取的情况下，利用目标领域的无标签样本进行漂移补偿。MIDA 通过最大化子空间中样本与背景特征的独立性来学习领域不变子空间，同时保留数据的方差和标签信息等有用信息。我们还提出了一种特征增强策略，以使算法可以对不同领域的样本差异化投影。该方法的优点是使用灵活，求解简单。

我们利用本文研究的呼气分析系统实地采集了同时包含设备差异和时变漂移的呼气数据集，在该数据集和另外两个公开的漂移数据集上验证了上述算法的有效性。上述算法利用迁移学习的思想解决漂移问题，将多任务学习、基于自编码器的特征学习、领域不变子空间学习等算法框架应用到设备差异和时变漂移问题中，并针对存在标定样本、领域分布连续变化等情况做出了改进，拓展了前人算法的适用范围。

1.5　论文结构安排

本文章节组织安排如下：

第 1 章为引言，总体介绍电子鼻的应用场合和工作原理，引出本文研究内容和意义。

第 2 章介绍基于电子鼻的呼气分析系统的设计，首先调研一些典型疾病与呼气标志物的关系以及现有呼气分析系统；之后介绍本文系统的传感器阵列、硬件设计和采样流程，以及糖尿病筛查和血糖预测实验的算法、数据集和结果。

第 3 章研究针对相关特征的迭代特征删除算法，指出已有的 SVM-RFE 算法存在的相关偏差问题，研究其改进方案，并且进行实验和讨论。

第 4 章至第 6 章致力于解决电子鼻的设备差异和时变漂移问题。

第 4 章讨论基于多任务学习和加权正则项的漂移补偿，首先总体调研现有的漂移补偿算法，并对它们进行分析和总结；之后描述本文采用的漂移数据集；接着提出两个模型层面的算法：结合标准化误差的模型改进、基于标定样本的多任务学习；最后进行实验和对比。

第 5 章提出漂移补偿自编码器，首先回顾已有的基于自编码器的迁移学习算法；接着提出背景特征，描述本文算法的网络结构和目标函数；最后进行实验和讨论。

第 6 章探讨最大独立领域适配算法，首先调研典型的特征层面的无监督领域适配方法；接着提出本文算法的优化目标和特征增强策略；最后在人工合成数据

集和漂移数据集上验证算法的有效性。

第 7 章总结全文，列举了几种漂移补偿算法的异同点和全文的创新点，并且展望了未来可能的研究方向。

第2章　基于电子鼻的呼气分析系统

电子鼻相关的研究中，研究者通常针对要分析的气体选择适合的气敏传感器组成阵列，然后制作成电子鼻。在本章中，我们重点关注利用呼气分析辅助糖尿病筛查和血糖预测，为此调研了相关文献，之后进行了硬件和软件系统的设计与优化、样本采集，以及分类与回归实验。

2.1　相关工作

2.1.1　疾病与呼气标志物

人体呼出气体中，绝大部分成分是氮气、氧气、二氧化碳、水蒸气和惰性气体[22]。剩余的不到 100 ppm（parts per million，每百万分之一体积）的部分是超过 500 种成分组成的混合物，其中浓度在 ppm 级别的主要有丙酮、一氧化碳、甲烷、氢气等[22]。已经有大量文献对特定疾病与气体的对应关系进行了研究[12,15,22-36]。表 2.1 总结了部分疾病或医学指标与呼气标志物对应关系。

表 2.1　部分疾病或医学指标与呼气标志物对应关系总结

疾病或指标	呼气标志物
糖尿病，血糖，糖化血红蛋白	丙酮，乙醇，一氧化碳，硝酸甲酯等[22-30]
肾衰竭，血尿素氮	氨气[31,32]
心脏病	戊烷[33]
肺癌	丁烷等多种有机化合物[34]
乳腺癌	壬烷等多种有机化合物[35]
功能性肠病	氢气[36]

以糖尿病为例，其呼气中升高的丙酮浓度是由新陈代谢产生的[28]。糖尿病人由于胰岛素分泌不足或自身无法有效利用胰岛素，葡萄糖难以进入身体细胞，导致血液中葡萄糖含量升高。同时由于身体细胞缺少能量来源，肝脏增加脂肪分解，产生酮体。血液经过肺时，酮体中的丙酮随呼气挥发。较严重的糖尿病人呼气中的丙酮甚至可以当面闻到（类似烂苹果的气味）。20 世纪 60 年代以来，中外科学家开始对人体气味（以呼吸为主）中气体标志物的浓度和糖尿病之间的联系进行定量分析。Deng 等人[23]分析了 15 位 II 型糖尿病患者的呼气丙酮浓度，发现其范围在 1.76 到 3.73 ppm。与此相比，参加测试的正常人的呼气丙酮浓度

只有 0.22 到 0.80 ppm，明显小于糖尿病人。Wang 等人[27]统计了 12 位 I 型糖尿病人的呼气丙酮浓度，发现其均值为 2.19 ppm，而正常人为 0.48 ppm。该文同时测量了病人的血糖水平和糖化血红蛋白水平（缩写 HbA1c，一种反映长期血糖水平的指标），与呼气丙酮浓度进行比较后，发现呼气丙酮浓度和血糖以及 HbA1c 之间呈现近似的正相关关系。Ueta 等人[25]的实验指出，呼气丙酮浓度和 HbA1c 之间的相关系数为 0.776。与血糖相关的呼吸标志物中，丙酮是浓度较高、容易分析的一种，浓度较低的还包括乙醇、硝酸甲酯等[28]。利用多种气体浓度作为自变量进行血糖的拟合可以获得更好的结果[26]。值得一提的是，尽管丙酮等气体已被证实与糖尿病、血糖和 HbA1c 相关，但也有一些干扰因素会影响这种相关性，例如患者自身的饮食结构、年龄等[29]。不同个体的相关规律也可能有所不同[24]。

糖尿病是一种慢性疾病，发病率高且会带来多种严重并发症，及早诊断和血糖监测对病人来说非常重要。诊断和监测常用的手段是血糖测试，其结果较准确，但操作较复杂，同时也是一种有创检测，会带来疼痛。不少患者因此忽视血糖检查，错过了最佳的治疗时间[28]。因此，社会迫切需要一种无创和操作简便的糖尿病筛查和血糖预测方法。呼气分析就是其中较有潜力的方案之一。

2.1.2　基于电子鼻的呼气分析

国内外多个实验室已经对基于电子鼻的呼气分析进行了探索。1997 年，浙江大学的 Wang 等人[37]利用对丙酮敏感的半导体气体传感器开发了一个电子鼻，测量了 18 个糖尿病人和 14 个健康人的呼气样本。结果显示该系统能够以 100%的正确率区分病人和健康人。2001 年，台湾工业技术研究院的 Lin 等人[38]开发了以 6 个石英晶体气敏传感器为核心的电子鼻。他们测试了 30 个正常人、61 个慢性肾功能不全/肾衰竭患者以及 83 个尿毒症患者的呼气，并用判别分析算法得到了总体 86.78%的正确率。Dragonieri 等人[39]在 2007 年使用商业化电子鼻 Cyranose 320 尝试对哮喘病人和健康人进行分类，分类算法是主成分分析+线性判别分析+Mahalanobis 距离分类。他们发现年轻的健康人和年轻的轻度哮喘患者很容易分开（正确率 100%），年老的健康人和年老的重度哮喘患者稍难些（90%），轻度和重度哮喘患者较难分开（65%），年轻和年老的健康人则完全无法区分（50%）。

Yu 等人[40]在 2005 年使用 4 个导电聚合物气体传感器成功分类 3 个健康人和 3 个糖尿病人。香港理工大学的 Guo 等人[14]在 2010 年研发了基于 12 个商业化半导体气敏传感器的呼气分析系统。该系统的传感器阵列能够检测多种疾病的呼气标志物。他们测试了健康人、糖尿病人、肾病病人和呼吸道炎症病人各约 100

人的呼气。PCA+*K* 近邻分类的结果表明，该系统区分这三种疾病和健康人呼气的正确率各约为 87%，85%和 73%. 另外，该系统还能以 81%的正确率区分透析治疗前后的肾衰竭病人。Guo 等人在 2012 年进一步用该系统和支持向量序贯回归算法预测糖尿病人的血糖水平[41]。他们将 192 个样本的呼气数据集按血糖水平分为四个级别，最终预测平均正确率为 68.66%. Saraoğlu 等人[42]尝试利用电子鼻来预测 30 个糖尿病人的血糖值和 HbA1c 值，其平均相对正确率分别是 74.76%和 83.03%. 上述文献表明利用电子鼻进行特定疾病的分析是可行的，并且也提出了一些行之有效的传感器阵列构造和信号分析算法，为我们构建自己的系统提供了参考。然而上述文献中并没有选择最优的电子鼻特征，也没有针对漂移问题做出补偿，因此在系统精度和实用性上尚有改进空间。我们将在接下来的第 3 至 6 章研究这些问题。

2.2　系统设计与优化

2.2.1　气敏传感器阵列

传感器阵列是电子鼻的核心部件，直接决定了气体检测的范围和精度。我们主要针对糖尿病筛查和血糖预测筛选传感器，兼顾其他疾病的呼气标志物检测。由于商业化气敏传感器发展较成熟、成本较低，我们收集了日本 Figaro、FIS，韩国 Ogam，中国炜盛，美国 Baseline-MOCON，英国 city 等公司的传感器产品资料，依据 2.1.1 节介绍的呼气标志物文献调研结果进行了传感器筛选和实验。我们首先选择了 16 个传感器组成阵列[43]，它们的厂家、敏感范围和量程各不相同，包含半导体（MOS）、电化学、光离子化[16]等种类。之后利用这个阵列制作了电子鼻，采集了 167 个健康人和 151 个糖尿病呼气样本进行糖尿病筛查实验。我们测试了每一种可能的传感器组合的分类正确率，综合考虑了传感器的稳定性、气体覆盖范围等因素后，最终筛选出了由 10 个气敏传感器组成的阵列，如表 2.2 所示。

表 2.2 中，VOCs 代表包括丙酮在内的有机挥发物（volatile organic compounds）。10 个传感器中，1 号传感器为固体电解质二氧化碳传感器，可以用来指示实际呼气的量[44]。其他均为 MOS 传感器。糖尿病主要的呼气标志物是丙酮。该阵列中 TGS826、GSBT11、TGS2602 等传感器对丙酮较为敏感，而其他传感器的敏感性稍差，这种互补性有利于在后续的信号分析中利用模式识别算法进行预测。虽然有些传感器的量程标称值较高（如 TGS2610-D00），但在实际实验中，我们发现它们同样能对较低浓度的气体产生响应。事实上很多传感器的最低检测限都比标称值要低[45]。由于 MOS 传感器的交叉敏感性，阵列中的传感

器都或多或少都对多种 VOC 敏感。这使得该阵列可以检测多种呼气标志物，因此也可以用来分析与糖尿病有关的其他标志物和一些其他的疾病。

表 2.2　本章呼气分析系统采用的气敏传感器阵列

传感器编号	型号	主要敏感气体	量程（ppm）	生产厂家
1	TGS4161	CO_2	350-10000	日本 Figaro Inc.
2	TGS822	VOCs, H_2, CO	50-5000	日本 Figaro Inc.
3	TGS826	丙酮, NH_3	30-300	日本 Figaro Inc.
4	TGS2610-D00	H_2, VOCs	500-10000	日本 Figaro Inc.
5	SP3S-AQ2	VOCs, H_2, CO	1-1000	日本 FIS Inc.
6	GSBT11	VOCs, 丙酮	1-1000	韩国 Ogam Inc.
7	WSP2111	VOCs, H_2	5-40	中国炜盛
8	TGS2600-TM	H_2, VOCs, CO	1-100	日本 Figaro Inc.
9	TGS2602-TM	VOCs, 丙酮, NH_3, H_2S	1-30	日本 Figaro Inc.
10	WSP2111-TM	VOCs, H_2	5-40	中国炜盛

表 2.2 中传感器 8、9、10 型号中的“-TM”表示温度调制（temperature modulation）。MOS 传感器的敏感特性与温度密切相关，不同温度下的敏感特性不同。一般的 MOS 传感器工作在固定的温度下，而如果使它们工作在一个周期性变化的温度下，不同气体形成的响应曲线形状就会有差别，从而带来更丰富的信息，提高传感器的选择性[18,46]。在尝试多种候选传感器和温度调制波形的基础上，我们选择了阵列中 TGS2600、TGS2602 和 WSP2111 三个传感器，给它们施以阶梯型的加热电压。其他传感器仍采用传统恒压加热方式。具体加热电压将在 2.2.3 节介绍。根据传感器组合穷举实验，加入温度调制技术后的传感器对分类正确率的贡献明显高于以恒温方式工作的传感器[43]。第 3 章的特征选择实验也证明了这一点。据我们所知，这是温度调制技术首次被应用在呼气分析系统中。

2.2.2　电子鼻结构设计

图 2.1 是本章电子鼻的结构框图，其中主要包含气泵、气室、传感器阵列、信号处理电路、数据采集卡、上位机等模块。微型真空泵将呼气抽入到气室，气室中的嵌入的传感器阵列随即产生响应。信号处理电路中利用分压电路将电阻信号转换为电压信号，再进行 RC 低通滤波，过滤掉高频噪声。对于 TGS4161，我们还采用运算放大电路对信号进行了放大。处理后信号随后传送到数据采集卡，由后者将模拟信号数字化，并经由 USB 端口传送到上位机进行储存。另一方面，

上位机端的自动控制程序给数据采集卡发送控制信号，控制气泵的开闭和传感器的温度调制。由于 MOS 传感器的工作温度较高，我们还加入了一个风扇用来散热。电子鼻实物的内部结构见图 2.2。

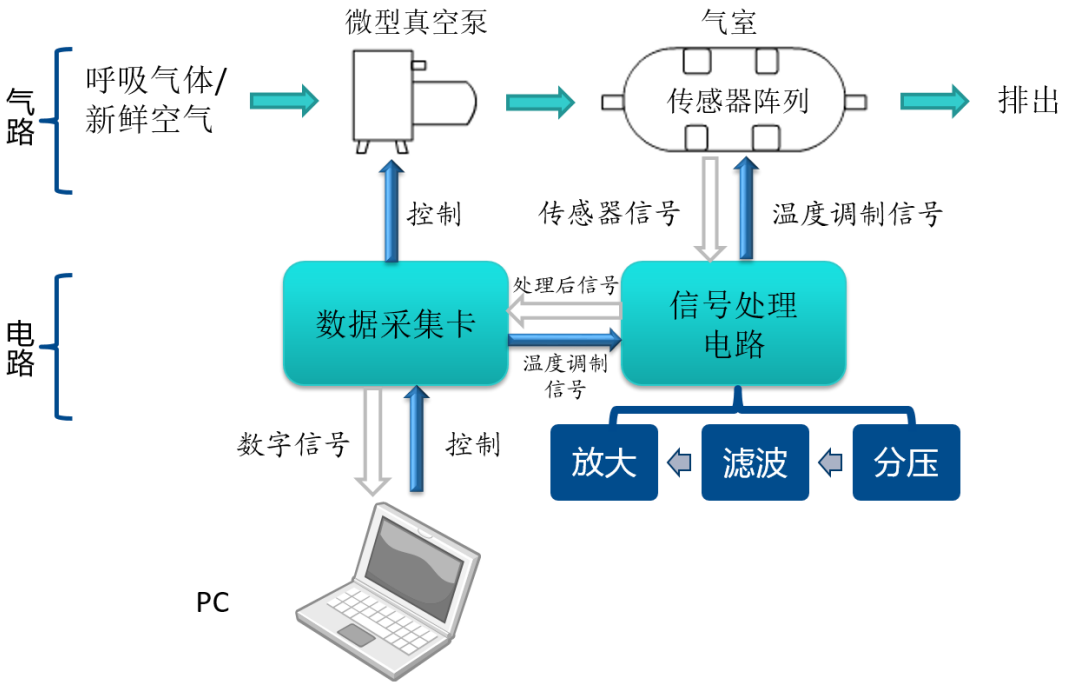


图2.1　电子鼻结构框图

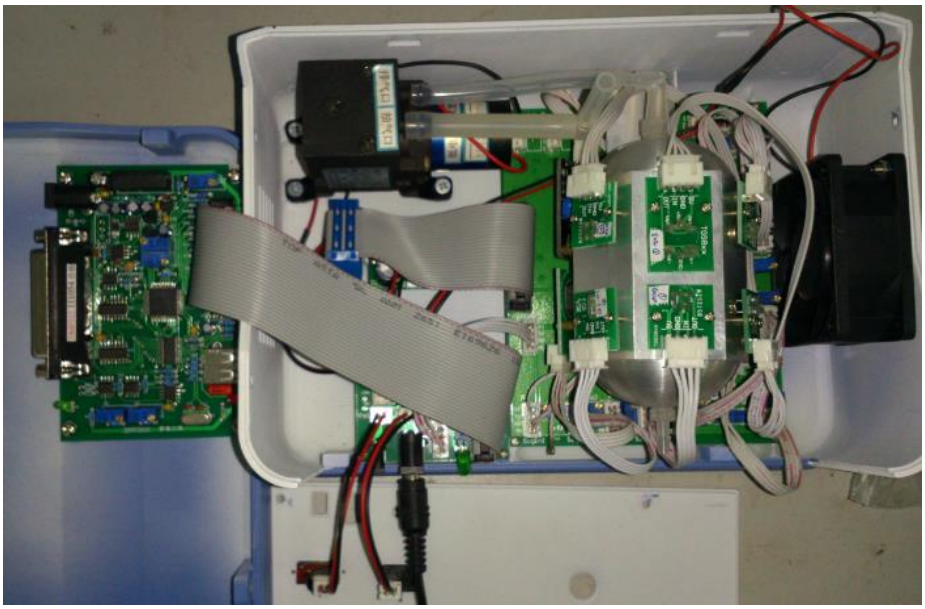


图2.2　电子鼻内部结构快照

气室是气体分子与传感器接触的场所，气室的形状、大小还有传感器分布的位置对于传感器的响应有较大影响，因此需要对其进行优化。我们设计了如图 2.3 所示的气室，其外壁为六边形，内壁为圆柱形。每个传感器镶嵌在气室壁的

一个孔洞上，方便进行单独更换。气室前后各有一个半球形的盖子，两个盖子上面分别开有进气口和出气口。气室内壁平滑，有利于气体的流动，避免洗气死角和气体残留引起的记忆效应[19]。传感器对称均匀分布，接触到的气体浓度均匀，有利于精确测量。同时气室还能保持较小的体积，提高仪器的便携性。

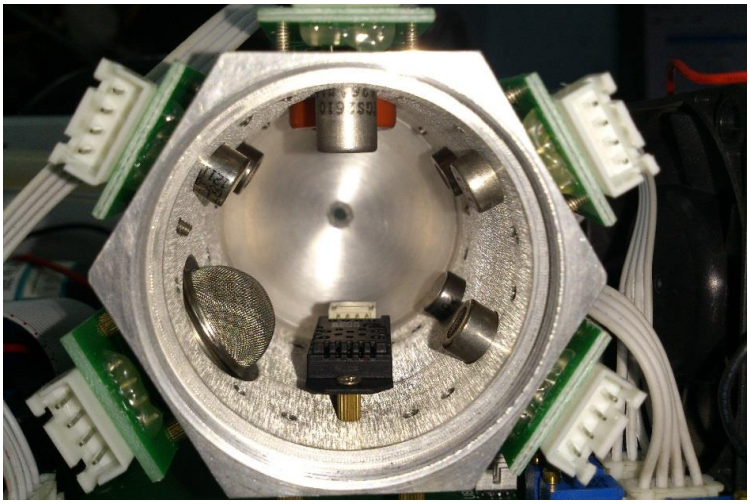


图2.3　气室内部结构快照

2.2.3　采样流程

为了便于大规模采样，我们参考前人的方案[14]制定了一套标准化、便捷的操作流程。使用者首先需要使用一个一次性的吹嘴吹满一个 600mL 的 Tedlar 气袋，如图 2.4 所示。吹嘴与气袋中间接有装着棉垫的吸湿盒，用来过滤掉一部分水蒸气。随后，将气袋插到设备的进气口，点击上位机数据采集软件的开始按钮，采样软件就会自动完成采样流程，将呼气转换为数字化的传感器响应，如图 2.5。

图2.4　呼气收集方法

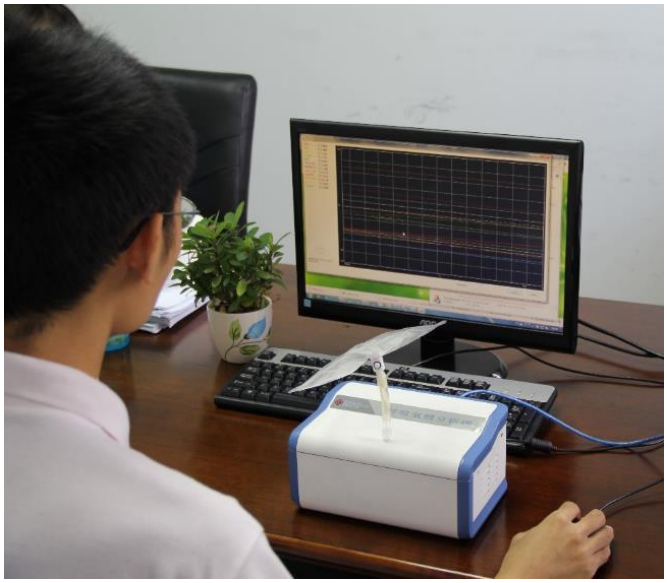


图2.5　自动采样中的呼气分析系统图示

呼气分析系统对一个呼气样本的采样流程包含 4 个阶段：

1) 基线阶段（0-1 s），真空泵关闭，各传感器的基线值被记录；

2) 进气阶段（1-8 s），真空泵开启，气袋中的呼出气体被抽取到气室中与传感器阵列接触，气敏传感器产生响应；

3) 反应阶段（8-64 s），真空泵关闭，呼出气体在气室中与传感器阵列充分接触，各传感器响应趋近稳态值；

4) 洗气阶段（64-144 s），真空泵开启，抽取外界洁净空气清洗气室，将气室中的呼出气体排出。各传感器响应逐渐恢复到基线值。

表 2.3 列出了呼气分析系统的基本参数。图 2.6 展示了普通和温度调制传感器的响应示例。前者被施以 5 V 恒定加热电压，而后者则是由 3-7 V 周期变化的加热电压控制的。可以明显看出后者的曲线随着加热电压的变化而呈阶梯型，细节较多，因此能够包含更多关于气体的信息。

表 2.3　呼气分析系统基本参数

参数名称	参数值
尺寸	22*15*11 cm
气敏传感器	10 个
检测用时	144 s
采样频率	8 Hz
进气速率	50 mL/s
气室容积	100 mL

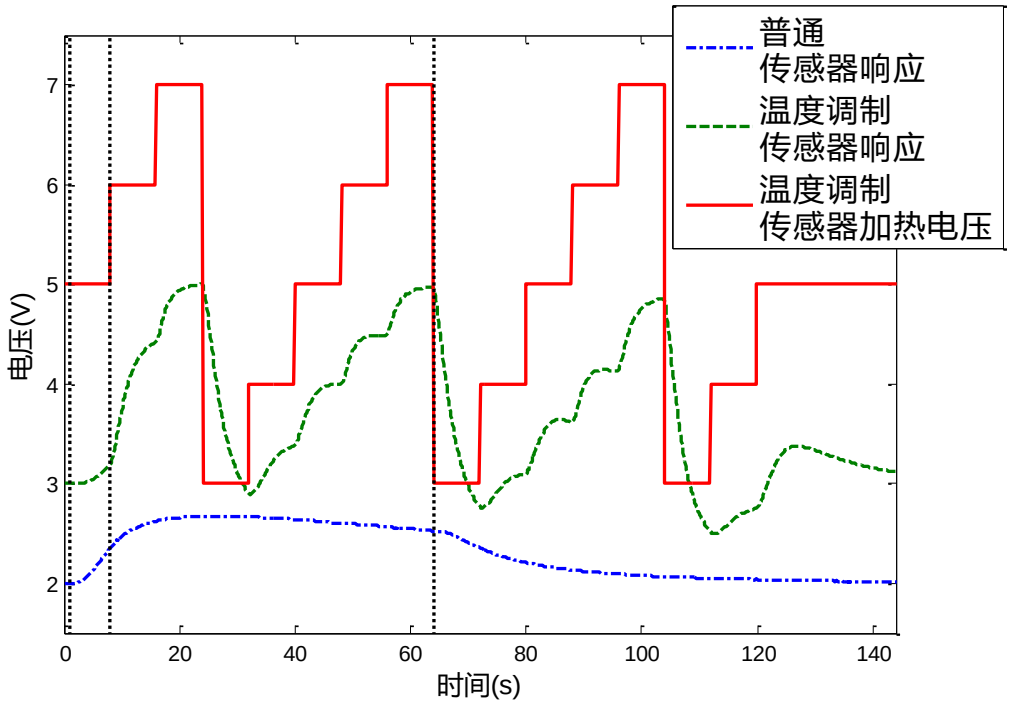


图2.6　温度调制传感器加热电压与传感器响应示例。竖直虚线分割了采样的四个阶段：基线阶段、进气阶段、反应阶段、洗气阶段

2.2.4　信号分析与特征增强

每个呼气样本都包含 10 条气敏传感器的响应曲线。在本章的糖尿病筛查和血糖预测实验中，信号分析包含预处理、特征提取和预测三步，特征选择将在下一章考察。本章数据集中漂移较小，因此暂不考虑漂移补偿。预处理时，对每条曲线进行基线校正，即整条曲线减去该曲线基线阶段的平均值，目的是去除基线漂移。接着，将 10 条曲线串接成一个特征向量，一共包含 8 Hz×144 s×10 = 11520 个数据点。由于特征维度较高，我们采用主成分分析（PCA）进行降维。降维后的特征被归一化为零均值、单位方差。最后，我们采用支持向量机（SVM）对糖尿病和健康人进行分类，采用支持向量回归（SVR）对血糖值进行回归。出于系统验证的目的，我们采用的是较常用的特征和预测算法。

根据文献[24]，血糖水平与呼气标志物的关系存在个体差异。因此，如果对所有病人都训练统一的回归模型，将会损失精度。而如果对每一个病人都训练一个回归模型，又会有很多病人的训练样本数不足。为此，我们提出了一种特征增强策略，用来在回归模型中部分补偿个体差异。假设在数据集中有 n 个病人，则在降维后的特征向量中增加 n 维特征。对于一个样本，如果它来自第 i 个病人，则增加的特征中第 i 维为 1，其余为 0。当使用线性回归算法时，这种做法相当于在模型中增加了一个与个体相关的偏置，从而补偿了个体差异中的常数差异部分。该策略应用的前提是，要预测的样本来自多个个体，不同个体的模型有所差

异，但每个个体的训练样本数又不够训练单独的模型。

根据文献调研，目前国内外利用呼气标志物进行疾病监控（如血糖预测）的研究中，对于个体差异问题的考虑较少。前人或者是直接对所有个体的样本训练一个统一的模型，如文献[26,41,42]；或者是对每个个体单独训练一个模型，如文献[24]。由于我们的数据集中很多个体只有一个呼气样本，因此无法采用对每个个体单独训练模型的方法。上述特征增强策略相当于在训练统一模型的基础上的一个改进，我们将在 2.4.3 节的实验部分对比加入该策略前后的血糖预测精度。

2.3　糖尿病呼气数据集

我们与广东省中医院合作，利用上面介绍的电子鼻实地采集了健康人和糖尿病人的呼气样本，样本的详细情况如表 2.4 所示。所有样本都由医院提供了诊断标签。279 个糖尿病呼气样本是由 87 个糖尿病病人提供的，每个人提供了 1 至 11 个不等的样本。这些病人的具体情况见表 2.5。我们将在后文研究血糖预测准确率与个人训练样本数之间的关系。在采集糖尿病人呼气样本的同时，我们还记录了病人的实时血糖，以便后续对血糖预测进行实验。这些样本的血糖分布如图 2.7 所示。

表 2.4　糖尿病呼气数据集

样本类型	采样地点	样本数量
健康人	广东省中医院体检中心	295
糖尿病人	广东省中医院内分泌病房	279

表 2.5　糖尿病病人统计

项目	值
人数	87
性别	男 39 人，女 48 人
年龄	39 - 91
糖尿病类别	I 型 1 人，II 型 86 人
病程（年）	0.5 - 19
血糖值（mmol/L）	4.4 - 23.1

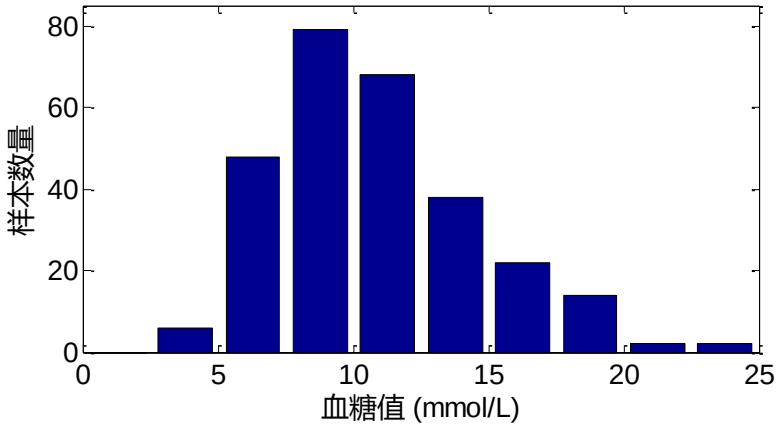


图2.7　279个糖尿病样本的血糖值分布

2.4　实验结果和讨论

2.4.1　丙酮浓度预测实验

为了检验系统的有效性，我们首先测试其能否预测糖尿病的主要呼气标志物——丙酮气体的浓度。有文献[24]指出，I 型糖尿病患者的呼气丙酮浓度可高达 21 ppm，因此我们准备了 8 种浓度的丙酮样本（0.1, 0.2, 0.5, 1, 2, 5, 10, 20 ppm），每种浓度两个样本。利用上文提到的采样流程得到响应曲线，然后用基线补偿+PCA+SVR 算法进行留一法（leave-one-out）回归实验。最后，用平均绝对误差（Mean Absolute Error，MAE）来评价回归的精度，其定义为

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \hat{x}_i\right| \tag{2-1}$$

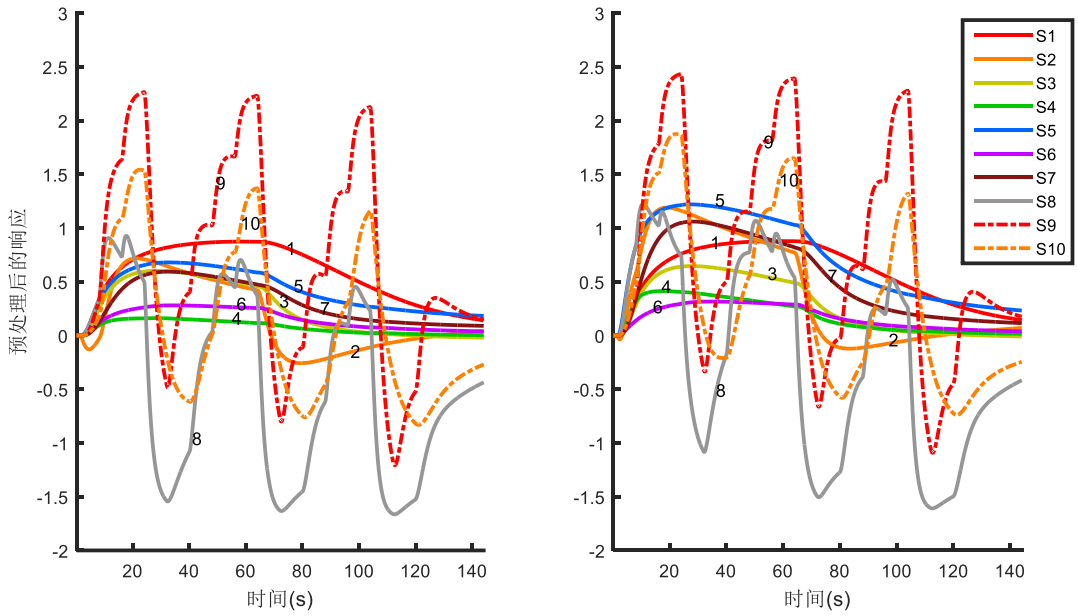
其中 n 是样本数，x 是真实值，$\hat{x}$ 是预测值。这个实验得到的 MAE 是 0.16 ppm。

实际呼气中存在多种挥发性有机物，会干扰丙酮浓度的预测，因此有必要测试本系统在存在干扰气体的情况下的效果。为了模拟真实呼气样本中存在的干扰，我们从 5 个健康志愿者处获得了 40 个呼气样本，之后向其中添加丙酮蒸汽，使得最终混合样本中加入的丙酮浓度分别为 0, 0.2, 0.3, 0.7, 1.7, 3.3, 5.0, 和 6.7 ppm。随后用基线补偿+PCA+SVR 算法进行留一法回归实验，并且使用了 2.2.4 节提到的特征增强策略。最终得到的 MAE 为 0.22 ppm。这个实验的结果证明，本系统能够在存在干扰气体的情况下较准确地预测丙酮的浓度。

2.4.2　糖尿病筛查

图 2.8 展示了健康人和糖尿病样本的平均响应曲线。可以发现，后者的多条传感器曲线幅值高于前者。本系统中较高的响应曲线幅值对应较高的气体浓度，因此图 2.8 验证了相关文献的结论，即糖尿病人呼气中丙酮含量高于健康人。

在信号分析过程中，我们设置 PCA 降维后的能量保留比例为 99.99%；SVM 采用高斯核函数。随机抽取两类样本各 140 个用来训练，两类样本各 139 个用来测试，重复 50 次，计算平均正确率。在医学研究中，常用敏感性和特异性两个指标来评价诊断方法的好坏。敏感性是被正确判断为患病的人在所有病人中的比例，特异性是被正确判断为健康的人在所有健康人中的比例。最终计算得到的敏感性和特异性分别为 91.51%和 90.77%，证明了本系统在糖尿病筛查实验中的有效性。



(a) 健康　　(b) 糖尿病

图2.8　两类样本的平均响应。每条曲线代表一个气敏传感器，具体型号见表2.2

2.4.3　血糖预测

文献调研表明，呼气中丙酮浓度和血糖值有一定的相关性。Wang 等人[27]将呼气样本按照血糖值高低分为几个组后，观察到组内平均丙酮浓度随血糖水平的升高而升高。我们进行了类似的实验，结果如图 2.9 所示。可以看出，大部分传感器的平均值随着血糖水平的上升而上升，验证了文献中的发现。S6 和 S9 是两个例外的传感器，可能是由于它们对呼气中其他干扰气体敏感度更高。

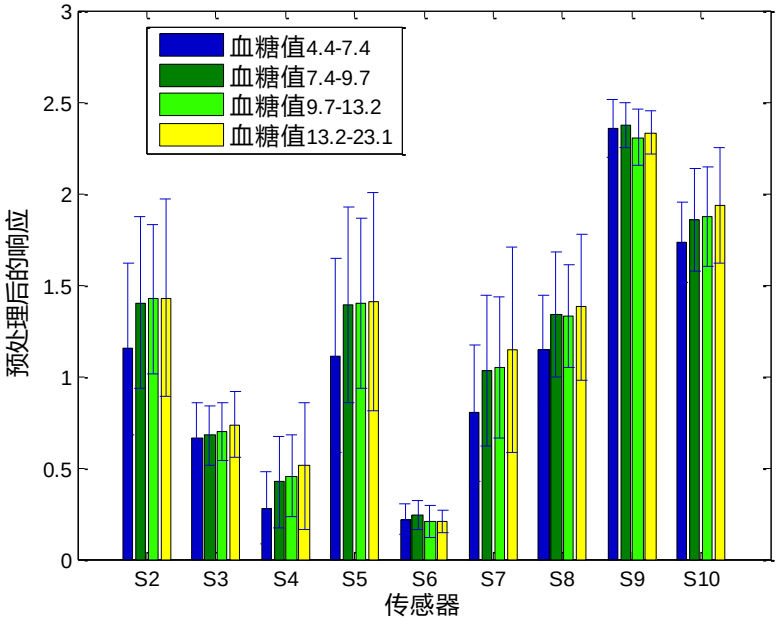


图2.9　不同血糖级别的糖尿病样本中，各传感器稳态响应的平均值和方差

从图 2.9 还可以看出，每组样本传感器响应的方差也较大，说明利用呼气进行血糖预测具有一定挑战性。我们设置 PCA 降维后保留的能量比例为 99.1%；为了防止过拟合，SVR 采用线性核函数。用三种指标来评价预测的准确度：MAE、平均绝对误差率（Mean Absolute Error Ratio，MAER）和相关系数 r，后两者的计算公式是：

$$\mathrm{MAER}=\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right|\right)\times 100\% \tag{2-2}$$

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\hat{x}_i-\bar{\hat{x}})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(\hat{x}_i-\bar{\hat{x}})^2}} \tag{2-3}$$

