DOI: 10.13382/j.jemi.B1902505

气体传感器阵列混合气体检测算法研究*

谭光韬[1] 张文文[2] 王 磊[2]
(1. 同济大学 中德学院 上海 201804; 2. 同济大学 电子信息与工程学院 上海 201804)

摘 要:针对传统模式识别算法对混合气体定性和定量检测准确率较低的问题,提出了一种基于机器学习的新型混合气体定性识别和浓度定量检测算法。算法首先构造传感器阵列数据特征图,然后利用卷积神经网络(CNN)提取特征,根据特征提取后的特征图,使用不同分支网络对不同气体进行定性识别,得到气体种类和相应气体所处浓度区间;根据前面的气体识别结果,使用核主成分分析(KPCA)与梯度提升树(GBDT)对混合气体的组成成分进行定量估计;最后采用加州大学机器学习数据库的动态混合气体气体传感器阵列数据集进行对比验证。实验结果表明,算法在乙烯和甲烷定性识别中准确率达到了 98.7%,定量检测平均相对误差小于 4.1%。通过与传统模式识别算法对比,所提出的基于机器学习的混合气体检测算法具有更高的精度和泛化能力。

关键词: 传感器阵列;卷积神经网络;核主成分分析;梯度提升树
中图分类号: TP212.9;TN911.72 **文献标识码**:A **国家标准学科分类代码**: 510.10

Research on mixed gas detection algorithm of gas sensor array

Tan Guangtao[1] Zhang Wenwen[2] Wang Lei[2]
(1. Sino-German College, Tongji University, Shanghai 201804, China; 2. College of Electronic and Information, Tongji University, Shanghai 201804, China)

Abstract: In view of the low accuracy of the traditional pattern recognition algorithm for qualitative and quantitative detection of mixed gases, a novel algorithm of hybrid gas qualitative identification and concentration quantitative detection based on machine learning is proposed. The algorithm constructs the feature map of sensor array data first, then uses the convolutional neural network (CNN) to extract features from feature maps. According to the feature map after feature extraction, different branches are used to identify different gases, then the species of gases and their concentration range were obtained; based on the results of gas identification, the kernel principal component analysis (KPCA) and gradient boosting decision tree (GBDT) were used to estimate the composition of the mixed gases quantitatively. Finally, this paper used the dataset of sensor array of mixed gases of Machine Learning Database of the University of California to verify the results. Experimental results show that the accuracy of the algorithm in the qualitative recognition of ethylene and methane reaches 98.7% and the average relative error of quantitative detection was less than 4.1%. Compared with the traditional pattern recognition algorithm, the machine learning based mixed gas detection algorithm that proposed has higher accuracy and stronger generalization ability.

Keywords: sensor array; convolutional neural networks (CNN); kernel principal component analysis (KPCA); gradient boosting decision tree (GBDT)

收稿日期: 2019-08-26 Received Date: 2019-08-26
*基金项目:国家重点研发计划(2018YFE0105000)、国家重点研发计划(2017YFE0100900)、国家留学基金(201906260029)资助项目

