基于动态特征重要度的电子鼻传感器阵列优化方法

魏广芬^{1*}, 赵捷², 冯烟利², 何爱香¹, 余隽³, 唐祯安³

1. 山东工商学院信息与电子工程学院, 烟台 264005

2. 山东工商学院计算机科学与技术学院, 烟台 264005

3. 大连理工大学生物医学工程学院, 大连 116024

* 通信作者. E-mail: guangfen.wei@yahoo.com

收稿日期: 2018-11-12; 修回日期: 2019-01-09; 接受日期: 2019-03-21; 网络出版日期: 2020-04-27

国家自然科学基金 (批准号: 61174007) 和烟台市科技发展计划 (批准号: 2016ZH053, 2017ZH063) 资助项目

摘要 气体传感器阵列的优化是电子鼻领域亟需解决的关键问题之一, 同时也是一种特殊的特征选择问题. 本文结合特征相关性和特征重要度, 提出了一种有效的新型传感器 (特征) 重要性衡量方法——动态特征重要度, 并在此基础上提出了一种新的基于动态特征重要度的电子鼻传感器阵列优化算法 SAO_DFI. 通过对两种不同的气体环境下采集的数据进行分析, 测试了重复传感器、传感器 (特征) 重要度、传感器 (特征) 相关性以及传感器特征参数对 SAO_DFI 算法的影响, 其优化结果证明了该阵列优化算法的有效性、鲁棒性和可解释性.

关键词 阵列优化, 气体传感器, 特征选择, 特征相关性, 电子鼻, 动态特征重要度

1 引言

电子鼻 (electronic nose, EN) 在 20 世纪末被正式提出^[1], 作为一种人工嗅觉系统, 它由适当的模式识别系统和具有交叉选择性的气体传感器组成的传感器阵列构成^[1,2], 是一种能够识别单一或复杂气味/气体的仪器, 其应用领域包括食品工业、环境监测、军事安全等, 是当前气体检测领域的研究热点. 基于气体传感器阵列的电子鼻系统的提出主要是为了解决气体传感器选择性差的问题. 因为实际测量环境中的气体往往是以多种成分混合的状态存在, 而气体传感器选择性差、交叉敏感的特点使得它在工作容易受到环境中其他气体甚至是温湿度的干扰. 有学者通过改善传感器结构、制备新型气敏材料以及研究新的传感器控制模式如温度调制等方法来克服气体传感器存在的问题, 如 Evans 等^[3]提出的基于碳纳米管与沸石结构的传感器降低了对水蒸气的交叉敏感性; Huang 等^[4,5]测试了凝胶溶胶法研制的厚膜传感器在不同频率、波形, 以及温度下对低浓度丙酮的响应; Kabir 等^[6]设计的基于声表面波的气体传感器对 Hg^0 在 35°C 和 75°C 时的检测下限分别在 2.6 和 1 ppbv (parts per billion

引用格式: 魏广芬, 赵捷, 冯烟利, 等. 基于动态特征重要度的电子鼻传感器阵列优化方法. 中国科学: 信息科学, 2020, 50: 743–765, doi: 10.1360/N112018-00302
Wei G F, Zhao J, Feng Y L, et al. An effective gas sensor array optimization method based on dynamic feature importance (in Chinese). Sci Sin Inform, 2020, 50: 743–765, doi: 10.1360/N112018-00302

by volume, 每体积十亿分之一); Zhu 等^[7]揭示了 SnO_2 基乙醇传感器的传感反应, 研究不同湿度下传感器的交叉敏感性, 为湿度对 SnO_2 传感器的争议影响提供了新的见解。

然而从基础研究到实际应用, 气体传感器的性能差别依然很大, 通过单一的气体传感器仍然难以完成复杂的任务需求, 而气体传感器阵列与模式识别算法的结合有效地改善了气体传感器选择性差、交叉敏感的问题^[2,8~10]。因此, 基于气体传感器阵列的电子鼻技术成为当前气体检测领域的研究热点, 而如何选择或优化传感器阵列也就成为了目前电子鼻系统中亟需解决的关键问题之一。

由于交叉敏感性, 在对待测对象进行测量之前, 往往难以获知某个传感器对分类判别结果是否起作用, 因此在为具体应用选择初始阵列时, 通常会尽可能多地选择能够对环境响应的不同类型的传感器。气体传感器阵列优化一般是通过对实验数据的处理来筛选传感器, 消除初始阵列中的冗余信息^[11,12], 提高信噪比, 从而得到优化阵列, 提高系统的识别率, 并且以期同时降低接口电路的复杂性、功耗和成本等。阵列优化的一般意义有: 其一, 可以在最大程度上减小传感器阵列的规模, 避免由于传感器数量过多而导致后续数据处理而产生的“维数灾难”^[13]; 其二, 有利于剔除阵列中的冗余传感器, 用尽量少的传感器来完成指定的测试任务, 以满足传感器阵列的微型化, 乃至电子鼻系统微型化的需要。在当前数据处理能力逐渐提高的情况下, 第二个意义更加突出。

随着当前计算能力和数据处理能力的提高, 数据维度也逐渐升高, 特征选择 (feature selection, FS) 已经成为模式识别、数据挖掘领域的重要研究问题。它通过一定的策略发现特征与类别的内在联系, 选择原始特征集合中的有效特征构建特征子集, 达到降低数据维数, 提高或保持系统分类能力的目的^[14~16]。特征选择通常包括 4 个基本步骤: 特征子集的搜索、特征子集评价、终止条件判断和特征子集验证。特征选择算法按照子集的搜索策略可分为 3 类: 全局最优、随机搜索和启发式搜索。全局最优往往需要穷举或分支定界, 这种策略在处理高维数据时, 尤其是难以确定特征数量时, 受到许多限制无法广泛应用, 即便它可以获得全局最优解; 相比之下, 随机搜索的复杂度低但搜索效率不稳定, 依赖于参数, 常用的随机搜索有模拟退火、遗传算法; 启发式搜索方法利用问题拥有的启发信息来引导搜索, 其特点是效率高, 不过要以牺牲全局最优为代价。当前主要的特征选择方法包括顺序前向选择方法 (sequential forward selection, SFS)^[17], 顺序浮动前向选择方法 (sequential floating forward selection, SFFS)^[18], 顺序后向选择方法 (sequential backward selection, SBS)^[19] 等, 这些方法一般需要评估函数对特征进行排序。

有学者研究表明, 无关特征和冗余特征会影响学习模型的精度和性能^[20], 但多数特征选择算法将重点放在特征对分类的贡献, 而对特征的冗余性关注不足。传感器阵列优化 (sensor array optimization, SAO) 的目的是优化传感器的组成, 删除阵列中影响电子鼻检测精度和性能的无关或冗余传感器。因此冗余传感器 (特征) 的分析在 SAO 领域恰恰是一个不可忽视的问题。针对特征冗余性, Hall^[21]提出了基于特征相关性的方法 (correlation-based feature selection, CFS), 通过皮尔逊相关系数 (Pearson correlation coefficient, PCC) 和对称不确定性评价特征与特征之间、特征与类别之间的相关性来构建特征子集; Peng 等^[22]提出的最小冗余最大相关方法 (minimum redundancy maximum relevance, mRMR) 使用互信息和 F 统计量定义了最小冗余 - 最大相关性标准。这些方法以特征相关性为核心, 其对冗余特征的讨论在很多应用中显出优势^[15,23]。

基于此, 本文提出了一种新的特征评价函数——动态特征重要度 (dynamic feature importance, DFI) 用于评估特征的重要性。将 DFI 应用于电子鼻传感器阵列优化, 提出了基于 DFI 的气体传感器阵列优化算法 SAO.DFI。通过在两种不同气体环境下采集的数据上进行实验, 分析了传感器的重要性衡量结果, 研究了 SAO.DFI 算法的性能。本文第 1 节先介绍电子鼻的结构和气体传感器阵列优化的一般问题; 第 2 节介绍常用的特征重要度 (feature importance, FI) 和特征相关性 (feature-feature

图 1 (网络版彩图) 人体嗅觉系统与电子鼻系统对照示意图

Figure 1 (Color online) Schematic diagram of human olfaction system and electronic nose system

correlation, FFC) 并提出 DFI 和 SAO_DFI 算法; 第 3 节介绍实验平台及数据的获取; 第 4 节通过对数据的处理和分析, 讨论不同的参数设置对 SAO_DFI 算法的影响; 最后给出结论.