其中 $\bar{x}$ 是真实值均值，$\bar{\hat{x}}$ 是预测值均值。最终计算得到的血糖预测准确度为 MAE = 2.09 mmol/L，MAER = 20.58%，r = 0.658。可见本系统对于血糖预测具有一定的有效性。在 2013 年的一篇文献[42]中，作者对 30 个呼气样本进行预测，得到 MRAE = 25.24%。相比之下，我们采用的样本规模更大，而误差更小。预测的散点图见图 2.10，可以看出预测值与真实值的相关性。然而，预测仍有一定误差，这可能是由血糖与呼气标志物关系的个体内差异[29]、个体间差异[24]以及呼气中干扰气体等因素造成的。未来可以考虑将部分个体内差异（饮食规律等）和个体间差异（年龄、疾病类型等）因素作为变量加入到预测模型中，并且采集更多样化的呼气样本来学习相关规律。

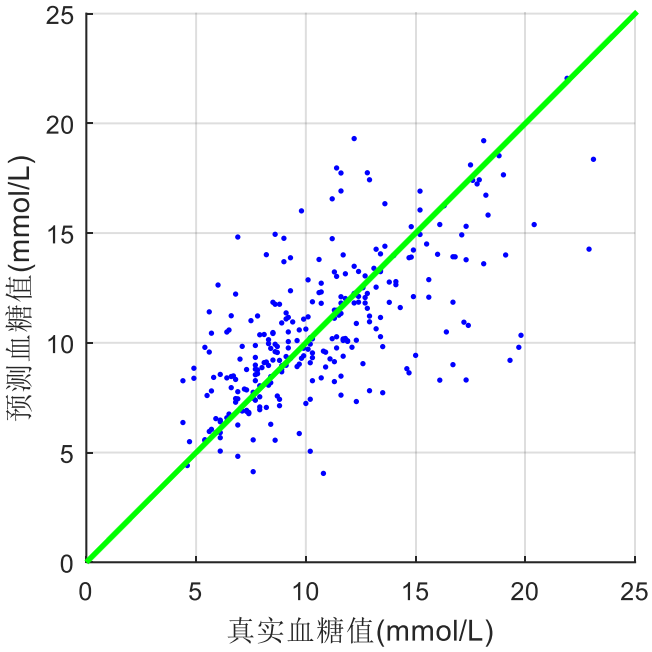


图2.10　血糖预测散点图，每个点代表一个糖尿病呼气样本

上述血糖预测结果采用了 2.4 节介绍的特征增强策略。当不使用特征增强时，结果为 MAE = 2.82 mmol/L，MAER = 28.16%，r = 0.310，可以发现精度（尤其是相关系数）有明显下降。应用特征增强策略使得 MAE 降低了 25.9%，MAER 降低了 7.58 个百分点，这是因为该策略缩小了个体差异带来的影响，提高了预测精度。

2.4.4　分析：个体训练样本数与血糖预测误差的关系

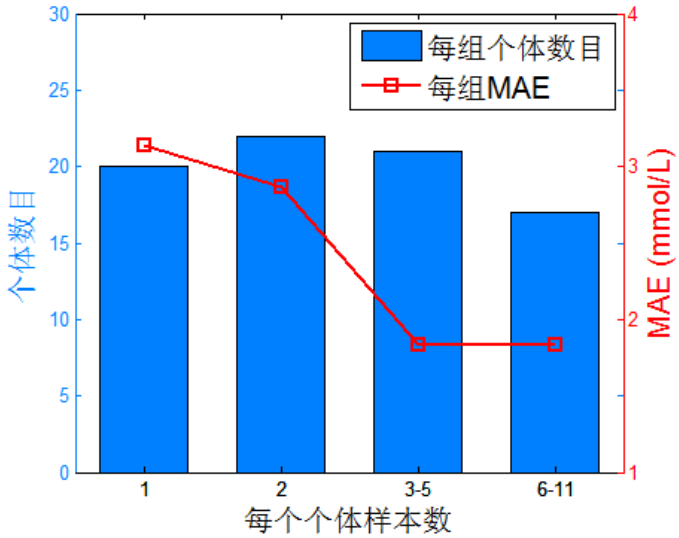


图2.11　每个个体的训练样本数与血糖预测误差的关系。将所有个体按照训练样本数分为4组。柱状图是每一组的个体数，曲线是该组的MAE

在血糖预测实验中，我们使用了所有个体（病人）的呼气样本来对每个个体训练模型。直观上讲，在全部训练样本中，某个个体的训练样本数越多，对这个个体的预测应当越精确。为了验证这一点，我们对预测结果进行了统计，结果如图 2.11 所示。可以发现训练样本数大于 2 的个体的 MAE 明显较小于只有 1-2 个训练样本的个体。这提示我们，为了提高血糖预测的精度，可以对每个个体采集更多的训练样本。

2.5　本章小结

本章设计了用于糖尿病筛查和血糖预测的呼气分析系统。根据糖尿病相关的呼气标志物，我们选择了气敏传感器阵列，设计了非侵入式、易于操作、便携的电子鼻。为补偿个体间差异，提出了一种特征增强策略。最后，在丙酮气体样本以及超过 500 个真实呼气样本上的实验证明了系统的有效性。

第3章　针对相关特征的迭代特征删除

在上一章中，我们提取了传感器信号的主成分作为特征。事实上，气敏传感器的信号曲线上包含了较丰富的信息，可以提取多种特征。通过在候选特征集中选取最优的特征组合，可以进一步提高准确率。同时，分析各维特征的重要性可以帮助我们更好地理解数据。由于气敏传感器的交叉敏感性和其信号的低频特性，某些候选特征之间的相关性可能较大。这种情况下，需要些已有的特征选择算法做出一些改进。

3.1　相关工作

3.1.1　特征选择概述

特征选择是模式识别领域常用的技术。有效的特征选择可以去除与标签不相关的、噪声大的、冗余的特征，从而减轻过拟合风险，提高预测模型的性能，同时缩小其时空复杂度。更重要的是，通过对特征的重要性进行分析，我们能对数据本身有更深入的了解[47-49]。

特征选择算法通常可以分为过滤式（filters），封装式（wrappers）和嵌入式（embedded）三种[49]，分类的依据是它们与后续预测模型的关系。过滤式算法对每一个或一组特征用一个预定义的准则进行评价，然后选择得分最高的一组特征。评价准则例如相关准则、Fisher 准则、互信息准则等，一般与预测算法无关。封装式算法将后续预测模型当作一个黑盒，通过特定的搜索策略将不同的特征组合输入到这个黑盒中，然后直接使用预测性能作为特征评价的准则。这类算法包括前/后向选择，遗传算法、模拟退火等。嵌入式方法将预测模型训练和特征选择有机结合在一起，可以大体分为两类，其中一类的预测模型本身就能选择出一部分重要的特征，例如决策树、基于 L1 范数正则项的方法等等；另一类从预测模型训练的中间值中估计特征的重要性，例如支持向量机迭代特征删除（Support Vector Machine Recursive Feature Elimination，SVM-RFE）[47-49]。

过滤式算法不需要使用预测模型来评价特征，因此一般速度较快，但选择的精度较低。单变量的过滤式算法无法考虑特征间的依赖性。封装式算法能够考虑特征依赖，精度较高，但速度较慢，尤其是在特征维度较高、样本数较大的情况下。嵌入式方法结合了前两类方法的优点。由于与预测模型结合，因此精度较高；由于不需要训练集上进行交叉验证，因此比封装式算法更快，能够更充分地利用数据。但同封装式的算法一样，特征选择结果较依赖预测模型及其参数的选择

[47-49]。

3.1.2　SVM-RFE

SVM-RFE 是一种在生物信息学中常用的经典的特征选择算法[50-52]。作为一种嵌入式算法，它比封装式算法更适用于候选特征维度较高的情况。我们希望能在一个较大的电子鼻候选特征集上进行选择，因此使用该算法进行实验。SVM-RFE 基于 SVM 分类器中的系数对特征进行评分，然后用迭代特征删除策略给特征排序，排名最高的若干个特征即可作为选择出的特征子集。根据 SVM 算法核函数的不同，SVM-RFE 可分为线性和非线性两个版本。

线性 SVM-RFE：直观上，SVM 的目标是寻找一个最大间隔分类面。在样本线性可分的情况下，间隔的定义是距离分类面最近的样本与分类面之间的距离的两倍[53]。给定一组训练样本$\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in \mathbf{R}^m, y_i \in \{-1,1\}$，$n$ 为样本数，m 为特征维数，设线性 SVM 的决策函数为

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{3-1}$$

可证明[53]在一定约束下，间隔$M = 2/\|\mathbf{w}\|$。因此最大化间隔等价于在一定约束下最小化$\|\mathbf{w}\|^2$。这里我们直接给出 SVM 的目标函数的对偶（dual）拉格朗日形式：

$$L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{3-2}$$

其中α_i是拉格朗日乘子，通过在$\alpha_i \ge 0$和$\sum_{i=1}^n \alpha_i y_i = 0$的约束下最大化上式即可求得。详细推导见文献[53]。对应着非零α的那些训练样本通常称为支持向量[53]。权重向量$\mathbf{w}$可由下式求得：

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \tag{3-3}$$

线性 SVM-RFE 按照$\mathbf{w}$赋予每个特征一个得分，第 k 个特征的评分就是$\mathbf{w}$的第 k 个分量的平方，即

$$J(k) = w_k^2 \tag{3-4}$$

得分越低说明该特征对于分类的影响越小，因此越不重要。

非线性 SVM-RFE：当候选特征维度比样本数大得多时，为了防止过拟合，线性的 SVM-RFE 更为合适。然而在一些样本较充足的场合，应用非线性 SVM-RFE 可能会得到更加精确的结果。非线性 SVM 将特征映射到一个高维空间：$\mathbf{x} \in \mathbf{R}^m \mapsto \Phi(\mathbf{x}) \in \mathbf{R}^h$，期望样本在其中线性可分。式（3-2）因此变为

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \tag{3-5}$$

目标函数中只存在样本在高维空间的内积，因此并不需要知道Φ的具体形式，只需要定义一个核函数$K(\mathbf{x}_i, \mathbf{x}_j)$，也就是所谓的核技巧。在本章中我们采用一种常用的非线性核函数：高斯 RBF 核$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2\right)$。

由于Φ的形式未知，权重向量$\mathbf{w}$无法像线性 SVM-RFE 那样直接求出。为了得到每个特征的得分，需要采用一种稍复杂的策略[50,51]。如果去掉一个特征导致式（3-5）产生较小的改变，则说明该特征的贡献较小。因此可以用式（3-5）在去除特征 k 之前和之后的改变值来给特征评分，即

$$J(k) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i^{(-k)}) \cdot \Phi(\mathbf{x}_j^{(-k)}) \tag{3-6}$$

其中“(-k)”表示将特征 k 删除，即$\mathbf{x}^{(-k)} \in \mathbf{R}^{m-1}$。为了减小计算量，可以假设去掉特征前后$\alpha$不变，因此无需重新求解 SVM 模型[50]。这个准则适用于所有核函数。当使用线性核函数$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$时，可以证明这个准则与式（3-4）等价。

迭代特征删除：由于特征间依赖的存在，如果去掉排在末尾的若干特征，前面的特征排序很可能会发生变化。因此直接用每个特征的得分作为特征的最后排序并非最佳方案。Guyon 等人[50]因此提出了迭代特征删除策略。首先用所有特征训练一个 SVM 模型，计算每个特征的得分，得分最小的特征被删除；然后用剩余的特征重新训练 SVM 模型；如此循环直到所有特征都被删除[50]。之后，用删除的先后顺序对特征进行排序，越晚删除的特征越重要。

当特征维度很高时，每次删除一个特征耗时过长，因此可以每次删除多个特征（例如剩余特征数的一半）。然而，这种策略会对排序精度产生影响，并且导致相关偏差问题。我们将在下一节进行详述。

3.2　改进SVM-RFE：相关偏差缩减

3.2.1　相关偏差

在一些应用中，特征存在较大相关性。例如，基因分类中，具有相似功能或位置的基因对应的特征具有高度相关性[54]。由于气敏传感器的交叉敏感性和其信号的低频特性，电子鼻信号中提取的部分特征的相关性也较大。对于一些特征选择算法，这种相关性会带来负面影响。Toloşi 等人[54]考察了基于 Lasso 逻辑回归和随机森林的两种特征选择算法后发现，相关性较大的若干个特征由于在分类模型中的起着相近的作用，其在特征选择算法中的评分会被分摊，因此单个特征

的重要性会被低估，即使它们都是比较重要的特征。这种效应被称为“相关偏差”（correlation bias），相关的特征数越多，偏差越明显。

现有文献还没有研究过相关偏差对 SVM-RFE 的影响。事实上，SVM-RFE 本质上是从 SVM 的模型系数去估计特征得分，所以同样会受到相关偏差的影响。可以通过一个模拟实验来证实这一点。我们构造了一个人工合成数据集，其中包含 500 个样本和 100 个特征，每个样本都对应一个类标签 y。y 是 22 个从正态分布 $\mathcal{N}(0,1)$ 中采样的隐变量 $z_1,\ldots,z_{22}$ 的函数：

$$y=\operatorname{sign}\left(\sum_{i=1}^{22} w_i z_i+\epsilon\right) \tag{3-7}$$

其中 ϵ 是一个从 $\mathcal{N}(0,0.01)$ 中采样得到的噪声项，$w_1,\ldots,w_{22}$ 是一组预定义的权重。将前 80 维特征分为 22 组，每组特征由同一个隐变量生成，即

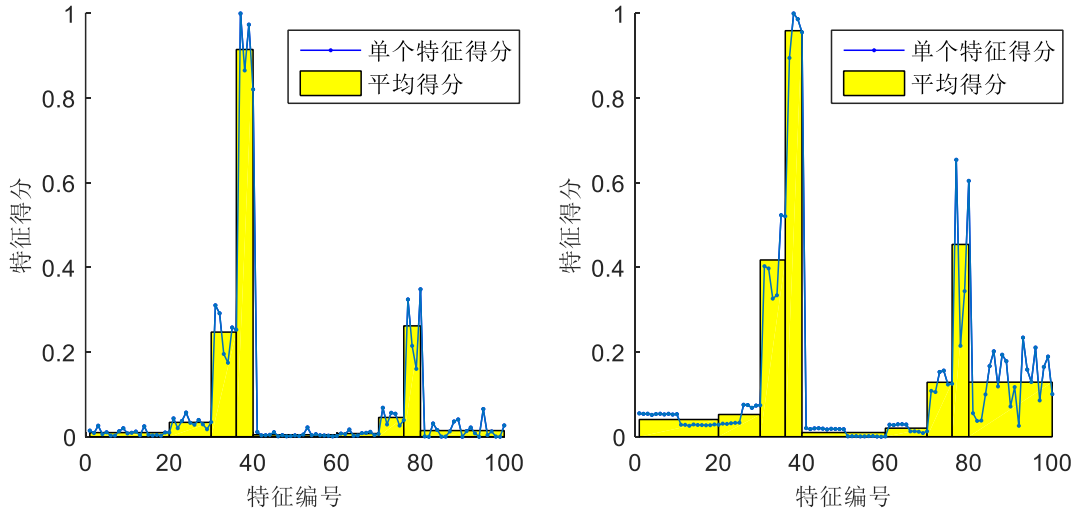
$$x_{ij}=z_i+\epsilon_{ij},\ j=1,\ldots,k_i,\ i=1,\ldots,22 \tag{3-8}$$

x_{ij} 是第 i 组的第 j 个特征，k_i 是第 i 个组的特征数，$\epsilon_{ij}\sim\mathcal{N}(0,0.01)$。因此，每个组内的特征是高度相关的。将这 22 个组的特征串接得到前 80 维特征，再加上从 $\mathcal{N}(0,1)$ 中采样的 20 个无关特征，得到数据集中的 42 组共 100 个特征。每组特征对应的特征数 k 和隐变量权重 w 列于表 3.1。

表 3.1　合成数据集中的隐变量权重 w 和每组特征数 k

变量	组号								
	1-2	**3-4**	**5-7**	**8-11**	**12-13**	**14-15**	**16-18**	**19-22**	**23-42**
特征数 k	10	5	2	1	10	5	2	1	1
权重 w	1	1	1	1	0.5	0.5	0.5	0.5	0

用上一节介绍的 SVM-RFE 的特征评分准则对这 100 个特征进行评分（不用 RFE 排序），结果展示在图 3.1。对于一个可靠的特征评分准则来说，得分应当只与隐变量权重 w 有关，即特征 1-40, 41-80, 81-100 的得分应当分别为 1, 0.5, 0。然而在图 3.1 中，k 和 w 都会影响特征的得分。由于相关偏差的影响，特征数 k 较大的组内的特征得分更小。在这种情况下进行 RFE 时，如果设定一次删除多个特征，那么 k 较大的组内的特征很有可能会被错误地全部删除。



(a) 线性SVM-RFE　　(b) 非线性SVM-RFE

图3.1　线性和非线性SVM-RFE对特征的一次评分（归一化到[0, 1]）。曲线代表每个特征的得分，柱状图代表相同的k和w对应的特征的平均得分

3.2.2　相关偏差缩减

相关偏差现象不只会影响特征选择，还会影响回归算法，使回归模型方差增大，精度下降[55]。为了解决这个问题，Park 等人[55]提出对特征进行层次聚类，然后使用每类中心作为新的特征进行回归。Sharma 等人[56]提出用一种惩罚策略来自动对相关特征进行分组和选择，然而这种方法只适用于线性模型。

我们提出了一种不同于前人的方法。在一次 RFE 迭代中，先将高度相关的特征分在一组。当需要删除多个特征时，如果我们发现一个组内的特征被完全删除，可以将其中的一个代表特征放回被保留的特征列表中。在下一次特征评分时，由于和它高度相关的特征已被删除，该特征将会得到更公正的评分。这个代表特征可以是上次评分中组内的分最高的特征。这种策略被称为相关偏差缩减（Correlation Bias Reduction，CBR）。它能够监督并修正潜在的错误删除，而不用改变特征集合或评价准则。

在一次 RFE 迭代中，我们定义$\mathcal{F}^{\text{out}}$为将要被删除的特征列表，$\mathcal{F}^{\text{in}}$为被保留的特征列表。CBR 的目的是将潜在的被错误删除的特征由$\mathcal{F}^{\text{out}}$移回到$\mathcal{F}^{\text{in}}$。为了检测高度相关的特征组，定义两个阈值 T_c 和 T_g。CBR 的具体步骤如下：按照得分从高到低检查$\mathcal{F}^{\text{out}}$中的每一个特征。对一个特征 k，如果$\mathcal{F}^{\text{in}} \cup \mathcal{F}^{\text{out}}$中存在**超过** T_g 个特征（包括 k）与 k 的 Pearson 相关系数的绝对值大于 T_c，则这些特征被当做一个组。如果组内的特征全部在$\mathcal{F}^{\text{out}}$，说明这个组可能受到了相关偏差的影响被错误地全部删除了，因此可以将 k 移回$\mathcal{F}^{\text{in}}$。重复这个步骤直到$\mathcal{F}^{\text{out}}$中每个特征都被检查过。算法总结见算法 1，其中$|\mathcal{F}|$表示$\mathcal{F}$内的元素数，$\mathcal{F}-\{k\}$表示

从 $\mathcal{F}$ 中去掉 k。|corr(a,b)|表示特征 a 和 b 的相关系数的绝对值。

算法 1：相关偏差缩减 CBR

输入：特征列表 $\mathcal{F}^{\text{in}}$ 和 $\mathcal{F}^{\text{out}}$，阈值 T_c 和 T_g。

输出：修正后的 $\mathcal{F}^{\text{in}}$ 和 $\mathcal{F}^{\text{out}}$。

1. 按照得分由高到低对 $\mathcal{F}^{\text{out}}$ 中的特征进行排序。
2. **For** p = 1 **to** $\left|\mathcal{F}^{\text{out}}\right|$：

   2.1. 设 $\mathcal{F}^{\text{out}}$ 中第 p 个特征为 k，令

$$\mathcal{G}^{\text{out}} \leftarrow \left\{ l_1 \in \mathcal{F}^{\text{out}} \,\middle|\, \left|\text{corr}(l_1, k)\right| > T_c \right\}$$

$$\mathcal{G}^{\text{in}} \leftarrow \left\{ l_2 \in \mathcal{F}^{\text{in}} \,\middle|\, \left|\text{corr}(l_2, k)\right| > T_c \right\}$$

   如果 $\left|\mathcal{G}^{\text{out}}\right| > T_g$ 且 $\left|\mathcal{G}^{\text{in}}\right| == 0$，则

$$\mathcal{F}^{\text{out}} \leftarrow \mathcal{F}^{\text{out}} - \{k\}$$

$$\mathcal{F}^{\text{in}} \leftarrow \mathcal{F}^{\text{in}} \cup \{k\}$$

   End

T_g 越大，被检测出的组数越少。在实验中，我们发现设置 T_g 为 1 或 2 能够得到较好的效果，如果继续提高 T_g，则会使一些特征组被漏检。T_c 是相关系数阈值，我们将会在实验中进行考察。

在多次 RFE 迭代中，每次删除的特征数量关系到算法的时间复杂度和精度。我们采用一种同时考虑了时间复杂度和精度的方案[51]：在算法初始阶段，每次删除剩余特征的一半；当剩余特征数少于一个删除数目阈值 T_e 时，改为每次只删除一个特征，从而提高最重要的一部分特征的排序精度。

3.2.3　特征选择的稳定性和集成策略

特征选择的稳定性是近年来研究者比较关心的问题，它描述了特征选择算法对于训练集的变动的敏感性[57]。在训练集有较小变动时，被选择的特征与训练集变动前越相似，说明算法越稳定。这一点对于生物信息学等数据挖掘领域非常重要。Jaccard 指数常被用来衡量两个特征集合的相似度[57,58]，其定义为

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3-9}$$

其取值范围为[0, 1]，0 表示两个集合没有交叠，1 表示两个集合相同。在特征选

择的过程中，为了估测稳定性，可以采用 N 折交叉验证，对每一折的训练集选择出一个特征子集，然后用 Jaccard 指数计算 $N(N-1)/2$ 个子集对的相似度，最后用它们的平均作为算法的稳定性[57]。

为了提高 SVM-RFE 算法的稳定性，我们采用集成策略[58]。随机选择部分训练样本（本章中为十分之九的样本）利用 SVM-RFE 进行特征排序，然后重复 M 次，最后根据每个特征的平均排名来决定该特征的最终排名。这种策略可以使得那些排名较高且较稳定的特征处在最终排名的前列，从而提高选择出的特征子集的稳定性。

3.3　候选特征集

本章采用 3.2.1 节使用的人工合成数据集和上一章介绍的糖尿病呼气数据集来测试特征选择算法。对于后一个数据集，首先需要对电子鼻信号提取候选特征集。传统电子鼻信号分析常用稳态响应作为特征。稳态响应是气敏传感器的信号达到稳定值时的幅度，与被测气体的浓度直接相关。然而，气敏传感器信号的瞬态响应中也包含了很多有用信息[18,19]。瞬态响应指信号的瞬时值，与气体的种类、浓度和传感器周围的气流（进气/洗气）、温度（如温度调制传感器）等都有关系。

本节从 10 个气敏传感器的信号曲线上提取了 1712 个瞬态特征，包括幅度、差分、导数、二阶导数、积分、时间常数和相位特征 7 种。这是一个较大的瞬态特征集合，提取这些特征的目的是为了尽可能地挖掘信号曲线上的信息，从而提高特征选择后的分类精度。对于每个普通传感器（无温度调制），我们从响应曲线上提取 98 个特征，详细描述见表 3.2。温度调制传感器的响应曲线更复杂，包含 18 个“阶梯”，因此可以从每个阶梯上提取特征。特征的具体定义与表 3.2 基本相同，但不包含洗气阶段二阶导数的特征和有关幅度的时间常数特征。对每个温度调制传感器提取的特征总数为 18×19=342。

表 3.2　普通传感器的瞬态特征描述

特征类型	描述	特征数量
幅度	曲线 M 的降采样	21
	$\max(M)$	1
	归一化曲线 $M / \max(M)$的降采样	21
差分	$M(t_{i+1})- M(t_i)$, t = 0, 8, 36, 64, 92, 120 s, i=1, …, 5	5
导数	曲线导数 D 的降采样	21
	$\max(D)$和 $\min(D)$	2

二阶导数	进气和洗气阶段的最大和最小二阶导数	4
积分	$\int_{t_i}^{t_{i+1}} M(t)\mathrm{d}t$，$t$ 的定义同上	5
时间常数	进气阶段 M 到达 max(M)的 30%, 60%, 90%和 100%的时间，以及洗气阶段到达 max(M)的 90%, 60%, 30%的时间	7
	D 到达 max(D)和 min(D)的时间	2
	二阶导数到达其进气和洗气阶段的最大和最小值的时间	4
相位特征	相位特征由 Martinelli 等人提出[59]，首先将响应曲线变换到由 M 和 D 张成的相位空间，然后相位特征可由 $\int_{M(t_i)}^{M(t_{i+1})} D\,\mathrm{d}M$ 得到，t 的定义同上	5

图 3.2 展示了 1712 个特征两两之间的 Pearson 相关系数的绝对值的分布状况。尽管大部分特征的相关性较小，但相关系数绝对值大于 0.9 的特征对从数量上来看仍大于 10^4，因此该数据集存在大量相关特征，需要考虑相关偏差的影响。

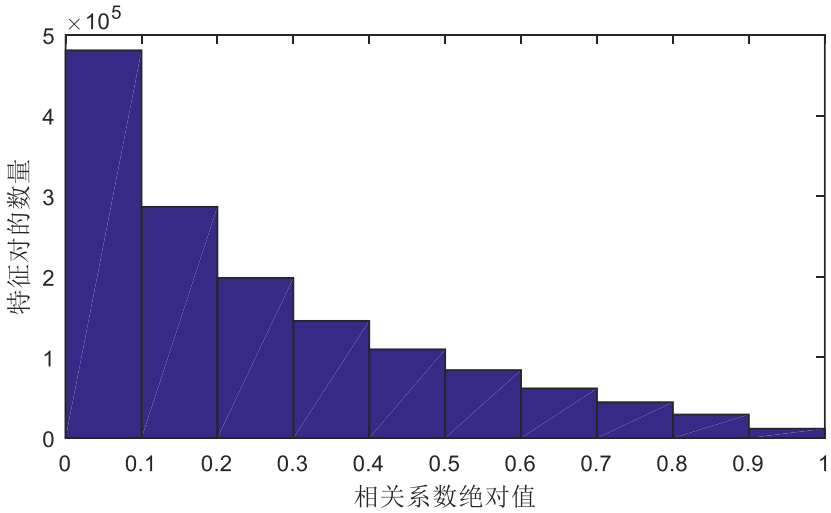


图3.2　瞬态特征之间的相关性

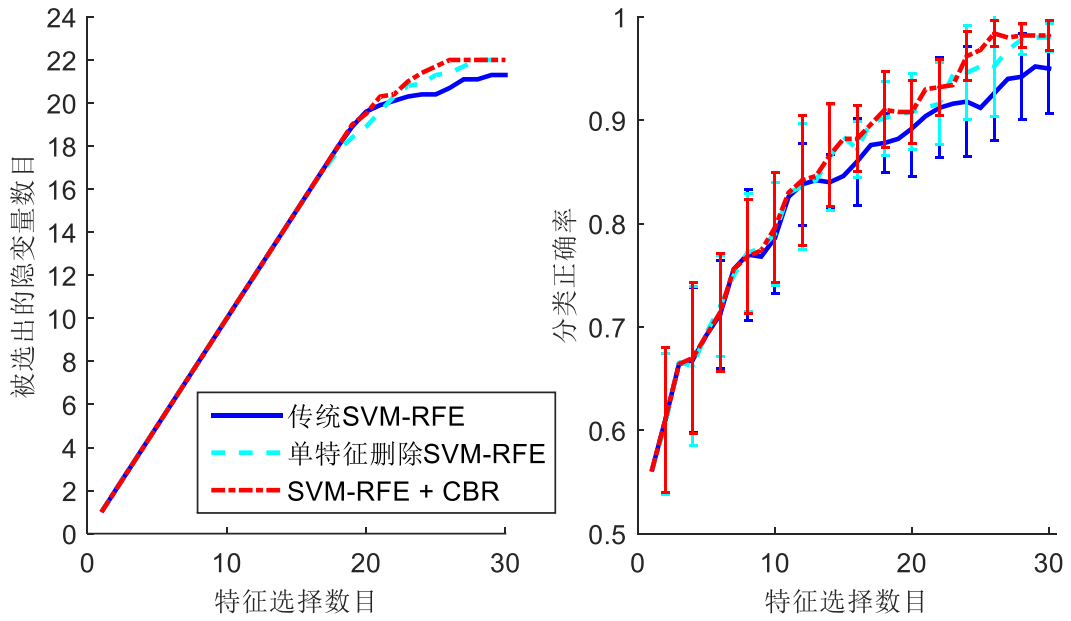
3.4　实验结果和讨论

3.4.1　人工合成数据集

3.2.1 节使用了一个人工合成数据集在来展示相关偏差对 SVM-RFE 的影响，这里仍然使用这个数据集来验证本章提出的相关偏差缩减策略 CBR 的效果。我们采用 10 折交叉验证，首先在每个训练集上进行特征排序，然后基于排名靠前的特征用 SVM 对测试集进行分类，最后计算平均正确率和标准差。类标签是用

隐变量线性组合得到的，因此采用线性的 SVM-RFE 和 SVM 分类器。SVM-RFE 和 SVM 分类器中的惩罚参数 C 均设为 2^3。对于 RFE 迭代，删除数目阈值 T_e 设为 22；对于 CBR，组内特征数阈值 $T_g = 1$，相关系数阈值 $T_c = 0.9$。

图 3.3 中比较了三种方法：无 CBR 的传统 SVM-RFE，每次只删除一个特征的 SVM-RFE，和加入 CBR 的 SVM-RFE。其中第二个方法理论上不会受到相关偏差的影响，但速度也是最慢的。被测试的数据集中包含 100 个特征，其中的 80 个相关特征是由 22 个隐变量生成的。由分图(a)可以看出，传统 SVM-RFE 排序得到的前 30 个特征中没有包含全部隐变量。由于相关偏差的影响，有些相关特征组被过早地删除了。单特征删除 SVM-RFE 和 SVM-RFE+CBR 都能够在前 30 个特征中包含所有隐变量。由分图(b)可以看出，单特征删除 SVM-RFE 和 SVM-RFE+CBR 的正确率相当，都高于传统 SVM-RFE。这证明了 CBR 策略的效果。不过，单特征删除 SVM-RFE 在特征数较多的数据集中并不实用。在下一节用到的糖尿病呼气数据集中，SVM-RFE+CBR 的一次排序仅耗时 100 s，而它耗时 9 h 仍未完成特征排序，因此我们没有记录它的结果。



(a) 排名在前的特征中包含的隐变量数目　(b) 排名在前的特征的分类正确率

图3.3　人工合成数据集的特征选择结果，误差线代表标准差

3.4.2　糖尿病呼气数据集

仍然采用 10 折交叉验证。RBF 核 SVM 被用来作为最终的分类器，参数为 $C = 2^3$，$\gamma = 2^{-6}$。在线性和非线性 SVM-RFE 中均设置 $C = 2^3$，而在非线性 SVM-RFE 中设置 $\gamma = 2^{-8}$。值得注意的是我们在特征选择算法和分类模型中使用了不同的核参数 γ，这是因为特征选择时需要从多到少对多组不同数目的特征建模，而分类

时只需要对最终选择出的特征子集建模。另外，设置 $T_e = 60$，$T_g = 2$。暂不使用 3.2.3 节的集成策略。

图 3.4 展示了线性和非线性 SVM-RFE+CBR 的平均正确率。表 3.3 对比了多种特征抽取/选择算法的结果。对于每个算法，我们对训练集中排序得到的第 1 名、前 2 名，……，前 60 名特征分别在测试集进行测试，选择出最好的结果进行展示。算法的参数都进行了寻优。对于 PCA，把保留能量比例从 80%到 99.9%进行了搜索。应用最小冗余最大相关算法（min-Redundancy Max-Relevance，mRMR）[60]时，对特征进行了三值离散化以计算互信息。为了去除相关偏差的影响，除了 CBR 外，我们还尝试在特征选择之前先对特征进行层次聚类（Hierarchical Clustering，HC），并且对类别数从 800 到 1700 进行了搜索。

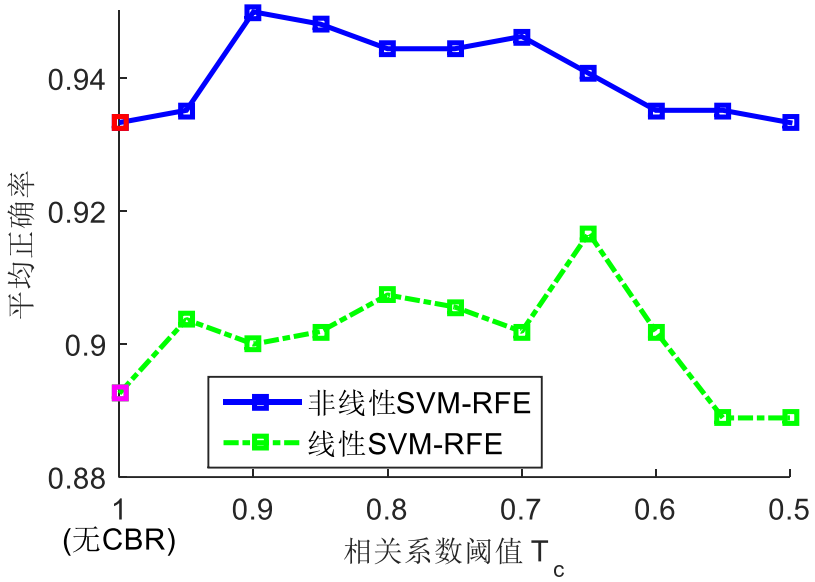


图3.4　相关系数阈值T_c变化时，SVM-RFE+CBR在糖尿病呼气数据集的平均正确率。$T_c = 1$相当于传统SVM-RFE

表 3.3　算法性能比较。稳定性由 Jaccard 指数度量。粗体表示最佳结果

算法	特征数	敏感性(%)	特异性(%)	平均正确率(%)	稳定性
所有瞬态特征	1712	87.04±6.36	90.74±5.59	88.89±4.86	–
PCA+瞬态特征	58	90.74±5.01	90.00±4.95	90.37±2.44	–
PCA+原始信号（第 2 章）	40	92.59±4.62	90.74±6.11	91.67±4.56	–
mRMR[60]	40	90.37±4.68	91.85±4.55	91.11±3.24	**0.5183**
前向选择	28	83.70±6.1	88.15±10.88	85.93±6.25	0.1088
传统线性	57	90.37±7.03	88.15±4.20	89.26±3.79	0.2201

SVM-RFE					
线性 SVM-RFE+CBR (T_c = 0.65)	17	90.74±7.66	92.60±3.51	91.67±3.24	0.2238
传统非线性 SVM-RFE	30	93.33±5.74	93.33±4.20	93.33±3.40	0.4996
非线性 SVM-RFE+HC	36	94.07±5.3	94.07±4.95	94.07±3.50	0.4228
非线性 SVM-RFE+CBR (T_c = 0.9)	31	**94.44**±6.11	**95.56**±3.40	**95.00**±3.39	0.4572

由图 3.4 可以看出，当 T_c 不小于 0.6 时，CBR 都策略能够改善传统 SVM-RFE 的效果。在表 3.3 中，SVM-RFE+CBR 的敏感度、特异性以及平均正确率都高于传统 SVM-RFE。因此，从正确率角度来讲，CBR 策略是有效的。另外，RBF 核的 SVM-RFE 始终优于线性版本，这是由于数据本身线性不可分导致的。在表 3.3 中，当所有瞬态特征都被用于分类时，效果并非最优。尽管其中包含了所有的有用特征，但是也包含了大量不相关和冗余特征，因此会造成过拟合，影响分类器训练的精度。PCA 能够抽取互相正交的特征表达，因此对提高正确率有所帮助。我们也列出了直接对传感器信号幅度进行 PCA 的结果，也就是第 2 章中采用的特征。由于随机选取训练样本的因素，结果与第 2 章略有不同。这个结果优于 PCA+瞬态特征，可能是由于瞬态特征中存在更多的无关特征（与类标签无关），而 PCA 无法将它们去除。

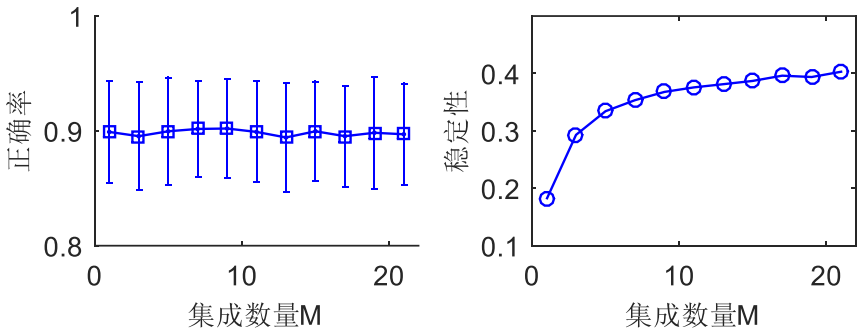
相比起来，一个较好的特征选择算法既能去除冗余特征也能去除无关特征。我们对比的算法都能做到这两点。mRMR 算法是一种多变量过滤式算法，它基于互信息准则选择既和类标签相关性最大又与已选特征相关性最小的特征。前向选择算法迭代式地对每个特征进行检查，选取和已选特征组合后正确率最高的特征。如果候选特征中存在冗余，会增加该算法的耗时，但并不会影响特征的选择。这两种方法的缺点在于它们都是贪婪搜索算法，容易陷入局部最优。作为封装式算法，前向选择还容易对训练集过拟合，尤其是样本数较少时，因此在表中其结果较差。另外，在高维数据集中运行类似的封装式算法的时间复杂度也较高。

SVM-RFE+HC 算法采用了与 Park 等人[55]类似的策略，首先对特征进行聚类，保留每组中最靠近组中心的特征。之后用这个过滤了冗余特征的特征集（大约 1500 个特征）进行特征选择。结果优于传统 SVM-RFE，但我们提出的 SVM-RFE+CBR 用更少的特征获得了更好的准确率。

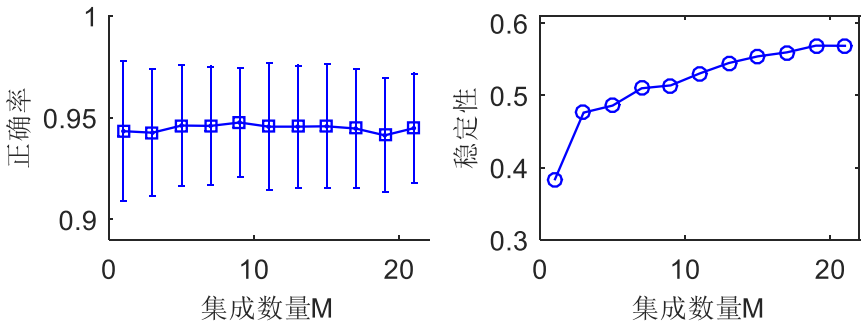
3.4.3　分析：稳定性和集成策略

表 3.3 中列出了每种特征选择算法选出的前 60 个特征在 10 折交叉验证中的平均稳定性。总体来说，由于大量高度相关特征的存在，特征选择的稳定性并不高。mRMR 获得了最高的稳定性，可能是由于其他算法均依赖 SVM 模型的训练，因此对训练数据的变动更为敏感，这与 Kalousis 等人[57]的结果是一致的。并且他们也指出，较高的正确率可能同时会带来较低的稳定性，尤其是存在大量冗余特征时[57]。非线性 SVM-RFE 的稳定性优于线性，可能是由于数据本身线性不可分造成的。CBR 策略在每次 RFE 迭代中需要将一些特征从删除列表移回保留列表，而高度相关的特征的评分很可能会由于训练数据的变动而变动，因此会改变移回的特征，从而增加了算法的不确定性，这就导致 SVM-RFE+CBR 的稳定性有时略低于传统 SVM-RFE。

为了提高 SVM-RFE+CBR 的稳定性，我们测试了 3.2.3 节提到的集成策略，结果如图 3.5。可以发现，随着集成数量（运行特征排序的次数）的增长，正确率的平均值和标准差差并没有明显改变，但稳定性有明显提高。当集成数量大于 9 时，非线性 SVM-RFE+CBR 的稳定性超过了 mRMR。因此，集成策略在增加耗时的代价下提高了稳定性。在对稳定性要求较高的特征选择应用里可以考虑使用该策略。



(a) 线性SVM-RFE+CBR



(b) 非线性SVM-RFE+CBR

图3.5　加入集成策略的SVM-RFE+CBR的正确率和稳定性。误差线代表标准差

3.4.4　分析：特征选择结果

本节对糖尿病呼气数据集特征选择的结果进行进一步的分析，期望能获得以下问题的答案：哪些传感器比较重要？哪种特征比较适合普通/温度调制传感器？温度调制传感器上施加多大的加热电压时获得的特征较有区分性？

SVM-RFE+CBR 的输出结果是一个特征排名。为了解答以上问题，还需要对这个排名进行一些统计。例如，我们从一个传感器上提取了若干个特征，如何从这些特征的排名估计这个传感器的重要性呢？简单地对排名进行平均并非最佳方案，因为这些特征中可能会有一些表现较差的特征拉低平均排名，而我们更关心的是这些特征中间表现较好的部分。因此，我们只计算这些特征中排名最靠前的五个特征的平均排名，作为该传感器的排名，越小表示越重要。将 10 折交叉验证得到的 10 个排名列表放在一起并按这种方法统计后，我们得出以下结论：

- **传感器**：我们使用的电子鼻的传感器阵列见表 2.2。在糖尿病筛查实验中，10 个气敏传感器的重要性由大到小为 TGS2600-TM, TGS2602-TM, TGS2610-D00, TGS826, WSP2111-TM, TGS4161, GSBT11, SP3S-AQ2, TGS822 和 WSP2111。最重要的两个传感器均为温度调制传感器。WSP2111-TM 和 WSP2111 是同一个型号的传感器，前者应用了温度调制，排名第五，而后者没有应用，排名最后。这些发现充分证明了温度调制技术在这个实验中的有效性。
- **特征提取算法**：表 3.4 列出了 7 种瞬态特征的平均排名。从温度调制传感器上提取的相位特征最为重要；时间常数特征和导数特征对两种传感器都比较重要；而二阶导数和积分特征则是最不重要的两种。

表 3.4　7 种瞬态特征的平均排名，数值越小排名越高，越重要

特征类型	普通传感器	温度调制传感器
幅度	7.2	6.4
差分	20.4	4.6
导数	6.2	2.6
二阶导数	80.4	35.4
积分	46.0	15.6
时间常数	**4.0**	2.6
相位	11.8	**1.0**

- **温度调制加热电压**：我们发现当加热电压在 6V-7V-3V 区间内变化时（见图 2.6），特征的平均排名最高。这表明在检测糖尿病的呼气标志物时，当加热

电压靠近或高于普通加热电压（5V）时，温度调制传感器的响应更具有区分性。根据文献[46]，当加热温度较低时，气敏传感器的响应常常不具有区分性。这与我们的发现相吻合。

3.5　本章小结

本章工作可以分为两部分：对 SVM-RFE 相关偏差问题的研究，和电子鼻信号瞬态特征的选择。首先介绍了线性和非线性 SVM-RFE 算法的特征评分准则以及迭代特征删除策略；随后指出候选特征高度相关时，SVM-RFE 存在相关偏差问题，并在合成数据集上进行了验证；接着提出了相关偏差缩减策略，通过检查被删除的特征的相关性来减轻这个问题，并在合成数据集和糖尿病呼气数据集上确认了该策略的有效性。经过特征选择后，糖尿病筛查的平均正确率从上一章的 91.67%提高到了 93.33%；加入本章提出的 CBR 策略后，进一步提高到了 95.00%. 集成策略可以改进特征选择的稳定性。

我们从电子鼻信号中提取了一个由多种瞬态特征组成的 1712 维的候选特征集合，进行了特征选择，并利用特征选择的结果分析了不同传感器、特征提取算法以及温度调制加热电压的有效性，这些结论可以给其他电子鼻研究者在选择传感器阵列、设计硬件和特征选择算法时提供参考。另外，本章提出的改进的 SVM-RFE 算法[61]可以应用于其他需要特征选择的领域。

第4章　基于加权正则化和多任务学习的漂移补偿

本章和接下来的两章将重点研究电子鼻的设备差异和时变漂移问题。为了简洁起见，我们将这两种影响电子鼻鲁棒性的因素统称为漂移，因为它们都对应了特征空间中分布的移动。漂移是电子鼻实际使用中经常遇到的、具有一定挑战性的问题。我们将首先对现有漂移补偿算法进行综述和总结，之后介绍这三章的实验中将要使用的漂移数据集。在本章中，我们提出了两种漂移补偿算法：结合标准化误差的模型改进、基于标定样本的多任务学习。它们都是基于标定样本在模型层面进行补偿的算法。不同之处在于，前者在已有的模型中加入加权正则项以提高模型对于领域差异的鲁棒性；而后者对模型进行显式迁移，对不同领域的样本学习不同的模型。

4.1　相关工作

4.1.1　设备差异补偿

同样型号的电子鼻在测试同样的气体样本时，响应通常不会完全相同（图 4.1）。造成这个问题的主要原因是气敏传感器和电子鼻在制作过程中无法确保完全一致。每制作一台新设备都采集大量有标签样本训练预测模型往往是不现实的。我们希望能够只在一台电子鼻（主设备）上采集足够的有标签样本并训练预测模型，然后将模型应用到其他所有电子鼻（从设备）上[19,62]。因此，对设备差异进行补偿就显得尤为重要。

设备差异并非电子鼻独有的问题。例如，化学计量学里的光谱分析学（spectroscopy）中也存在这个问题[21]。光谱信号与气敏传感器信号类似，都是频率不高的一维信号，因此大部分漂移补偿算法在这两个领域中都可以通用。很多补偿方法都需要用到标定样本。标定样本是被测样本（用来训练、预测的样本）之外的一组用于辅助漂移补偿的样本，需要在主、从设备上都进行测量。透过它们可以得到主从设备特征空间分布之间的一一对应的关系，从而为设备差异补偿提供依据。图 4.2 展示了两台电子鼻采集的被测样本以及标定样本的分布，可以看出从设备样本的分布相对主设备有一个向右上方的漂移，这个漂移可以利用两组标定样本的对应关系进行估测。

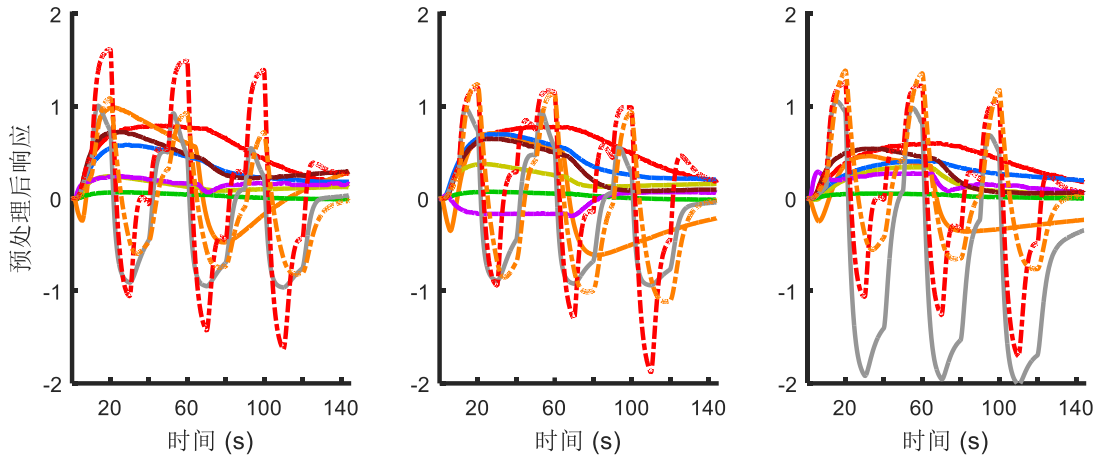


图4.1　设备差异示例：三台同样型号的电子鼻对同一个呼气样本的响应曲线

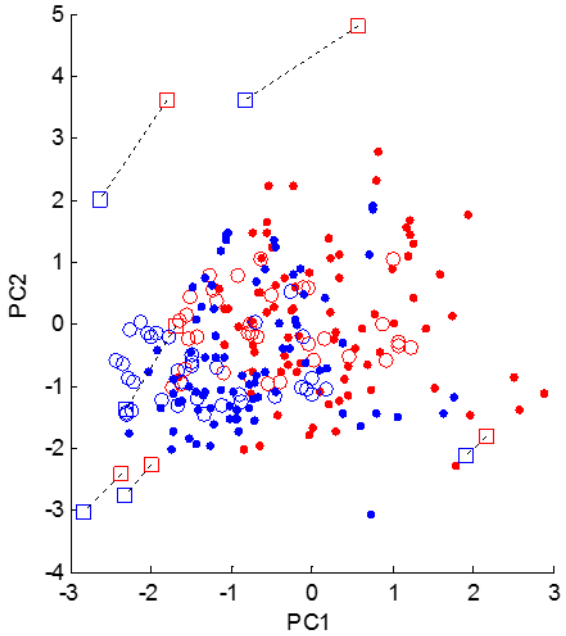


图4.2　两台电子鼻上采集的被测样本和标定样本图示。蓝色和红色分别代表主设备和从设备，点和圈为两类被测样本，用虚线连接的方形为一一对应的标定样本。坐标轴为被测样本的前两维主成分

为了缩小设备差异，可以对特征进行预处理，例如使用标准正态变换（Standard Normal Variate，SNV）[19]。这种方法根据每台设备的标定样本的均值和方差将相应设备的各维特征均值置零、方差归一，从而去除简单的加性和乘性漂移。事实上，如果设备差异不复杂，SNV 已经能够取得较好的补偿效果。

特征标准化（feature standardization）是电子鼻和光谱学领域中应用较多的一类算法。该方法依据标定样本将从设备的每一维特征向主设备的对应特征进行拟合，从而缩小两台设备样本的特征差异。拟合后的从设备特征即可利用主设备

的模型进行预测。根据拟合自变量个数的不同，可以分为单变量标准化、多变量标准化和分段标准化等[21,63]。研究者们还尝试了多种拟合算法，如鲁棒回归[62,65]、人工神经网络[66]、偏最小二乘回归[67]等。

Fan 等人[64]提出了基于典型相关分析的方法。该方法将主、从设备的特征都进行投影，使得投影后的变量彼此相关系数最大化；之后用多变量标准化将从设备投影后变量向主设备投影后变量进行拟合；最后用拟合后变量重构从设备的原始特征，重构出的特征向量即可用主设备的预测模型进行预测。实验表明该方法在标定样本数较多时优于分段特征标准化。但当特征维度高且标定样本数小时，有陷入过拟合的风险。

上述方法均为特征层面的补偿方法，仅仅考虑消除主从设备特征间的差异。模型层面的补偿方法直接在预测模型系数上进行操作，借助标定样本或来自从设备的少量有标签被测样本来修正主设备的预测模型，从而使得修正后的模型能够用于预测从设备的样本。在鲁棒回归算法[68]中，特征标准化得到的误差协方差矩阵被加入岭回归的系数求解公式中，使得求得的模型系数对标准化误差较大的特征鲁棒。Kalivas 等人提出一种模型更新算法[69]，通过对岭回归（ridge regression）的目标函数进行修改，同时最小化主设备和从设备的有标签被测样本的训练误差，从而得到一个可以同时应用于两台设备的预测模型。然而当两台设备的差异较大时，更好的方案是对每台设备学习一个预测模型。因此有研究者提出了基于支持向量回归的正则化多任务学习算法[70]，对主设备和从设备的有标签被测样本学习两组模型系数，在最小化两台设备的训练误差的同时约束两组系数尽量相似。

4.1.2　时变漂移补偿

在气敏传感器的长期使用过程中，气敏元件常常会发生性质的改变[71]，导致其敏感特性变化。另外，采样环境、甚至样本本身的构成也可能会随时间变化，这些因素导致样本的分布会随着时间变化。这种变化通常是较复杂的，随传感器的种类型号、使用方式、使用环境的不同而不同，难以建立统一的补偿模型[18]。图 4.3 展示了我们采集的呼气分析数据集中部分样本包含的时变漂移。可以发现第一个传感器的响应总体来讲缓慢下降，表明其对气体的敏感度逐渐降低（老化）。第二个传感器老化的速度较快，因此我们在第 200 天左右给两台设备都更换了一个新的传感器，导致其响应上升，之后又逐渐下降。另外，从图中（例如传感器 2 在第 200 天之后的响应）也可以观察出两台电子鼻的设备差异。因此，时变漂移的补偿对于电子鼻也非常重要。

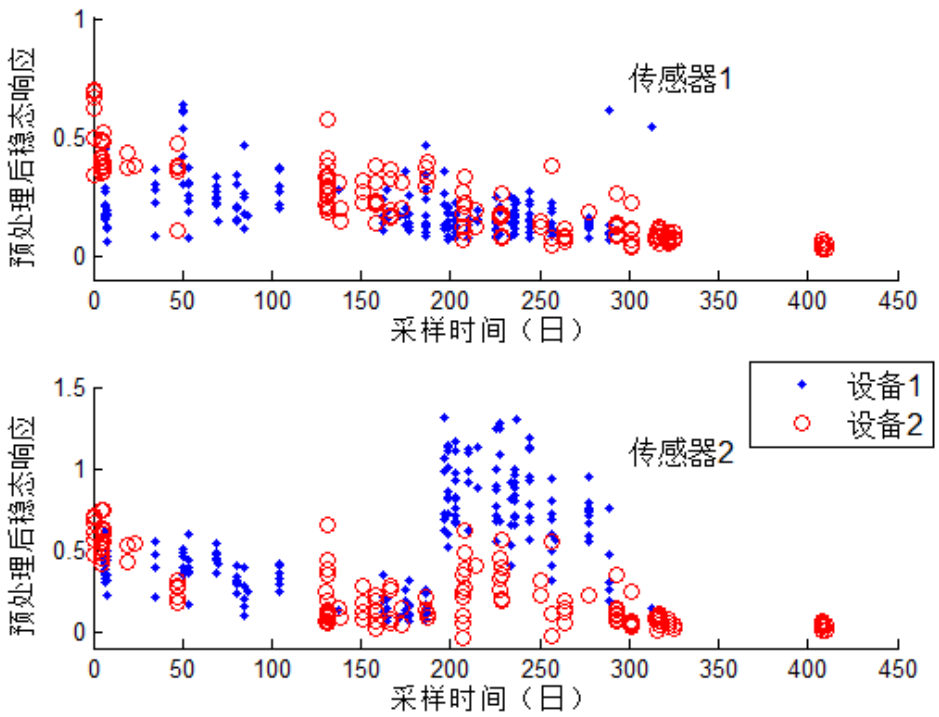


图4.3　时变漂移示例：呼气分析数据集中传感器的稳态响应随采样时间的变化。上下两个图代表两个传感器。每个点代表一个慢性肾脏病的呼气样本

首先介绍特征层面的补偿算法。时变漂移比设备差异复杂的地方在于它是一个动态变化的过程，因此需要周期性地采集标定样本，并且对不同时间采集的被测样本采用不同的补偿系数。如果我们将不同时间段采集的样本按照采集时间分成多个批次，并且假定每个批次内的漂移较小、可以忽略，那么用于补偿设备差异的很多方法也可以用在时变漂移问题中，例如 SNV 预处理，特征标准化等。只不过这里的“主设备”和“从设备”变成了“较早批次”和“较晚批次”。研究者们还提出了多种其他的时变漂移补偿算法。一种较简单的方法是乘性漂移校正（Multiplicative Drift Correction，MDC）[73]。MDC 将每个传感器在标定样本上的响应拟合成一条相对于时间的曲线，与初始时间段的响应比较后得到每个时间点的校正系数；然后对每个样本乘以相应时间的校正系数。该方法被应用在一些商用电子鼻中[74]。

成分校正（Component Correction，CC）是一类经典的方法，至今还被很多文章作为基准算法进行比较。这类方法假设漂移存在于一个子空间中，需要将样本在此子空间的投影减去。例如，基于 PCA 的成分校正（CC-PCA）[74]利用标定样本和 PCA 求出此子空间 P 以及被测气体 X 在 P 上的投影 T，然后利用公式 $X_{\text{corrected}} = X - TP$ 对样本进行补偿。该方法的一个问题是各类样本的漂移对应的子空间可能不同。正交信号校正（Orthogonal Signal Correction，OSC）[75]则通过偏最小二乘回归算法（Partial Least Squares，PLS）迭代减除 X 中与其标签正交

的子空间，从而去除掉与标签无关的所有成分，包括漂移。OSC 不需要标定样本，但可能需要较多的漂移后有标签样本才能准确地估测出需要去除的子空间。另外，成分校正算法共同的问题在于，漂移和有用信息所在的子空间不一定能够完美分离。

如果我们假设有用信号与漂移是两组互相独立的随机变量，那么独立成分分析（Independent Component Analysis，ICA）就可以用来检测信号漂移。Di Natale 等人[76]使用 ICA 来分析电子鼻信号时，发现一个特定的独立成分与温湿度干扰相关，去掉这个成分后类间区分度得到了提高。这种方法可以被用来去除时变漂移，但如果时变漂移的成因复杂，则难以确定哪个分量应当被去除。另一类漂移补偿方法被称为自适应补偿方法。此类方法的核心在于利用漂移过程中的新样本动态的更新数据模型，例如基于自组织图网络（Self Organizing Map，SOM）的方法[71,72]。该类方法的缺点是数据模型通常较为复杂，需要较多训练样本，且对于学习率的设定较为敏感[71]。

在模型层面，如果将样本按采集时间分成批次，那么前一节介绍的模型更新等算法也可以用来补偿时变漂移。Zhang 等人[77]提出领域适配极限学习机（Domain Adaptation Extreme Learning Machine，DAELM），基于漂移后的有标签和无标签样本更新极限学习机模型，并且在一个公开的电子鼻时变漂移数据集上取得了较好的结果。该方法不需要标定样本，但需要较多的漂移后有标签样本才能准确的对模型做出更新。还有研究者将概念漂移适配（concept drift adaptation）、迁移学习、半监督学习等领域的算法引入漂移补偿。Vergara 等人[78]不对漂移进行显式补偿，而是对不同批次的样本训练得到的分类器进行加权集成，从而提高预测模型鲁棒性。Liu 等人[79]提出利用漂移过程中的无标签样本来改进预测模型。他们提出了一种组合测地线流核学习算法，可以用于适配具有时序关系的多个领域（即不同批次的样本），并采用半监督分类器——拉普拉斯 SVM 进行分类。同样利用半监督学习算法来改进预测模型的还有 De Vito 等人[80]。

4.1.3　标定样本选择

在基于标定样本的漂移补偿算法中，标定样本的选择也是比较重要的一环。为了准确补偿，标定样本应能代表大多数被测样本。同时，标定样本的数量应当尽量小，从而减轻采集时的工作量[21,81]。在模型训练阶段一般只能获取源领域的样本，因此研究者通常先构造一个源领域的候选标定样本集合，然后从中选择出一组最具代表性的样本。这些样本即可作为标定样本，在每台新设备以及新时间段进行采集。

在电子鼻及光谱学领域，应用最广的标定样本选择算法是 Kennard & Stone（KS）算法[82,21,62]。它要求选出的样本之间彼此距离较远。给定一组候选样本，KS 算法首先选择相距最远的两个样本，接下来不断选择剩余样本中与已选样本的最近距离最大的那个样本；重复这个步骤直到已选样本数目达到预定值。这个算法能保证标定样本覆盖整个候选样本空间。但因为野点距离其他样本很远，所以很有可能选择到野点。

我们将机器学习领域的主动学习算法引入标定样本选择问题，考察了两种主动学习算法。直推式实验设计（Transductive Experimental Design，TED）[83]选择样本的标准是，被选出的样本应该能够较好地重构所有候选样本。局部线性重构（Locally Linear Reconstruction，LLR）[84]则进一步将候选样本的局部流形结构纳入考量。LLR 利用每个样本的近邻来重构该样本，然后选择能够较好地重构整个候选样本集合的一组样本。通过实验，我们发现 LLR 在这三种样本选择算法中效果最好，因此采用它作为后续实验中的标定样本选择算法。

为了更好地补偿时变漂移，需要定期采集标定样本，因此一个相关的问题就是标定时间的选择。有文献推荐在电子鼻投入实用的头一个月应当每周标定，之后可以改为每月标定[85]。还有一种思路是用算法检测漂移的显著程度，从而自动确定标定时间，这方面的算法可以参考概念漂移适配领域的变化检测算法（change detection）[91]，我们留待今后进行研究。

4.2　算法总结与面临的挑战

4.1.1 和 4.1.2 节介绍了典型的设备差异与时变漂移补偿算法。从机器学习的视角出发，我们将漂移补偿问题视为迁移学习问题，按照迁移层面、迁移方法以及迁移依据三个分类标准对现有算法进行简单总结：

- **迁移层面**：主要分为特征表达和模型系数两个层面。前者例如标准正态变换、特征标准化、成分校正等，后者例如模型更新、正则化多任务学习、分类器集成等。从现有文献来看，模型层面迁移精度较高，但如果更换损失函数需要重新求解；特征层面的算法则可搭配各种模型使用，比较灵活。
- **迁移方法**：主要分为显式迁移以及领域不变两种。前者对不同的领域学习不同的特征变换或模型；而后者则对所有领域学习统一的、对领域差异鲁棒的特征变换或模型，或从特征中去掉与领域差异有关的成分。前者例如标准正态变换、特征标准化、正则化多任务学习，后者例如成分校正、模型更新、鲁棒回归。
- **迁移依据**：为了将知识从源领域迁移到目标领域，需要知道目标领域的相关信息，这种信息可以从标定样本、有标签目标领域样本、无标签目标领域样

本这三类依据中获得。基于标定样本的方法包括标准正态变换、特征标准化、基于 PCA 的成分校正等，基于有标签目标领域样本的方法如正交信号补偿、模型更新、正则化多任务学习、分类器集成等，基于无标签目标领域样本的方法包括组合测地线流核学习和基于半监督学习的方法[80]。三类方法中，标定样本包含领域间对应信息，信息量大，因此要达到同样的迁移精度需要的样本数量一般也较少。但标定样本的配制和采集需要一定工作量。有标签目标领域样本不包含领域间对应信息，通常要达到同样的迁移精度需要的样本数量多于标定样本。另外，目标领域的有标签样本有时较难获得。无标签目标领域样本最容易获得，但包含的信息量比前两种样本都少，因此迁移精度可能不够理想。

漂移补偿算法面临的挑战可以总结为欠迁移和过迁移两种。当补偿算法无法充分描述和消除领域间差异时，会造成**欠迁移**。这类算法通常较为简单，自由度不足，例如特征层面的标准正态变换、单变量特征标准化，和模型层面的模型更新算法。缓解欠迁移问题的方法是提高算法的迁移能力，例如将单变量特征标准化升级多变量特征标准化，或将单一模型的模型更新算法升级为多个模型的多任务学习[70]。但此时还要防范**过迁移**，即过度强调消除迁移依据中体现的领域差异。由于标定样本或有标签目标领域样本通常数量较少，不能完全刻画领域间的迁移关系；由于采集过程中的不确定性，迁移依据中还常常存在噪声。因此，能较好地消除迁移依据中领域差异的算法不一定能准确地消除其他被测样本的领域差异，也就是泛化性能不佳。这一点类似机器学习算法中的过拟合现象。此外，有些算法仅仅考虑消除领域间差异，可能导致有用信息丢失，例如成分校正算法。缓解过迁移问题的方法是增加约束，限制算法的迁移能力，例如限制多变量特征标准化中的自变量数目从而得到分段特征标准化，或在多任务学习算法中加入模型相似性约束及其他先验约束。

总之，一个较好的漂移补偿算法应当具有足够的迁移能力，能够充分描述和消除领域间差异；同时利用数据中固有的先验约束防止过迁移。通常可以在算法中加入可调的参数来平衡这两者。

4.3　漂移数据集

在本章及接下来的两章中，我们将使用三个包含漂移的数据集来评价补偿算法，分别是阵列漂移数据集、Corn 数据集和呼气分析数据集。

阵列漂移数据集（gas sensor array drift dataset）主要包含时变漂移。它由

Vergara 等人[78]采集并公开在 UCI 机器学习仓库网站①。它由一个包含 16 个气敏传感器的电子鼻采集，采集时间跨度 36 个月。数据集中包含不同浓度的 6 种气体样本（氨气、乙醛、丙酮、乙烯、乙醇和甲苯），样本总数 13910，在电子鼻领域属于规模较大的数据集。每个样本对应一个 128 维的特征向量，是响应曲线的指数滑动平均特征[78]。Vergara 等人根据样本采集时间将所有样本分为 10 个批次，每个批次对应的时间段和样本数可见表 4.1，其中采集时间为月份编号（如第 9 批样本在第 24 个月和第 30 个月采集）。为了测试漂移补偿算法的效果，我们选择第一批次的样本作为源领域，第 2 至 10 批次分别为 9 个目标领域，利用源领域的有标签样本训练气体分类模型，然后在每个目标领域进行测试。这种评价方法与现实应用中电子鼻的使用方法类似，同时也是多篇文献[77-79]中采用的方法。

表 4.1　阵列漂移数据集中每个批次的采集时间和样本分布。气体名称后括号中数字代表浓度范围（单位：ppm）

批次序号	采集时间(月份编号)	氨气 (2.5-1000)	乙醛 (2.5-300)	丙酮 (10-1000)	乙烯 (2.5-300)	乙醇 (2.5-600)	甲苯 (1-230)	总数
1	1,2	83	30	70	98	90	74	445
2	3,4,8,9,10	100	109	532	334	164	5	1244
3	11,12,13	216	240	275	490	365	0	1586
4	14,15	12	30	12	43	64	0	161
5	16	20	46	63	40	28	0	197
6	17,18,19,20	110	29	606	574	514	467	2300
7	21	360	744	630	662	649	568	3613
8	22,23	40	33	143	30	30	18	294
9	24,30	100	75	78	55	61	101	470
10	36	600	600	600	600	600	600	3600

样本集包含的时变漂移可以由图 4.4 直观地看出。批次 1-6 中的两类样本被投影到前两维主成分方向。可以看出氨气样本大体先向上（批次 2）又向右下方（批次 3-6）漂移，而乙醛样本的漂移不甚明显。同时还存在一些野点样本，如乙醛批次 6 的部分样本。图中反映出了该数据集漂移的复杂性。

① http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations。

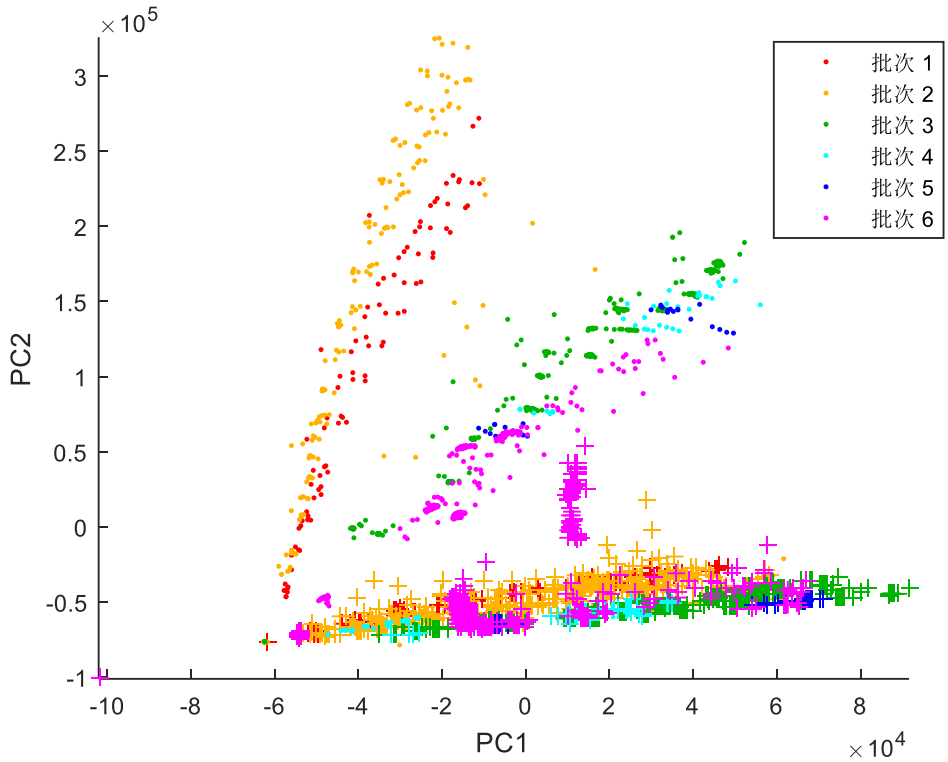


图4.4　阵列漂移数据集里批次1-6的样本分布。点和加号分别代表氨气和乙醛样本；不同颜色代表不同批次

Corn 数据集主要包含设备差异。它也是一个公开的数据集①，由三台编号分别为 m5, mp5, mp6 的近红外光谱仪采集。前文提到，光谱仪与电子鼻的信号类似，并且也存在设备差异问题，补偿算法可以通用。因此我们采用这个数据集来验证本文算法的有效性和通用性。数据集中，每台光谱仪被用来测量 80 个玉米样本的水分、油脂、蛋白和淀粉含量。这四种含量的数值范围分别为 9.377 至 10.993，3.088 至 3.832，7.654 至 9.711，和 62.826 至 66.472。每个样本对应一个 700 维的特征向量，特征为 1100-2498 nm、间隔 2 nm 的光谱。从图 4.5 中可以很明显地看出不同设备的样本分布差异。

我们采用 Yu 等人[70]提出的 4 折交叉验证测试方法。在第 i 折中，选择第 i, i+4, …, i+76 个样本作为测试样本，其余作为训练样本。用设备 m5 采集的训练样本训练回归模型，mp5 和 mp6 采集的测试样本进行测试。也就是说，m5 为主设备，mp5 和 mp6 为从设备。对四种含量值分别训练回归模型。

① http://www.eigenvector.com/data/Corn/。

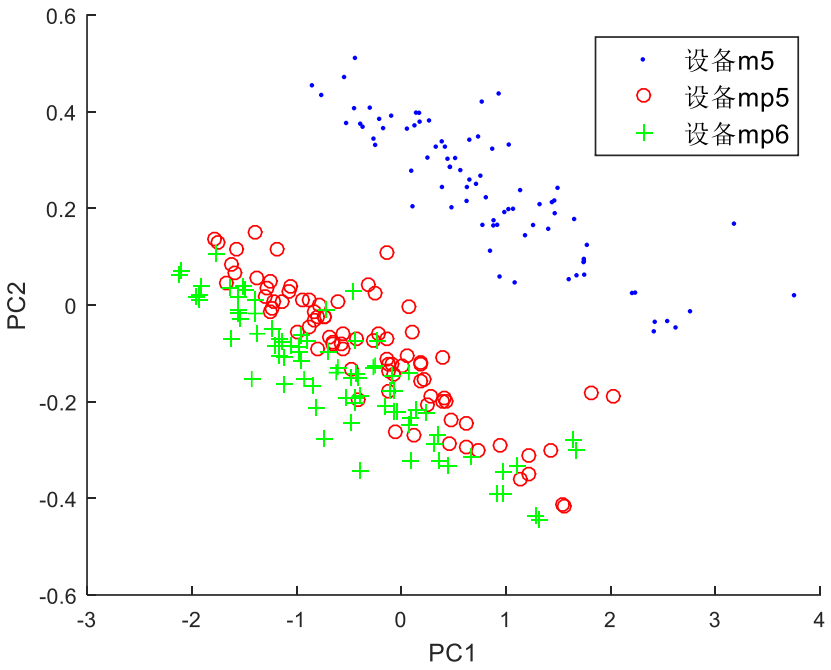


图4.5　Corn数据集里三台设备的样本分布

呼气分析数据集同时包含设备差异和时变漂移。第 2 章介绍的糖尿病呼气数据集中漂移较小，不适合用来评价漂移补偿算法。这里使用的呼气分析数据集是我们与广东省中医院合作采集的一个规模更大、疾病种类更多的数据集。由于采集周期更长，数据中的漂移也更复杂。采集过程中使用了两台同样型号的电子鼻，采集时间为 2014-2015 年，跨度大约两年。在数据集中，我们选取了对应五种疾病的呼气样本，分别是糖尿病、慢性肾病、心脏病、肺癌、乳腺癌。这些疾病已被证明与特定呼气标志物相关，见表 2.1。由于气敏传感器的交叉敏感性，我们的设备可以测量这些疾病对应的呼气标志物。我们还用两台设备周期性地采集了标定样本。每组标定样本包含 8 种标准气体，分别是 1%和 5%浓度的二氧化碳，2 ppm 和 50 ppm 的氢气，1 ppm 和 10 ppm 的丙酮，1 ppm 和 10 ppm 的氨气。二氧化碳样本用来标定电子鼻中的二氧化碳传感器；其他三种气体在呼气中含量较高或与疾病相关。气体浓度的选择也考虑到了呼气中该气体的浓度范围。经实验[86]，这 8 种标定样本与呼气样本的漂移趋势比较相似。图 4.6 总结了标定样本、健康样本与五类疾病样本的采集时间、设备号和样本数。

呼气分析数据集的漂移也比较显著，示例可见 4.1.2 节的图 4.3。我们对健康样本和每种疾病样本进行二分类。由于每类样本数不同，使用 F-score 作为准确率的评价指标：

$$\text{F-score} = \frac{2}{\frac{1}{\text{敏感性}} + \frac{1}{\text{特异性}}} \tag{4-1}$$