2 气体传感器阵列优化

2.1 电子鼻结构与工作原理

电子鼻是一种模拟人和其他哺乳动物的嗅觉系统的仿生学仪器^[24], 主要由气体传感器阵列、信号预处理和模式识别 3 个部分组成, 实现对气味或气体的感知、处理和判别. 图 1 为电子鼻与人类嗅觉系统的示意图, 可以看出气体传感器阵列就像人体嗅觉系统中的嗅觉受体细胞和鼻上皮细胞等初级嗅觉细胞, 完成对气体分子的吸附和反应; 信号预处理部分用于采集、传输和调节信号, 其功能与人体中嗅球的功能类似; 所有信号最终通过电子鼻的“大脑”, 即模式识别系统对气味或气体信息进行判别和分析并输出结果^[24, 25].

2.2 气体传感器阵列优化的一般问题

假设选择 M 个传感器组成初始传感器阵列 A , 现将 $A(s_1, s_2, \dots, s_M)$ 中任一传感器的响应信号记作 $s_i(t)$, $i = 1, 2, \dots, M$, 则传感器阵列在 t 时刻的输出为响应向量 $S(t) = [s_1(t), s_2(t), \dots, s_M(t)]$.

假设某一具体应用是对被测环境中的 K 种气体及其混合气体测试, 用 $G = \{g_j | j = 1, 2, \dots, K\}$

表示, 则气体传感器阵列 $\mathbf{A}$ 在 $t_1, t_2, \dots, t_d$ 时刻对 $g_j \in G$ 的响应可表示为

$$\mathbf{S}(g_j) = \begin{pmatrix} S(t_1|g_j) \\ S(t_2|g_j) \\ \vdots \\ S(t_d|g_j) \end{pmatrix} = \begin{pmatrix} s_1(t_1|g_j) & s_2(t_1|g_j) & \cdots & s_M(t_1|g_j) \\ s_1(t_2|g_j) & s_2(t_2|g_j) & \cdots & s_M(t_2|g_j) \\ \vdots & \vdots & \ddots & \vdots \\ s_1(t_d|g_j) & s_2(t_d|g_j) & \cdots & s_M(t_d|g_j) \end{pmatrix}, \quad (1)$$

$\mathbf{S}(g_j)$ 即为用于识别气体 g_j 的原始阵列数据. 在实际应用中由于原始数据的数据量较大, 通常提取响应曲线的特征参数, 作为气体识别量化算法的输入. 若从传感器响应中提取 c 个特征参数, 则获得的特征空间可表示为

$$\mathbf{X}(g_j) = \begin{pmatrix} x_1^1(g_j) & x_2^1(g_j) & \cdots & x_M^1(g_j) \\ x_1^2(g_j) & x_2^2(g_j) & \cdots & x_M^2(g_j) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^c(g_j) & x_2^c(g_j) & \cdots & x_M^c(g_j) \end{pmatrix}, \quad (2)$$

其中 $x_i^a(g_j)$ 是第 i 个传感器对气体 g_j 响应的第 a 个特征参数, 也可以将特征空间表示为 $[x_i^a(g_j)]_{1 \times Mc}$, 后者在此后的数据处理和模式识别中更方便. 假设通过气体/气味识别算法 f 对传感器阵列 $\mathbf{A}$ 获得的特征空间 $\mathbf{X}$ 进行映射, 预测结果和真实类别分别用 $f(\mathbf{X}|\mathbf{A})$ 和 Y 表示, 则初始传感器阵列 $\mathbf{A}$ 的识别率 $p(\mathbf{A})$ 为

$$p(\mathbf{A}) = \frac{\sum I(f(\mathbf{X}|\mathbf{A}) = Y)}{N}, \quad (3)$$

其中 N 为数据集中样本的数量, I 是示性函数, 当预测结果与真实类别相同时示性函数结果为 1, 不同时结果为 0.

实际应用中, 为了满足传感器阵列微型化的需要, 通常需要选择使用最少的传感器组成阵列来完成指定的测试任务. 因此, 气体传感器阵列优化问题, 简化为 SAO, 就是基于初始阵列 $\mathbf{A}$ 如何选择传感器以形成优化阵列 $\mathbf{A}^*$ 的问题.

3 传感器阵列优化算法 SAO_DFI

本文提出的动态特征重要度包含特征相关性和特征重要性两部分, 因此在定义 DFI 之前, 先对 FFC 和 FI 进行定义, 然后提出 DFI 的概念, 最后给出 SAO_DFI 算法的具体流程.

3.1 特征相关性

定义1 (特征相关性) 对于特征空间 $\mathbf{X}$ 中任意两个特征 x_i 和 x_j ($i \neq j$), 通过特征相关性描述两特征之间的关系, 定义特征相关性系数 ρ_{ij} 表示两特征之间的相关性的强度.

特征相关性系数 ρ_{ij} 可由皮尔逊相关系数、互信息和最大信息系数等方法获得, 不同的 ρ_{ij} 的基本原理、计算方式和适用范围各异, 但可对所有的 ρ_{ij} 进行归一化处理, 使其取值范围为 $[0, 1]$. 一般 ρ_{ij} 越大, 特征 x_i 和 x_j 的相关性越强, 当 ρ_{ij} 为 1 时, 特征 x_i 和 x_j 完全相关; 当 ρ_{ij} 为 0 时, 特征 x_i 和 x_j 相互独立.

3.1.1 皮尔逊相关系数

皮尔逊相关系数是最早由统计学家卡尔·皮尔逊 (Karl Pearson) 设计的统计指标, 是研究连续变量之间线性相关程度的量. 其定义如下:

$$r(E, F) = \frac{\text{Cov}(E, F)}{\sqrt{\text{var}(E)\text{var}(F)}}, \quad (4)$$

其中 $\text{Cov}(E, F)$ 表示变量 E 与 F 之间的协方差, 而分母中的 $\text{var}(E)$ 和 $\text{var}(F)$ 分别表示变量 E 与 F 的方差. 在自然科学领域, PCC 是最常用的相关性系数, 其值介于 -1 到 1 之间, 当其等于 1 (或 -1) 时表示正 (负) 相关; 当等于 0 时则表示两变量之间不相关. 本文在对 PCC 归一化时取其绝对值.

3.1.2 互信息

互信息 (mutual information, MI) ^[22] 起源于信息论, 用于衡量一个随机变量 E 对另一个随机变量 F 的信息量, 可表示为

$$\text{MI}(E, F) = H(E) + H(F) - H(E, F), \quad (5)$$

其中 $H(E, F)$ 是 E 和 F 的交叉熵, $H(E)$ 和 $H(F)$ 分别表示随机变量 E 和 F 的信息熵.

$$H(E) = - \sum_e P(e) \log_2 P(e), \quad (6)$$

$$H(F) = - \sum_f P(f) \log_2 P(f), \quad (7)$$

$$H(E, F) = - \sum_{e,f} P(e, f) \log_2 P(e, f), \quad (8)$$

其中 $P(e)$ 和 $P(f)$ 分别表示随机变量 E 和 F 的边缘分布, $P(e, f)$ 则是变量 E 和 F 的联合分布. 互信息可用于表征两个随机变量的统计相关性, 多用于离散变量, 当变量是连续型时需要计算或估计其概率分布.