为了测试漂移补偿算法的有效性，我们只采用设备 1 采集的每类前 50 个样本作为有标签训练样本（图 4.6 中加号），其他样本均为测试样本。

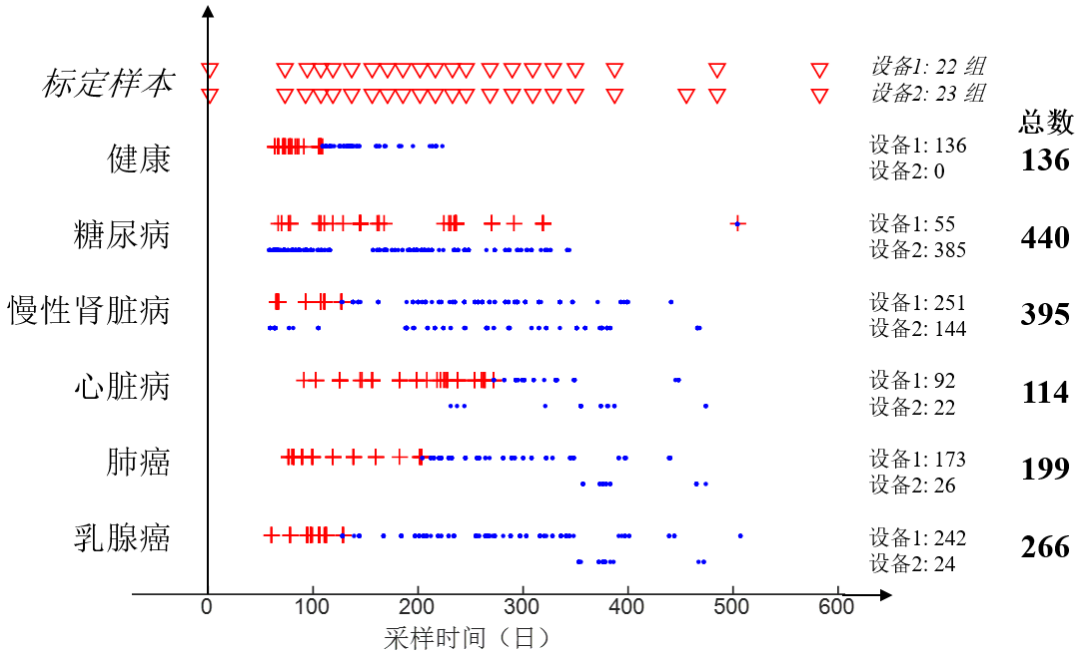


图4.6　呼气分析数据集样本概览。每个加号或点代表一个呼气样本，加号为训练样本；每个三角形代表一组标定样本。x坐标为采样时间。每类样本的第1、2行分别由设备1、2采集。右侧数字为样本数

由于厂家停产等因素，该数据集使用的电子鼻传感器阵列与第 2 章略有不同，WSP2111-TM 被去除。因此，新电子鼻中包含 9 个气敏传感器。另外，该数据集为与医院方面合作采集，由于采样者非专业以及实际采集环境中的干扰等因素，样本信号中存在较多噪声，造成第 3 章介绍的多种瞬态特征无法准确提取。因此我们并没有使用第 3 章选择出的最优特征组合，而是采用了传统的传感器稳态响应作为特征，9 个传感器对应 9 维特征向量。由于疾病与呼气标志物的浓度相关，而传感器的稳态响应直接反映气体的浓度，因此该特征向量可以包含大部分的疾病分类信息。另外，阵列漂移数据集和 Corn 数据集是公开数据集，为了与前人算法对比，我们也没有采用第 3 章提出的算法进行特征选择。

4.4　基于加权正则化的漂移补偿

特征层面的补偿算法，如标准正态变换（SNV）和特征标准化等，可以基于标定样本对特征进行匹配。我们将这种匹配统称为标准化。标准化过后领域间的

特征差异会被缩小，然而并不能完全消除。有些特征的标准化效果可能较好而有些较差，这是由于不同气敏传感器和特征的稳定性和信噪比不同造成的。直观上讲，如果能让后续的预测模型更依赖的标准化效果好的特征，就可以提高模型在目标领域样本上的泛化性能。这就是结合标准化误差的模型改进算法（Standardization-Error-based Model Improvement，SEMI）的基本思路。

为此，我们首先定义标准化误差（SE）。第 j 个特征的标准化误差定义为标准化后的源领域和目标领域的标定样本 $\hat{\mathbf{t}}_S$ 和 $\hat{\mathbf{t}}_T$ 在该特征上的均方根偏差，即：

$$\mathrm{SE}_j = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}(\hat{t}_{S,ij} - \hat{t}_{T,ij})^2} \tag{4-2}$$

上式中 n_t 为标定样本的数量，下标 ij 表示第 i 个样本的第 j 个特征。然后，在现有的预测模型的目标函数中加入一个加权的 Tikhonov 正则项

$$\mu\sum_{j=1}^{m}\left(\mathrm{SE}_j \cdot \beta_j\right)^2 \tag{4-3}$$

其中 $\boldsymbol{\beta}$ 是模型中待求的系数，m 是特征总数，μ 是正则项的权重系数。这样使得标准化误差越大的特征，其惩罚系数越大，因此在模型中的权重就会越小，从而提高模型对领域间特征差异的鲁棒性。

SEMI 是一种普适的正则化策略，可以应用于多种分类或回归模型，甚至其他的漂移补偿算法中。SEMI 仅包含模型系数的平方项。对于大部分预测模型，如逻辑回归、SVM、岭回归等，只需要对原有求解方法简单修改就可以得到加入 SEMI 的模型的求解方法。SEMI 将特征层面的补偿与预测模型训练有机结合，首先由标定样本得到了关于特征的稳定性和信噪比的先验信息，然后将这种先验信息引入到了模型中。本章第 6 节的实验将会证明其有效性。

4.5　基于多任务学习的漂移补偿

多任务学习是一种归纳式（inductive）的迁移学习方法[20]。它对不同的领域学习不同的预测模型，并将学习这些模型当作一系列相关的任务同时进行。同时学习多个模型的好处是这些任务可以共享各自的信息，从而提高所有模型的精度和泛化能力，尤其有利于有标签样本数较少的领域[87]。信息共享的方式通常是构造特定的目标函数将关于模型和特征的先验知识加入其中。最常见的模型先验知识是模型之间的相似性。例如，正则化多任务学习（RMTL）[88]在算法目标函数中加入正则项，惩罚不同模型系数之间的差异。Yu 等人将这个算法应用在光谱仪的设备差异补偿中，获得了很好的效果[70]。Zhou 等人将疾病进展预测表述为一个多任务回归问题[89]，提出了 Temporal Group Lasso 算法。该算法对每一个

时间段学习一个预测模型，并要求相邻时间段的模型相似。目标函数中还加入了 group Lasso 正则项用于特征选择。

在本节中，我们将多任务学习框架应用于设备差异和时变漂移补偿。首先考虑一个源领域和一个目标领域的情况，提出了基于标定样本的双任务学习算法（Transfer-sample-based Coupled Task Learning，TCTL）。TCTL 与 Yu 的算法[70]的最大差别在于，后者利用目标领域的有标签样本进行学习，而 TCTL 利用标定样本进行学习。接着我们将 TCTL 拓展到多个目标领域的情况，提出了基于标定样本的多任务学习算法（Transfer-sample-based MultiTask Learning，TMTL），并针对设备差异和时变漂移的不同先验设计了并行和串行两种迁移策略。最后，为了更精确地补偿时变漂移，提出了一种动态模型策略，使得每个设备的模型随着采样时间动态变化。

4.5.1　双任务学习

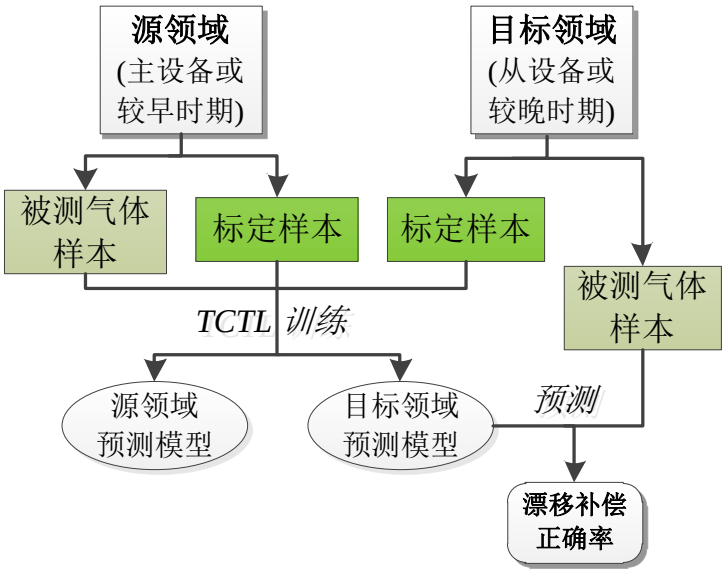


图4.7　TCTL应用流程图

基于 TCTL 的漂移补偿流程见图 4.7。给定源领域（主设备或初始时间段）的有标签样本和一组标定样本，TCTL 的目的是学习适用于目标领域（从设备或后续时间段）的预测模型。设 $X_S \in \mathbf{R}^{n\times m}$ 为源领域样本矩阵，每行为一个特征向量，n 和 m 分别为样本数和特征维数。$\mathbf{y}_S \in \mathbf{R}^n$ 是对应的标签向量。$T_S, T_T \in \mathbf{R}^{n_t\times m}$ 分别是源领域和目标领域的标定样本矩阵，两者每行的样本一一对应。n_t 是标定样本总数。$\boldsymbol{\beta}_S, \boldsymbol{\beta}_T \in \mathbf{R}^m$ 分别是待学习的源领域和目标领域的预测模型系数。TCTL 的目标函数为：

$$\begin{aligned}\min_{\boldsymbol{\beta}_S,\boldsymbol{\beta}_T} & \mathcal{L}(X_S,\mathbf{y}_S,\boldsymbol{\beta}_S)+\lambda_1\left\|T_S\boldsymbol{\beta}_S-T_T\boldsymbol{\beta}_T\right\|_2^2 \\ & +\lambda_2\left\|X_S\boldsymbol{\beta}_S-X_S\boldsymbol{\beta}_T\right\|_2^2+\mu\sum_{j=1}^{m}w_j{}^2\left(\beta_{S,j}{}^2+\beta_{T,j}{}^2\right).\end{aligned} \tag{4-4}$$

通过最小化上式，可以同时得到两个领域的模型。在式（4-4）中，第一项是源领域有标签样本的**预测损失函数**；第二项可以被称为**标定样本项**，它要求源领域和目标领域的标定样本在经过各自领域的模型投影后较好地对齐；第三项可以被称为**模型相似约束项**，它要求两个模型将源领域样本投影到相近的位置；第四项是上一节提到的 **SEMI 正则项**。$\lambda_1, \lambda_2, \mu > 0$为正则项权重。下面分别对三个正则项进行一些解释。

标定样本项是 TCTL 中信息迁移的关键。它通过将两个领域的标定样本在模型层面对齐，消除了模型层面的领域间差异，从而使得源领域的鉴别信息可以应用在目标领域。这一项还可以视为传统的特征标准化方法的改进。在线性情况下，特征标准化本质上是求解一个变换矩阵 M 将目标领域的特征变换到源领域，并满足$T_T M \approx T_S$。在预测模型训练阶段，需要进一步将特征按照$\boldsymbol{\beta}_S$投影。因此我们实际上可以直接在投影方向上消除领域间差异，即最小化$\|T_S\boldsymbol{\beta}_S - T_T M\,\boldsymbol{\beta}_S\|_2^2$。如果设$\boldsymbol{\beta}_T = M\boldsymbol{\beta}_S$，就得到了标定样本项。比起特征标准化，TCTL 不需要再求解 M，从而更加高效；需要求解的系数减少，更不容易过拟合。

为了求解$\boldsymbol{\beta}_T$，仅仅依赖标定样本项是不够的，因为标定样本数量通常较少，能够最小化标定样本项的$\boldsymbol{\beta}_T$将有无穷多个，容易造成过迁移。为了加强对$\boldsymbol{\beta}_T$的控制，需要加入**模型相似约束项**。该项反映了我们的先验知识，即两个领域是相关的，因此两个模型应当相似。为了在模型训练之前尽量缩小两个领域的差异，可以先用标准正态变换（SNV）对特征进行初步的标准化。在一些多任务学习算法中[88,89]，模型相似性约束采用的是$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|_2^2$。这种约束要求两个模型严格相似，因此在领域间差异大时较不容易满足。我们采用的模型相似性约束$\|X_S\boldsymbol{\beta}_S - X_S\boldsymbol{\beta}_T\|_2^2$则有所放松，通过对样本的投影间接要求模型相似。在本章的实验中，该约束的效果优于上述传统约束。

加入 **SEMI 正则项**的目的是减小模型对标准化误差较大的特征的依赖。式中的 w 是归一化后的式（4-2）中的标准化误差 SE，归一化是为了便于后续参数调整。SE 是经过 SNV 处理后的两个领域标定样本的均方根偏差，将 SE 的均值归一后即得到 w。

最后讨论式（4-4）中的**预测损失函数**$\mathcal{L}$。该函数可以是各种分类或回归模型中的损失函数。不失一般性，这里简介两种常用的函数：逻辑回归分类器中的对数损失函数，以及岭回归模型中的平方损失函数。对于**对数损失函数**，记

$\mathbf{x}^{(i)} \in \mathbf{R}^m$ 为第 i 个训练样本，$y^{(i)} \in \{0,1\}$ 为其标签。$X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}]^T$，$\mathbf{y} = [y^{(1)}, \ldots, y^{(n)}]^T$。在二分类问题中，逻辑回归分类器的决策函数为 sigmoid 函数 $h_{\boldsymbol{\beta}}(\mathbf{x}) = 1/(1+e^{-\boldsymbol{\beta}^T\mathbf{x}})$，若 $h_{\boldsymbol{\beta}}(\mathbf{x}) \geq 0.5$，则判断 $\mathbf{x}$ 为正类，否则判断为负类。对数损失函数的形式为[90]：

$$\mathcal{L}_L(X, \mathbf{y}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log h_{\boldsymbol{\beta}}\left(\mathbf{x}^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_{\boldsymbol{\beta}}\left(\mathbf{x}^{(i)}\right)\right)\right] \tag{4-5}$$

将上式代入 TCTL 的目标函数，可以得到对数损失的 TCTL 目标函数：

$$\begin{aligned} J_L(\boldsymbol{\beta}_S, \boldsymbol{\beta}_T) &= \mathcal{L}_L(X_S, y_S, \boldsymbol{\beta}_S) + \frac{\lambda_1}{2n_t}\left\|T_S\boldsymbol{\beta}_S - T_T\boldsymbol{\beta}_T\right\|_2^2 \\ &+ \frac{\lambda_2}{2n}\left\|X_S\boldsymbol{\beta}_S - X_S\boldsymbol{\beta}_T\right\|_2^2 + \frac{\mu}{2}\sum_{j=1}^{m} w_j^{\ 2}\left(\beta_{S,j}^{\ \ 2} + \beta_{T,j}^{\ \ 2}\right), \end{aligned} \tag{4-6}$$

其导数为：

$$\begin{aligned} &\frac{\partial J_L}{\partial \boldsymbol{\beta}_S} = \frac{1}{n}X_S^T\left(h_{\boldsymbol{\beta}}(X_S) - \mathbf{y}_S\right) + \frac{\lambda_1}{n_t}T_S^T\left(T_S\boldsymbol{\beta}_S - T_T\boldsymbol{\beta}_T\right) + \frac{\lambda_2}{n}X_S^T X_S\left(\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\right) + \mu W\boldsymbol{\beta}_S, \\ &\frac{\partial J_L}{\partial \boldsymbol{\beta}_T} = -\frac{\lambda_1}{n_t}T_T^T\left(T_S\boldsymbol{\beta}_S - T_T\boldsymbol{\beta}_T\right) - \frac{\lambda_2}{n}X_S^T X_S\left(\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\right) + \mu W\boldsymbol{\beta}_T, \\ &W = \mathrm{diag}(w_1^2, \ldots, w_m^2). \end{aligned} \tag{4-7}$$

该问题可以用数值优化算法来求解，例如共轭梯度法等。对于多类别分类问题，可以使用一对多的策略将其转化为多个二分类问题，然后比较测试样本在每个模型中的决策函数的大小来判断其类别。

对于回归问题，本章采用**平方损失函数**，代入 TCTL 的目标函数后可得新的目标函数为：

$$\begin{aligned} J_S(\boldsymbol{\beta}_S, \boldsymbol{\beta}_T) &= \frac{1}{2n}\left\|X_S\boldsymbol{\beta}_S - \mathbf{y}_S\right\|_2^2 + \frac{\lambda_1}{2n_t}\left\|T_S\boldsymbol{\beta}_S - T_T\boldsymbol{\beta}_T\right\|_2^2 \\ &+ \frac{\lambda_2}{2n}\left\|X_S\boldsymbol{\beta}_S - X_S\boldsymbol{\beta}_T\right\|_2^2 + \frac{\mu}{2}\sum_{j=1}^{m} w_j^{\ 2}\left(\beta_{S,j}^{\ \ 2} + \beta_{T,j}^{\ \ 2}\right) \end{aligned} \tag{4-8}$$

令上式的导数为 0，可以得到其解析解为

$$\begin{aligned} &\begin{pmatrix}\boldsymbol{\beta}_S \\ \boldsymbol{\beta}_T\end{pmatrix} = (A_1 + A_2 + A_3)^{-1}\mathbf{b}，\text{其中} \\ &A_1 = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix},\ A_2 = \frac{\lambda_1}{n_t}\begin{pmatrix} T_S^T T_S & -T_S^T T_T \\ -T_T^T T_S & T_T^T T_T \end{pmatrix}, A_3 = \begin{pmatrix} \lambda_2 P + \mu W & -\lambda_2 P \\ -\lambda_2 P & \lambda_2 P + \mu W \end{pmatrix}, \\ &\mathbf{b} = \frac{1}{n}\begin{pmatrix} X_S^T\mathbf{y}_S \\ 0 \end{pmatrix};\ P = \frac{1}{n}X_S^T X_S,\ W = \mathrm{diag}(w_1^2, \ldots, w_m^2) \end{aligned} \tag{4-9}$$

上面的讨论没有涉及模型中的常数项。实际试验中我们可以在特征中加入一维常数 1，从而使得学习到的模型中包含常数偏置。但需要注意 SEMI 正则项不应对该常数偏置进行惩罚。

4.5.2　多任务学习

TCTL 仅能够利用两个领域的信息。在实际应用中，当电子鼻被批量制造时，每个新设备都代表了一个新领域，每个领域与其他领域数据分布不同但相关。当电子鼻被长时间使用时，每一个使用时间段也代表了一个领域，前提是时间段内的数据漂移较小、可以忽略。如果假设时变漂移较平缓的话，可以认为每个领域的数据分布都与其前一个领域比较相似。为了利用多个领域的信息，我们将双领域的 TCTL 拓展到多领域的 TMTL。TMTL 比 TCTL 更有优势：由于一个目标领域的标定样本通常数量较少且存在噪声，因此 TCTL 更容易过迁移；而 TMTL 同时利用多个领域的标定样本，增加了信息量；多个领域的模型可以互相约束，从而稀释了标定噪声的影响，缓解了过迁移问题，提高了所有模型的泛化性能。另外，同时学习多个模型，也提高了效率。

由于设备差异和时变漂移的先验信息不同，我们设计了两种模型迁移策略：**并行 TMTL** 和**串行 TMTL**。**并行 TMTL** 假设多个领域彼此相似，适用于设备差异的情况。这里用下标 0 标记源领域的变量，下标 k 标记第 k 个目标领域的变量，例如 T_k 就代表第 k 个新设备或时间段采集的标定样本。设目标领域共有 d 个。则并行 TMTL 的目标函数可以表示为：

$$\begin{aligned}\min_{\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d} & \mathcal{L}(X_0,\mathbf{y}_0,\boldsymbol{\beta}_0)+\lambda_1\sum_{k=1}^{d}\left\|T_0\boldsymbol{\beta}_0-T_k\boldsymbol{\beta}_k\right\|_2^2 \\ & +\lambda_2\sum_{k=0}^{d}\left\|X_0\left(\boldsymbol{\beta}_k-\frac{1}{d+1}\sum_{r=0}^{d}\boldsymbol{\beta}_r\right)\right\|_2^2+\mu\sum_{k=0}^{d}\sum_{j=1}^{m}w_{k,j}^2\cdot\beta_{k,j}^2\end{aligned} \tag{4-10}$$

这是 TCTL 在多目标领域情况下的自然推广。上式的标定样本项要求每个目标领域的标定样本与源领域的标定样本在被各自的模型投影后互相对齐；模型相似约束项要求每个模型都与平均模型相似[88]。在 SEMI 正则项中，第 k 个目标领域中特征 j 的惩罚系数 $w_{k,j}$ 与 TCTL 中求法相同，源领域中特征 j 的惩罚系数 $w_{0,j}$ 是所有目标领域该特征惩罚系数的平均。

串行 TMTL 适用于时变漂移，它与并行 TMTL 的主要区别在于，在模型相似约束项中，它要求每个模型与其前一个时间段的模型相似，即：

$$\min_{\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d} \mathcal{L}(X_0,\mathbf{y}_0,\boldsymbol{\beta}_0)+\lambda_1\sum_{k=1}^{d}\left\|T_0\boldsymbol{\beta}_0-T_k\boldsymbol{\beta}_k\right\|_2^2 \\ +\lambda_2\sum_{k=0}^{d}\left\|X_0\left(\boldsymbol{\beta}_k-\boldsymbol{\beta}_{k-1}\right)\right\|_2^2+\mu\sum_{k=0}^{d}\sum_{j=1}^{m}w_{k,j}^2\cdot\beta_{k,j}^2 \tag{4-11}$$

这种约束反映了时变漂移的平缓先验，它使得学习得到的模型不会“变化太快”，因为快速变化的模型很可能是由于受到了标定噪声的干扰。标定噪声的一个例子见图 4.8，可以发现随着传感器的老化，标定样本的响应并非单调下降，而是有一定程度的振荡。此时如果使用 TCTL 进行源领域和目标领域一对一的迁移，这种噪声就会带来很大的影响。而使用串行 TMTL 对前后模型相似性进行约束后，由标定噪声引起的模型跳变就可以被平滑。

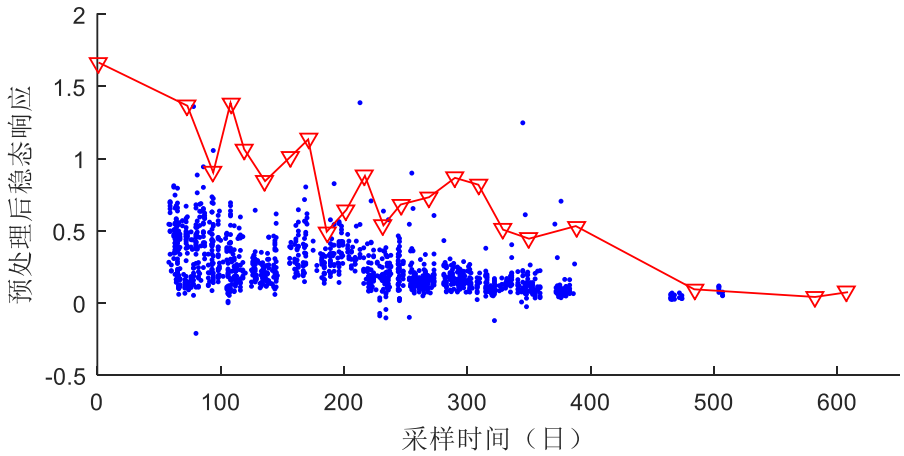


图4.8　呼气分析数据集中的标定样本噪声示例，蓝色为呼气样本，红色为10ppm丙酮的标定样本，图示为TGS826传感器的响应

在实际应用中，可能有两种补偿时变漂移的方式：离线补偿和在线补偿。在离线补偿中，所有时间段的样本被放在一起进行补偿，这意味着后一时间段采集的标定样本可以为前一时间段的模型提供信息，因此可以获得更好的补偿效果。在这种方式中，通过优化式（4-11），可以同时获得所有时间段的模型。而在线方式要求对样本进行实时分析，在训练当前时间段的模型时，只能利用之前时间段采集的标定样本。在这种方式中，可以在每次采集了标定样本后重新优化式（4-11），获得最新的模型。

最后，我们讨论设备差异和时变漂移同时存在的情况，如图 4.9 所示。每台设备都在不同的时刻采集了若干组标定样本。在 TMTL 中，每组标定样本事实上都对应了一个模型。为了学习这些模型，可以把串行和并行 TMTL 相结合。首先，在标定样本项中，将最初始的设备采集的第一组标定样本作为参考，使其他组标定样本经过各自模型投影后向其对齐。之后，在模型相似约束项中，使每个模型与其同设备的前一个模型相似，而每个设备的第一个模型应与它们的均值

$\mathrm{mean}(\boldsymbol{\beta}_{k,1})$ 相似，如图 4.9 中箭头所示。

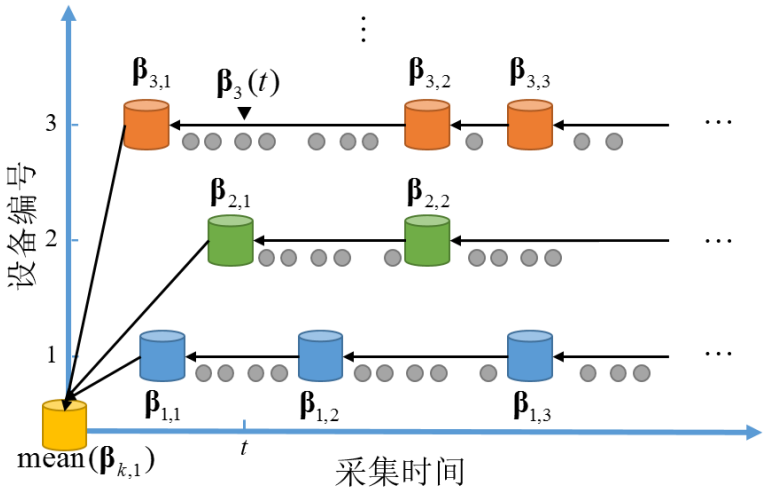


图4.9　多个设备长时间样本采集图示。每行代表一个设备采集的样本，样本的横坐标为采集时间。每个圆柱形代表一组标定样本，同时也对应了一个模型 $\boldsymbol{\beta}$。圆点代表普通样本。箭头代表模型的相似性关系，箭尾的模型应与箭头指向的模型相似。

4.5.3　动态模型策略

在上一节的串行 TMTL 算法中，虽然利用多个时间点采集的标定样本学习了多个模型，但是并没有指定在某一时刻 t 采集的样本应该使用哪一个模型。在一些传统方法中[77-79,89]，需要将不同时间采集的样本分成若干个批次，对不同批次应用不同的模型。然而，这种策略会带来如何划分批次的问题，还丢失了同一批次内样本的时间先后顺序信息，忽略了批次内的漂移。为了解决这些问题，我们提出一种动态模型策略。假设时变漂移平缓变化，那么有理由让模型也随着时间动态平滑变化。具体来说，可以令第 i 个设备在 t 时刻的模型 $\boldsymbol{\beta}_i(t)$ 为该设备在所有标定样本处的模型的加权组合，即

$$\begin{aligned}\boldsymbol{\beta}_i(t) &= \sum_j c_{i,j}(t)\boldsymbol{\beta}_{i,j},\\ c_{i,j}(t) &= \frac{\exp\left(-\gamma(t-t_{i,j})^2\right)}{\sum_j \exp\left(-\gamma(t-t_{i,j})^2\right)}.\end{aligned} \tag{4-12}$$

其中 $t_{i,j}$ 为设备 i 的第 j 组标定样本的采集时间，$c_{i,j}(t)$为第 j 个模型在 t 时刻的权重。t 与 $t_{i,j}$ 越靠近，第 j 个模型的权重越大。$c_{i,j}(t)$中的分母起到了归一化的作用。应用上式，即可以对某一时刻采集的测试样本应用针对该时刻的模型进行预测。通过这种策略，传统的多任务学习框架被灵活地拓展到领域分布连续变化的场合。另外，这种策略对模型进行了平滑，进一步缓解了过迁移的问题。调整上式中的带宽系数 γ 可以改变平滑的程度。

利用相似的策略，我们还可以将 SNV 也拓展到动态 SNV，使得其能够对连续的样本流进行特征预处理。具体来说就是将每组标定样本的均值和方差利用式（4-12）进行加权组合，然后用来归一化特定时刻采集的被测样本。

在应用动态模型策略后，串行 TMTL（包括串并行结合的 TMTL）的目标函数中源领域有标签样本也可以使用式（4-12）的动态模型进行投影。这样就可以允许有标签训练样本中也包含漂移，也就是不要求有标签样本一定要在第一台设备的初始阶段采集。使用者可以随时将新采集的有标签样本加入模型重新训练。

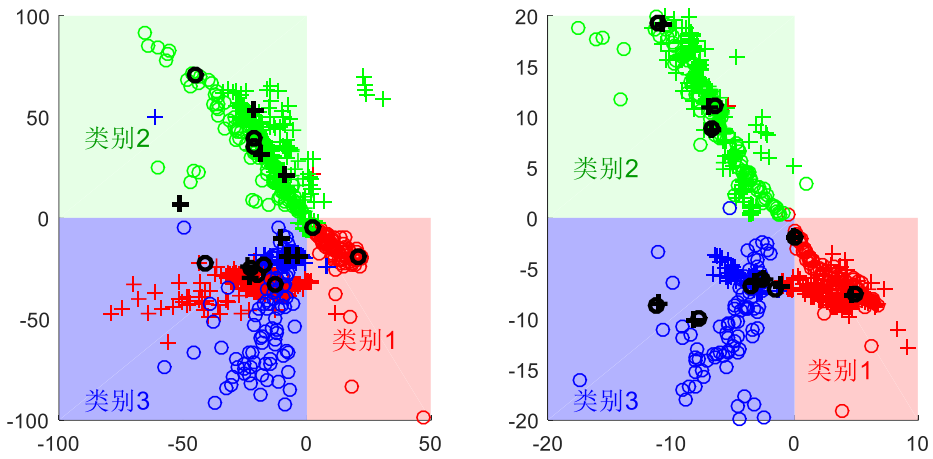
4.6　实验结果和讨论

本节将在 4.3 节介绍的三个漂移数据集上测试本章提出的算法，并与典型的漂移补偿和多任务学习算法进行比较，同时对比本章算法的不同实现方式的效果。

4.6.1　阵列漂移数据集

本数据集的构成及测试方法已于 4.3 节介绍。由于本数据集并没有直接提供每个批次的标定样本，因此需要手工构造各批次的候选标定样本集，然后利用 4.1.3 节介绍的 LLR 算法进行选择。第 k 批次（$k = 2, \ldots, 10$）的候选标定样本集定义为批次 1 和 k 之间的重叠样本，即气体种类和浓度都相同的样本。然后，利用标准正态变换（SNV）对每个批次的样本根据该批次的标定样本进行预处理。最后，对批次 2 至 10 的样本应用漂移补偿算法并计算分类正确率。TMTL 可以利用多个批次的样本辅助某一个批次的补偿。我们采用之前提到的在线补偿方式，即对批次 k 进行补偿时，利用批次 1 的有标签样本和批次 1 至 k 的标定样本训练 TMTL 模型。

为了直观地展示本章算法的效果，我们选取本数据集的批次 1 和 2 中三个类别的样本进行了两组实验，结果绘于图 4.10。分图(a)的实验中，用逻辑回归（LR）在批次 1 上训练两个分类模型，然后利用它们对批次 1 和 2 的样本进行投影。由于漂移的存在，批次 1 的模型不适用于批次 2 的样本，部分批次 2 的样本落在了错误的区域（类别 1 的样本落在了类别 3 的区域）。可以发现两个批次的标定样本也并没有对齐。分图(b)的实验基于批次 1 的有标签样本和两个批次的标定样本用 TMTL 同时学习两个批次的模型。可以发现两个批次的标定样本在模型层面较好地对齐了，而大部分被测样本也都落在了正确的区域。



(a) 批次1、2的样本均由批次1的LR模型投影

(b) 用TMTL同时学习两个批次的模型，每个批次的样本由各自的模型投影

图4.10　TMTL效果图示。除黑色粗线样本代表标定样本外，不同颜色的样本代表不同类别。圆圈代表批次1的样本（源领域），加号代表批次2的样本（目标领域）。不同着色区域代表每个类别样本的正确分布区域

图 4.11 对比了 TCTL、TMTL 与仅 SNV（仅用 SNV 对特征进行预处理）[19]、特征标准化[92]、Temporal Group Lasso（TGL）[89,93]、和正则化多任务学习（RMTL）[88]等算法在不同数量的辅助样本（标定样本或有标签目标领域样本）时的性能。图中每个数据点都是批次 2 至 10 的平均分类正确率。所有算法的参数都进行了网格法搜索优化。可以发现所有算法的效果在辅助样本数量 n_t 增加时都有所改善，而在 $n_t > 10$ 后趋于平稳。仅 SNV 和特征标准化这两种方法采用逻辑回归对标准化后的目标领域样本进行分类。这两种特征层面的漂移补偿算法在漂移较复杂时存在欠迁移问题，补偿精度有限。TGL 和 RMTL 是两种基于目标领域有标签样本的多任务学习算法，其效果与 TCTL 相当。本数据集包含时变漂移，数据已被分为批次，因此采用串行 TMTL 而不包含动态模型策略。串行 TMTL 在所有 n_t 值时都取得了最好的正确率。TMTL 优于 TCTL，体现出多任务相对于双任务的优势。在对批次 k 进行漂移补偿时，TCTL 仅用到了批次 1 和 k 的标定样本，而 TMTL 能够利用批次 1 至 k 的所有标定样本，因此引入了更多信息。串行 TMTL 能够有效利用时序平缓先验，从而减轻补偿过程中的过迁移问题。

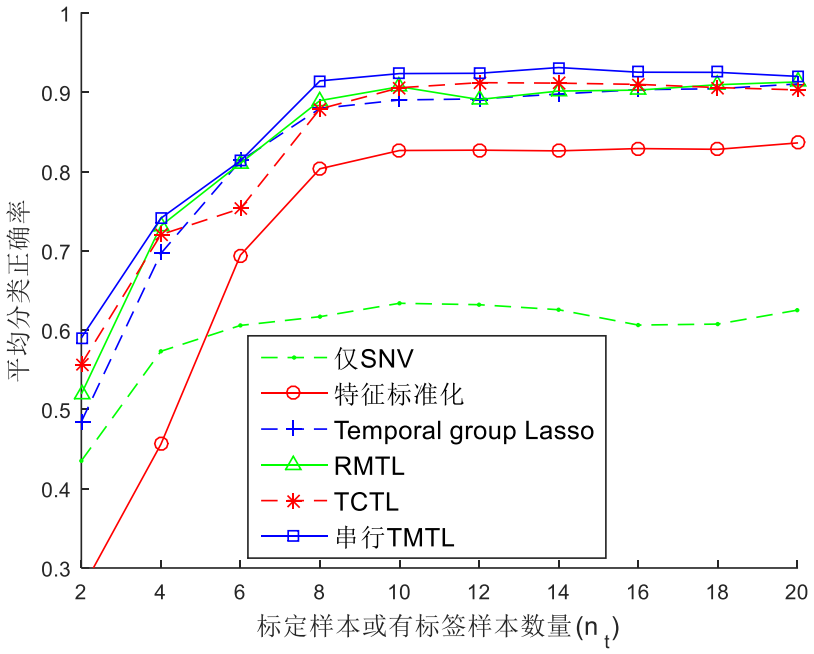


图4.11　阵列漂移数据集上的算法性能比较

表 4.2　阵列漂移数据集各目标批次上的算法性能比较（%）

算法	批次 2	3	4	5	6	7	8	9	10	平均
无补偿	88.59	66.96	40.99	54.82	43.22	44.40	31.63	45.74	39.11	50.61
CC-PCA[74]	90.92	40.86	47.20	59.39	56.74	56.71	36.39	45.32	37.72	52.36
OSC[75]	88.10	66.71	54.66	53.81	65.13	63.71	36.05	40.21	40.08	56.50
ML-comGFK[79]	80.25	74.99	78.79	67.41	77.82	71.68	49.96	50.79	53.79	67.28
Ensemble[78]	74.36	87.83	93.79	95.43	69.17	69.72	**91.84**	76.38	65.50	80.45
DAELM-S[77]	87.98	95.74	85.16	95.99	94.14	83.51	86.90	**100.0**	53.62	87.00
TCTL	97.35	95.46	90.68	**98.48**	93.22	**93.91**	89.12	87.02	69.97	90.58
TMTL(直接相似约束)	**97.51**	98.74	**93.79**	96.95	95.04	90.51	90.14	92.55	69.72	91.66
并行 TMTL	97.35	97.16	93.17	97.46	**95.96**	91.70	90.14	90.85	**73.28**	91.90
串行 TMTL	97.35	**98.80**	90.06	**98.48**	95.35	91.50	**91.84**	96.38	71.56	**92.37**

更多的实验结果列于表 4.2，其中详细列出了每个目标领域批次上的分类正确率。“无补偿”算法直接将批次 1 上训练的模型应用于批次 2 至 10，其正确率较低（尤其是批次号较大时），体现了时变漂移的带来的严重干扰。基于 PCA 的成分校正（CC-PCA）和正交信号补偿（OSC）这两种算法的结果同样较差，可能是因为这两种基于成分校正的算法无法准确分开包含漂移和有用信息的数据

成分。我们直接从原文引用了组合测地线流核学习（ML-comGFK）、分类器集成（ensemble）和领域适配极限学习机（DAELM-S）这三种算法的结果。后两者尽管正确率较高，但都需要较多目标领域的有标签样本：分类器集成要求目标批次之前的批次全部为有标签样本，而 DAELM-S 需要在每个目标批次选择 30 个有标签样本。

在表格的后 4 行列出了本章算法的结果，每个目标领域使用了 10 个标定样本（n_t = 10）。可以发现，相对于前人算法，本章的算法使用了更少的辅助样本获得了更高的正确率。在“TMTL(直接相似约束)”中，将串行 TMTL 的模型相似约束项由本章采用的$\|X_S\boldsymbol{\beta}_S - X_S\boldsymbol{\beta}_T\|_2^2$换为传统的$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|_2^2$，见 4.5.1 节。我们发现其效果略差于本章采用的约束项。另外，串行 TMTL 略优于并行 TMTL，验证了串行迁移策略的有效性。

为了观察结合标准化误差的模型改进策略（SEMI）的效果，表 4.3 对比了特征标准化算法和 TMTL 在加入 SEMI 前后的正确率。在无 SEMI 的算法中，采用常数值的 Tikhonov 正则项代替 SEMI 正则项。可以看出加入 SEMI 后，两种算法在几乎各个目标批次的正确率都有所提高。

表 4.3　阵列漂移数据集上有无 SEMI 策略的比较（%），1 表示无 SEMI，2 表示有 SEMI

算法	批次 2	3	4	5	6	7	8	9	10	平均
特征标准化 1	96.30	**98.87**	**94.41**	96.45	64.52	75.64	57.14	95.32	62.61	82.36
特征标准化 2	**96.46**	97.54	88.82	**96.95**	**67.04**	**78.80**	**60.20**	**97.66**	**66.64**	**83.35**
TMTL 1	96.46	97.35	**95.65**	97.97	95.04	84.83	82.31	93.19	70.78	90.40
TMTL 2	**97.35**	**98.80**	90.06	**98.48**	**95.35**	**91.50**	**91.84**	**96.38**	**71.56**	**92.37**

图 4.12 展示了目标函数（4-11）中正则项权重λ_1，λ_2的选择对平均正确率的影响。式（4-11）中的权重μ固定为10^{-3}。可以发现 TMTL 在较大参数范围内都可以获得较好的结果，但当λ_1过小或λ_2过大时效果变差。这是因为标定样本项的权重过小或过度强调模型间的相似性时，漂移无法得到有效补偿。

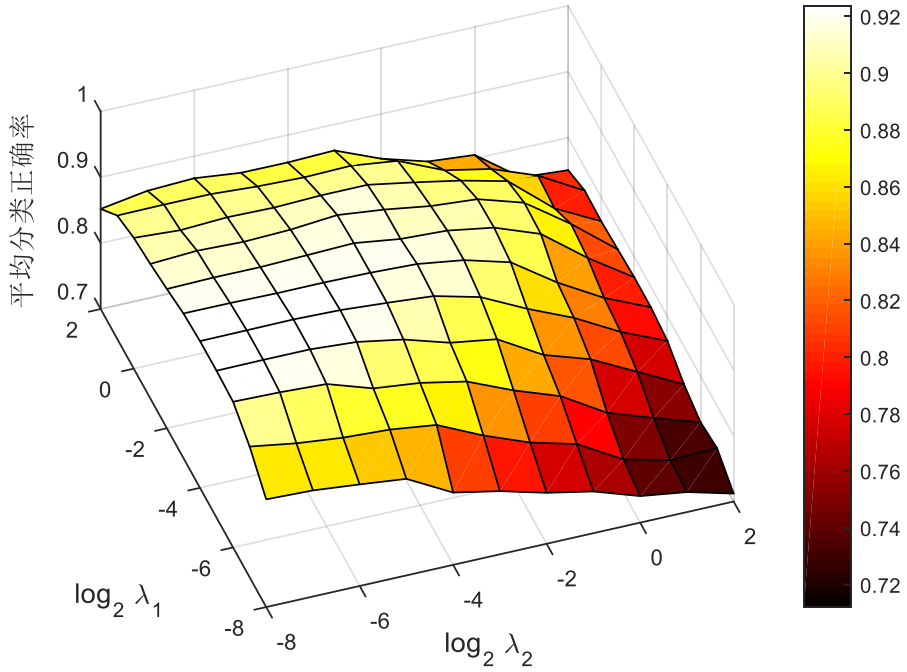


图4.12　参数选择对串行TMTL正确率的影响

4.6.2　Corn数据集

按照 4.3 节介绍的测试方式，我们对两台从设备分别测试四项回归任务的 RMSE。在每一折交叉验证中，从训练集中用 LLR 选择 7 个标定样本或有标签样本来辅助漂移补偿。所有算法的参数都经过了网格法寻优。结果如表 4.4 所示。可以发现当不进行漂移补偿时，预测误差较大，反映出设备差异的影响。所有漂移补偿算法都能明显降低 RMSE。模型层面的 DAELM 和 RMTL 的平均 RMSE 低于特征层面的 SNV 和特征标准化。对于 RMTL，我们采用了两种损失函数：平方损失和 SVR 算法中的ε-insensitive 损失。后者的结果从 Yu 等人的论文[70]中直接引用（该文献只提供了 Mp6 为从设备的结果）。基于平方损失的 TCTL 和 TMTL 能够获得最低的漂移补偿平均 RMSE。并行 TMTL 同时学习三个模型（一台主设备两台从设备），其结果优于同时学习两个模型的 TCTL，体现出多任务相对于双任务的优势。表格最后一行的“同设备训练”使用同一台从设备的样本训练和测试，因此不受设备差异的影响，可以被视为漂移补偿的目标结果。我们发现 TMTL 只使用 7 个标定样本就可以获得与该目标结果相当的精度。

最后，类似上一个数据集，我们对比了特征标准化算法和并行 TMTL 在加入 SEMI 前后的 RMSE，如表 4.5 所示。可以看出加入 SEMI 后，两种算法在几乎所有预测任务上的精度都有所改善。

表 4.4　Corn 数据集中各预测任务的 RMSE 比较

算法	Mp5 为从设备					Mp6 为从设备				
	水分	油脂	蛋白	淀粉	平均	水分	油脂	蛋白	淀粉	平均
无补偿	1.499	0.288	1.097	2.086	1.242	1.632	0.343	1.350	2.063	1.347
仅 SNV [19]	0.252	0.105	0.172	0.414	0.236	0.292	0.104	0.164	0.402	0.240
特征标准化[86]	0.253	0.106	0.173	0.418	0.238	0.294	0.103	0.165	0.407	0.242
DAELM-S [77]	0.223	0.111	0.153	0.366	0.213	0.249	0.115	0.157	0.386	0.227
RMTL(平方)	0.205	0.119	0.157	**0.322**	0.201	0.202	0.118	0.158	**0.335**	0.204
RMTL(SVR) [70]			-			0.235	**0.089**	0.160	0.357	0.210
TCTL	0.182	0.104	0.148	0.350	0.196	0.193	0.104	0.144	0.338	0.195
并行 TMTL	**0.182**	**0.098**	**0.127**	0.339	**0.186**	**0.190**	0.100	**0.130**	0.338	**0.189**
同设备训练	0.137	0.101	0.140	0.363	0.185	0.139	0.099	0.149	0.369	0.189

表 4.5　Corn 数据集上有无 SEMI 策略的比较（RMSE），1 表示无 SEMI，2 表示有 SEMI

算法	Mp5 为从设备					Mp6 为从设备				
	水分	油脂	蛋白	淀粉	平均	水分	油脂	蛋白	淀粉	平均
特征标准化 1	**0.253**	0.106	0.173	0.418	0.238	**0.294**	0.103	0.165	0.407	0.242
特征标准化 2	0.270	**0.097**	**0.148**	**0.409**	**0.231**	0.309	**0.092**	**0.139**	**0.399**	**0.235**
TMTL 1	0.196	0.100	0.145	0.373	0.204	0.219	0.103	0.145	0.367	0.208
TMTL 2	**0.182**	**0.098**	**0.127**	**0.339**	**0.186**	**0.190**	**0.100**	**0.130**	**0.338**	**0.189**

4.6.3　呼气分析数据集

本数据集的特点是同时包含设备差异和时变漂移，并且样本没有按照采样时间分成批次，因此我们只比较能够处理连续样本流的算法。本数据集的标定样本与被测样本种类不同（前者是标准气体，后者是呼吸气体），为了公平起见，我们只测试了基于标定样本的算法。按照 4.3 节介绍的实验方式，采用设备 1 采集的每类前 50 个样本作为有标签训练样本，基于在两台设备上定期采集的 45 组标定样本进行漂移补偿。考虑到本数据集漂移的复杂性，采用离线方式进行补偿，即将所有标定样本加入训练，同时获得所有时间段的模型。结果见表 4.6。

对于乘性差异校正（MDC），我们首先将标定样本相对于采样时间进行了平滑以减轻标定噪声，然后建立一个乘性校正系数的分段线性模型，以求出每个时

间点的校正系数。TCTL 只能处理离散的两个领域，因此没有包含在实验中。对于 TMTL，首先使用 4.5.3 节介绍的动态 SNV 对特征进行预处理，然后应用动态模型策略训练和测试 TMTL 模型。由表 4.6 可以看出，TMTL 获得了最高的平均 F-score，证明了该算法的有效性。加入 SEMI 的 TMTL 略优于不加入 SEMI 的 TMTL。

最后，我们还考虑了随机选取训练样本的实验方案，即在所有被测样本中随机选取每类的 50 个训练样本。训练样本不局限在设备 1 的前 50 个样本，而是包含各种采集设备和时间，因此样本的漂移可以在模型中被隐式考虑，不做漂移补偿即可得到较好的正确率，可以作为本数据集漂移补偿的一个目标结果。而在这种方式下如果再应用 TMTL，可以进一步提高正确率。这一点反映出 TMTL 可以利用标定样本中包含的信息有效地去除漂移的影响。

表 4.6　呼气分析数据集中各预测任务的 F-score 比较（%）

算法	**任务 1**	**2**	**3**	**4**	**5**	**平均**
无补偿	34.34	63.67	73.71	43.17	42.93	51.57
仅 SNV	63.20	71.49	75.87	64.36	66.55	68.30
MDC[73]	**83.33**	66.41	**87.58**	66.22	52.71	71.25
CC-PCA[74]	70.82	79.24	84.59	74.19	76.24	77.02
TMTL 无 SEMI	63.50	81.10	84.85	**88.75**	**85.48**	80.73
TMTL 有 SEMI	73.81	**86.35**	84.42	81.85	82.24	**81.73**
随机选取训练样本	87.85	85.44	90.10	85.09	**84.54**	86.60
随机选取训练样本+TMTL	**94.41**	**89.27**	**91.88**	**85.63**	83.48	**88.93**

4.7　本章小结

本章提出了两种漂移补偿算法：结合标准化误差的模型改进（SEMI）和基于标定样本的多任务学习（TMTL）。SEMI 利用加权正则项给标准化误差不同的特征以不同的惩罚系数，从而提高已有模型对漂移的鲁棒性。实验证明，将 SEMI 与特征标准化算法或 TMTL 结合后，能够提高补偿精度。TMTL 对不同设备或时间段采集的样本学习不同的模型。我们针对不同的漂移类型设计了不同的迁移策略：并行 TMTL 适用于设备差异，串行 TMTL 适用于时变漂移。为了更精确地补偿已知采样时间的样本，设计了一种动态模型策略，用已有模型的加权组合作为特定时刻的模型。

在两个包含复杂漂移的电子鼻数据集和一个包含设备差异的光谱数据集上，

本章提出的算法超越了被测的典型漂移补偿和多任务学习算法，且仅使用了较少数量的辅助样本（不多于 10 个标定样本）。本章算法的特点可以总结如下：

- 利用了漂移数据的特点。SEMI 利用了不同特征的稳定性和信噪比不同这个特点；TMTL 利用了不同设备以及相邻时间段数据分布的相似性；
- 通用性较好。SEMI 可以与多种模型结合；TMTL 是一个算法框架，可以搭配多种分类或回归损失函数；
- TMTL 与大部分现有多任务学习算法的不同之处在于，它依赖标定样本而非有标签样本将源领域的知识传递到目标领域，因此在漂移补偿问题中更加实用。以呼气分析系统为例，每制作一台新电子鼻都要采集健康人和病人的呼气样本用来训练显然是不现实的，而标定样本的采集就相对容易一些；
- 从迁移学习理论的角度来讲，多任务框架以及模型加权组合的应用，可以缓解由标定样本数量少和存在噪声带来的过迁移问题，提高模型的泛化性能；动态模型策略将多任务学习框架从离散领域之间的迁移拓展到了连续样本流的情况。

第5章　漂移补偿自编码器

在上一章中，我们提出了两种模型层面的漂移补偿算法，其精度较高，但需要对每种损失函数设计求解方法。在特征层面的漂移补偿算法中，经过补偿后的特征可以应用到多种预测模型，因此更容易使用。但传统的算法如特征标准化、成分校正等补偿精度有限。能否将结合两个层面方法的优点结合到一起呢？

在本章中，我们提出一种基于自编码器（autoencoder）的特征学习和漂移补偿算法，该算法同时具有模型层面算法的精度和特征层面算法的易用性。它在标定样本的辅助下对设备差异和时变漂移进行显式建模和补偿，学习得到的特征中漂移被减小、且具有更好的鉴别能力。

本章常用的符号列于表 5.1。

表 5.1　本章常用符号列表

符号	描述	符号	描述
m	原始特征维度、自编码器输入层单元数	m_d	背景特征维度
h	学习得到的特征维度、自编码器隐层单元数	n	样本总数
n_L	有标签样本数	c	标签数目
$n_{t,total}$	标定样本总对数	n_{dev}	设备数目
$\mathbf{x}$	被测样本的原始特征向量	$\tilde{\mathbf{x}}$	损坏后的 $\mathbf{x}$
$\mathbf{z}$	编码后的 $\mathbf{x}$、隐层输出	$\hat{\mathbf{x}}$	重构后的 $\mathbf{x}$
$\mathbf{t}$	标定样本的原始特征向量	$\mathbf{d}$	背景特征向量
σ	编码、解码阶段的激活函数	σ_{cor}	补偿层激活函数
$W,\mathbf{b}$	编码阶段的权值矩阵和偏置向量	f	编码器
$W',\mathbf{b}'$	解码阶段的权值矩阵和偏置向量	g	解码器
$W_S,\mathbf{b}_S$	有监督项中的权值矩阵和偏置向量	λ_1	有监督项的权重
W_D	基本 DCAE 中补偿项的权值矩阵	λ_2	标定样本项的权重
W_{D0},W_{D1}	含补偿层的 DCAE 中补偿项的权值矩阵	h_{cor}	补偿层单元数

5.1　相关工作

近年来，基于深度学习的特征学习算法在图像、语音、自然语言处理等领域都获得了很大的成功[94-96]。在电子鼻领域，Längkvist 等人尝试利用自编码器和受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）从电子鼻信号中学习特征，获得了比传统特征更好的效果[97-99]。然而，目前还没有文献利用这类算法同时进行特征学习和漂移补偿。

自编码器是一种被广泛应用的深度学习算法[94,95,98-100]。它通过对无标签样本进行重构，使得学习到的特征能够较好地代表原始特征；然后利用有监督微调（fine-tuning）提高特征的鉴别能力。非线性激活函数和多层堆叠策略的应用让特征能够描述数据中复杂、抽象的结构。在本节中，我们首先对自编码器算法进行简单介绍，然后回顾已有的基于自编码器的迁移学习算法。

5.1.1　自编码器

浅层自编码器的结构类似于一个单隐层的前馈神经网络（feed-forward neural network）。它接受原始特征向量 $\mathbf{x} \in \mathbf{R}^m$ 作为输入后，将其编码得到一个新的特征表达 $\mathbf{z} \in \mathbf{R}^h$，再将其解码得到重构的特征向量 $\hat{\mathbf{x}} \in \mathbf{R}^m$。编解码的过程可以表示为：

$$\mathbf{z} = f(\mathbf{x}) = \sigma(W\mathbf{x} + \mathbf{b}) \tag{5-1}$$

$$\hat{\mathbf{x}} = g(\mathbf{z}) = \sigma(W'\mathbf{z} + \mathbf{b}') \tag{5-2}$$

其中 $W \in \mathbf{R}^{h\times m}, W' \in \mathbf{R}^{m\times h}$ 为权值矩阵，$\mathbf{b} \in \mathbf{R}^h, \mathbf{b}' \in \mathbf{R}^m$ 为偏置向量。σ 为网络中的激活函数，可以为 sigmoid 函数、双曲正切 $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$、线性函数（即恒等函数）等。自编码器的目标函数为最小化重构误差[94]：

$$J_{AE}(W, \mathbf{b}, W', \mathbf{b}') = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i - \hat{\mathbf{x}}_i\right\|^2 \tag{5-3}$$

利用自编码器进行特征学习时，可以将网络的隐层输出当做学习到的特征，因为它能够较好地重构原始特征。当隐层节点数小于原始特征维度时，隐层即成为网络的一个“瓶颈”，隐层输出可以作为原始特征的一个压缩表达。将浅层自编码器进行堆叠可以得到栈式自编码器（stacked autoencoder），从而学习到抽象程度更高的特征。堆叠时可以采用逐层训练策略，将外层自编码器的隐层输出当作内层的输入，最后将最内层自编码器的隐层输出作为最终学习到的特征。在训练浅层自编码器时，权重绑定（weight tying）是一种常用的策略，即令 $W' = W^T$，从而减少需要学习的参数数目[94]。

上述训练过程没有用到样本的标签，因此常被称为无监督预训练。有监督微调通常在预训练结束后进行，即将网络的编码器部分当作一个前馈神经网络，并

且在其输出端再加一个输出层，该层的功能是预测有标签数据的标签。将预训练得到编码器的权值矩阵当作初始值，对编码器和输出层进行微调[100]。

去噪自编码器（denoising autoencoder）[100]是自编码器的一个改进算法。它首先随机将原始特征向量 $\mathbf{x}$ 的某些维度置零，得到其损坏版本 $\tilde{\mathbf{x}}$，然后利用式（5-1）和（5-2）得到 $\tilde{\mathbf{x}}$ 的重构，接着最小化该重构与未被损坏的 $\mathbf{x}$ 的差异。在这种策略下，学习到的特征会更加鲁棒，更能够描述数据分布中的有用信息[100]。

5.1.2　基于自编码器的迁移学习

当数据涉及多个领域时，有必要将自编码器与迁移学习结合以使学习到的特征更适用于目标领域。Kandaswamy 等人[101]提出首先用源领域的样本训练网络，然后保持一部分权值不变，再用目标领域的有标签样本微调另一部分网络权值，从而将源领域样本的部分信息与目标领域共享。另外一些算法不需要目标领域的有标签样本。Glorot[102]和 Bengio[95]等指出可以用多个领域的无标签样本共同训练自编码器。Chopra 等人[103]提出了一种领域插值方法，通过将源领域和目标领域样本按不同比例混合，在两个领域之间构造中间领域；然后对每个中间领域利用自编码器学习特征，最后将所有特征串接。这样学习到的特征中包含有领域间分布移动的信息。Glorot[102]和 Chopra[103]等人的方法尽管可以使学习到的特征具有描述多个领域数据的能力，但并不能显式缩小不同领域间特征分布的差异。

一些文献[104-108]显式考虑了领域间差异。Deng 等人[104]的思路与上一章涉及的多任务学习相似，利用不同领域的权值矩阵的相似性约束在领域间传递信息。Zhuang 等人[105]则修改自编码器的目标函数，使得不同领域学习到的特征的均值的 Kullback-Leibler（KL）散度较小，从而缩小学习到的特征的领域间差异。还有几种方法的思路与漂移补偿中的特征标准化算法有相通之处，都是采用与标定样本类似的样本对来辅助特征迁移。Zhou 等人[106]对每个领域分别利用自编码器学习特征，然后用“跨域样本对”学习两组特征之间的变换矩阵。Chen 等人[107]提出极限学习机自编码器，用于光谱仪的设备差异补偿。其算法本质是利用随机提取的从设备特征重构主设备特征，相当于基于极限学习机的特征标准化算法。与之相似的是 Kan 等人[108]的算法，他们设计了一个多层自编码器用来对人脸特征进行姿态补偿，在训练过程中，有姿态变化的人脸图像被用来重构无姿态变化的人脸图像，其训练样本对也可以视为标定样本。

5.2　背景特征

包括上一小节介绍的算法在内的多数迁移学习算法假设样本来自两个或更多的离散领域。这些算法可以被用来处理设备差异，因为每个设备代表了一个离

散的领域。然而在时变漂移问题中，样本流里特征分布的漂移是连续的。概念漂移适配算法可以处理这种问题，然而绝大多数该类算法[91,110]依赖于新采集的有标签样本更新现有模型，在实际应用中，获取数据的标签通常需要较大工作量。上一章提出的动态模型策略利用模型加权平均生成随时间变化的模型，在一定程度上解决了这个问题。那么，能否设计一种随时间变化特征学习算法呢？

更具体地来说，我们希望能够设计一种特征学习算法，学习到的特征中设备差异和时变漂移都得到了补偿。采集设备和采集时间都是样本的背景属性，这种属性与样本本身的标签无关，因此被称为“背景”。我们设计的算法中应当对学习到的特征进行补偿，补偿幅度的大小应该根据样本的背景属性来确定，从而实现设备差异和时变漂移的精确校正。

为了根据背景属性确定补偿幅度，首先需要定义一组“背景特征”来具体描述背景属性，即样本的采集设备和时间。如果仅仅考虑采集设备，可以使用哑变量对设备编号进行编码。假设有 n_{dev} 台设备，则背景特征 $\mathbf{d} \in \mathbf{R}^{n_{dev}}$ 。当样本由第 p 台设备采集时，令 $\mathbf{d}$ 的第 p 个分量为 1，其余分量为 0。如果将采集时间也纳入考虑的话，可以进一步把采样时间加入背景特征向量。当样本由第 p 台设备在 t 时刻采集时，令 $\mathbf{d} \in \mathbf{R}^{2n_{dev}}$ ，其中

$$d_q = \begin{cases} 1, & q = 2p-1; \\ t, & q = 2p; \\ 0, & \text{其他} \end{cases} \tag{5-4}$$

通过这种方式，就可以方便地将样本的背景属性都包含在 $\mathbf{d}$ 中。事实上，背景属性不仅限于采集设备和采集时间，还可以包括采集环境等，例如气体样本的温湿度。我们可以很容易地对背景特征进行拓展，使其包含更多的信息。配合合适的补偿算法和标定样本，就可以对这些干扰因素进行统一补偿，而不需要再对不同因素分别处理。背景特征也可以应用于普通的迁移学习问题中，例如样本的领域标签就可以用哑变量进行编码。合理地应用背景特征，将能使迁移学习的应用范围得到很大扩展，例如拓展到分布连续变化的情况。本章及下一章的漂移补偿算法都利用了背景特征。

5.3　漂移补偿自编码器

本节提出漂移补偿自编码器（Drift Compensation AutoEncoder，DCAE），旨在学习能代表原始特征、有鉴别能力、同时显式补偿了漂移的特征。DCAE 的学习需要四组数据：$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ 包含了所有有标签和无标签样本；$\mathcal{D}_L = \{(\mathbf{x}_j^{(L)}, \mathbf{y}_j^{(L)})\}_{j=1}^{n_L}$ 包含了所有有标签样本及其标签；$\mathcal{D}_S = \{\mathbf{t}_p^{(S)}\}_{p=1}^{n_{t,total}}$ 和 $\mathcal{D}_T = \{\mathbf{t}_q^{(T)}\}_{q=1}^{n_{t,total}}$ 分别是参考组标

定样本和所有标定样本的集合。$\mathbf{x},\mathbf{x}^{(L)},\mathbf{t}^{(S)},\mathbf{t}^{(T)}\in\mathbf{R}^{m}$，$\mathbf{y}^{(L)}\in\mathbf{R}^{c}$，$c$ 是标签数目。每个样本（$\mathbf{x}_i,\mathbf{x}_j^{(L)},\mathbf{t}_p^{(S)},\mathbf{t}_q^{(T)}$）都对应一个背景特征向量（$\mathbf{d}_i,\mathbf{d}_j^{(L)},\mathbf{d}_p^{(S)},\mathbf{d}_q^{(T)}$）。

需要指出的是，在应用了背景特征之后，我们并不需要像一般的迁移学习算法那样，指明每个样本 $\mathbf{x}$ 属于哪个领域，而只需要在该样本的背景特征中加入相关信息。但为了补偿漂移，仍需要指定一组标定样本作为参考。与上一章相似，我们指定初始的设备和时间段采集的一组标定样本作为参考，使其他组的标定样本都向参考组对齐。这些标定样本包含在 $\mathcal{D}_S$ 和 $\mathcal{D}_T$ 中。假设每组标定样本包含 $n_{t,gas}$ 个样本，一共采集了 $n_{t,group}$ 组。$\mathcal{D}_T$ 被定义为所有这些样本的集合，而 $\mathcal{D}_S$ 是参考组重复 $n_{t,group}$ 次。每个 $\mathcal{D}_T$ 中的样本 $\mathbf{t}_p^{(T)}$ 都和一个 $\mathcal{D}_S$ 中的 $\mathbf{t}_p^{(S)}$ 组成了一对，标定样本总对数 $n_{t,total}=n_{t,gas}n_{t,group}$。DCAE 可以同时利用所有标定样本中包含的信息。

5.3.1　网络结构与目标函数

本小节介绍基本版本的 DCAE，一种更适合处理复杂时变漂移的 DCAE 将在下一小节介绍。首先考虑浅层自编码器。我们希望隐层输出的特征中没有漂移，但自编码器以特征重构为优化目标，去除漂移后的隐层输出可能无法很好地重构带有漂移的原始特征。为此我们设计了如下方案：在原始自编码器的基础上加入一个新的权值矩阵 $W_D\in\mathbf{R}^{h\times m_d}$，其中 h 是隐层单元数，m_d 是背景特征维度。将 $W_D\mathbf{d}$ 作为补偿项，在编码阶段向隐层输入中加入，从而显式地对漂移进行补偿；之后在隐层向输出层的输出中对称地减去该补偿项，重新将漂移成分补回。式（5-1）、（5-2）因此变为了：

$$\mathbf{z}=f(\mathbf{x},\mathbf{d})=\sigma(W\mathbf{x}+W_D\mathbf{d}+\mathbf{b}) \tag{5-5}$$

$$\hat{\mathbf{x}}=g(\mathbf{z},\mathbf{d})=\sigma(W'(\mathbf{z}-W_D\mathbf{d})+\mathbf{b}') \tag{5-6}$$

这种设计使得隐层输出的特征中的漂移得到了补偿，而输出层的输出中又包含了漂移，方便进行特征重构。如果网络中采用线性激活函数，输出层输出将与原始自编码器相同。W_D 的第 i 行第 j 列相当于第 j 个背景特征对第 i 个隐层输出的补偿系数，因此 W_D 相当于对设备差异和时变漂移的一个显式建模。当使用深层自编码器时，我们向第一个隐层输入加入补偿项，从最后一个隐层的输出中将其减除，从而保证整个网络的隐层学习到的特征表达都是补偿了漂移的。

图 5.1 展示了一个由两个浅层自编码器堆叠而成的基本 DCAE 的网络结构。其中左侧部分与原始自编码器相同，右侧是我们加入的补偿项。样本的原始特征和背景特征（式（5-4））分别由最下方的输入层输入，学习到的特征由中间隐层输出。补偿项被加入第一个隐层的输入，并被从最后一个隐层的输出中减除。

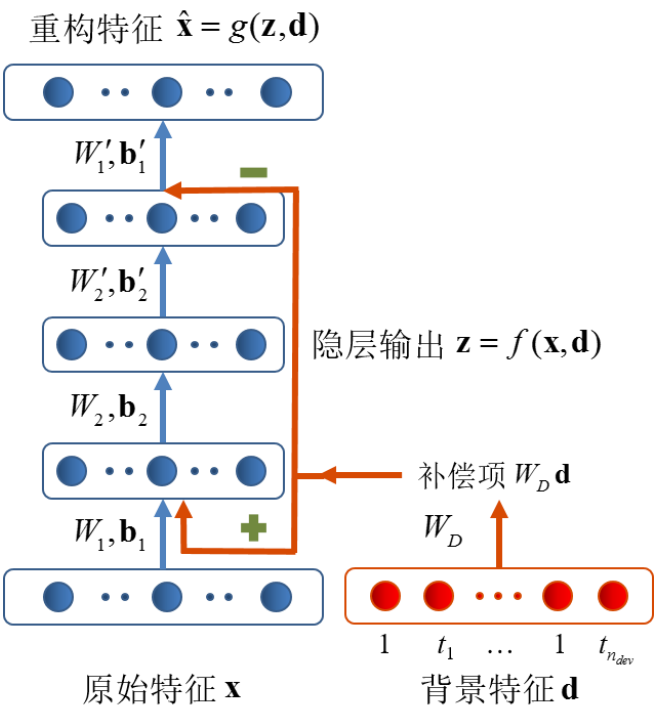


图5.1　基本DCAE的网络结构

为了学习 W_D 以及 DCAE 中包含的其他权值矩阵和偏置向量，我们设计了如下的目标函数：

$$
\begin{aligned}
& J_{DCAE}(W,\mathbf{b},W',\mathbf{b}',W_D,W_S,\mathbf{b}_S) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i - g(f(\mathbf{x}_i,\mathbf{d}_i),\mathbf{d}_i)\right\|^2 \\
& + \frac{\lambda_1}{n_L}\sum_{j=1}^{n_L}\mathcal{L}\left(f(\mathbf{x}_j^{(L)},\mathbf{d}_j^{(L)}),\mathbf{y}_j^{(L)};W_S,\mathbf{b}_S\right) \\
& + \frac{\lambda_2}{n_{t,total}}\sum_{p=1}^{n_{t,total}}\left\|f(\mathbf{t}_p^{(S)},\mathbf{d}_p^{(S)}) - f(\mathbf{t}_p^{(T)},\mathbf{d}_p^{(T)})\right\|^2
\end{aligned}
\tag{5-7}
$$

上式中，$W,\mathbf{b}$ 和 $W',\mathbf{b}'$ 分别代表编码和解码阶段网络中的权值矩阵和偏置向量；$W_S \in \mathbf{R}^{c\times h},\mathbf{b}_S \in \mathbf{R}^c$ 为有监督项的权值矩阵和偏置向量；f 和 g 分别代表编码和解码函数，见图 5.1；$\lambda_1, \lambda_2 > 0$ 为权重系数。通过最小化上式，网络中的所有系数可以被同时学习得到。

上式同时考虑了样本的重构、有监督预测和漂移补偿，因此可以学习到能够代表原始特征、具有鉴别能力以及补偿了漂移的特征。该式可以被分成三个部分。第一部分是全部样本的**重构误差**，该项与原始自编码器相似，不同之处在于 f 和 g 由式（5-5）和（5-6）定义，其中考虑到了漂移。第二部分是**有监督项**，它利用 $W_S,\mathbf{b}_S$ 对有标签样本的隐层输出进行预测，并统计预测误差，其中损失函数 $\mathcal{L}$ 需要根据问题来确定，例如分类问题通常使用 softmax 函数[105]，而回归问题可

以使用平方损失函数。

第三部分是**标定样本项**，是 DCAE 中漂移补偿的关键。它要求源领域和目标领域的对应标定样本具有相似的隐层输出。W_D 的学习主要依赖这一项；$W,\mathbf{b}$ 也会因为这一项的约束而更偏向于抽取原始特征中对漂移鲁棒的成分。还有一种可行的学习策略是使用补偿后的标定样本的隐层输出去重构参考组的标定样本。这种策略与有些文献[107,108]类似，但在实验中效果不如我们提出的策略。这可能是由于这种重构策略涉及到编码和解码两个阶段，需要学习的系数更多，因此不容易找到最优解。另外，训练好的自编码器只在与训练样本相似的数据中有较小的重构误差[94]，但标定样本有时与训练样本有差异，因此不一定能被较好地重构。

5.3.2　处理复杂时变漂移

DCAE 的第一隐层中的每个单元都会得到一个补偿值。假设样本的背景特征为 $\mathbf{d}$，那么第 i 个隐层单元得到的补偿值就是 $W_D\mathbf{d}$ 的第 i 个元素。由于 $\mathbf{d}$ 中包含了样本的采集设备和时间，因此补偿值都与设备有关并且与采集时间线性相关。然而，在大部分现实应用中，由于时变漂移的成因复杂，时间与时变漂移的关系往往是非线性的，甚至不是单调的（图 4.3，图 4.4）。线性的时变漂移补偿项可能不足以处理这种漂移。因此，我们提出一种含补偿层的 DCAE 以更好地处理复杂时变漂移。

图 5.2 展示了含补偿层的 DCAE 的网络结构。可以看出，我们在补偿项之前插入了一个补偿层，补偿项因此变成了 $W_{D1}\sigma_{cor}(W_{D0}\mathbf{d})$，式（5-5）、（5-6）中的编码和解码表达式也需要进行相应改动。$W_{D0}\in\mathbf{R}^{h_{cor}\times m_d}$ 是从背景特征到补偿层的权值矩阵，$W_{D1}\in\mathbf{R}^{h\times h_{cor}}$ 是从补偿层到 DCAE 第一隐层的权值矩阵，h_{cor} 为补偿层单元数。补偿层激活函数 σ_{cor} 应该是非线性函数，因为线性函数会导致补偿项与时间线性相关，从而与基本 DCAE 等价。背景特征中的常数 1（式（5-4））可以视为与设备相关的常数项，因此补偿层不包含偏置向量。非线性的 σ_{cor} 和一分为二的补偿权值矩阵给 DCAE 带来了更强的时变漂移补偿能力。h_{cor} 应根据时变漂移的复杂程度设置，理论上更多的单元能够拟合（相对时间）更复杂的补偿函数。

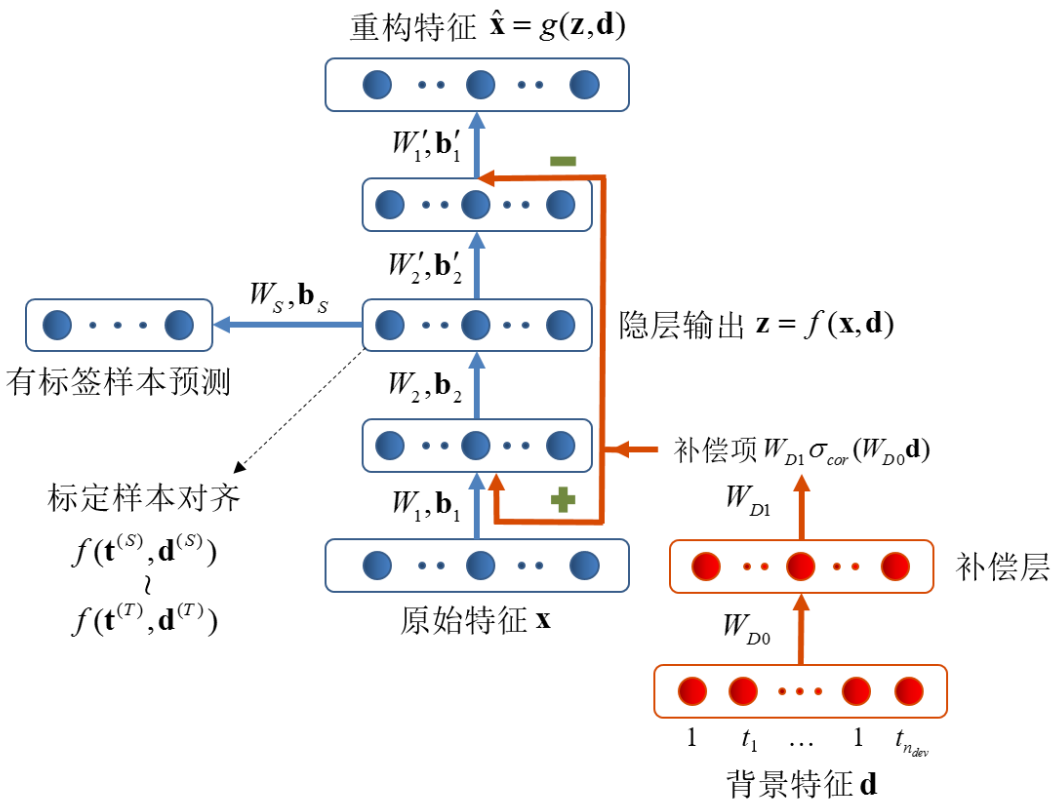


图5.2　含补偿层的DCAE的网络结构图示。左侧是优化目标中的有监督项和标定样本项示意

5.3.3　训练流程

无监督预训练+有监督微调是自编码器应用中经常采用的训练流程。对于DCAE，我们也采用类似的流程。首先，用所有样本逐层预训练一个栈式去噪自编码器[100]。本章中将特征损坏概率设为 10%，并且应用权重绑定策略以减少需要学习的参数数目。然后，利用所有有标签样本对网络的编码器部分进行微调。这两步都与一般的自编码器应用相同，没有考虑补偿项。它们的目的是将网络中的权值矩阵和偏置向量初始化到一个离 DCAE 最优解更近的区域，这里最优指的是泛化能力最优而不一定是目标函数值最小[94]。最后，利用前两步得到的网络初始化 DCAE 的 $W, \mathbf{b}, W', \mathbf{b}'$，随机初始化 W_S, W_{D1}，将 $\mathbf{b}_S', W_{D0}$ 的元素初始化为 0，然后利用梯度下降或共轭梯度等数值优化算法[109]最小化 DCAE 的目标函数式（5-7）。此时不使用去噪和权重绑定策略。这就是说，我们事实上采用了两步预训练，一步无监督，一步有监督。该训练流程的作用将在 5.4.5 节进行分析。