3.1.3 最大信息系数

Reshef 等 ^[26] 于 2011 年在互信息的基础上提出了最大信息系数 (maximal information coefficient, MIC), 又一次引起了学者们对于特征相关性的关注. MIC 不仅可以对变量间的线性与非线性关系进行测度, 还能发掘变量之间的非函数依赖. 值得一提的是 MIC 的一般性和公平性, 一般性指其不限制变量间的函数类型, 能识别多种关系类型; 公平性指其对相同噪声水平的不同关系给出相同的得分.

MIC 基本原理是: 利用散点图表示随机变量 E 与 F 组成的二维空间, 通过对变量 E 与 F 的划分, 将散点图划分为多个区域, 统计每个区域内的变量 E 和 F 的分布情况, 以此作为联合概率的估计, 解决了在互信息中的联合概率难以计算的问题.

MIC 的计算过程如下: 首先确定划分数 a 与 b , 将随机变量 E 与 F 的散点图进行 a 行 b 列的网格化, 并求出最大的互信息值, 记作 $I^*(E, F, a, b)$. 值得注意的是相同划分数可以得出不同的网格化方案, 因为不一定是均匀划分.

$$I^*(E, F, a, b) = \max \text{MI}(E, F)_{a,b}. \quad (9)$$

然后, 将其得到的最大的互信息值标准化, 并作为划分数为 a 与 b 时的最大信息系数.

$$\text{MIC}(E, F)_{a,b} = \frac{I^*(E, F, a, b)}{\log_2 \min(a, b)} \quad \text{s.t. } ab < B(n), \quad (10)$$

其中 n 是数据量, $B(n)$ 是一个关于 n 的函数, 文献 [26] 建议将 $B(n)$ 设置为 $n^{0.6}$. 之后计算不同划分数下的最大信息系数, 取其最大值作为变量 E 和 F 的最大信息系数.

$$\text{MIC}(E, F) = \max_{ab < B(n)} [\text{MIC}(E, F)_{a,b}]. \quad (11)$$

从本质上来讲, MIC 是一种归一化的最大互信息值, 其取值区间为 $[0, 1]$.

3.1.4 其他特征相关性度量方法

上述 3 种方法为常用的度量 FFC 的方法, 除此之外, 还有一些其他的 FFC, 如 spearman 相关系数, 它是根据变量的秩次大小来衡量定序变量之间相关性的, 对变量的原始分布不作要求; kendal 相关系数, 是通过同序对与逆序对的数量判断分类变量的相关性. 这些方法也可用作本文中 ρ_{ij} 的度量.

3.2 特征重要度

定义2 (特征重要度) 对于特征空间 $\mathbf{X}$ 中任意一个特征 x 与类别 Y , 通过特征重要度描述特征 x 对分类的贡献, 定义 $\text{FI}(x)$ 表示特征 x 的重要程度.

与 ρ_{ij} 类似的是, 不同分类算法得到的 $\text{FI}(x)$ 的特点不同, 但 $\text{FI}(x)$ 可不作归一化处理, 由各特征得到的 $\text{FI}(x)$ 的数值相对大小可知其重要程度.

3.2.1 基于学习模型的特征重要度

机器学习 (machine learning, ML) 是人工智能的核心. 它是一种通过数据或经验对计算机算法进行优化, 使机器模拟人类学习活动的研究. 自 20 世纪 80 年代, 机器学习进入了全新时期, 如今机器学习算法在各行各业都有着广泛的应用.

基于学习模型的特征重要度是指根据特定的机器学习算法, 为每个特征和响应变量建立预测模型, 预测结果可以作为该特征的重要度, 其定义如下:

定义3 (基于学习模型的特征重要度) 对于特征空间 $\mathbf{X}$ 中任意一个特征 x 与类别 Y , 通过某种学习算法建立特征 x 到类别 Y 的学习模型 f , 并将模型的预测精度定义为特征 x 的基于模型 f 的特征重要度:

$$\text{FI}(x)_{\text{ML}} = \frac{\sum I(f(x) = Y)}{N}, \quad (12)$$

其中 N 为样本数量, I 是示性函数, 用于判断预测结果 $f(x)$ 与真实类别 Y 是否相同, 相同时为 1, 不同时为 0.

与其他相关性不同的是, 基于学习模型的重要度更加简单、直接地表达在特定的学习算法中特征与类别之间存在的关系, 同时对数据的分布和类型没有过多的要求; 当然该重要度一般情况下仅适用于特定的学习算法. 常用的机器学习算法有支持向量机 (support vector machine, SVM) [27]、误差逆传播神经网络 (back propagation neural network, BP) [28]、随机森林 (random forest, RF) [29] 和 K 近邻 (K-nearest neighbor, KNN) [30] 等, 通过这些算法可分别计算特征重要度. 本文选取了这些方法进行测试, 其中基于 SVM 的特征重要度记为 SVMI, 基于 BPNN 的特征重要度记为 BPI, 基于 KNN 的特征重要度记为 KNNI, 基于 RF 的特征重要度记为 RFI.

3.2.2 基于随机森林的特征重要度

基于学习模型的特征重要度是一种直接的衡量方法,而不同的机器学习算法往往有其独特的实现方法,有些算法在训练或识别的过程中可以完成隐式的特征选择,如 SVM^[31,32], RF 等.

以 RF 为例, RF 是以决策树为基分类器的集成分类器,根据其算法特点,也可以用于特征选择.以分类回归树 (classification and regression tree, CART)^[33] 为例, CART 决策树使用基尼指数 (Gini index) 分割节点,假设某节点凭借特征 x 分裂,其基尼指数可表示为

$$\text{Gini}(x) = \sum_y^{|Y|} p_y(1 - p_y), \quad (13)$$

其中 $|Y|$ 表示所有的类别, p_y 表示其中类别 y 的概率.在决策树生成过程中,父节点每次分裂产生的两个子节点的基尼指数之和小于该父节点的基尼指数,并将该父节点基尼指数与其两子节点基尼指数之和的差称为基尼下降量 (Gini decrease),记为 $\text{GD}(x)$:

$$\text{GD}(x) = \text{Gini}(x) - \text{Gini}_l(x) - \text{Gini}_r(x), \quad (14)$$

其中 $\text{Gini}_l(x)$ 和 $\text{Gini}_r(x)$ 分别表示左右子节点的基尼系数,而特征 x 的基尼重要度 (Gini importance, GI)^[29,34,35] 就是指特征 x 的所有特征值置换前后 $\text{GD}(x)$ 差的平均值.

$$\text{GI}(x) = \frac{\sum_{tr=1}^{\text{Ntree}} \text{GD}_{tr}(x) - \text{GD}_{tr}(x)'}{\text{Ntree}}, \quad (15)$$

其中 Ntree 是随机森林中决策树的数量, $\text{GD}_{tr}(x)'$ 是某树中置换后的特征 x 的基尼下降量.

RF 中还可与基尼重要度类似的置换重要度 (permutation importance)^[29,36] 来衡量特征的重要度,区别在于置换重要度计算的是置换后袋外数据准确率之差.

3.2.3 其他特征重要度

常用的特征重要度除了基于学习模型的重要度、基尼重要度和置换重要度以外,还有一些回归模型的权系数也可以作为特征重要度,如回归、正则化方法的权值系数.此外,3.1 小节中提到的 PCC, MI 和 MIC 等也可以用于表征特征与类别的关系,特征与类别的相关性越强,表示特征重要度越高.

3.3 动态特征重要度

在实际的特征选择 (或 SAO) 问题中,特征 (或传感器) 的重要性和相关性需要同时考虑,因此本文提出了动态特征重要度的概念.

定义4 (动态特征重要度) 特征选择中,特征 x 的动态特征重要度定义为

$$\text{DFI}(x)_i = \begin{cases} \text{FI}(x), & i = 1, \\ \prod_{k=1}^{i-1} (1 - \rho_{xk}) \times \text{FI}(x), & i > 1, \end{cases} \quad (16)$$

其中 i 是当前选择的轮数; ρ_{xk} 则是该特征与第 k 轮选择的特征之间的特征相关性; $\text{FI}(x)$ 是特征 x 的特征重要度.

动态特征重要度的核心思想是通过待选特征与已选特征的相关性来修正待选特征在当前特征子集中对分类的贡献,从而得到既考虑特征有效性又考虑特征冗余性的特征排序.在选择过程中,当 $i = 1$

Algorithm 1 The sensor array optimization algorithm based on Dynamic Feature Importance SAO_DFI**Input:** Raw data of initial array $\mathbf{A}$, variance threshold v_{th} , type of FI, FFC and sensor feature parameters;**Process:**

```

1: Phase 1:
2: for  $s_i$  in  $\mathbf{A}$  do
3:    $vs \leftarrow 0$ ;
4:   for  $g_j$  in  $G$  do
5:     Calculate the variance of the response curve for  $s_i$  to  $g_j$  and record it as  $v_{ij}$ ;
6:      $vs \leftarrow vs + v_{ij}$ ;
7:   end for
8:   if  $vs > v_{th}$  then
9:     Add the sensor  $s_i$  to the array  $\mathbf{A}'$ ;
10:  end if
11: end for
12: Select the data corresponding to the sensors in array  $\mathbf{A}'$  from the raw data and extract the feature parameters to form
    a new dataset  $D_{n \times (m+1)}$ ;
13: Phase 2:
14: Initialize  $A_{sub} \leftarrow \emptyset$ ;
15: Calculate the accuracy of the current array by training on the dataset  $D$ , denoted as  $Acc_{th}$ ;
16: Calculate the FI of each feature;
17: Calculate the matrix of FFC;
18: Select the feature with the largest FI and add it to  $A_{sub}$ ;
19: for  $i$  in  $1 : m - 1$  do
20:   Calculate the accuracy of the array  $A_{sub}$  by training on the subset  $D[A_{sub}]$ , denoted as  $Acc_{sub}$ ;
21:   if  $Acc_{sub} \geq Acc_{th}$  then
22:     break;
23:   else
24:     Update DFI of candidate features according to 16;
25:     Select the feature with the largest DFI and add it to  $A_{sub}$ ;
26:   end if
27: end for
28: Denote the optimized array as  $\mathbf{A}^*$  that consists of the corresponding sensors in the  $A_{sub}$ , and delete the repetitive
    sensors in  $\mathbf{A}^*$ ;
Output: the optimized array  $\mathbf{A}^*$ .

```

时, 特征子集为空, 此时特征 x 的 DFI 等于其 FI; 当 $i > 1$ 时, DFI 会受到前 $i - 1$ 轮选择的特征的影响, 若待选择的特征与已选择的特征之间存在较强的相关性, 该待选特征的重要度会下降, 通过这种方式尽可能使每轮选择的特征有效且非冗余, 因此在选择过程中 DFI 是动态计算的。

在 SAO 问题中, 传感器等效于特征, 因此可用 FI 衡量传感器对气体识别的贡献大小; 用 FFC 判断传感器在优化阵列中是否冗余. 通过动态特征重要度 DFI 选择传感器, 促使“功能”不同的有效的传感器组成优化阵列, 保证优化阵列数据的多样性. 至于相关系数和特征重要度的确定, 可根据实际应用和数据类型从前述方法中选择。

3.4 SAO_DFI 算法描述

基于提出的动态特征重要度的概念, 进一步研究设计了相应的传感器阵列优化算法, 具体过程如算法 1 所示。

算法输入包括初始阵列 $\mathbf{A}$ 的原始数据、方差阈值 v_{th} 、传感器特征参数、FI 和 FFC 的类型; 输出的是优化阵列 $\mathbf{A}^*$, 其具体过程可分为: 数据前处理阶段和特征选择阶段.

3.4.1 数据前处理阶段

该阶段主要作用是初步优化传感器阵列, 快速过滤掉阵列中的无效传感器, 并提取传感器响应曲线的特征参数以进行后续处理.

无效传感器的过滤可通过方差法实现, 通过计算传感器所有样本的方差之和表征, 其计算公式为

$$v_i = \sum_{g_j \in G} \frac{\sum_{t \in T} [s_i(t|g_j) - \bar{s}_i(g_j)]^2}{T-1}, \quad (17)$$

其中 v_i 是第 i 个传感器的方差; $s_i(t|g_j)$ 和 $\bar{s}_i(g_j)$ 分别是第 i 个传感器关于气体 g_j 的响应曲线在 t 时刻值和均值; T 是总采样时间, 当采样频率为 1 Hz 时, T 也是采样总数. 该方法不同于简单的计算特征方差, 而是先计算每个传感器对每种气体响应曲线的方差之和, 再将低于阈值 v_{th} 的传感器剔除, 完成阵列的初步优化. 方差可以描述传感器响应曲线的偏离程度, 而方差过小意味着在传感器对于目标环境与标准空气环境的响应程度接近, 无法对目标环境中的气体做出有效的反应, 这样的传感器明显不是优化阵列所需要的.

在方差选择法初步优化后, 提取响应曲线的特征参数, 结合类别组成数据集. 传感器的特征参数在阵列优化的一般问题中曾提到, 可以选择部分信号特征或者从整体曲线中提取关键信息. 本文后面使用的稳态响应值是气体传感器信号处理中最常用的特征参数之一, 并在提取稳态响应值之后, 对其进行差分处理, 其具体处理方式如下:

$$x_i = \max s_i - s_i(t_0), \quad (18)$$

其中 s_i 和 x_i 分别是第 i 个传感器关于某个气体样本的信号和特征参数, 而 $s_i(t_0)$ 是 s_i 的基线值, 以此抑制基线漂移对特征参数的影响.

通过数据前处理阶段后, 当前阵列中保留的传感器应该是在目标环境中响应的, 或者是对分类是有效的, 即阶段一实质是传感器有效性分析. 除此之外, 若将该算法拓展至其他领域, 可考虑数据特点和要求等因素, 使用其他过滤方法筛选特征.

3.4.2 特征选择阶段

特征选择阶段是 SAO_DFI 算法的核心, 通过算法输入指定的 FI 和 FFC 的类型, 在每一轮选择中计算特征的 DFI, 将当前 DFI 最大的特征添加到特征子集中, 直到特征子集可以满足准则函数, 否则重复这一过程; 最终将特征子集中的每个特征所对应的传感器选入优化阵列, 并删除阵列中重复传感器. 若前一阶段传感器提取了多个特征参数, 可能导致某传感器的多个特征参数同属于特征子集, 因此重复传感器应剔除.

不同于一般的特征选择算法, 阶段二中特征的有效性分析和冗余性分析是通过 DFI 的计算同时进行的, FI 和 FFC 分别判断特征的有效性和冗余性. 在每一轮选择时, 待选特征对当前已选特征子集的贡献是动态的, 不是一成不变的; 在选择后, 需判断已选特征子集是否满足准则函数. 本文所需要的优化阵列的规模较小, 故使用所选特征子集的预测精度作为准则函数, 其计算方法与式 (12) 相似, 只需将单一特征改为特征子集即可. 当该特征子集的预测精度大于或等于精度阈值时, 停止选择. 这种方

图 2 (网络版彩图) 测试系统
Figure 2 (Color online) Test system diagram

法的优点在于结果直观, 而且当特征子集满足条件时, 模型的训练也在同一时间完成; 缺点是时间成本可能较大, 这取决于分类算法的时间复杂度.

4 实验测试

4.1 测试平台

本次实验所用数据由一套气体传感器及电子鼻系统测试平台获得. 该平台包括高精度自动配气系统和气体传感器及电子鼻测试系统两部分, 如图 2 所示. 其中自动配气系统包括 4 路高精度质量控制器 (mass flow controller, MFC) (Alicat 公司), 可实现同时配置 3 种不同浓度的气体及其混合气体. 测试所用气源由烟台飞鸢特种气体公司提供, 并经过标准测试仪器标定. 气体传感器及电子鼻系统由可处理 16 路传感器的接口电路、数据采集卡 (NI6341) 以及控制电脑组成. 测试平台软件基于 LabView 环境开发, 实现测试环境所需的气体的自动配气和传感器动态控制及测试. 通过采集到的数据可以进一步获得气体传感器响应曲线以及最大响应值、响应与恢复时间等特征参数, 为后续的传感器阵列优化工作提供了必要的参考数据.