当网络优化完成后，可以将有标签样本和它们的背景特征输入到 DCAE，用得到的隐层输出和样本标签训练一个预测模型，用来对无标签样本的隐层输出进行预测。完整的训练流程总结见算法 2。

算法 2：漂移补偿自编码器 DCAE

输入：所有样本集合$\mathcal{D}$，有标签样本集合$\mathcal{D}_L$，标定样本集合$\mathcal{D}_S$和$\mathcal{D}_T$。所有样本的采样设备和时间已知。

输出：$\mathcal{D}$中样本的标签预测。

1. 使用$\mathcal{D}$无监督逐层预训练一个栈式去噪自编码器，然后使用$\mathcal{D}_L$对其进行微调；
2. 利用上一步得到的网络初始化 DCAE；
3. 按式（5-4）构造每个样本的背景特征；
4. 训练 DCAE，即使用$\mathcal{D}, \mathcal{D}_L, \mathcal{D}_S$和$\mathcal{D}_T$和它们的背景特征优化式（5-7）；
5. 使用$\mathcal{D}_L$的隐层输出训练预测模型，然后对$\mathcal{D}$的隐层输出进行预测，得到标签。

5.4　实验结果和讨论

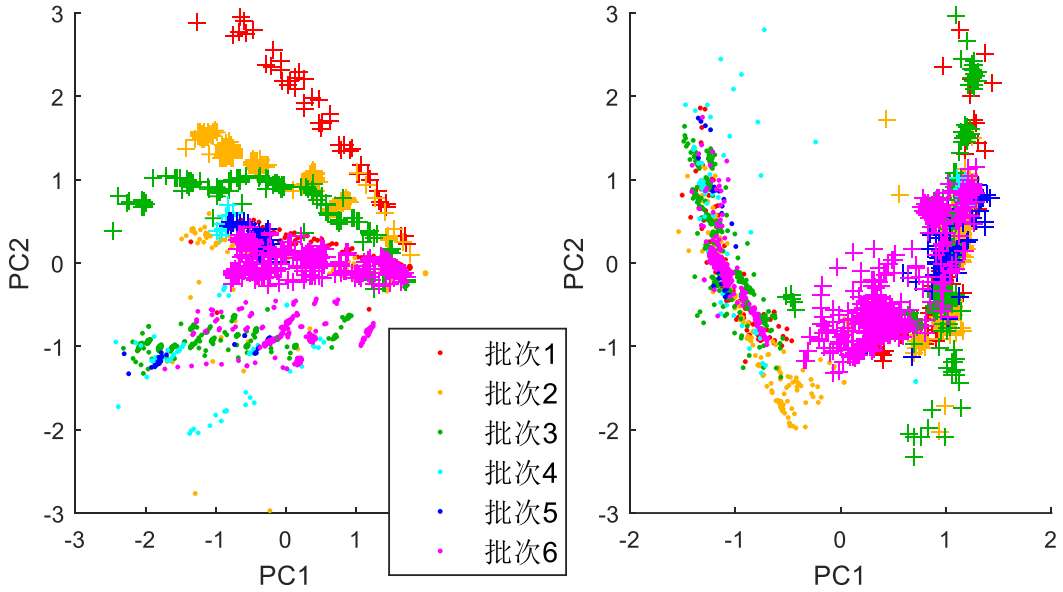
本节将在 4.3 节介绍的三个漂移数据集上测试本章提出的算法，并与典型的漂移补偿算法和基于自编码器的迁移学习算法进行比较。DCAE 算法基于 Theano 库[111]用 Python 语言实现，采用共轭梯度法优化，最大迭代次数设为 1000。

5.4.1　阵列漂移数据集

本数据集的原始特征维度为 128。我们训练了一个包含三个隐层的 DCAE（由两个浅层自编码器堆叠而成的），隐层的单元数分别为 30, 20, 30。我们发现将隐层数由 1 增加到 3 时可以带来较大改进，体现了深层网络相对于浅层的优势。但在上述配置的基础上进一步增加层数或单元数无法带来更大改进。网络中采用双曲正切函数作为激活函数，这是一个有界函数，值域为(-1, 1)。因此，为了使网络能够较好重构原始特征，需要对原始特征的动态范围进行缩放。一个直接的办法是将每个特征归一化到-1 和 1 之间，但由于数据集中存在很多野点，会给这种方式会带来较大干扰。因此我们首先统计批次 1 中每个特征的均值和方差，然后用这些数据将所有特征的均值归零、方差归一，之后再把所有特征除以 2，以使大部分特征的值落在-1 和 1 之间。每个样本的背景特征由式（5-4）构造。由于数据集中并没有给出样本的具体采集时间，只有批次号，因此定义式（5-4）中的 t 为该样本的批次号减 1，例如批次 6 的 $t = 5$。每个批次的标定样本的选择方法同上一章。采用类似上一章的在线补偿方式，即对批次 k 进行补偿时，利用批次 1 的有标签样本、批次 1 至 k 的无标签样本和标定样本训练 DCAE。最终的分类算法为逻辑回归。

图 5.3 比较了在原始特征表达下和 DCAE 学习到的特征表达下、不同批次的

样本的分布情况。可以发现在原始特征表达下，两个类别的样本都有较严重的漂移；而在 DCAE 学习到的特征表达下，漂移被较好地去除了，而且两个类别的样本分别聚集在一起。这体现了 DCAE 目标函数中的标定样本项和有监督项的作用。



(a) 原始特征　　(b) 含补偿层的DCAE学习得到的特征

图5.3　阵列漂移数据集中批次1-6的样本分布。点和加号代表两个类别的样本；不同颜色代表不同批次

DCAE 有两个版本：基本版本（DCAE-基本）和含补偿层的版本（DCAE-补偿层）。在 DCAE-补偿层中，我们对补偿层也采用双曲正切作为激活函数。DCAE-补偿层有三个主要参数：λ_1，λ_2，和补偿层单元数 h_{cor}。图 5.4 评估了它们在本数据集中对于算法性能的影响，其中每批次使用了 10 个标定样本。由于网络初始化存在随机性，我们对每个实验使用 10 个随机种子重复 10 遍，然后报告它们的平均正确率和标准差。可以发现有监督项权重 λ_1 较小时正确率下降较多。h_{cor} 的影响不大，大于 2 后正确率就趋于平稳。最终我们取 $\lambda_1 = 2^{-4}$，$\lambda_2 = 2^{-3}$，$h_{cor} = 6$。

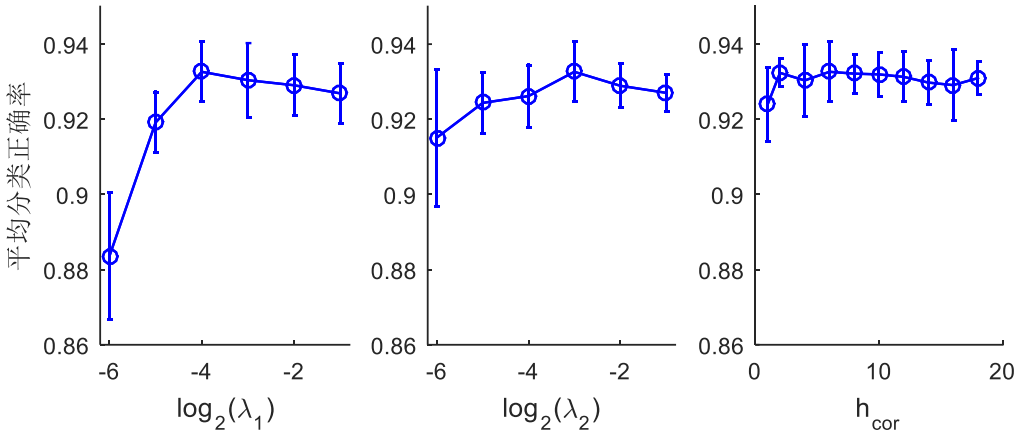


图5.4　阵列漂移数据集中，DCAE的参数对平均分类正确率的影响。改变一个参数时固定其他参数。误差线代表标准差

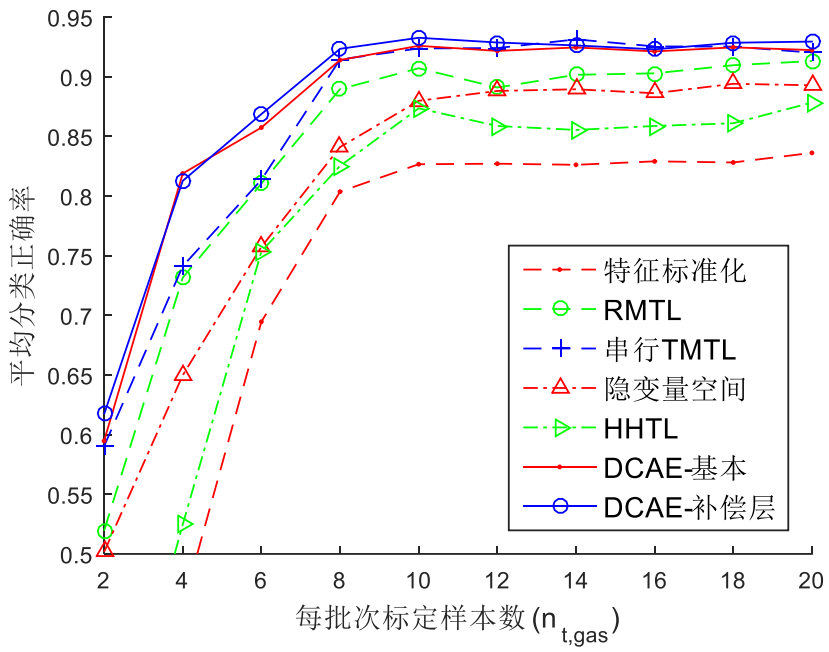


图5.5　阵列漂移数据集上的算法性能比较

图 5.5 对比了 DCAE-基本、DCAE-补偿层和特征标准化[92]、正则化多任务学习（RMTL）[88]、串行 TMTL（上一章），以及另外两个基于自编码器的算法。“隐变量空间”由 Zhuang 的算法[105]改进而来，原始算法中要求对齐两个领域的均值的隐层输出，我们将均值改为标定样本，因为其包含了更多信息。该算法本质上是寻找一个隐变量空间，在其中不同领域的样本有相似的分布，并且隐变量具有较好的鉴别能力。“隐变量空间”算法和 DCAE 的唯一区别在于前者没有应用背景特征，而是将不同领域的样本映射到同一个隐变量空间；而后者对不同背景的样本加入不同的补偿项。混合异质迁移学习算法（Hybrid Heterogeneous

Transfer Learning，HHTL）[106]首先使用边缘化栈式去噪自编码器（marginalized stacked denoising autoencoder）对每个领域（批次）学习特征，然后利用特征标准化将目标领域的特征变换到源领域，最后将栈式自编码器每一层学习到的特征进行串接。对于基于自编码器的算法，图 5.5 中的每个数据点是 10 次随机实验的均值。

从图 5.5 可以发现，平均正确率随标定样本数上升而上升，在标定样本数超过 10 后趋于平稳。特征标准化和 HHTL 的效果不佳，主要是由于它们在特征层面进行漂移补偿，补偿精度有限。而 HHTL 优于特征标准化，则表明深度网络学习到的特征比原始特征更加有效。隐变量空间算法期望寻找一个不包含漂移的隐变量空间，但有时漂移与有用信息并不能完全分开，因此该算法与成分校正算法面临相似的问题。DCAE 不寻求一个统一的隐变量空间，而是对不同背景的样本分别补偿，因此效果比隐变量空间更胜一筹。DCAE、TMTL 和 RMTL 的效果较好。与模型层面的 RMTL 和 TMTL 类似，DCAE 同时考虑了漂移补偿和标签信息。DCAE 的独有优势在于使用了深度网络架构，能够学习到数据中非线性和抽象的结构，因此效果最好，尤其是在标定样本数较少时。同时，DCAE 是特征层面的算法，通用性更好。

更多的实验结果列于表 5.2，其中漂移补偿算法都使用 10 个标定样本。为简洁起见，一些在表 4.2 中已经列出的漂移补偿算法这里就不再重复。“联合训练”算法使用批次 1 和 k 的样本联合训练栈式去噪自编码器。根据 Ben-David 等人的理论[112]，这种策略会实际会增加学习到的特征的领域间差异[102]，因此并非适用于所有场合。实验结果证明这种策略不适用于漂移补偿。Liu 等人[113]使用深度置信网络（Deep Belief Network，DBN）在批次 1 上学习特征后在后续批次上测试，表中为直接从原文献引用的结果。由于漂移的影响，该特征的效果不佳，但平均正确率优于原始特征，体现出深度网络学习到的特征的优势。DCAE 超越了上一章提出的 TMTL。对比 DCAE-基本和 DCAE-补偿层，可以发现前者在较早批次的效果普遍较好，而后者在较晚批次正确率更高。原因在于，实验中对批次 k 进行补偿时利用了批次 1 至 k 的标定样本，因此 k 越大，批次 1 至 k 的整体漂移状况越复杂，而补偿层在时变漂移较复杂时较有优势。

最后，我们尝试将 4.3 节介绍的 SEMI 正则项加入 DCAE 的目标函数。由于 DCAE 不是模型层面的补偿算法，无法直接对模型系数进行加权惩罚，因此我们改为操作 DCAE 的输入层到第一个隐层的权值矩阵 W_1，正则项的形式为：

$$\mu\sum_{j=1}^{m}\mathrm{SE}_j^2\cdot\left\|\mathbf{w}_j\right\|^2 \tag{5-8}$$

其中 $\mathbf{w}_j$ 是 W_1 的第 j 列，对应第 j 个原始特征。我们发现当 μ 设为10^{-6}时，平均

正确率略有改善（结果列于表中），而再增大 μ 会导致正确率下降。如果将正则项中的权重改为常数，效果不如 SEMI 加权。这说明加入 Tikhonov 正则项对 DCAE 的作用不大，可能是由于其现有目标函数已经可以学习到较优的模型，不需要再对系数进行收缩。后两个数据集情况类似，加入 SEMI 后性能变化不大，因此不再展示该策略的效果。

表 5.2　阵列漂移数据集各目标批次上的算法性能比较（%）

算法	2	3	4	5	6	7	8	9	10	平均
无补偿	88.59	66.96	40.99	54.82	43.22	44.40	31.63	45.74	39.11	50.61
串行 TMTL	97.35	**98.80**	90.06	98.48	95.35	91.50	91.84	96.38	71.56	92.37
联合训练[102]	59.41	56.33	58.63	37.26	44.77	43.17	20.72	34.26	35.55	43.34±1.14
DBN[113]	78.46	83.98	73.91	71.07	70.57	50.15	74.49	67.32	29.94	66.65
DCAE-基本	**97.58**	96.82	88.57	**98.73**	95.40	94.76	91.19	95.66	74.60	92.59±0.61
DCAE-补偿层	97.47	96.13	90.50	98.58	94.94	**95.47**	**92.62**	96.40	76.82	93.22±0.52
DCAE+SEMI	97.28	96.61	**90.50**	98.48	**95.81**	95.02	92.01	**97.15**	**77.45**	**93.37**±0.68

5.4.2　Corn数据集

考虑到本数据集分布的复杂度和样本量相对较小，我们采用一个线性激活函数的浅层网络进行特征学习，隐层中单元数设为 15。数据集由三台设备采集，只包含设备差异，因此背景特征的构造采用 5.2 节介绍的哑变量的方式，$n_{dev} = 3$。由于四个回归标签值的动态范围不同，我们在训练 DCAE 前首先将每个标签值均值归零、方差归一，在计算 RMSE 时再将预测值反变换回原来的均值、方差。回归算法使用岭回归。

本数据集不包含时变漂移，因此只对基本 DCAE 进行测试。实验中我们发现参数取 $\lambda_1 = 2^1$，$\lambda_2 = 2^4$ 时 RMSE 较小。该取值大于上一个数据集，这是由于本数据集包含的噪声较小，不易产生过拟合和过迁移，因此可以对有监督项和标定样本项赋予更大的权重。

表 5.3 比较了多种算法在本数据集上的 RMSE，其中漂移补偿算法都使用 7 个标定样本。为简洁起见，一些在表 4.4 中已经列出的漂移补偿算法这里就不再重复。我们发现，DCAE 优于另外两种基于自编码器的迁移学习算法：隐变量空间和 HHTL。但 DCAE 的效果与 TCTL 相当，尚不如并行 TMTL。这可能是因为基于自编码器的算法中需要优化的系数较多，而本数据集样本量较少。另外，DCAE 中的深度学习框架和补偿层比较擅长处理分布和时变漂移较复杂的数据，而本数据集不具有这些特性，因此体现不出优势。

表 5.3　Corn 数据集中各预测任务的 RMSE 比较

算法	Mp5 为从设备					Mp6 为从设备				
	水分	油脂	蛋白	淀粉	平均	水分	油脂	蛋白	淀粉	平均
无补偿	1.499	0.288	1.097	2.086	1.242	1.632	0.343	1.350	2.063	1.347
TCTL	0.182	0.104	0.148	0.350	0.196	0.193	0.104	0.144	**0.338**	0.195
并行 TMTL	**0.182**	0.098	**0.127**	**0.339**	**0.186**	**0.190**	0.100	**0.130**	**0.338**	**0.189**
隐变量空间[105]	0.191	0.095	0.162	0.370	0.205	0.197	0.097	0.166	0.385	0.211
HHTL[106]	0.262	0.092	0.146	0.404	0.226	0.300	**0.091**	0.140	0.377	0.227
DCAE-基本	0.183	**0.096**	0.149	0.348	0.194	0.199	0.096	0.148	0.350	0.198
同设备训练	0.137	0.101	0.140	0.363	0.185	0.139	0.099	0.149	0.369	0.189

5.4.3　呼气分析数据集

本数据集的原始特征维度为 9，维度较低，因此我们采用一个浅层网络进行特征学习，隐层中单元数设为 4。本数据集的特点在于同时具有设备差异和时变漂移，且时变漂移较复杂。每个样本都有采集设备和时间两个背景属性，可以根据式（5-4）构造样本的背景特征，其中 $n_{dev}=2$，t 是样本的确切采集时间（见图 4.6，以年为单位）。自编码器和补偿层的激活函数均使用双曲正切函数。通过对 DCAE 的参数λ_1、λ_2、h_{cor} 进行调优，我们发现λ_1对本数据集的影响不大，λ_2在 2^{-6} 时较优，$h_{cor}\geq 4$ 后正确率趋于稳定。因此最终采用$\lambda_1=2^{-4}$，$\lambda_2=2^{-6}$，$h_{cor}=4$。

表 5.4　呼气分析数据集中各预测任务的 F-score 比较（%）

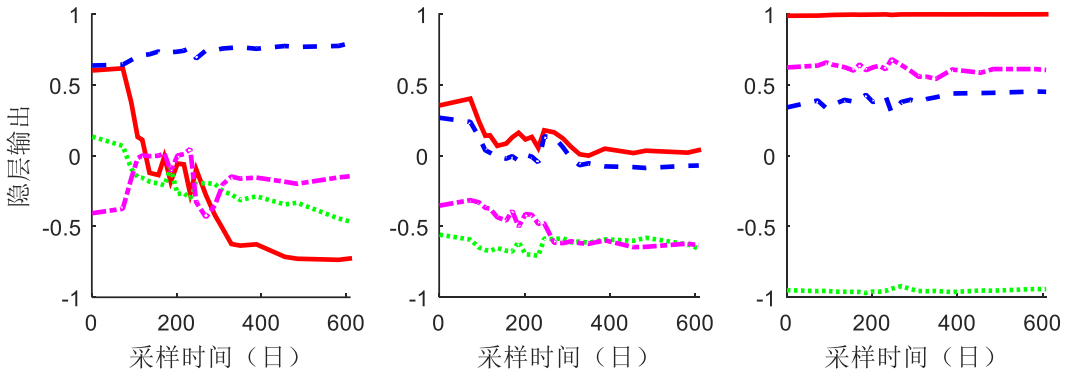
算法	任务 1	2	3	4	5	平均
无补偿	34.34	63.67	73.71	43.17	42.93	51.57
TMTL	73.81	**86.35**	84.42	**81.85**	82.24	81.73
隐变量空间[105]	57.49	68.73	73.27	68.04	74.12	68.33±2.37
DCAE-基本	74.44	82.20	89.36	81.15	82.04	81.84±0.67
DCAE-补偿层	**82.16**	84.27	**89.94**	81.34	**82.92**	**84.13**±0.82
随机选取训练样本	87.85	85.44	90.10	85.09	84.54	86.60

表 5.4 比较了多种算法在本数据集上的 F-score。为简洁起见，一些在表 4.6 中已经列出的漂移补偿算法这里就不再重复。HHTL 算法只能处理离散的领域，因此不能应用在本数据集。根据表 5.4，DCAE 超过了上一章的 TMTL 算法，获得了最好的平均准确率。含补偿层的 DCAE 更胜一筹，证明了其在处理复杂时

变漂移问题上的有效性。可以发现其正确率已经接近“随机选取训练样本”（在所有被测样本中随机选取每类的 50 个训练样本，包含各种采集设备和时间，可以作为本数据集漂移补偿的一个目标结果）。隐变量空间算法在本数据集上效果不佳，原因可能是本数据集中采用传感器的稳态响应作为特征，而漂移和气体浓度的变化都会影响传感器的稳态响应，因此漂移与有用信息较难分开。DCAE 不寻求一个统一的隐变量空间，而是借助背景特征对不同背景的样本分别补偿，因而能够在补偿漂移的同时保留有用信息。

5.4.4　分析：欠迁移与过迁移

DCAE 根据样本的背景计算补偿项，取得了较好的补偿效果。然而理论上该算法仍有改进的空间：样本的漂移不仅与其背景有关，还可能与气体本身的种类和浓度相关（见图 4.4 中阵列漂移数据集的例子，不同类别的样本漂移情况可能不同）。一个可能的改进办法是在补偿层的输入中加入原始特征，从而让补偿项是背景特征和原始特征两者的函数。我们对这种方法进行了实验，并将补偿效果和平均正确率展示于图 5.6。



(a) 欠迁移($\lambda_2 = 0$), 正确率70.94%　　(b) 效果较好的迁移($\lambda_2 = 2^{-6}$), 正确率84.13%　　(c) 过迁移, 正确率57.58%

图5.6　一个标定样本（设备2采集的10 ppm丙酮）的隐层输出随时间的变化

图中的 4 条曲线为 DCAE 学习到的 4 个隐层输出。如果漂移被完美补偿，同一个标定样本的 4 个隐层输出应当不随采样时间变化。分图(a)中，由于将标定样本项权重 λ_2 设为 0，没有漂移补偿，因此隐层输出随时间变化较明显，可以归因为欠迁移。(b)中采用了上小节实验结果用到的 λ_2 值，可以发现虽然正确率较高，但曲线仍有波动，漂移未被完全去除。而(c)采用了上面提到的改进方法，利用样本和背景共同确定补偿项，可以发现标定样本中的漂移几乎被完全去除了，但分类的平均正确率很低。改进方法的迁移能力强大，对标定样本的补偿效

果很好，但由于每组标定样本只有 8 个且存在噪声，过于追求标定样本的补偿造成了过迁移。此外，欠迁移和过迁移还可能与 DCAE 中的标定样本项权重 λ_2 的设置有关，过小有可能造成欠迁移，过大则可能造成过迁移。总之，在标定样本较少的情况下，需要在欠迁移和过迁移中间寻找平衡，不应过于强调标定样本的补偿。

5.4.5　分析：训练流程

DCAE 的训练由三个步骤组成：无监督预训练、有监督预训练、DCAE 训练。为了分析两个预训练步骤的作用，我们在表 5.5 中对不同训练方案进行了比较。步骤后面括号中的数字代表预设的最大迭代次数。为公平起见，四种方案的总最大迭代次数都是 3000。

表 5.5　不同的训练步骤在三个数据集上的平均分类正确率或回归 RMSE

训练步骤	阵列漂移数据集	Corn 数据集	呼气分析数据集
DCAE(3000)	76.56 ±2.36	0.2096 ±0.0014	81.28 ±2.11
无监督(1000)+DCAE(2000)	75.83 ±2.15	0.2055 ±0.0014	83.41 ±2.94
有监督(1000)+DCAE(2000)	93.04 ±0.44	0.1964 ±0.0017	83.44 ±0.89
无监督(1000)+有监督(1000)+DCAE(1000)	**93.22** ±0.52	**0.1961** ±0.0022	**84.13** ±0.82

由上表可以看出，本章使用的三步方案在三个数据集上都获得了最优的结果。另一方面，直接训练 DCAE 的效果不佳。与普通自编码器的预训练类似，本章使用的预训练方案将网络中的权值矩阵初始化到一个好的局部最优值附近，这里的好是泛化性能意义上的好，而非简单的目标函数值最小[94]。直接训练 DCAE 得到的目标函数值通常更小，但网络的泛化性能较差。三个训练步骤中使用的目标函数由简单到复杂，其寻优难度也逐渐增加。可以认为，一个恰当的简单的模型的泛化性能通常较好，因此利用简单模型初始化复杂模型，能够提高复杂模型的泛化性能。

另外，三步方案使用的实际迭代次数在上表的四个方案中也是最小的，这是因为两步预训练的模型较简单，收敛较快，实际训练次数通常小于 1000。比较上表的最后两行可以发现无监督预训练起到的作用相对较小，但其耗时也较小，因此我们保留了这个步骤。在阵列漂移数据集批次 2 的漂移补偿实验上，使用一个配置为 Intel Core 2 (2.66 GHz) CPU 和 4GB RAM 的 PC 时，三个步骤的耗时分别是 1.7，7.3 和 48.1 s。

5.5　本章小结

本章提出了漂移补偿自编码器（DCAE），一种具有漂移补偿功能的特征学习算法。它与原始自编码器的区别在于：

1. DCAE 中应用了背景特征和补偿项、补偿层，以便对设备差异和时变漂移进行显式建模和补偿；
2. DCAE 分别在编码阶段和解码阶段加入和减除了补偿项，使得其隐层学习到的特征不受漂移影响；
3. DCAE 利用标定样本学习补偿项的权值矩阵。

相比其他漂移补偿算法，DCAE 的优点在于：

- 它能够自然地处理训练、测试样本中的离散和连续的漂移，不需要将样本划分成领域、批次；
- 在漂移补偿的同时，它利用深度学习框架学习更具有鉴别能力的特征，缓解欠拟合问题；
- 它的网络结构能够缓解过迁移的问题：首先，补偿项仅与样本背景有关；其次，对于时变漂移，补偿项是时间的线性（基本 DCAE 中为$W_D\mathbf{d}$）或非线性（含补偿层的 DCAE 中为$W_{D1}\sigma_{cor}(W_{D0}\mathbf{d})$）函数，相对于时间一般是缓慢变化的，符合时变漂移的平缓先验；
- 它具有较强的灵活性，只需要输入样本的原始特征和采集设备、时间，以及若干有标签样本和标定样本，没有对有标签样本和标定样本的采集设备、时间做出任何限制。

由于应用了深度学习框架和补偿层，DCAE 更适用于分布和时变漂移较复杂的数据，在样本量较小、较简单的数据集可能不具有明显的优势。因此，它在阵列漂移数据集和呼气分析数据集上取得了比上一章的 TMTL 更好的效果，而在 Corn 数据集上则是 TMTL 领先。另外，DCAE 擅长处理复杂数据的代价是更大的训练时空复杂度。

第6章　最大独立领域适配

在前两章中，我们提出了三种基于标定样本的漂移补偿方法，在三个漂移数据集上取得了较高的精度。然而它们都需要在每台新设备上周期性地采集标定样本。尽管我们已经使用样本选择算法选择了尽量少的标定样本以减少标定的工作量，但标定仍然是电子鼻使用过程中比较耗费精力的一环，尤其是对于非专业的使用者而言。另外，使用者可能缺乏采集标定样本的工具，例如标准气体、配气装置等。能否在没有标定样本的情况下，仅利用目标领域的无标签被测气体对设备差异和时变漂移做出补偿呢？

本章将提出一种基于领域适配的漂移补偿算法，利用目标领域的无标签样本学习领域不变特征。为了解决时变漂移的适配问题，该算法沿用了前一章提出的背景特征，然后利用 Hilbert-Schmidt 独立性准则来最大化学习到的特征与背景特征之间的独立性，从而缩小特征的领域差异。为了对不同背景的样本差异化地投影，本章还提出了一种特征增强策略。

6.1　相关工作

6.1.1　无监督领域适配

术语“领域适配”（domain adaptation）与“迁移学习”含义相似[114,115]，也是研究如何在分布不同的领域之间进行知识迁移的问题。在本章中，我们遵循相关算法研究者的习惯，使用领域适配一词。根据**目标领域**是否存在少量有标签样本，可以将领域适配算法分为半监督和无监督两类[116]。本章中，我们主要关注目标领域只存在无标签样本的无监督领域适配。Pan 等人[20]和 Patel 等人[117]的论文是两篇很好的针对领域适配问题的文献综述。

学习领域不变特征（domain-invariant features）是领域适配的重要课题[112]，其目的在于寻找方法缩小领域间的特征分布差异，同时保留有用信息。一些算法将各领域样本投影到一个公共的隐变量空间[118-120]。例如，迁移成分分析（Transfer Component Analysis，TCA）[118]利用再生核希尔伯特空间（Reproducing Kernel Hilbert Space，RKHS）中的最大均值差异准则（Maximum Mean Discrepancy，MMD）学习领域间的迁移成分。为了引入标签信息、同时保留数据的流形结构，该文献又提出了半监督迁移成分分析（Semi-supervised TCA，SSTCA）。Shi 等人[119]用样本与其二值领域标签的互信息作为领域差异的度量准则，用目标领域样本与其聚类标签的互信息的负值作为期望分类误差的估计值，并同时最小化这两

个值。低秩迁移子空间学习（Low-rank Transfer Subspace Learning，LTSL）[120]在子空间中用局部的源领域样本重构每个目标领域样本，借此适配两个领域的样本分布。标签信息和流形结构可以通过在 LTSL 中嵌入不同的子空间学习算法来得到保持。

另外一些算法首先将不同领域的样本投影到各自的子空间中，然后在这些子空间之间建立联系[121,122,79,123]。Fernando 等人[121]提出子空间对齐算法，用 PCA 投影的主方向代表子空间，然后用一个变换矩阵将源领域的子空间变换到目标领域。测地线流核学习（Geodesic Flow Kernel，GFK）[122]对 Grassmann 流形中两个领域的测地线上的无穷多个子空间进行积分，从而得到一个领域不变的核矩阵。Liu 等人[79]针对电子鼻的时变漂移补偿问题对 GFK 进行了改进，设计了组合测地线流核学习算法。该算法对多个批次的无标签样本进行领域适配时，利用 GFK 将每个中间批次和首末批次联系起来，最后将得到的所有核矩阵进行平均。移动协方差领域适配（Domain Adaptation by Shifting Covariance，DASC）[123]也是 GFK 算法的一个改进。在描述每个领域时，该算法用协方差矩阵来代替 GFK 中的子空间。

现有无监督领域适配算法的缺点在于，它们大多只能处理离散的领域，而在时变漂移问题中需要处理领域分布连续变化的样本流。其次，基于隐变量空间的方法对所有样本采用统一的投影，这种方式不够灵活，不能确保领域差异在每个投影维度都被消除。另外，以气敏传感器老化为例，老化后的传感器对气体的敏感度可能会下降，造成同样的样本在老化前可能代表低浓度，而老化后就代表高浓度，换言之就是目标领域的条件分布发生了变化。基于隐变量空间的方法将同样的样本投影到同样的位置，从而无法区分老化前后特征相似但本质不同的样本。前一章的算法中，我们对不同背景的样本加上不同的补偿项，从而解决了这个问题。在本章的算法中，我们也希望能对不同背景的特征有区别地进行投影。

6.1.2　Hilbert-Schmidt独立性准则

Hilbert-Schmidt 独立性准则（Hilbert-Schmidt Independence Criterion，HSIC）是一种用来度量两个随机变量的独立性的准则。设两个随机变量的联合分布是 p_{xy}，它们的核函数分别是 k_x 和 k_y，对应着两个 RKHS：$\mathcal{F}$ 和 $\mathcal{G}$。HSIC 定义为互协方差运算符 $\mathcal{C}_{xy}$ 的 Hilbert-Schmidt 范数的平方[124]：

$$\begin{aligned}&\mathrm{HSIC}(p_{xy},\mathcal{F},\mathcal{G})=\left\|\mathcal{C}_{xy}\right\|_{\mathrm{HS}}^2\\&=\mathbf{E}_{xx'yy'}[k_x(x,x')k_y(y,y')]+\mathbf{E}_{xx'}[k_x(x,x')]\mathbf{E}_{yy'}[k_y(y,y')]\\&\quad-2\mathbf{E}_{xy}\{\mathbf{E}_{x'}[k_x(x,x')]\mathbf{E}_{y'}[k_y(y,y')]\}\end{aligned}\tag{6-1}$$

这里 $\mathbf{E}_{xx'yy'}$ 是对从 p_{xy} 中抽取的独立样本组合 (x,y) 和 (x',y') 求取的期望。可以证

明，对于典型核函数（characteristic kernels）k_x 和 k_y，HSIC($p_{xy},\mathcal{F},\mathcal{G}$) 为 0 当且仅当两个随机变量独立[125]。HSIC 越大，两个随机变量的依赖性越强（在所选核函数意义下）。

HSIC 有一个有偏经验估计。设 X 和 Y 是从 p_{xy} 中抽取的两组观测样本，$Z = X \times Y = \{(x_1, y_1),\ldots,(x_n, y_n)\}$，$K_x, K_y \in \mathbf{R}^{n\times n}$ 分别是 X 和 Y 的核矩阵，则[124]：

$$\mathrm{HSIC}(Z,\mathcal{F},\mathcal{G}) = (n-1)^{-2}\,\mathrm{tr}(K_x H K_y H) \tag{6-2}$$

其中 $H = I - n^{-1}\mathbf{1}_n\mathbf{1}_n^T \in \mathbf{R}^{n\times n}$ 为中心化矩阵。由于上式有效且应用方便，HSIC 已被应用在特征抽取[126,118,127]和特征选择[125]等领域。通常的方法是最大化抽取或选择的特征与标签之间的依赖，即最小化独立性。然而，据我们所知，HSIC 还没有在领域适配问题中被用来最大化子空间中特征与背景特征之间的独立性。

6.2　最大独立领域适配

6.2.1　无监督情况

最大独立领域适配（Maximum Independence Domain Adaptation，MIDA）可以被视为一种特征抽取（feature extraction）算法，目标在于学习一个领域不变子空间，利用 Hilbert-Schmidt 独立性准则来最大化子空间中的特征与背景特征之间的独立性。直观来讲，子空间中的特征与背景特征独立意味着我们不能通过一个样本在子空间中的位置来推断它的背景，也就是说不同背景的样本不存在分布差异。根据 Ben-David 等人的理论[112]，为了取得较好的领域适配效果，需要使得不同领域的样本变得难以区分，这与我们的思路是一致的。通过这种方式，我们不仅可以对两个离散领域进行适配，还可以处理多个离散领域和连续分布变化的情况。下面介绍这一思路的实现细节。

设 $X \in \mathbf{R}^{n\times m}$ 为原始特征向量的样本矩阵，原始特征维度为 m，样本数为 n，其中既包含训练也包含测试样本。与 DCAE 类似，这里我们不需要指明每个样本 $\mathbf{x}$ 属于哪个领域，相关信息蕴含在这些样本的背景特征中。我们首先用一个非线性的映射函数 Φ 将原始特征映射到一个高维特征空间。基于核技巧（kernel trick），Φ 的具体形式并不需要给出，只需要将 $\Phi(X)$ 的内积用一个核矩阵代替 $K_x = \Phi(X)\Phi(X)^T$。然后，定义一个投影矩阵 $\tilde{W}$ 将 $\Phi(X)$ 投影到一个维度为 h 的子空间，得到投影后的样本 $Z = \Phi(X)\tilde{W} \in \mathbf{R}^{n\times h}$。在核子空间学习中[128,129]，可以用高维空间中的样本的线性组合来表达该投影矩阵，即 $\tilde{W} = \Phi(X)^T W$。$W \in \mathbf{R}^{n\times h}$ 才是真正需要学习的投影矩阵。这时，子空间中的样本可表达为：

$$Z = \Phi(X)\Phi(X)^T W = K_x W \tag{6-3}$$

其核矩阵为：

$$K_z = K_x W W^T K_x \tag{6-4}$$

设背景特征矩阵为 $D \in \mathbf{R}^{n \times m_d}$，$m_d$ 为背景特征维度。使用线性核函数得到其核矩阵

$$K_d = D D^T \tag{6-5}$$

根据背景特征的定义式（5-4），若样本 i 和 j 来自不同设备，$(K_d)_{ij} = 0$；若来自相同设备，$(K_d)_{ij} = 1$（不考虑时变漂移）或 $1 + t_i t_j$（考虑时变漂移）。将式（6-4）和 K_d 代入 HSIC 的经验估计式（6-2）并忽略掉常数因子后，我们得到需要最小化的量：$\mathrm{tr}(K_z H K_d H) = \mathrm{tr}(K_x W W^T K_x H K_d H)$。