4.2 电子鼻数据采集

为了验证算法的有效性, 本文针对两种实际环境的气体测试需求, 依据环境被测气体成分和已有的商用传感器特性, 分别选取了不同数量的气体传感器组成初始阵列.

目标环境 I 针对空气中的 CO₂ 和 SO₂ 及其混合气体进行测试, 使用了 12 个传感器, 分别是 Figaro 公司的 TGS2602, TGS2600, TGS2610, TGS2603 和 TGS2620 气体传感器, 各 2 个; TGS2611 及霍尼韦尔公司的 4SO2-2000 电化学传感器各 1 个. 各传感器的特性说明见表 1.

根据实际需求和数据分析的需要, 分别设定不同浓度的待测气体, 每种气体测试 4 或 8 个浓度点, 其中 CO₂ 气体测试 4 个浓度点, 如测试 1%, 4%, 6% 和 8% 的 CO₂ 气体; 而 SO₂ 气体和混合气体均测试 8 个浓度点, 为了确保特征的重复性和稳定性, 在相同的实验条件下每组测试重复 5 次, 具体设置如表 2 所示, 表中 ppm 为每体积百万分之一 (parts per million).

在上述实验环境与测试气体的设置下, 获取目标环境 I 中初始传感器阵列对不同浓度气体的稳态

表 1 气体传感器型号及其参数

Table 1 Type and parameter of gas sensors

Type	Sensor	Main response gases	Range	Accuracy/Sensitivity
Metal oxide	TGS2600	H ₂ S, C ₂ H ₅ OH, CO, et al	1~30 ppm	0.3~0.6 ppm
Metal oxide	TGS2602	VOC, NH ₃ , H ₂ S, et al	1~30 ppm	0.08~0.5 ppm
Metal oxide	TGS2603	Trimethylamine, Methyl mercaptan, H ₂ S, et al	1~10 ppm	< 0.5 ppm
Metal oxide	TGS2610	C ₃ H ₈ , C ₄ H ₁₀ , et al	1%~25%	0.56±0.06 ppm
Metal oxide	TGS2611	CH ₄ , Natural gas, et al	1%~25%	0.6±0.06 ppm
Metal oxide	TGS2620	C ₂ H ₅ OH, Organic solvent, et al	50~500 ppm	0.3~0.5 ppm
Electrochemical	4SO2-2000	SO ₂	0~2000 ppm	0.02 ± 0.08 μA/ppm

表 2 目标环境 I 测试气体成分及浓度设置

Table 2 Test gas composition and concentration settings in environment I

Gas	Target gas configuration		Time control		Batch
	CO ₂ (%)	SO ₂ (ppm)	Inject synthetic air (min)	Inject target gases (min)	
CO ₂	1, 4, 6, 8	—	10	6	5
SO ₂ low	—	10, 30, 50, 80	10	6	5
SO ₂ high	—	100~400 interval 100	10	6	5
Mixture	1, 4, 6, 8	30, 50	10	6	5

表 3 目标环境 II 测试气体成分及浓度设置

Table 3 Test gas composition and concentration settings in environment II

Gas	Target gas configuration		Time control		Batch
	CH ₄ (ppm)	CO (ppm)	Inject synthetic air (min)	Inject target gases (min)	
CH ₄	500~2000 interval 500	—	20	8	5
CO	—	50~200 interval 50	20	8	5
Mixture	500~2000 interval 500	50~150 interval 50	20	8	5

响应, 典型曲线如图 3 所示. 可以看出每个传感器与被测气体之间都会产生与时间相关的响应曲线, 图 3 的每个子图中包含 4 种不同浓度的同类气体的稳态响应曲线. 通常, 响应电压随气体浓度增加而增加, 但每个传感器对不同气体的响应都有其自身的差异. 如 4SO₂-2000 对 SO₂ 的浓度变化非常敏感, 但对 CO₂ 几乎没有反应; TGS2610 和 TGS2611 的响应曲线的变化趋势相似; TGS2602 和 TGS2620 的响应值与基线值相比几乎没有差异.

目标环境 II 针对空气中的 CO, CH₄ 及其混合气体进行测试, 使用了 6 种传感器, 分别是 Figaro 公司的 TGS2602, TGS2600, TGS2610, TGS2603, TGS2620 和 TGS2611 传感器. 气体的配置设置见表 3, 一些典型的稳态响应曲线如图 4 所示. 从图 4 看出, 相同的实验环境下 TGS2610, TGS2611 和 TGS2620 传感器的响应趋势相似, 幅值也相近, 对 CH₄ 有较大响应, 且各浓度之间的差异较小; TGS2600 情况相反, 对 CO 的灵敏度较高.

图 3 (网络版彩图) 目标环境 I 传感器阵列的稳态响应曲线

Figure 3 (Color online) Steady responses of sensor array in environment I. (a) 1%, 4%, 6% and 8% CO_2 ; (b) 10, 30, 50, 80 ppm SO_2 ; (c) 100~400 ppm SO_2 ; (d) 30 ppm SO_2 + 1%, 4%, 6% and 8% CO_2 ; (e) 50 ppm SO_2 + 1%, 4%, 6% and 8% CO_2

图 4 (网络版彩图) 目标环境 II 传感器阵列的稳态响应曲线

Figure 4 (Color online) Steady responses of sensor array in environment II. (a) 500~2000 ppm CH_4 ; (b) 50~200 ppm CO ; (c) 50 ppm CO + 500~2000 ppm CH_4 ; (d) 100 ppm CO + 500~2000 ppm CH_4 ; (e) 150 ppm CO + 500~2000 ppm CH_4

5 算法分析及验证

根据上述两种实验环境的数据, 基于 SAO_DFI 算法, 分别讨论了阵列中传感器冗余性、各传感器重要性和相关性、各特征参数对传感器阵列优化的影响, 最终获得优化阵列并进行验证。

5.1 传感器冗余性分析

在构造气体传感器阵列时, 为了避免因为传感器损坏或电路故障等问题影响优化结果, 通常会在

图 5 (网络版彩图) 目标环境 I 稳态响应值分布情况

Figure 5 (Color online) Distribution of steady state response values in environment I. Feature value of (a) TGS2600; (b) TGS2600.L; (c) TGS2610; (d) TGS2610.L; (e) TGS2611; (f) TGS2603; (g) TGS2603.L; (h) 4SO2-2000

初始阵列中加入重复的传感器. 然而在后面的处理中若不删除, 可能会面临优化阵列中包含重复传感器的问题. 本部分主要讨论重复传感器对 SAO_DFI 算法的影响.

根据 SAO_DFI 算法, 在阶段一剔除了 TGS2602 和 TGS2620, 从图 3 中也可以看出: 与其他传感器相比, 这两种传感器在目标环境中几乎没有响应. 接下来提取稳态响应值作为传感器的特征参数. 图 5 是目标环境 I 下提取的稳态响应值, 可以看到不同传感器响应的特点. 电化学传感器 4SO2-2000 的稳定性好于其他金属氧化物传感器, 而且选择性强几乎只对 SO_2 反应, 因此仅靠 4SO2-2000 无法分辨混合气体与 SO_2 ; 其他 4 种传感器对环境 I 中的气体均有响应, 同型号传感器的数据分布相似度高, 但彼此之间仍有明显的差异, 这也是使用重复传感器以及本部分分析的原因.

进一步分析各传感器之间的相关性 FFC, 各传感器的稳态响应值之间的皮尔逊相关系数 PCC 如图 6 所示, 通过图 5 和 6 可以发现 TGS2610 和 TGS2611 的稳态响应值之间呈现出高度线性相关. 各传感器的重要性 FI 通过基尼重要性 GI 进行衡量, 图 7 中可以看到初始阵列中的每个传感器的 GI. 通过 SAO_DFI 算法, 以 RF 为模式识别算法, 以 PCC 为相关系数, 以 GI 为 FI 对该环境下传感器阵列进行优化, 得到优化阵列为 TGS2610.L, TGS2603.L 和 4SO2-2000, 识别率为 94.8% (本文中的识别率都是相同条件下 50 次试验的平均值), 初始阵列的识别率为 91.1%, 若仅考虑 GI, 按照 GI 从大到小排序的话, 前三分别是 4SO2-2000, TGS2610.L 和 TGS2600.L, 而该阵列的识别率为 90.3%, 明显低于算法选择的优化阵列.

根据式 (16), 在 FI 值接近的情况下 SAO_DFI 算法几乎不会选到同型的传感器, 而且重要度相近且相关性强的相似传感器也很难被同时选到, 因此在进行阵列优化时不需要手动从重复的传感器中挑

图 6 (网络版彩图) 目标环境 I 中各传感器稳态响应值之间的皮尔逊相关系数

Figure 6 (Color online) Pearson correlation coefficient between steady state response values of sensors in environment I

图 7 (网络版彩图) 目标环境 I 中各传感器稳态响应值的基尼重要度

Figure 7 (Color online) Gini importance of steady state response values of sensors in environment I

选, SAO_DFI 算法几乎不会受到传感器的冗余性的影响, 同时选择重复传感器中性能最好的.

5.2 特征重要度和相关性分析

本部分重点讨论不同的 FFC, FI 和模式识别算法对 SAO_DFI 算法的影响.

为了简化分析, 本部分的优化中删去了重复的传感器, 相关性选择了 PCC 和 MIC (因为气体传感器的数据是连续型且概率分布难以确定, 不建议使用 MI), 模式识别算法使用了 SVM, BP, KNN 和 RF, 特征重要度包括这 4 种机器学习算法的 FI 以及 GI. 计算得到各传感器之间的 PCC 如图 6 所示, MIC 如图 8 所示. 对比图 6 和 8, 传感器之间的 MIC 与 PCC 差别明显, 大多 PCC 的值要高于 MIC, 但 TGS2610, TGS2611 与 4SO2-2000 之间的 PCC 低于 MIC. 各传感器的 FI 如表 4 所示, 不同的算法下, 4SO2-2000 的 FI 总是最高的, 在只使用 4SO2-2000 的情况下, 识别率甚至可以达到 88.6%, 而其他传感器的 FI 排名浮动较大, 如 TGS2611 在前 4 种 FI 中排名第 2, 但在 GI 中却成了 FI 最低传感器. 有趣的是, 通过对比图 7 和表 4 中 GI 的值, 可以发现 TGS2603 的排名发生变化, 这是因为此时的初始阵列中没有了重复传感器和基尼指数的特点导致的.

SAO_DFI 算法在不同条件下的优化阵列比较相似, 表 5 记录了每一轮的选择. 观察表 5 发现, 相关性相同时, 各种情况的选择基本相同, 唯一不同的是使用 MIC 时, SVM 进行了 3 轮选择, 而其他重要度进行到第 2 轮就停止了. 此外, 值得注意的是使用 MIC 时, 算法在第 2 轮选择了 TGS2603; 而使用 PCC 时, 算法却选择了 TGS2610.

分别计算了优化阵列和相同规模下的其他阵列的识别效果, 如图 9 和 10 所示, 其中图 9 是 3 元子阵列的识别率, 图 10 是 2 元子阵列和初始阵列的识别率. 图 9(a) 空白的对角线区域上方为 BP 算法的结果, 下方是 SVM 算法; 图 9(b) 空白的对角线区域上方是 RF 算法, 下方是 KNN 算法且每个阵列都包含 4SO2-2000. 如表 4 所示, 4SO2-2000 传感器的每种重要度都是最高的, 所以 SAO_DFI 算

图 8 目标环境 I 中各传感器稳态响应值之间的最大信息系数

Figure 8 Maximal information coefficient between steady state response values of sensors in environment I

表 4 目标环境 I 中传感器的特征重要度

Table 4 Feature importance of sensors in environment I

Sensor	SVMI	BPI	KNNI	RFI	GI
TGS2600	0.602	0.556	0.684	0.602	0.1618
TGS2610	0.746	0.782	0.654	0.632	0.1559
TGS2611	0.774	0.786	0.722	0.726	0.1017
TGS2603	0.550	0.768	0.620	0.546	0.1501
4SO2-2000	0.872	0.886	0.862	0.826	0.4305

表 5 SAO_DFI 算法在不同设置下的优化过程

Table 5 Optimization process of SAO_DFI algorithm with different settings

FFC	Epoch	SVMI	BPI	KNNI	RFI	GI
PCC	First	4SO2-2000	4SO2-2000	4SO2-2000	4SO2-2000	4SO2-2000
	Second	TGS2610	TGS2610	TGS2610	TGS2610	TGS2610
	Third	TGS2603	TGS2603	TGS2603	TGS2603	TGS2603
MIC	First	4SO2-2000	4SO2-2000	4SO2-2000	4SO2-2000	4SO2-2000
	Second	TGS2603	TGS2603	TGS2603	TGS2603	TGS2603
	Third	TGS2610				

法在第一轮都选择了 4SO2-2000, 而且经过验证不含 4SO2-2000 的阵列效果都非常差, 故没有列出. 图 10 中 4SO2-2000 和 TGS2603 的阵列明显优于 4SO2-2000 和 TGS2610 的阵列, 同时除了 SVM 算法, 4SO2-2000 和 TGS2603 的阵列的效果也优于大多数 3 元阵列. 换言之, 当 SAO_DFI 算法使用 MIC 作为 FFC 优化环境 I 的初始阵列时, 无论 FI 和气体识别算法如何确定, 总能作出最佳的判断, 选出最好的阵列.

图 9 (网络版彩图) 不同算法下 3 元阵列识别率 (含 4SO2-2000). (a) 中空白的对角线区域上方为 BP 算法的结果, 下方是 SVM 算法; (b) 中空白的对角线区域上方是 RF 算法, 下方是 KNN 算法

Figure 9 (Color online) 3D array recognition rate with different algorithms (including 4SO2-2000). (a) Above the blank diagonal area is the result of BP algorithm, under is the result of SVM algorithm; (b) above the blank diagonal area is the result of RF algorithm, and under is the result of KNN algorithm

图 10 (网络版彩图) 不同情况下 2 元阵列识别率 (含 4SO2-2000)

Figure 10 (Color online) 2D array recognition rate with different conditions (including 4SO2-2000)

图 11 (网络版彩图) 目标环境 I 中各传感器稳态响应值和一阶导数值的基尼重要度

Figure 11 (Color online) Gini importance of steady state response values and derivative values of sensors in environment I

相关性上, MIC 的效果好于 PCC; 算法上, BP, SVM 和 KNN 算法的最大识别率都达到 98% 以上, 但 BP 和 KNN 的规模更小. 阵列优化的过程会受到模式识别算法和 FFC 的影响, 尤其是 FFC, 所以根据实验数据的特点选择合适的相关性系数显得尤为重要; 同时在传感器对分类都有贡献的情况下, FI 对优化过程的影响相对较小. 此外, 对于某些三元阵列的识别率大于二元阵列的情况, 可将算法的停止准则设置为 100%, 针对不同的应用需求, 可以选择在最小规模和最大识别率之间权衡.

图 12 (网络版彩图) 目标环境 I 中各传感器稳两种特征参数之间的最大信息系数

Figure 12 (Color online) Maximal information coefficient between two characteristic parameters of sensors in environment I

5.3 特征参数重要性实验

在之前的实验分析中, 响应曲线的特征参数都是稳态响应值, 实际上从传感器响应曲线上不但可以取其他特征参数还可以同时使用多个特征参数, 本部分主要讨论响应曲线的不同特征参数对 SAO_DFI 算法的影响.