领域适配问题中，不仅需要缩小领域差异，还需要保留数据中的有用信息，例如方差[118]。可以通过最大化 Z 的协方差矩阵的迹的方式来最大化方差：

$$\begin{aligned}\mathrm{cov}(Z) &= \mathrm{cov}(K_x W) \\ &= \frac{1}{n}\left(K_x W - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n{}^T K_x W\right)^T \left(K_x W - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n{}^T K_x W\right) \\ &= W^T K_x H K_x W\end{aligned} \tag{6-6}$$

其中 $H = I - n^{-1}\mathbf{1}_n \mathbf{1}_n^T \in \mathbf{R}^{n \times n}$ 与（6-2）式中相同。为了避免平凡解，我们限制 W 中投影方向单位正交，得到最终的 MIDA 算法目标函数：

$$\begin{aligned}&\max_W \ -\mathrm{tr}(W^T K_x H K_d H K_x W) + \mu \mathrm{tr}(W^T K_x H K_x W), \\ &\text{s.t. } W^T W = I\end{aligned} \tag{6-7}$$

其中 $\mu > 0$ 为权重系数。为了求解上式，利用拉格朗日乘子法构造函数：

$$\mathrm{tr}(W^T K_x(-H K_d H + \mu H) K_x W) - \mathrm{tr}((W^T W - I)\Lambda) \tag{6-8}$$

Λ 为拉格朗日乘子矩阵。令上式对 W 的导数为 0，即得：

$$K_x(-H K_d H + \mu H) K_x W = W \Lambda \tag{6-9}$$

因此 W 是矩阵 $K_x(-H K_d H + \mu H) K_x$ 的 h 个最大特征值对应的特征向量。另一种常见的约束项是限制 $\tilde{W}$ 中投影方向单位正交[127]，在该情况下求解将得到一个广义特征值问题。然而我们在试验中发现这种约束在领域适配精度和训练速度上都不如式（6-8），因此并未采用。

在计算 K_x 时，选择一个合适的核函数很重要。常见的核函数包括线性（$k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T \mathbf{y}$），多项式（$k(\mathbf{x},\mathbf{y}) = (\sigma \mathbf{x}^T \mathbf{y} + 1)^d$），高斯径向基函数（Radial Basis Function，RBF，$k(\mathbf{x},\mathbf{y}) = e^{\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2)}$）等等[125]。根据文献[125]，在 HSIC 中使用不

同的核函数蕴含了不同的独立性假设。多项式和 RBF 核将原始特征映射到高维甚至无穷维，因此能够检测更多种特征依赖，但选择合适的核参数σ也很重要。

6.2.2　半监督情况

MIDA 在缩小不同背景样本的分布差异时没有考虑样本的标签。如果部分样本标签已知，将其结合到学习过程中有利于提高子空间中特征的鉴别能力。提高特征的鉴别能力也可以降低领域适配误差上界[112]。因此，我们将无监督情况下的 MIDA 拓展到半监督情况，并命名为 SMIDA（Semi-supervised MIDA）。在 SMIDA 中，无标签和有标签样本都可以来自任何领域。SMIDA 利用 HSIC 最大化子空间中有标签样本的特征与标签的依赖性[118]。用这种方式利用样本标签的好处在于，不管是分类问题中的离散标签还是回归问题中的连续标签，都可以被统一地集成到算法中。

首先定义样本的标签矩阵 Y。对于 c 类分类问题使用哑变量编码方式，即 $Y \in \mathbf{R}^{n\times c}$，若 $\mathbf{x}_i$ 为有标签样本且属于第 j 类，则 $y_{ij} = 1$；否则 $y_{ij} = 0$。对于回归问题，首先将标签的均值归零、方差归一，然后定义 $Y \in \mathbf{R}^{n}$。若 $\mathbf{x}_i$ 为有标签样本，则 y_i 等于该标签值；否则 $y_i = 0$。使用线性核函数得到其核矩阵：

$$K_y = YY^T \tag{6-10}$$

SMIDA 的目标函数可以表达为：

$$\begin{aligned} &\max_{W} \ \ \mathrm{tr}(W^T K_x(-HK_dH + \mu H + \gamma HK_yH)K_xH), \\ &\text{s.t. } W^TW = I \end{aligned} \tag{6-11}$$

其中$\gamma > 0$ 为权重系数。上式的解为矩阵 $K_x(-HK_dH + \mu H + \gamma HK_yH)K_x$ 的 h 个最大特征值对应的特征向量。与上两章算法相似，SMIDA 允许有标签样本出现在任何领域，因此也可以用于半监督领域适配问题。

6.2.3　与其他算法的联系

一种常用的度量两个分布之间差异的准则是最大均值差异准则（MMD）[130,118]。Song 等人[125]证明，在二分类问题中，如果采用特定形式的标签核矩阵，则用 MMD 估计两类样本的分布差异以及用 HSIC 估计特征与类标签的独立性等价（只差一个常数因子）。这一证明体现了 MMD 与 HSIC 之间的联系。然而 MMD 只适用于衡量两个离散的领域之间的分布差异，而利用了 HSIC 的 MIDA 可以处理多个离散领域以及连续分布变化的情况。

平稳子空间分析（Stationary Subspace Analysis，SSA）[131]能够在多维时序信号中找到平稳的成分，将每个样本投影到平稳子空间，因此可以应用于样本流的

领域适配问题。然而，SSA 仅能保证找到均值和方差不变的成分，不能保证这些成分保持了数据中的有用信息。事实上，某些成分可能仅仅是平稳的噪声[118]。概念漂移适配算法也可以处理样本流的领域适配问题，然而绝大多数该类算法[91,110]依赖于新采集的有标签样本更新现有模型，而 MIDA 能够在没有样本标签的情况下进行适配。

除了方差和鉴别信息，数据的流形结构也是一种重要的属性信息。保持数据的流形结构可以使用流形正则化（Manifold Regularization，MR）[132]。MR 采用邻接图来描述流形结构。为了将 MR 引入 MIDA，需要在目标函数中加入一个新的正则项 $-\lambda \mathrm{tr}(W^T K_x L K_x W)$，其中 L 是图拉普拉斯矩阵[132]，$\lambda > 0$ 为权重系数。在我们的实验中，加入 MR 通常可以轻微改进领域适配的效果，但同时也引入了三个需要调整的参数，包括λ、计算邻接图时的近邻数和带宽系数。因此我们在后续的实验中并没有使用MR，但对于一些流形结构比较重要的数据仍可以采用它。

6.3　特征增强

在 6.1.1 节我们曾经提到，希望算法对不同背景的特征有区别地进行投影。实现这一点的一种方案是特征增强。如果对不同背景的特征采用不同的增强方式，后续仍只需要学习一个投影矩阵，但事实上可以对不同背景的样本分别投影。一个例子是 Daumé等人[133]的特征增强策略。在该策略中，如果 $\mathbf{x} \in \mathbf{R}^m$ 属于源领域，则增强后的特征向量为 $\mathbf{x} = [\mathbf{x}^T, \mathbf{x}^T, \mathbf{0}_m^T]^T \in \mathbf{R}^{3m}$；若属于目标领域，则 $\mathbf{x} = [\mathbf{x}^T, \mathbf{0}_m^T, \mathbf{x}^T]^T \in \mathbf{R}^{3m}$。用增强后的源领域和目标领域的所有样本训练一个预测模型。对不同的领域，该模型可以视为不同的 $2m$ 维的模型，前 m 维和后 m 维分别为领域无关和领域相关的模型成分。

然而，上述策略只能处理离散的领域，不能处理时变漂移。我们提出一个更通用且高效的特征增强策略：将原始特征和背景特征进行串接，即：

$$\mathbf{x} = \begin{bmatrix} \mathbf{x} \\ \mathbf{d} \end{bmatrix} \in \mathbf{R}^{m+m_d} \tag{6-12}$$

我们用一个线性子空间投影的例子来说明这个策略的作用。假设对于增强后的特征，已经学习好了投影矩阵 $W \in \mathbf{R}^{(m+m_d)\times h}$，$h$ 是子空间维度。W 可以被分成两部分：$W = [W_x^T, W_d^T]^T$，其中 $W_x \in \mathbf{R}^{m\times h}$ 是背景无关的部分，$W_d \in \mathbf{R}^{m_d \times h}$ 是背景相关的部分。投影后的样本是 $W^T \hat{\mathbf{x}} = W_x^T \mathbf{x} + W_d^T \mathbf{d} \in \mathbf{R}^h$，也就是说，在特征增强后，子空间中的每一维特征 j 都被加上了一个与样本背景相关的偏置 $(W_d^T \mathbf{d})_j$，从而在平移的意义下实现了不同背景的差异化投影。而 W_x 部分对所有样本相同，使得不同背景的样本的子空间又具有相似性。这种策略其实与上一章中 DCAE 的

补偿项策略有共通之处，不过 DCAE 使用标定样本来学习补偿项和整个网络，补偿精度更高。

从另一个分析角度来看，特征增强策略在学习投影矩阵之前首先将样本映射到一个高维空间。在高维空间中，寻找能够对齐样本的投影矩阵会更加容易。该策略的有效性将在下一节的人工合成数据集和漂移数据集的实验上得到验证。

我们把 MIDA 和 SMIDA 的算法流程总结在算法 3，中括号标注的部分仅适用于 SMIDA。

算法 3：最大独立领域适配 MIDA[或 SMIDA]

输入：所有样本的矩阵 X，其采样设备和时间已知；[部分样本的标签]；X 的核函数；h, μ, [和 γ]。

输出：子空间中的样本 Z。

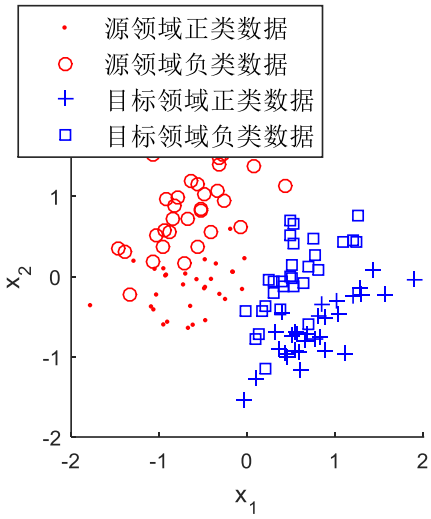
1. 按式（5-4）构造每个样本的背景特征；
2. 按式（6-12）对所有样本进行特征增强；
3. 计算核矩阵 K_x，K_d（式（6-5）），[和 K_y（式（6-10））]；
4. 求解 W，即矩阵 $K_x(-HK_dH+\mu H)K_x$ [或 $K_x(-HK_dH+\mu H+\gamma HK_yH)K_x$] 的 h 个最大特征值对应的特征向量；
5. $Z = K_x W$ 。

6.4　实验结果和讨论

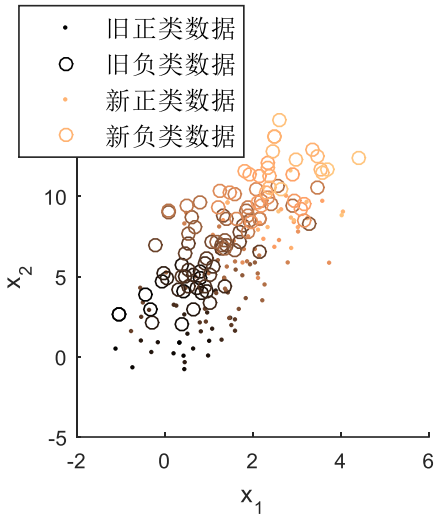
为了直观观察本章提出的算法的效果，我们首先在几个人工合成数据集上进行领域适配实验。接着考察 4.3 节介绍的三个漂移数据集，并与近年来典型的无监督特征适配算法进行比较。

6.4.1　人工合成数据集

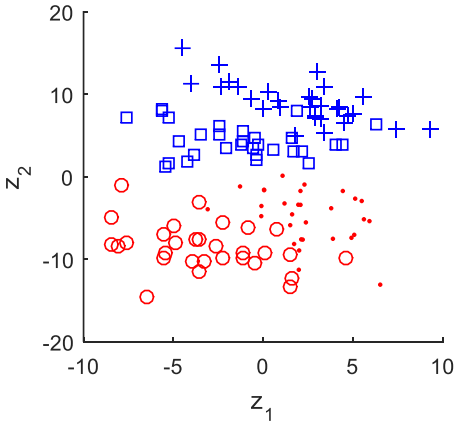
图 6.1 的分图(a)-(c)使用了一个包含两个离散领域的 2D 数据集。对于迁移成分分析（TCA）和 MIDA 两种算法，我们都采用线性核函数，设权重系数$\mu = 1$。为了定量评价算法效果，我们还在源领域训练了逻辑回归分类器并应用在目标领域，正确率展示在分图标题中。根据图示和正确率，可以看出 MIDA 的领域适配效果优于 TCA。TCA 仅能在第一个投影维度 z_1 上对齐两个领域，因此我们在分类时也只使用了该维度。但两类样本在该维度重叠较大，导致正确率虽比原始特征有所改善，但仍不如 MIDA。这是因为在本数据集中，对齐两个领域与分开



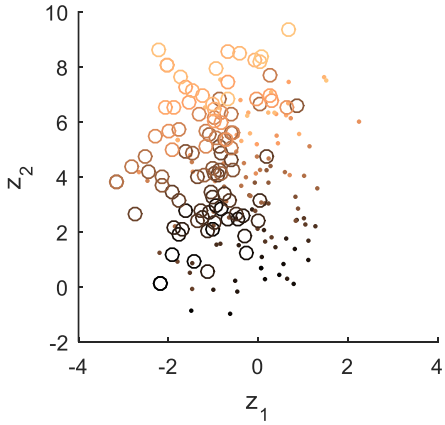
(a) 原始特征，53%



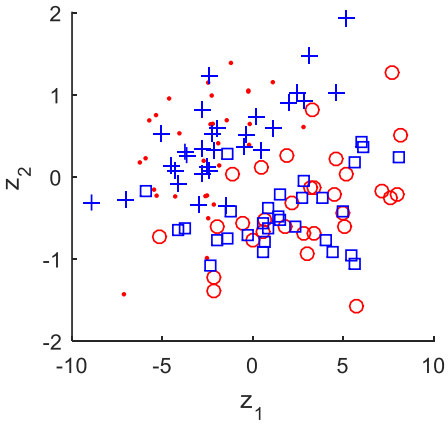
(d) 原始特征，55%



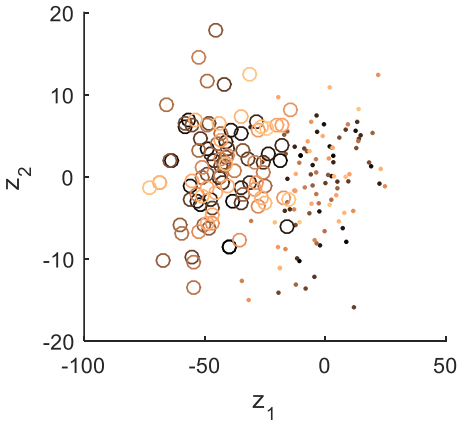
(b) TCA[118]，73%（仅用z_1分类）



(e) SSA[131]，74%（仅用z_1分类）



(c) MIDA，88%



(f) MIDA，90%

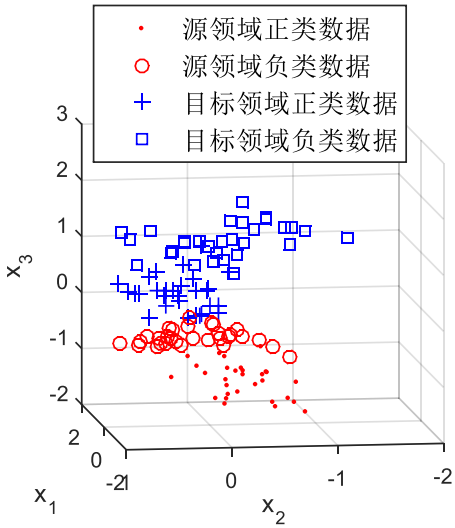
图6.1　人工合成数据集一（分图(a)-(c)）和二（分图(d)-(f)）上的算法效果和领域适配正确率比较，图例见分图(a)和(d)

两个类别需要的投影方向不一致。使用 SSTCA 引入源领域标签信息也不能带来很大帮助。相反地，MIDA 在两个投影方向上都成功适配，这主要是由于特征增强策略可以给不同领域的每一维度加上不同的偏置，使得两个领域在 z_2 方向上的差异被消除了。

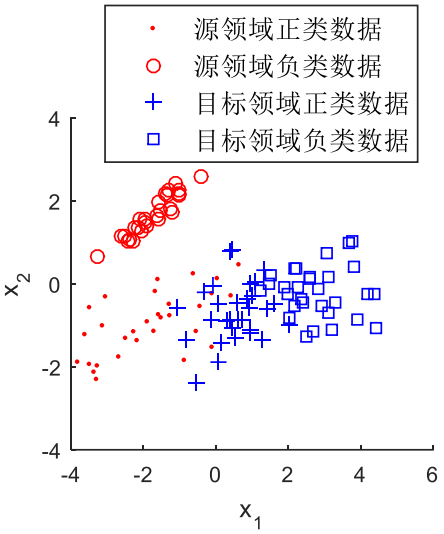
图 6.1 的分图(d)-(f)使用了一个包含连续分布变化的 2D 数据集来模拟时变漂移。图中样本点颜色由深到浅，代表采集时间由早到晚。原始特征中两个类别的样本都向右上方漂移。MIDA 的参数配置同合成数据集一。平稳子空间分析（SSA）算法中设置平稳成分数为 1。我们使用两类的较早采集的一半样本训练，较晚采集的一半样本测试，得到了分图标题中的正确率。分图(e)中，SSA 能够找到一个不含时变漂移的分量 z_1，但两类样本在该分量上并不能很好分开。分图(f)中，不同采集时间的样本错杂分布，证明时变漂移在两个投影方向 z_1 和 z_2 上都被 MIDA 消除了。MIDA 先利用特征增强策略将数据转换到一个 3D 空间中（2D 数据+时间），然后将 3D 数据投影到一个与时变漂移方向正交的 2D 平面，在新平面上样本与时间的独立性被最大化。先升维再降维也提高了类别可分性。

上面两个实验都没有利用样本的标签信息。如果需要在低维子空间中尽可能提高样本与标签的依赖性，可以使用 SMIDA 代替 MIDA。图 6.2 的分图(a)-(c)使用了一个包含两个离散领域的 3D 数据集，目标是找到一个消除领域差异的 2D 子空间。不考虑样本标签的话，能够对齐两个领域的最佳投影方向是 x_3，但投影后两类样本会重叠到一起，如分图(b)所示。分图(c)中，SMIDA 利用了源领域的标签信息，算法参数 μ 和 γ 都设为1。可以发现不同类别在2D 子空间能够被分开。事实上，MIDA 学习的第 3 维投影方向也具有类别可分性，但我们通常希望数据的重要信息保留在前几维特征中，在这种情况下 SMIDA 优于 MIDA。

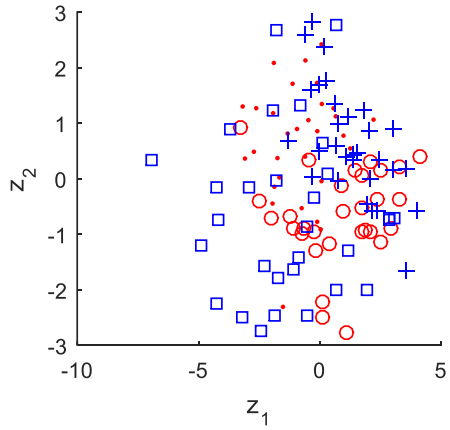
在机器学习算法中，非线性核常常在数据线性不可分时使用。此外，合成数据集四表明，领域适配问题中非线性核对于难以用线性投影对齐的数据可能有所帮助。从图 6.2 分图(d)中可以看出，两个类别的领域差异方向（目标领域相对于源领域 “漂移”的方向）不一致。此时，即使使用了特征增强，也无法找到一个线性投影方向将两个领域很好地对齐。特征增强仅能提供领域相关偏置，而本数据集需要在目标领域应用一个旋转矩阵，如分图(e)所示。但由于目标领域标签未知，学习变换矩阵的尝试很有可能造成过迁移。不过我们可以先用一个非线性核将原始特征映射到高维空间，在该空间中可能更容易找到一个适配效果较好的投影矩阵。分图(f)使用了一个 RBF 核，核参数 σ 设为 10。尽管两个领域仍不能完美对齐，但源领域训练的分类模型能更好地适配到目标领域，从而提高了领域适配后的分类正确率。



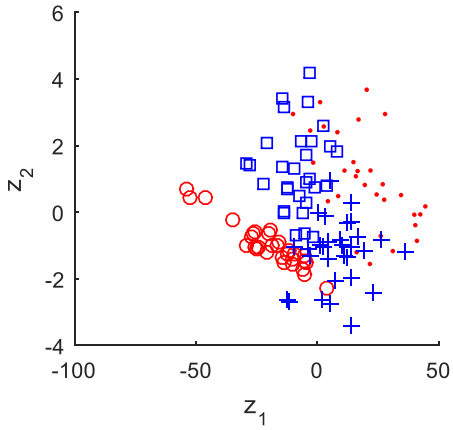
(a) 原始特征，50%



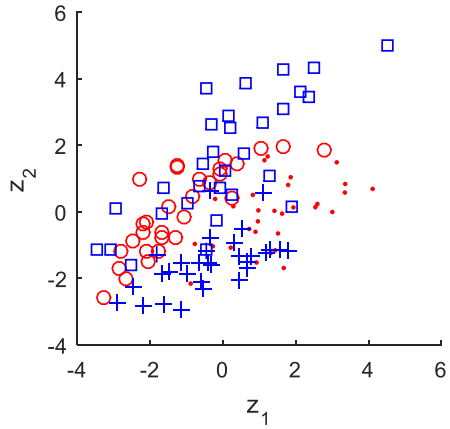
(d) 原始特征，50%



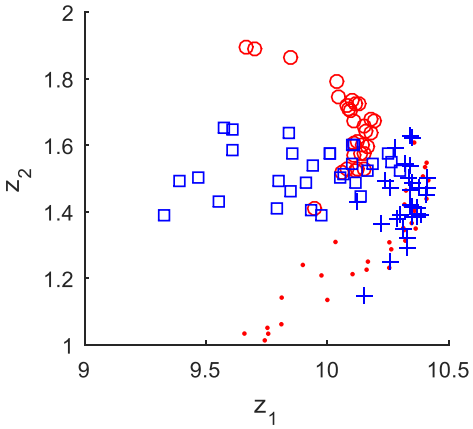
(b) MIDA，55%



(e) 线性核MIDA，57%



(c) SMIDA，82%



(f) RBF核MIDA，87%

图6.2　人工合成数据集三（分图(a)-(c)）和四（分图(d)-(f)）上的算法效果和领域适配正确率比较，图例见分图(a)和(d)

6.4.2　阵列漂移数据集

表 6.1　阵列漂移数据集各目标批次上的算法性能比较（%）

算法	2	3	4	5	6	7	8	9	10	平均
仅预处理	80.47	79.26	69.57	77.16	77.39	64.21	52.04	47.87	48.78	66.30
KPCA	75.88	69.04	49.07	57.87	62.65	52.26	37.07	47.66	49.97	55.72
TCA[118]	82.96	81.97	65.22	76.14	89.09	58.98	49.32	66.17	49.50	68.82
SSTCA[118]	**84.57**	80.90	**80.12**	75.63	87.26	66.37	54.76	61.28	54.44	71.70
SA[121]	80.79	80.01	71.43	75.63	78.35	64.68	52.04	48.51	49.58	66.78
GFK[122]	77.41	80.26	71.43	76.14	77.65	64.99	36.39	47.45	48.72	64.49
ML-comGFK[79]	80.25	74.99	78.79	67.41	77.82	**71.68**	49.96	50.79	53.79	67.28
ITL[119]	76.85	79.45	59.63	**96.45**	78.00	60.95	49.32	**77.02**	48.58	69.58
MIDA（离散）	81.03	85.62	60.25	75.63	87.61	62.44	48.30	67.87	48.36	68.57
SMIDA（离散）	80.47	**87.07**	65.22	75.63	90.04	59.20	50.00	62.77	44.81	68.36
MIDA（连续）	84.32	81.59	68.32	75.63	91.74	63.13	78.91	62.34	45.14	72.35
SMIDA（无增强）	82.23	83.17	67.70	75.13	85.22	61.67	51.02	61.49	**54.61**	69.14
SMIDA（连续）	83.68	82.28	73.91	75.63	**93.00**	63.49	**79.25**	62.34	45.50	**73.23**

由于本数据集每个特征的动态范围相差较大，我们首先对数据进行预处理，将每个批次的每个特征均值归零、方差归一。接着采用与前面两章中相似的实验方式，在对批次 k 进行测试时，利用批次 1 的有标签样本和批次 1 至 k 的无标签样本学习领域不变特征，但不使用标定样本。最后用批次 1 在新特征表达下的有标签样本训练逻辑回归分类器，用来分类批次 k 在新特征表达下的无标签样本。

表 6.1 列出了近年来典型的无监督领域适配算法的正确率。参加比较的算法包括核 PCA（KPCA）、迁移成分分析（TCA）、半监督迁移成分分析（SSTCA）、子空间对齐（Subspace Alignment，SA）、测地线流核学习（GFK）、组合测地线流核学习（ML-comGFK）和基于信息论的学习方法（Information-Theoretical Learning，ITL）。每种算法的参数都进行了调优以得到最好的结果。在 KPCA、TCA、SSTCA、MIDA 和 SMIDA 中，核函数为 2 阶多项式核。KPCA 利用批次 1 和 k 的数据学习一个核子空间。TCA、SSTCA、MIDA 和 SMIDA 需要进行特征值分解运算，为了降低运算量，我们在计算核矩阵时不使用全部的目标领域样本，而是随机抽取 n_t 个目标领域样本，n_t 为源领域样本数量的 2 倍。GFK 算法使用 PCA 对批次 1 和 k 分别计算子空间，其维度由原文献[122]中提出的子空间不

一致度量（subspace disagreement measure）计算得到。ML-comGFK 的结果从原文献中直接引用。

我们还比较了本章提出的算法的若干变种。在上表中，“（离散）”表明 MIDA 和 SMIDA 使用与其他算法相似的两个离散领域（批次 1 和 k）学习领域不变特征，这种情况下，批次 1 的背景特征是$[1,0]^T$，批次 k 是$[0,1]^T$。但这种策略无法利用中间批次的无标签样本提高领域适配的泛化性能。标有“（连续）”的算法利用到了中间批次的样本，我们从批次 2 至 k 中随机选出 n_t 个样本，使用样本的批次号作为其背景特征，并学习领域不变特征。另外，我们也测试了不使用特征增强的连续 SMIDA 算法，在表格中用“（无增强）”标出。

从表 6.1 中可以发现，SMIDA（连续）获得了最高的平均正确率。连续版本的 MIDA 和 SMIDA 优于离散版本，证明本章提出的算法能够有效利用样本流中蕴含的时序信息改善领域适配的效果。它们也超越了同样利用了时序信息的 ML-comGFK 算法。去除特征增强后 SMIDA 的平均正确率下降了 4 个百分点，表明特征增强对于本数据集有较大作用。另外也需要注意到，使用无监督领域适配算法进行漂移补偿后，正确率虽然相比无补偿有所改善，但比起使用标定样本的算法（表 4.2、表 5.2）还有较大差距。这是由于本数据集的时变漂移比较复杂，不同类别的漂移方向可能不一致，不同批次的样本构成也不完全相同。无标签样本包含的信息量小于标定样本，因此较难准确适配。图 6.3 展示了不同子空间维度 h 下各算法的平均正确率。MIDA 和 SMIDA 在 $h > 30$ 时优于其他算法。

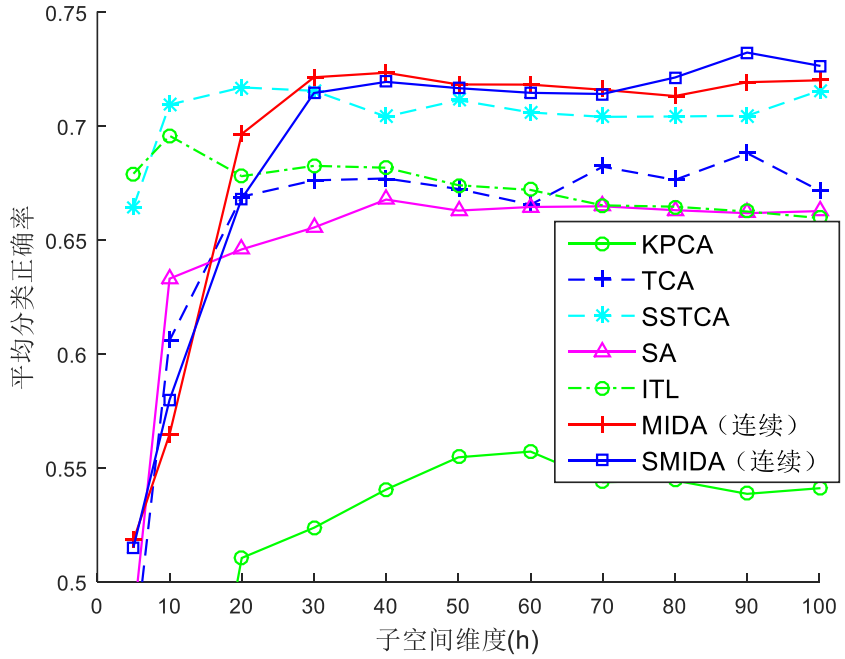


图6.3　阵列漂移数据集上的算法性能比较

SSTCA 的效果仅次于 MIDA、SMIDA。不过，从适用范围、参数数量和训练速度上来看，MIDA 和 SMIDA 比 SSTCA 更加实用。后者只能用于单一离散源领域和目标领域，而前两者可用于多领域和连续分布变化的情况；SSTCA 中使用了流形正则化来保留局部几何信息，从而比 SMIDA 多出 3 个参数；SSTCA 中的矩阵求逆和邻接图矩阵计算的时间复杂度较高。我们在一个配置了 Intel Core i5 CPU (2.20 GHz)和 8 GB RAM 的 PC 上对批次 1 和 2 学习领域不变子空间，编程环境为 Matlab 2014a，没有使用 GPU 和并行计算。最终，MIDA、SMIDA 和 SSTCA 的单次训练耗时分别为 1.25 s, 1.31 s 和 2.39 s。

6.4.3　Corn数据集

表 6.2 和图 6.4 展示了多种无监督领域适配算法在本数据集上的 RMSE。每种算法的参数都进行了寻优。对于每个从设备的每个预测任务的每一折交叉验证，我们利用主设备的有标签训练样本和从设备的无标签测试样本学习领域不变子空间。对于 MIDA 和 SMIDA，主设备的背景特征是$[1,0]^T$，从设备是$[0,1]^T$。在 KPCA、TCA、SSTCA、MIDA 和 SMIDA 中，核函数为线性核。基于信息论的学习方法 ITL[119]只适用于分类问题，因此没有在本数据集上测试。

表 6.2　Corn 数据集中各预测任务的 RMSE 比较

算法	Mp5 为从设备					Mp6 为从设备				
	水分	油脂	蛋白	淀粉	平均	水分	油脂	蛋白	淀粉	平均
无补偿	1.499	0.288	1.097	2.086	1.242	1.632	0.343	1.350	2.063	1.347
KPCA	0.278	0.123	0.258	0.477	0.284	0.267	0.102	0.322	0.499	0.298
TCA[118]	0.363	0.110	0.223	0.393	0.272	0.373	0.107	0.229	0.471	0.295
SSTCA[118]	0.299	0.100	0.148	0.339	0.221	0.314	0.107	0.162	0.357	0.235
SA[121]	0.276	0.099	0.182	0.382	0.235	0.307	0.096	0.168	0.383	0.239
GFK[122]	**0.263**	0.165	0.295	0.611	0.334	**0.266**	0.162	0.270	0.548	0.312
MIDA	0.287	**0.096**	**0.132**	**0.344**	**0.215**	0.318	0.099	**0.139**	0.360	0.229
SMIDA（无增强）	0.283	0.097	0.140	0.360	0.220	0.309	**0.093**	0.145	0.362	0.227
SMIDA	0.289	0.097	**0.132**	0.347	0.216	0.318	0.095	**0.139**	**0.350**	**0.225**
同设备训练	0.137	0.101	0.140	0.363	0.185	0.139	0.099	0.149	0.369	0.189

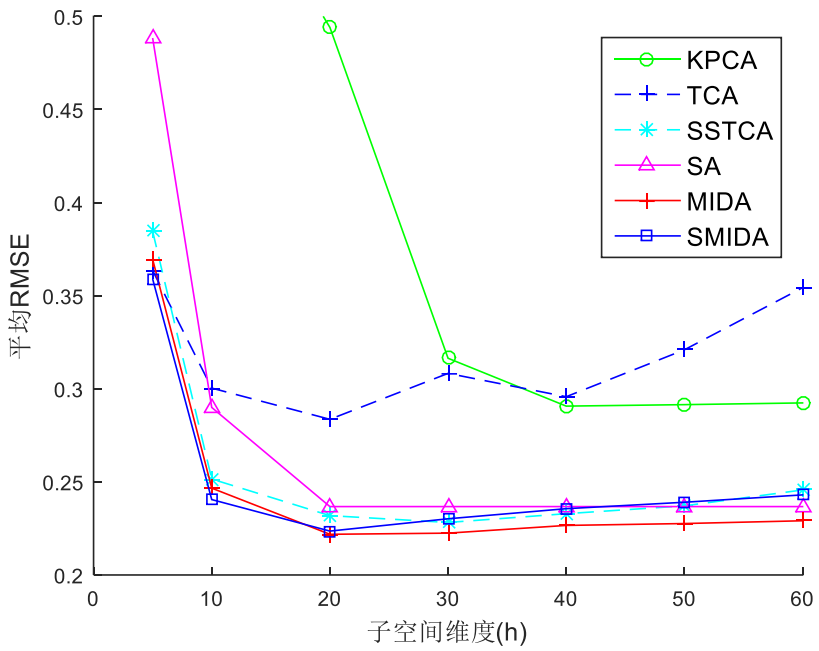


图6.4　Corn数据集上的上的算法性能比较

可以发现，相比“无补偿”，所有补偿算法都明显降低了 RMSE。MIDA 和 SMIDA 在两台从设备上都获得了最低的平均 RMSE。去掉特征增强的 SMIDA 在平均 RMSE 上略差于 SMIDA。与基于标定样本的算法相比，无需标定样本的 MIDA 和 SMIDA 虽然还不如前两章提出的 TMTL 和 DCAE（表 5.3），但已经超越了传统的特征标准化算法，与 DAELM 算法相近（表 4.4）。这证明了 MIDA 和 SMIDA 的有效性，也与本数据集漂移较简单有关。

6.4.4　呼气分析数据集

表 6.3　呼气分析数据集中各预测任务的 F-score 比较（%）

算法	任务 1	2	3	4	5	平均
无补偿	34.34	63.67	73.71	43.17	42.93	51.57
KPCA	57.17	71.31	84.74	51.36	41.89	61.29
SSA[131]	**91.64**	72.89	41.98	56.01	60.97	64.70
MIDA	75.71	67.37	76.77	**74.91**	70.14	72.98
SMIDA（无增强）	84.26	77.59	74.75	82.63	68.82	77.61
SMIDA	91.19	**81.53**	**92.09**	71.55	63.96	**80.07**
随机选取训练样本	87.85	85.44	90.10	85.09	84.54	86.60

在表 6.3 中，我们比较了能处理能够处理连续样本流的平稳子空间分析（SSA）、MIDA 和 SMIDA，以及简单利用所有样本学习核子空间的 KPCA。在 KPCA、MIDA 和 SMIDA 中，核函数为 RBF 核。所有算法的参数都经过了寻优。SSA 算法将所有样本视作以采集时间为自变量的多维时序信号，然后从中学习平稳信号子空间。MIDA 和 SMIDA 的背景特征构造方式与上一章相同，其中同时包含了每个样本的采集设备和采集时间信息。我们可以看出 SMIDA 算法明显优于其他算法。SMIDA 能够同时补偿设备差异和时变漂移，并且利用了有标签训练样本的标签信息。特征增强也对于提升正确率有所帮助。

6.5　本章小结

针对无标定样本情况下的漂移补偿问题，本章提出了最大独立领域适配算法（MIDA）。该算法可以利用目标领域的无标签样本学习领域不变特征。本章还提出了一种特征增强策略，用来对不同背景的样本进行差异化投影，以辅助领域对齐。实验证明，MIDA 以及其半监督版本 SMIDA 在漂移数据集上超越了被测的无监督领域适配算法。

与前面两章提出的 TMTL 和 DCAE 类似，MIDA 具有较强的灵活性。得益于背景特征以及 HSIC 准则的应用，MIDA 适用于各种领域适配问题，包括离散或连续的分布变化、单个或多个源领域或目标领域、无监督或半监督领域适配、分类或回归问题等。MIDA 的另一个优点是求解简单快速。因此，其具有较广阔的潜在应用前景。