与稳态响应值相对应, 本部分选择一个动态参数, 即一阶导数值作为第二个特征参数, 在计算曲线一阶导数之前对响应曲线做平稳滤波. SAO_DFI 算法选择 MIC 为 FFC, GI 为 FI, RF 算法作为模式识别算法, RF 算法的优势在于可以用袋外分数估计识别率. 特征之间的 GI 和 MIC 分别如图 11 和 12 所示, 其中前缀 “Diff.” 表示一阶导数值. 在图 11 中, 4SO2-2000 的一阶导数值的重要度位于第 2, 仅次于它自身的稳态响应值, 但其他传感器的一阶导数值却表现不佳, 整体上一阶导数值的重要度与稳态响应值相比有较大差距.

图 12 中右上 (左下) 部分是稳态响应值与一阶导数值之间的 MIC, 我们可以观察到同一传感器的特征参数之间的相关性往往大于其他传感器, 尤其是 4SO2-2000 的两特征参数完全相关, 因此即使图 11 中 4SO2-2000 传感器的一阶导数值的重要性排在第二, 在 SAO_DFI 算法中其重要性为 0. 此外, 经本文验证, 这种现象在 PCC 矩阵中会更加明显, 同型传感器之间的 PCC 都在 0.9 以上, 呈现出高度线性相关.

当同时使用稳态响应值和一阶导数值这两个特征参数时, 其结果与使用稳态响应值时无异, 算法挑选的优化阵列依旧是由 4SO2-2000 和 TGS2603 组成; 当仅使用一阶导数值时, 算法同样选择了

图 13 (网络版彩图) 目标环境 II 稳态响应值分布情况

Figure 13 (Color online) Distribution of steady state response values in environment II. Feature value of (a) TGS2602; (b) TGS2600; (c) TGS2610; (d) TGS2611; (e) TGS2603; (f) TGS2620

4SO₂-2000 和 TGS2603 这两个传感器, 但其识别率仅为 86.3%. 在优化阵列确定后, 在验证过程中使用每个传感器的两个特征参数进行训练, 发现其结果没有明显差异.

综上, SAO_DFI 算法在优化传感器阵列时, 不需要提取多个特征参数, 尤其是彼此相关性较强的特征参数, 另外建议使用稳态响应值, 采集方便且比较稳定.

5.4 算法验证

本部分主要测试 SAO_DFI 算法在目标环境 II 的数据上的效果, 并验证前 3 个实验得出的关于 SAO_DFI 算法的结论.

环境 II 的目标主要是检测并识别 CO, CH₄ 及其混合气体种类及其浓度, 详情可见表 3 或图 4, 由于 TGS2610, TGS2611 和 TGS2620 3 种传感器对不同浓度 CH₄ 之间的响应差异较小, 这对气体种类的识别有利. 为了进一步评价各传感器在环境 II 中的气体浓度识别和量化性能, 数据的类别按照不同气体不同浓度可分为 20 类. SAO_DFI 算法的参数仍设置为稳态响应值, MIC, GI 和 RF 算法. 目标环境 II 的稳态响应值分布、特征之间的 GI 和 MIC 分别如图 13~15 所示. 观察图 13 和 15, 从分布上 TGS2600, TGS2610, TGS2611 和 TGS2620 的稳态响应值具有较强的相关性, 其中 TGS2600 与另 3 个是负相关的关系. 值得一提的是 TGS2610 和 TGS2611 在环境 II 中完全相关, 而两传感器在环境

图 14 (网络版彩图) 目标环境 II 中各传感器稳态响应值的基尼重要性

Figure 14 (Color online) Gini importance of steady state response values of sensors in environment II

图 15 (网络版彩图) 目标环境 II 中各传感器稳态响应值之间的最大信息系数

Figure 15 (Color online) Maximal information coefficient between steady state response values of sensors in environment II

(a)

(b)

图 16 (网络版彩图) 不同算法下 2 元阵列识别率. (a) 中空白对角线区域上方为 BP 算法的结果, 下方是 SVM 算法; (b) 中空白对角线区域上方是 RF 算法, 下方是 KNN 算法

Figure 16 (Color online) 2D array recognition rate by different algorithms. (a) Above the blank diagonal area is the result of the BP algorithm, under is the result of SVM algorithm; (b) above the blank diagonal area is the result of RF algorithm, and under is the result of KNN algorithm

I 中也表现出了极高的相关性.

环境 II 初始阵列的识别率高达 99.3%, 通过 SAO_DFI 算法的优化, 将初始阵列中的 6 种传感器优化为 TGS2600 和 TGS2603 这两种, 而此优化阵列的识别率达到了令人满意的 100%, 略高于初始阵

列. 现计算所有 2 元传感器阵列在不同算法下的识别率, 并通过图 16 对比优化阵列与其余 2 元传感器阵列的性能. 图 16(a) 中空白的对角线区域上方为 BP 算法的结果, 下方是 SVM 算法; 图 16(b) 中空白的对角线区域上方是 RF 算法, 下方是 KNN 算法. 从图 16 中可见, RF 和 KNN 算法整体表现好于 SVM 和 BP 算法, 且图中凡是包含 TGS2603 的阵列的识别率都很高, 这种现象在图 16(a) 中尤为显著. 同时不难发现, 算法挑选的优化阵列并不能在所有算法上成为最优阵列, 当使用 BP 算法识别气体时, 由 TGS2603 和 TGS2600 组成的优化阵列的识别率比由 TGS2603 和 TGS2620 组成的阵列低 3%, 但整体表现良好, 除了 BP 算法以外, 识别率都可以达到 100%.

此外, 通过观察图 13~16, 本文发现一个有趣的现象: 当重要性最高的 TGS2603 与 TGS2600, TGS2610, TGS2611 和 TGS2620 这 4 个彼此之间强相关的传感器组成阵列时, 这些阵列的识别率很接近; 但当这 4 个传感器之间组成传感器阵列时, 阵列的识别率会变得非常低. 相关性较强的 TGS2602 和 TGS2603 组成的阵列的性能也较差, 这个现象进一步说明了阵列优化受 FFC 的影响较大, 而通过 SAO_DFI 算法可去掉相关性强的传感器, 提高阵列的识别率.

6 结论

本文论述了电子鼻系统中传感器阵列优化的一般问题, 并依据一般意义上的特征重要度和特征相关性的分析方法, 提出了一种新的特征 (传感器) 重要性的衡量方法——动态特征重要度, 并将其与电子鼻技术结合, 提出了一种基于动态特征重要度的有效的电子鼻气体传感器阵列优化算法 SAO_DFI. 通过对两种目标环境下的传感器阵列测试数据分析, 设计了具有针对性的实验, 证明了该算法的有效性.

结论如下:

- (1) 本文提出的 SAO_DFI 算法不受重复传感器的影响, 鲁棒性好;
- (2) 该算法会受到模式识别算法和特征相关性的影响, 尤其是特征相关性; 而在传感器对分类都有贡献的情况下, 特征重要度对优化过程的影响相对较小;
- (3) 该算法不需要提取大量的特征参数, 尤其是特征参数之间相关性较强的情况.

本文提出的 SAO_DFI 算法解决了电子鼻系统中依据经验设计传感器阵列或简单相关性优化阵列的低效问题, 算法选择有效且具有不同“功能”的传感器组成阵列, 模型的可解释性强; 数据前处理方法简单, 不需要提取大量的特征参数增加数据处理的难度; 算法优化的阵列对模式识别算法具有较好的鲁棒性.

SAO_DFI 算法同样存在缺陷, 在处理高维数据时, 需要先筛选掉无效特征, 否则当无效特征与其他有效特征相关性弱时, 可能会导致无效特征的 DFI 值过大, 从而选择无效特征.

下一步的工作将进一步研究每个传感器在优化阵列中的具体功能并完善基于 DFI 的特征选择算法, 使其能够应用在更多的领域之中.

参考文献

- 1 Gardner J W, Bartlett P N. A brief history of electronic noses. *Senss Actuators B-Chem*, 1994, 18: 210-211
- 2 Persaud K, Dodd G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature*, 1982, 299: 352-355
- 3 Evans G P, Buckley D J, Adedigba A L, et al. Controlling the cross-sensitivity of carbon nanotube-based gas sensors to water using zeolites. *ACS Appl Mater Interfaces*, 2016, 8: 28096-28104

- 4 Huang X, Meng F, Pi Z, et al. Gas sensing behavior of a single tin dioxide sensor under dynamic temperature modulation. *Senss Actuators B-Chem*, 2004, 99: 444–450
- 5 Huang X, Liu J, Shao D, et al. Rectangular mode of operation for detecting pesticide residue by using a single SnO₂-based gas sensor. *Senss Actuators B-Chem*, 2003, 96: 630–635
- 6 Kabir K M M, Sabri Y M, Matthews G I, et al. Cross sensitivity effects of volatile organic compounds on a SAW-based elemental mercury vapor sensor. *Senss Actuators B-Chem*, 2015, 212: 235–241
- 7 Zhu H, Li Q, Ren Y, et al. A new insight into cross-sensitivity to humidity of SnO₂ sensor. *Small*, 2018, 14: 1703974
- 8 Gardner J W. Detection of vapours and odours from a multisensor array using pattern recognition part 1. principal component and cluster analysis. *Senss Actuators B-Chem*, 1991, 4: 109–115
- 9 Zhang H, Wang J, Tian X, et al. Optimization of sensor array and detection of stored duration of wheat by electronic nose. *J Food Eng*, 2007, 82: 403–408
- 10 Zhang S, Xie C, Zeng D, et al. A sensor array optimization method for electronic noses with sub-arrays. *Senss Actuators B-Chem*, 2009, 142: 243–252
- 11 Zhang S P. Study on dynamic reaction of metal oxide gas sensor and sensor array optimization. Dissertation for Ph.D. Degree. Wuhan: Huazhong University of Science and Technology, 2009 [张顺平. 金属氧化物气体传感器响应动力学特性与阵列优化研究. 博士学位论文. 武汉: 华中科技大学, 2009]
- 12 Gardner J W, Boilot P, Hines E L. Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach. *Senss Actuators B-Chem*, 2005, 106: 114–121
- 13 Pearce T C, Gardner J W, Göpel W. Strategies for mimicking olfaction: the next generation of electronic noses? *Senss Update*, 1998, 3: 61–130
- 14 Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*, 2003, 3: 1157–1182
- 15 Xie J Y, Xie W X. Several feature selection algorithms based on the discernibility of a feature subset and support vector Machines. *Chin J Comput*, 2014, 37: 1704–1718 [谢娟英, 谢维信. 基于特征子集区分度与支持向量机的特征选择算法. *计算机学报*, 2014, 37: 1704–1718]
- 16 Sun G-L, Song Z-C, Liu J-L, et al. Feature selection method based on maximum information coefficient and approximate Markov blanket. *Act Autom Sin*, 2017, 43: 795–805 [孙广路, 宋智超, 刘金来, 等. 基于最大信息系数和近似马尔科夫毯的特征选择方法. *自动化学报*, 2017, 43: 795–805]
- 17 Whitney A W. A direct method of nonparametric measurement selection. *IEEE Trans Comput*, 1971, 20: 1100–1103
- 18 Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern Recogn Lett*, 1994, 15: 1119–1125
- 19 Marill T, Green D. On the effectiveness of receptors in recognition systems. *IEEE Trans Inform Theor*, 1963, 9: 11–17
- 20 Fleuret F. Fast binary feature selection with conditional mutual information. *J Mach Learn Res*, 2004, 5: 1531–1555
- 21 Hall M A. Correlation-based feature selection for machine learning. Dissertation for Ph.D. Degree. Hamilton: The university of Waikato, 1999
- 22 Peng H C, Long F H, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Machine Intell*, 2005, 27: 1226–1238
- 23 Mundra P A, Rajapakse J C. SVM-RFE with MRMR filter for gene selection. *IEEE Trans Nanobiosci*, 2010, 9: 31–37
- 24 Nagle H T, Gutierrez-Osuna R, Schiffman S S. The how and why of electronic noses. *IEEE Spectr*, 1998, 35: 22–31
- 25 Peng J. Study on optimization of sensor array for E-nose system. Dissertation for M.S. Degree. Wuhan: Huazhong University of Science and Technology, 2008 [彭婧. 气敏传感器阵列优化应用研究. 硕士学位论文. 武汉: 华中科技大学, 2008]
- 26 Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. *Science*, 2011, 334: 1518–1524
- 27 Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, 20: 273–297
- 28 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, 323: 533–536

- 29 Breiman L. Random forests. *Mach Learn*, 2001, 45: 5–32
- 30 Zhang M L, Zhou Z H. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recogn*, 2007, 40: 2038–2048
- 31 Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Mach Learn*, 2002, 46: 389–422
- 32 Xue Y, Zhang L, Wang B, et al. Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis. *Appl Intell*, 2018, 48: 3306–3331
- 33 Chou P A. Optimal partitioning for classification and regression trees. *IEEE Trans Pattern Anal Machine Intell*, 1991, 13: 340–354
- 34 Strobl C, Boulesteix A L, Zeileis A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf*, 2007, 8: 25–46
- 35 Janitza S, Tutz G, Boulesteix A L. Random forest for ordinal responses: prediction and variable selection. *Comput Stat Data Anal*, 2016, 96: 57–73
- 36 Altmann A, Tološi L, Sander O, et al. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 2010, 26: 1340–1347

An effective gas sensor array optimization method based on dynamic feature importance

Guangfen WEI^{1*}, Jie ZHAO², Yanli FENG², Aixiang HE¹, Jun YU³ & Zhen-An TANG³

1. *School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China;*

2. *School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China;*

3. *School of Biomedical Engineering, Dalian University of Technology, Dalian 116024, China*

* Corresponding author. E-mail: guangfen.wei@yahoo.com

Abstract Gas sensor array optimization is a key problem in the field of electronic noses, and it is also a special feature selection problem. In this paper, we propose a novel measure of sensor (or feature) importance, named dynamic feature importance, based on feature correlation and feature importance. Also, we propose an effective electronic nose sensor array optimization algorithm SAO_DFI based on the dynamic feature importance. We analyze the effects of repeated sensors, sensor (or feature) importance, sensor (or feature) correlation, and sensor characteristic parameters, based on the proposed SAO_DFI algorithm using data collected in two different gas environments. The optimization results demonstrate the effectiveness, robustness, and interpretability of the array optimization algorithm.

Keywords array optimization, gas sensor, feature selection, feature correlation, electronic nose, dynamic feature importance

Guangfen WEI was born in 1978. She received her Ph.D. degree in mechanical and electronic engineering from Dalian University of Technology, Liaoning, in 2006. Currently, she is a professor in the School of Information and Electronic Engineering at Shandong Technology and Business University. Her research interests include signal processing theory, electronic nose, and gas sensors and their detection methods.

Jie ZHAO was born in 1995. He received his B.S. degree in Electronic Science and Technology from Shandong Technology and Business University, Shandong, in 2017. Currently, he is an M.S. candidate in the School of Computer Science and Technology at Shandong Technology and Business University. His research interests include electronic nose, sensor arrays, machine learning, and feature selection.

Yanli FENG was born in 1962. He received his M.S. degree in computational mathematics from Dalian University of Technology, Liaoning, China in 1987. Currently, he is a professor in the School of Computer Science and Technology at Shandong Technology and Business University. His research interests include machine learning and feature selection.

Aixiang HE was born in 1979. She received her Ph.D. degree in microelectronics and solid-state electronics from Dalian University of Technology, Liaoning, in 2018. Currently, she is a lecturer in the School of Information and Electronic Engineering at Shandong Technology and Business University. Her research interests include electronic nose, signal processing, pattern recognition, feature extraction, and chemical sensor arrays.