第7章　总结和展望

7.1　总结

电子鼻应用场景丰富，可以给人们的生产生活带来诸多便利，因此正在引起越来越多研究者的兴趣。本文从系统搭建出发，利用特征选择算法提高了电子鼻的预测精度，设计漂移补偿算法提高了电子鼻的鲁棒性。

本文第 2 章设计了一套用于呼气分析的电子鼻系统，并采集了健康人和病人的呼气样本库，初步验证了该系统在糖尿病筛查和血糖预测问题上的有效性。该系统可以用于辅助医疗筛查和监控，具有无创、易用等优点。

特征选择是信号分析算法中重要的一环，可以提高预测准确率，同时帮助研究者更好地理解数据。本文第 3 章改进了 SVM 迭代特征删除算法（SVM-RFE）。提高了糖尿病呼气数据集的分类正确率，并得到了关于电子鼻软硬件的分析结论。

设备差异和时变漂移是影响电子鼻准确率和实用性的“短板”，也是本文重点研究的内容。本文提出了四种漂移补偿算法，其中第 4 章提出了两个模型层面的算法：结合标准化误差的模型改进（SEMI），和基于标定样本的多任务学习（TMTL）。SEMI 是一种加权正则项，可以加入其他的预测模型中，提高模型对于漂移的鲁棒性。TMTL 对于不同的领域（不同设备或时间段采集的样本）学习不同的预测模型。在第 5 章，我们描述了漂移补偿自编码器（DCAE），旨在学习无漂移且有较好鉴别能力的特征。针对无法获取标定样本的场合，我们在第 6 章提出了最大独立领域适配（MIDA），该算法能够基于目标领域的无标签样本学习领域不变子空间。

TMTL、DCAE 和 MIDA 这三种算法可以直接用于漂移补偿，其共同点在于：

- 能够同时用于设备差异和时变漂移的补偿，也就是能够处理样本分布离散和连续变化的情况；
- 同时考虑漂移补偿和鉴别能力提高：在 TMTL、DCAE 和 MIDA 的目标函数中都包含了有监督项；
- 使用灵活、可扩展性强，例如不限制样本中的有标签、无标签样本和标定样本的数目和所在领域；可以应用于分类或回归问题；TMTL 和 DCAE 可以与各种损失函数配合使用等。

TMTL、DCAE 和 MIDA 在算法设计层面的不同点总结于表 7.1。

表 7.1　TMTL、DCAE 和 MIDA 的算法差异

特点	TMTL	DCAE	MIDA
算法框架	多任务学习	自编码器	子空间学习
如何补偿？	在模型层面对齐标定样本	在学习到的特征层面对齐标定样本	最大独立准则，适配不同背景的无标签样本
如何避免欠迁移？	对不同背景的样本学习不同模型	对不同背景的样本应用不同补偿项	特征增强策略可以辅助领域适配
如何避免过迁移？	模型相似性约束；动态模型策略可以平滑模型	补偿项仅与背景有关且相对时间平缓变化	对不同背景的样本学习同一个子空间
如何处理样本流？	动态模型策略	利用背景特征	利用背景特征
适用场合	一般漂移补偿问题	时变漂移较复杂时；或样本分布较复杂、需要更有鉴别力的特征时	无法获取标定样本时

为了定量描述漂移补偿的效果，我们将三个数据集上的实验结果列于表 7.2，也就是第 4-6 章实验部分的总结。为简洁起见，只展示每个数据集的平均分类正确率或回归 RMSE。可以发现，本文提出的三种算法的精度相对于无补偿均有较大改善。相比之下，TMTL+SEMI 在仅包含设备差异的 Corn 数据集上获得了最好的效果，比无补偿的 RMSE 降低了 85.48%；DCAE 在包含复杂时变漂移的阵列漂移数据集和呼气分析数据集上效果更好，比起无补偿分别提高了 42.76 和 32.56 个百分点；在阵列漂移数据集上，TMTL 和 DCAE 都超越了现有漂移补偿文献中的方法。MIDA 没有使用标定样本，因此不适合与 TMTL+SEMI 和 DCAE 相比较，但也在三个数据集上相比无补偿的精度有大幅提高，并且超过了 6.4 节中测试的典型无监督领域适配算法。

表 7.2　漂移补偿算法精度比较

算法	阵列漂移数据集(%)	Corn 数据集(RMSE)	呼气分析数据集(%)
无补偿	50.61	1.295	51.57
TMTL+SEMI	92.37	**0.188**	81.73
DCAE	**93.22**	0.196	**84.13**
MIDA	73.23	0.221	80.07
目标结果	-	0.187	86.60

对于 Corn 和呼气分析数据集，我们还在表 7.2 中展示了“目标结果”，也就是通过一定方式得出的、消除漂移影响后理论上应该达到的精度。其中 Corn 数据集的目标结果是使用同一台从设备的样本训练和测试的，因此不受设备差异的影响；呼气分析数据集的目标结果是随机选取 50 个训练样本得到的，由于训练样本不局限在设备 1 的前 50 个样本，而是包含各种采集设备和时间，因此可以认为漂移的影响已在模型中被隐式考虑。我们发现，TMTL+SEMI 在 Corn 数据集上已经与目标结果不相上下，DCAE 在呼气分析数据集上也只与目标结果差 2.47 个百分点。总的来说，漂移补偿的效果与漂移的复杂程度密切相关。Corn 数据集的漂移比较简单（图 4.5），因此本文的算法在使用了 7 个标定样本后能够完全消除漂移的影响；呼气分析数据集的漂移稍复杂一些。而在阵列漂移数据集中，我们可以通过比较不同批次的分类正确率来观察补偿的效果。在 DCAE 的补偿结果中（表 5.2），批次 2-9 的正确率都 90%以上，可以认为漂移已经得到了较好的补偿。但批次 10 的正确率只有 76.82%，这是因为批次 10 距离批次 1 的时间最为久远，漂移最复杂[78]，无法通过简单的模型消除源领域和目标领域的分布差异，因此效果较差。这说明对于复杂漂移的场合，补偿算法仍有提升空间。在这种情况下，除了改进补偿算法，还可以通过重新采集训练样本、增加标定次数、控制工作环境等方法来提高精度。

7.2　创新点

本文的创新点主要包括：

1. 设计了实用的呼气分析系统，对其传感器阵列和气室等部分进行了优化。在血糖预测算法中，针对血糖与呼气标志物的关系存在个体差异的问题，提出了一种特征增强策略，用于将个体信息引入预测模型以提高预测精度。
2. 指出经典的特征选择算法 SVM-RFE 在特征高度相关时存在相关偏差问题，提出了相关偏差缩减策略；对电子鼻信号上提取的多种瞬态特征的排名结果进行了分析，得到了关于传感器、特征提取算法和温度调制电压的有用结论。
3. 提出结合标准化误差的模型改进（SEMI），对标准化误差大的特征进行惩罚，提高模型对于漂移的鲁棒性；提出基于标定样本的多任务学习算法（TMTL），针对设备差异和时变漂移的不同先验设计了不同的模型相似性约束；设计了基于模型加权平均的动态模型策略，使多任务学习算法可以更好地应用在分布连续变化的样本流场合。
4. 设计了背景特征，用来描述样本的采集设备和时间等背景属性，拓展了迁移学习中领域的概念；将漂移补偿与特征学习结合，提出了漂移补偿自编码器（DCAE），其中的补偿项可以对设备差异和时变漂移进行显式建模和补偿；

DCAE 在编码和解码阶段加入和减除补偿项，使得其隐层学习到的特征不受漂移影响；为了更灵活地补偿复杂时变漂移，还引入了非线性补偿层。

5. 将 Hilbert-Schmidt 独立性准则应用于领域适配问题，提出了最大独立准则，该准则通过最大化子空间中的特征与背景特征之间的独立性的方式来缩小领域差异。进而得到两种适用范围广、求解容易的领域适配算法：无监督和半监督最大独立领域适配（MIDA 和 SMIDA）。提出一种特征增强策略，将原始特征与背景特征串接，用来对不同背景的样本进行差异化投影，以辅助领域适配。

7.3　展望

无监督领域适配算法改进：虽然本文提出的 TMTL 和 DCAE 可以利用较少的标定样本达到较好的补偿精度，但不使用标定样本的 MIDA 算法的精度与 TMTL 和 DCAE 还有一定距离，仍有改进空间。现有大部分无监督领域适配算法只考虑适配样本的边缘分布 $P(X)$，但并不能保证不同领域的每一类样本都得到了对齐，也就是说还需要适配 $P(X|Y)$ [134]。在目标领域标签缺失的情况下，可以考虑首先对标签进行推测[134,135]，一种可能的研究方向是利用目标领域数据的聚类信息[119,136]。

算法适用范围拓展：我们设计的漂移补偿算法已经成功地应用在了电子鼻和光谱仪两个领域，未来可以考虑用其解决更多设备和传感器的漂移问题。算法可以补偿的漂移种类也不仅限于设备差异和时变漂移，可以对本文提出的背景特征进行拓展，将更多种类的背景属性包含在其中，例如采样地点、环境等，然后使用同样的算法进行补偿。本文算法还可以应用在更一般的特征选择和迁移学习问题中。

电子鼻应用探索：电子鼻的很多应用仍有待开发。以呼气分析为例，它具有无创无疼痛、样本易收集、操作简单成本低等优点，如果得到广泛应用，将会给病人带来极大的便利，同时带来社会和经济效益。推广的关键在于提高电子鼻和信号分析算法的精度和鲁棒性，这也是本文研究的主要目标。本文提出的呼气分析系统与实际应用的差距体现在：

- 本文中使用的数据集还不够大，每类只包含 114 至 440 个样本，这是由于采集病人的呼气样本比较困难，耗时费力且需要病人和医院配合。另外，数据集中的样本多样化程度还有待增强，病人应包含各种年龄、地域、疾病程度、并发症情况等。在投入实用之前，需要在更大规模和多样化的呼气样本库上验证系统的精度。

- 本文提出的漂移补偿算法 SEMI、TMTL 和 DCAE 均借助了标定样本。标定

样本是电子鼻领域常用的漂移补偿工具，但采集标定样本需要工具和专业知识。因此，如果在家庭中长期使用呼气分析系统的话，为了补偿长期漂移，就需要用户每隔一段时间找专业人员进行标定，这将带来不便。因此我们在第 6 章研究了无需标定样本的 MIDA 算法，但该算法的补偿精度尚不如 DCAE。如何兼顾易用性和精度，是系统投入实用前需要进一步研究的课题。将研究成果用于实践、服务国计民生，是本文的长远愿景。

参考文献

[1] Gardner J W, Bartlett P N. A brief history of electronic noses [J]. Sens Actuators B: Chem, 1994, 18(1): 211-220.

[2] Röck F, Barsan N, Weimar U. Electronic nose: current status and future trends [J]. Chem Rev, 2008, 108(2): 705-725.

[3] Chou J. Hazardous gas monitors: a practical guide to selection, operation and applications [M]. McGraw-Hill Professional Publishing, 2000.

[4] Zampolli S, Elmi I, Ahmed F, et al. An electronic nose based on solid state sensor arrays for low-cost indoor air quality monitoring applications [J]. Sens Actuators B: Chem, 2004, 101(1): 39-46.

[5] Romain A-C, Nicolas J. Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview [J]. Sens Actuators B: Chem, 2010, 146(2): 502-506.

[6] Zhang L, Tian F, Nie H, et al. Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine [J]. Sens Actuators B: Chem, 2012, 174: 114-125.

[7] Baietto M, Wilson A D. Electronic-nose applications for fruit identification, ripeness and quality grading [J]. Sensors, 2015, 15(1): 899-931.

[8] Qiu S, Wang J, Tang C, et al. Comparison of ELM, RF, and SVM on E-nose and E-tongue to trace the quality status of mandarin (Citrus unshiu Marc.) [J]. J Food Eng, 2015, 166: 193-203.

[9] Pardo M, Sberveglieri G. Coffee analysis with an electronic nose [J]. IEEE Trans Instrum Meas, 2002, 51(6): 1334-1339.

[10] García M, Fernández M, Fontecha J, et al. Differentiation of red wines using an electronic nose based on surface acoustic wave devices [J]. Talanta, 2006, 68(4): 1162-1165.

[11] Guo D. Automatic breath signal analysis and system design for medical applications [D]. Hong Kong: Department of Computing, The Hong Kong Polytechnic University, 2011.

[12] Turner A P, Magan N. Electronic noses and disease diagnostics [J]. Nature Reviews Microbiology, 2004, 2(2): 161-166.

[13] D'amico A, Di Natale C, Paolesse R, et al. Olfactory systems for medical applications [J]. Sens Actuators B: Chem, 2008, 130(1): 458-465.

[14] Guo D, Zhang D, Li N, et al. A novel breath analysis system based on electronic olfaction [J]. IEEE Trans Biomed Eng, 2010, 57(11): 2753-2763.

[15] Wilson A D, Baietto M. Advances in electronic-nose technologies developed for biomedical applications [J]. Sensors, 2011, 11(1): 1105-1176.

[16] Pearce T C, Schiffman S S, Nagle H T, et al. Handbook of machine olfaction: electronic nose technology [M]. John Wiley & Sons, 2006: 79-98.

[17] TGS 2600 datasheet [R/OL]. Figaro Inc., [May 2012]. http://www.figaro.co.jp/en/product/docs/tgs2600_product_information_rev02.pdf.

[18] Hierlemann A, Gutierrez-Osuna R. Higher-order chemical sensing [J]. Chem Rev, 2008, 108(2): 563-613.

[19] Marco S, Gutiérrez-Galvez A. Signal and data processing for machine olfaction and chemical sensing: a review [J]. IEEE Sens J, 2012, 12(11): 3189-3214.

[20] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Trans Knowl Data Eng, 2010, 22(10): 1345-1359.

[21] Feudale R N, Woody N A, Tan H, et al. Transfer of multivariate calibration models: a review [J]. Chemometr Intell Lab, 2002, 64(2): 181-192.

[22] Risby T H, Solga S. Current status of clinical breath analysis [J]. Appl Phys B: Lasers Opt, 2006, 85(2-3): 421-426.

[23] Deng C, Zhang J, Yu X, et al. Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization [J]. J Chromatogr B, 2004, 810(2): 269-275.

[24] Turner C, Walton C, Hoashi S, et al. Breath acetone concentration decreases with blood glucose concentration in type I diabetes mellitus patients during hypoglycaemic clamps [J]. J Breath Res, 2009, 3(4): 046004.

[25] Ueta I, Saito Y, Hosoe M, et al. Breath acetone analysis with miniaturized sample preparation device: In-needle preconcentration and subsequent determination by gas chromatography–mass spectroscopy [J]. J Chromatogr B, 2009, 877(24): 2551-2556.

[26] Lee J, Ngo J, Blake D, et al. Improved predictive models for plasma glucose estimation from multi-linear regression analysis of exhaled volatile organic compounds [J]. J Appl Physiol, 2009, 107(1): 155-160.

[27] Wang C, Mbi A, Shepherd M. A study on breath acetone in diabetic patients using a cavity ringdown breath analyzer: Exploring correlations of breath acetone with blood glucose and glycohemoglobin a1c [J]. IEEE Sens J, 2010, 10(1): 54-63.

[28] Turner C. Potential of breath and skin analysis for monitoring blood glucose concentration in diabetes [J]. Expert Rev Mol Diagn, 2011, 11(5): 497-503.

[29] Španěl P, Dryahina K, Rejšková A, et al. Breath acetone concentration; biological variability and the influence of diet [J]. Physiol Meas, 2011, 32(8): N23.

[30] Minh T D C, Blake D R, Galassetti P R. The clinical potential of exhaled breath analysis for diabetes mellitus [J]. Diabetes Res Clin Pr, 2012, 97(2): 195-205.

[31] Davies S, Spanel P, Smith D. Quantitative analysis of ammonia on the breath of patients in end-stage renal failure [J]. Kidney Int, 1997, 52: 223-228.

[32] Narasimhan L, Goodman W, Patel C K N. Correlation of breath ammonia with blood urea nitrogen and creatinine during hemodialysis [J]. Proceedings of the National Academy of Sciences, 2001, 98(8): 4617-4621.

[33] Sobotka A, Brottman D, Weitz E, et al. Elevated breath pentane in heart failure reduced by free radical scavenger [J]. Free Radical Biology and Medicine, 1993, 14(6): 643-647.

[34] Phillips M, Cataneo R N, Cummin A R, et al. Detection of lung cancer with volatile markers in the breath [J]. Chest Journal, 2003, 123(6): 2115-2123.

[35] Phillips M, Cataneo R N, Ditkoff B A, et al. Volatile markers of breast cancer in the breath [J]. The Breast Journal, 2003, 9(3): 184-191.

[36] Eisenmann A, Amann A, Said M, et al. Implementation and interpretation of hydrogen breath tests [J]. J Breath Res, 2008, 2(4): 046002.

[37] Wang P, Tan Y, Xie H, et al. A novel method for diabetes diagnosis based on electronic nose [J]. Biosens Bioelectron, 1997, 12(9): 1031-1036.

[38] Lin Y-J, Guo H-R, Chang Y-H, et al. Application of the electronic nose for uremia diagnosis [J]. Sens Actuators B: Chem, 2001, 76(1): 177-180.

[39] Dragonieri S, Schot R, Mertens B J, et al. An electronic nose in the discrimination of patients with asthma and controls [J]. J Allergy Clin Immunol, 2007, 120(4): 856-862.

[40] Yu J-B, Byun H-G, So M-S, et al. Analysis of diabetic patient's breath with conducting polymer sensor array [J]. Sens Actuators B: Chem, 2005, 108(1): 305-308.

[41] Guo D, Zhang D, Zhang L, et al. Non-invasive blood glucose monitoring for diabetics by means of breath signal analysis [J]. Sens Actuators B: Chem, 2012, 173(6): 106-113.

[42] Saraoğlu H M, Selvi A O, Ebeoğlu M A, et al. Electronic Nose System Based on Quartz Crystal Microbalance Sensor for Blood Glucose and HbA1c Levels From Exhaled Breath Odor [J]. IEEE Sens J, 2013, 13(11): 4229 - 4235.

[43] Yan K, Zhang D. Sensor Evaluation in a Breath Analysis System [C]. // Proceedings of the Medical Biometrics (ICMB), 2014 International Conference on. Shenzhen: IEEE, 2014: 35-40.

[44] Cao W, Duan Y. Current status of methods and techniques for breath analysis [J]. Crit Rev Anal Chem, 2007, 37(1): 3-13.

[45] Wolfrum E J, Meglen R M, Peterson D, et al. Metal oxide sensor arrays for the detection, differentiation, and quantification of volatile organic compounds at sub-parts-per-million concentration levels [J]. Sens Actuators B: Chem, 2006, 115(1): 322-329.

[46] Hosseini-Golgoo S, Hossein-Babaei F. Assessing the diagnostic information in the response patterns of a temperature-modulated tin oxide gas sensor [J]. Meas Sci Technol, 2011, 22(3): 035201.

[47] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. J Mach Learn Res, 2003, 3: 1157-1182.

[48] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23(19): 2507-2517.

[49] Tang J, Alelyani S, Liu H. Feature selection for classification: A review [M]//Aggarwal C C. Data Classification: Algorithms and Applications. Chapman and Hall/CRC. 2014: 37.

[50] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Mach Learn, 2002, 46(1-3): 389-422.

[51] Rakotomamonjy A. Variable selection using SVM based criteria [J]. J Mach Learn Res, 2003, 3: 1357-1370.

[52] Mundra P A, Rajapakse J C. SVM-RFE with MRMR filter for gene selection [J]. IEEE T NanoBiosci, 2010, 9(1): 31-37.

[53] Burges C J. A tutorial on support vector machines for pattern recognition [J]. Data Min Knowl Disc, 1998, 2(2): 121-167.

[54] Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions [J]. Bioinformatics, 2011, 27(14): 1986-1994.

[55] Park M Y, Hastie T, Tibshirani R. Averaged gene expressions for regression [J]. Biostatistics, 2007, 8(2): 212-227.

[56] Sharma D B, Bondell H D, Zhang H H. Consistent group identification and variable selection in regression with correlated predictors [J]. J Comput Graph Stat, 2013, 22(2): 319-340.

[57] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces [J]. Knowl Inf Syst, 2007, 12(1): 95-116.

[58] Awada W, Khoshgoftaar T M, Dittman D, et al. A review of the stability of feature selection techniques for bioinformatics data [C]. // Proceedings of the Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on. Las Vegas, USA: IEEE, 2012:356-363.

[59] Martinelli E, Falconi C, D'amico A, et al. Feature extraction of chemical sensors in phase space [J]. Sens Actuators B: Chem, 2003, 95(1): 132-139.

[60] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Trans Pattern Anal Mach Intell, 2005, 27(8): 1226-1238.

[61] Yan K. Feature selection with SVM-RFE [CP/OL]. 2015 [2016/04/10]. http://www.mathworks.com/matlabcentral/fileexchange/50701-feature-selection-with-svm-rfe.

[62] Zhang L, Tian F, Kadri C, et al. On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality [J]. Sens Actuators B: Chem, 2011, 160(1): 899-909.

[63] Wang Y, Veltkamp D J, Kowalski B R. Multivariate instrument standardization [J]. Anal Chem, 1991, 63(23): 2750-2756.

[64] Fan W, Liang Y, Yuan D, et al. Calibration model transfer for near-infrared spectra based on canonical correlation analysis [J]. Anal Chim Acta, 2008, 623(1): 22-29.

[65] Deshmukh S, Kamde K, Jana A, et al. Calibration transfer between electronic nose systems for rapid in situ measurement of pulp and paper industry emissions [J]. Anal Chim Acta, 2014, 841: 58-67.

[66] Zhang L, Tian F, Peng X, et al. Standardization of metal oxide sensor array using artificial neural networks through experimental design [J]. Sens Actuators B: Chem, 2013, 177: 947-955.

[67] Fernandez L, Guney S, Gutierrez-Galvez A, et al. Calibration transfer in temperature modulated gas sensor arrays [J]. Sens Actuators B: Chem, 2016, 231: 276-284.

[68] Galvão R K H, Soares S F C, Martins M N, et al. Calibration transfer employing univariate correction and robust regression [J]. Anal Chim Acta, 2015, 864: 1-8.

[69] Kalivas J H, Siano G G, Andries E, et al. Calibration maintenance and transfer using Tikhonov regularization approaches [J]. Appl Spectrosc, 2009, 63(7): 800-809.

[70] Yu B, Ji H. Near-infrared calibration transfer via support vector machine and transfer learning [J]. Anal Methods, 2015, 7(6): 2714-2725.

[71] Di Carlo S, Falasconi M. Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges [M]//WANG W. Advances in Chemical Sensors. InTech. 2012: 305-326.

[72] 刘涛. 电子鼻长期漂移抑制算法的研究 [博士学位论文]. 重庆: 重庆大学通信工程学院, 2012.

[73] Haugen J-E, Tomic O, Kvaal K. A calibration method for handling the temporal drift of solid state gas-sensors [J]. Anal Chim Acta, 2000, 407(1): 23-39.

[74] Artursson T, Eklöv T, Lundström I, et al. Drift correction for gas sensors using multivariate methods [J]. J Chemometr, 2000, 14(5-6): 711-723.

[75] Wold S, Antti H, Lindgren F, et al. Orthogonal signal correction of near-infrared spectra [J]. Chemometr Intell Lab, 1998, 44(1): 175-185.

[76] Di Natale C, Martinelli E, D'amico A. Counteraction of environmental disturbances of electronic nose data by independent component analysis [J]. Sens Actuators B: Chem, 2002, 82(2): 158-165.

[77] Zhang L, Zhang D. Domain Adaptation Extreme Learning Machines for Drift Compensation in E-Nose Systems [J]. IEEE Trans Instrum Meas, 2015, 64(7): 1790 - 1801.

[78] Vergara A, Vembu S, Ayhan T, et al. Chemical gas sensor drift compensation using classifier ensembles [J]. Sens Actuators B: Chem, 2012, 166: 320-329.

[79] Liu Q, Li X, Ye M, et al. Drift compensation for electronic nose by semi-supervised domain adaption [J]. IEEE Sens J, 2014, 14(3): 657-665.

[80] De Vito S, Fattoruso G, Pardo M, et al. Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction [J]. IEEE Sens J, 2012, 12(11): 3215-3224.

[81] Rodionova O Y, Pomerantsev A L. Subset selection strategy [J]. J Chemometr, 2008, 22(11-12): 674-685.

[82] Kennard R W, Stone L A. Computer aided design of experiments [J]. Technometrics, 1969, 11(1): 137-148.

[83] Yu K, Bi J, Tresp V. Active learning via transductive experimental design [C]. // Proceedings of the Intl Conf on Machine Learning (ICML). Pittsburgh: ACM, 2006: 1081-1088.

[84] Zhang L, Chen C, Bu J, et al. Active Learning Based on Locally Linear Reconstruction [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(10): 2026-2038.

[85] Chou J. Gas Sensor Calibration [M]//Chou J. Hazardous gas monitors: a practical guide to selection, operation and applications. McGraw-Hill Professional Publishing.1999: 161-173.

[86] Yan K, Zhang D. Improving the transfer ability of prediction models for electronic noses [J]. Sens Actuators B: Chem, 2015, 220: 115-124.

[87] Caruana R. Multitask learning [J]. Mach Learn, 1997, 28(1): 41-75.

[88] Evgeniou T, Pontil M. Regularized multi-task learning [C]. // Proceedings of the Proc ACM SIGKDD. Seattle, Washington: ACM, 2004:109-117.

[89] Zhou J, Yuan L, Liu J, et al. A multi-task learning formulation for predicting disease progression [C]. // Proceedings of the Proc ACM SIGKDD. San Diego, California: ACM, 2011:814-822.

[90] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning [M]. 2nd ed. New York: Springer, 2009.

[91] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation [J]. ACM Computing Surveys (CSUR), 2014, 46(4): 44.

[92] Yan K, Zhang D. Calibration transfer and drift compensation of e-noses via coupled task learning [J]. Sens Actuators B: Chem, 2016, 225: 288-297.

[93] Zhou J, Chen J, Ye J. MALSAR: Multi-tAsk Learning via StructurAl Regularization [CP/OL]. 2011 [2015-10-21]. http://www.public.asu.edu/~jye02/Software/MALSAR.

[94] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Trans Pattern Anal Mach Intell, 2013, 35(8): 1798-1828.

[95] Bengio Y. Deep learning of representations for unsupervised and transfer learning [J]. Unsupervised and Transfer Learning Challenges in Machine Learning, 2012, 7: 19.

[96] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[97] Längkvist M, Loutfi A. Unsupervised feature learning for electronic nose data applied to Bacteria Identification in Blood [C]. // Proceedings of the NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning. 2011

[98] Längkvist M, Coradeschi S, Loutfi A, et al. Fast classification of meat spoilage markers using nanostructured ZnO thin films and unsupervised feature learning [J]. Sensors, 2013, 13(2): 1578-1592.

[99] Längkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling [J]. Pattern Recogn Lett, 2014, 42: 11-24.

[100] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion [J]. J Mach Learn Res, 2010, 11: 3371-3408.

[101] Kandaswamy C, Silva L M, Alexandre L A, et al. Improving deep neural network performance by reusing features trained with transductive transference [C]. // Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2014. Springer, 2014: 265-272.

[102] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach [C]. // Proceedings of the Intl Conf on Machine Learning (ICML). 2011: 513-520.

[103] Chopra S, Balakrishnan S, Gopalan R. Dlid: Deep learning for domain adaptation by interpolating between domains [C]. // Proceedings of the ICML workshop on challenges in representation learning. 2013: 5.

[104] Deng J, Zhang Z, Eyben F, et al. Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition [J]. IEEE Signal Process Lett, 2014, 21(9): 1068-1072.

[105] Zhuang F, Cheng X, Luo P, et al. Supervised representation learning: transfer learning with deep autoencoders [C]. // Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press, 2015:4119-4125.

[106] Zhou J T, Pan S J, Tsang I W, et al. Hybrid heterogeneous transfer learning through deep learning [C]. // Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014

[107] Chen W-R, Bin J, Lu H-M, et al. Calibration transfer via an extreme learning machine auto-encoder [J]. Analyst, 2016, 141(6): 1973-1980.

[108] Kan M, Shan S, Chang H, et al. Stacked Progressive Auto-Encoders (SPAE) for Face Recognition across Poses [C]. // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[109] Bengio Y. Practical recommendations for gradient-based training of deep architectures [M] // Montavon G, Orr G B, Müller K-R. Neural Networks: Tricks of the Trade. Springer. 2012: 437-478.

[110] Kadlec P, Grbić R, Gabrys B. Review of adaptation mechanisms for data-driven soft sensors [J]. Comput Chem Eng, 2011, 35(1): 1-24.

[111] Bergstra J, Breuleux O, Bastien F, et al. Theano: a CPU and GPU math expression compiler [C]. // Proceedings of the Python for scientific computing conference (SciPy). Austin, TX, 2010:3.

[112] Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation [C]. // Proceedings of the Advances in neural information processing systems. 2007:137.

[113] Liu Q, Hu X, Ye M, et al. Gas Recognition under Sensor Drift by Using Deep Learning [J]. International Journal of Intelligent Systems, 2015, 30(8): 907-922.

[114] Jiang J. A literature survey on domain adaptation of statistical classifiers [R/OL]. 2008 [2014-6-25]. http://www.mysmu.edu/faculty/jingjiang/papers/da_survey.pdf.

[115] Lu J, Behbood V, Hao P, et al. Transfer learning using computational intelligence: a survey [J]. Knowledge-Based Systems, 2015, 80: 14-23.

[116] Gong B, Grauman K, Sha F. Learning kernels for unsupervised domain adaptation with applications to visual object recognition [J]. Int J Comput Vision, 2014, 109(1-2): 3-27.

[117] Patel V M, Gopalan R, Li R, et al. Visual domain adaptation: A survey of recent advances [J]. IEEE Signal Proc Mag, 2015, 32(3): 53-69.

[118] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis [J]. IEEE T Neural Networ, 2011, 22(2): 199-210.

[119] Shi Y, Sha F. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation [C]. // Proceedings of the Intl Conf on Machine Learning (ICML). Edinburgh: 2012.

[120] Shao M, Kit D, Fu Y. Generalized transfer subspace learning through low-rank constraint [J]. Int J Comput Vision, 2014, 109(1-2): 74-93.

[121] Fernando B, Habrard A, Sebban M, et al. Unsupervised visual domain adaptation using subspace alignment [C]. // Proceedings of the IEEE International Conference on Computer Vision. 2013: 2960-2967.

[122] Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation [C]. // Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012:2066-2073.

[123] Cui Z, Li W, Xu D, et al. Flowing on Riemannian manifold: Domain adaptation by shifting covariance [J]. IEEE Trans Cybern, 2014, 44(12): 2264-2273.

[124] Gretton A, Bousquet O, Smola A, et al. Measuring statistical dependence with Hilbert-Schmidt norms [C]. // Proceedings of the Algorithmic learning theory. Springer, 2005:63-77.

[125] Song L, Smola A, Gretton A, et al. Feature selection via dependence maximization [J]. J Mach Learn Res, 2012, 13(1): 1393-1434.

[126] Song L, Gretton A, Borgwardt K M, et al. Colored maximum variance unfolding [C]. // Proceedings of the Advances in neural information processing systems. 2007: 1385-1392.

[127] Barshan E, Ghodsi A, Azimifar Z, et al. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds [J]. Pattern Recogn, 2011, 44(7): 1357-1371.

[128] Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Comput, 1998, 10(5): 1299-1319.

[129] Schölkopf B, Mullert K-R. Fisher discriminant analysis with kernels [C]. // Proceedings of the IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing. Madison, WI: IEEE, 1999:41-48.

[130] Smola A, Gretton A, Song L, et al. A Hilbert space embedding for distributions [C]. // Proceedings of the Algorithmic learning theory. Springer, 2007:13-31.

[131] Von Bünau P, Meinecke F C, Király F C, et al. Finding stationary subspaces in multivariate time series [J]. Phys Rev Lett, 2009, 103(21): 214101.

[132] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. J Mach Learn Res, 2006, 7: 2399-2434.

[133] Daumé III H. Frustratingly easy domain adaptation [C]. // Proceedings of the 45th Ann Meeting of the Assoc for Computational Linguistics. Baltimore: 2007.

[134] 龙明盛. 迁移学习问题与方法研究 [博士学位论文]. 北京: 清华大学计算机科学与技术系, 2014: 37-41.

[135] Joachims T. Transductive inference for text classification using support vector machines [C]. // Proceedings of the Intl Conf on Machine Learning (ICML). 1999: 200-209.

[136] Hong X, Wang J, Qi G. Comparison of semi-supervised and supervised approaches for classification of e-nose datasets: Case studies of tomato juices [J]. Chemometr Intell Lab, 2015, 146: 457-463.

致　谢

衷心感谢我的导师张大鹏老师，本文的选题、实验、分析、撰写都离不开您的悉心指导。您不仅在学术上谆谆教诲，为我指引了前进方向，更用您的工作热情和人格魅力在我的人生中树立了学习的榜样。这个课题是一个较前瞻性的课题，研究过程中需要在硬件、采样、软件等方面解决种种问题。导师的鼓励和督促让我和整个实验室团队一起奋勇直前。在课题上的付出同时也让我在诸多方面获得了提高。

感谢深圳研究生院的郭振华老师，您在我的整个博士求学过程中提出了太多有益的建议，参加您的组会也让我受益匪浅。感谢哈尔滨工业大学的徐勇老师和卢光明老师在科研和生活方面给我的巨大帮助。感谢章毓晋、廖庆敏、陈友斌、王生进、林行刚等多位老师在我的课程、开题、答辩等方面给我的指导。感谢哈工大的骆南老师在硬件方面的指点。

感谢香港理工大学的郭东敏博士和重庆大学的张磊博士，你们给了我很多电子鼻方面的专业意见，帮助我在此领域入门和提高。感谢陈芳梅、王迪民、房小兆、曲晓峰、王兴政、武克斌等多位同门兄弟姐妹，同你们讨论问题总给我带来新知和启发。还有很多清华、哈工大、港理工的好友，在这六年中认识你们是我的幸运。谢谢宁宇婧、张霞和刘东旭等在设备制作、样本采集等方面的合作。

值得一提的是广东省中医院与我们合作采集呼气样本的医生们，你们的敬业精神让我感动。还有配合采样的志愿者、患者们，谢谢你们的付出。

最后，感谢我的父母，不管我遇到什么困难，你们总在背后默默关心、支持着我。由衷感谢本文的评阅人老师抽出宝贵时间评审本文。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1988 年 9 月 5 日出生于北京市。

2006 年 9 月考入清华大学电子工程系电子信息科学与技术专业，2010 年 7 月本科毕业并获得工学学士学位。

2010 年 9 月免试进入清华大学电子工程系信息与通信工程专业攻读工学博士学位至今。

发表的学术论文

[1] Yan Ke, Zhang David, Wu Darong, et al. Design of a breath analysis system for diabetes screening and blood glucose level prediction [J]. IEEE Transactions on Biomedical Engineering, 2014, 61(11): 2787-2795. (SCI 收录，检索号: AS2AS，影响因子: 2.347.)

[2] Yan Ke, Zhang David. Feature selection and analysis on correlated gas sensor data with recursive feature elimination [J]. Sensors and Actuators B: Chemical, 2015, 212: 353-363. (SCI 收录，检索号: CD3XS，影响因子: 4.097.)

[3] Yan Ke, Zhang David. Improving the transfer ability of prediction models for electronic noses [J]. Sensors and Actuators B: Chemical, 2015, 220: 115-124. (SCI 收录，检索号: CR3QX，影响因子: 4.097.)

[4] Yan Ke, Zhang David. Calibration transfer and drift compensation of e-noses via coupled task learning [J]. Sensors and Actuators B: Chemical, 2016, 225: 288-297. (SCI 收录，检索号: CY9XK，影响因子: 4.097.)

[5] Yan Ke, Zhang David. Blood glucose prediction by breath analysis system with feature selection and model fusion [C]. // Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Chicago: IEEE, 2014: 6406-6409. (EI 收录，检索号: 20152100873914. 生医系顶级会议.)

[6] Yan Ke, Zhang David. Sensor evaluation in a breath analysis system [C]. // Proceedings of the 2014 International Conference on Medical Biometrics (ICMB). Shenzhen: IEEE, 2014: 35-40. (EI 收录，检索号: 20143117993386.)

[7] Yan Ke, Zhang David. A novel breath analysis system for diabetes diagnosis [C]. // Proceedings of the 2012 International Conference on Computerized

Healthcare. Hong Kong: IEEE, 2012: 166-170. (EI 收录，检索号: 20140917372380.)

[8] Yan Ke, Chen Youbin, Zhang David. Gabor surface feature for face recognition [C]. // Proceedings of the 2011 First Asian Conference on Pattern Recognition (ACPR). IEEE, 2011: 288-292. (EI 收录，检索号: 20122715188495.)

[9] Yan Ke and Zhang David. Correcting instrumental variation and time-varying drift: A transfer learning approach with autoencoders [J]. IEEE Transactions on Instrumentation and Measurement. (已录用，SCI 收录，影响因子: 1.790.)

[10] Ning Yujing, Lu Guangming, Yan Ke, et al. Standardization of gas sensors in a breath analysis system [C]. // Proceedings of the 2014 9th Chinese Conference on Biometric Recognition (CCBR), 2014: 434-441. (EI 收录，检索号: 20144800259531.)