0 引 言

“电子鼻系统”也被称为人工嗅觉系统,是一种模仿生物鼻的系统,它的目的是为了实现对气体的识别。“电子鼻系统”结构分为气敏传感器阵列、信号预处理单元和模式识别单元[1]。目前,机器嗅觉的主要研究方向包括敏感材料的研制和相关模式识别算法的研究。金属氧化半导体(metal oxide semiconductor, MOS)气体传感器作为气敏传感器在气体检测领域运用广泛[2-4],具有成本较低、响应速度块、使用寿命长等优点。MOS 传感器在接触到氧化性或还原性气体后,其电阻特性会发生变化,故可对气体进行识别和检测。但是 MOS 传感器有显著的交叉敏感性[5]。由于化学反应的复杂性,人们认识到,只对一种气体敏感的气敏传感器是不存在的。解决这个问题的方法之一就是采用多个具有交叉敏感性不同类型的气敏传感器组成传感器阵列[6-7],可以分别对多种气体进行测量并给出整体的测量结果。但是,仅仅通过研制传感器和探索测量方式往往并不能带来令人满意的效果。由此,许多学者在信号处理和模式识别方面做了许多工作,并且取得了一定的研究成果。在信号预处理方面,文献[8-9]综述了近年来机器嗅觉在信号和数据处理方面取得的进展;文献[10]综述了近年来机器嗅觉使用的相空间(PS)、能量矢量(EV)和功率密度谱(PSD)等特征提取方法,并提出了一些建议和新的思路。在模式识别方面,文献[11]采用基于线性核主成分分析(KPCA)的线性判别分析(LDA)改善了气体传感器响应信号非线性对模式识别的影响,提出了每种传感器对不同气体检测的贡献度,然而并未讨论对气体浓度范围和浓度的测量;文献[12]提出使用独立成分分析(ICA)来分析 MOS 传感器阵列响应数据的方法,实现了对不同种类气体的分类,但是 ICA 算法有训练时间过长,目标函数在零点处不可微的缺点;文献[13]采用主成分分析(PCA)与支持向量机(SVM)的方法进行多组分气体检测,然而 PCA 算法属于线性特征提取和分类方法,在处理 MOS 传感器非线性响应信号时往往达不到较高的精度,基于 SVM 的分类方法对大规模训练样本难以实施,对常见多分类问题可能需要多个二分类支持向量机组合解决,并且 SVM 往往需要对参数进行优化,不同气体测量参数的选取可能发生变化,这些都会在一定程度下影响测量的精度;文献[14]采用改进的支持向量机(ISVMEN)来解决电子鼻中的多分类问题,将平均分类精度提高到 92.58%,达到了较好的分类和泛化性能;文献[15]提出了利用径向基神经网络(RBF)网络对氮氧化物浓度检测的方法;文献[16]提出了利用核主成分分析(KPCA)对多路非线性特征提取,后使用提取的特征进行 K 最近邻域(KNN)分类器建模识别目标气体的方案,气体识别准确率达到了 98.33%,但是 KNN 算法进行分类时计算量随着样本容量的增大而增大、各类样本不均衡时预测误差较大且有超参数 K 的选取问题。文献[17]在气体泄漏检测中利用改进人工鱼群算法对 SVM 算法中参数进行寻优,以消除温度对测量精度的影响。

本文针对 MOS 气体传感器对气体浓度响应曲线的非线性特性,提出一种气体识别卷积神经网络(gas recognition convolutional neural network, GRCNN),卷积神经网络本质是一个多层感知机,其采用的局部连接和权值共享方式,这种方式一方面减小了权值数量,使网络易于优化;另一方面降低了模型的复杂度,减小了过拟合的风险;同时,卷积神经网络具有一些传统机器学习方法所没有的优点,如具有良好的容错能力、泛化能力、避免人工进行特征提取等[18]。针对混合气体浓度估计的平均相对误差较大的问题,本文提出一种基于 KPCA[19] 和 GBDT[20] 的浓度估计方法。GBDT(gradient boosting decision tree)通过加法模型及不断减小训练过程产生的残差来将数据分类或回归,能够建立阵列式气体传感器与对应气体浓度复杂的非线性关系,具有鲁棒性高,解释性好等优点。本文算法首先使用 GRCNN 对气体分类,后根据混合气体情况使用不同 GBDT 模型进行浓度预测。一般而言,利用模式识别方法对气体进行定性识别和定量识别受到训练样本质量、数量及不同浓度配比影响较大,故采用加州大学动态混合气体气体传感器阵列数据集进行验证。数据集包含 95 种不同 CH_4 与 C_2H_4 混合气体浓度配比和近万条的稳态数据,有助于反映气体传感器阵列在测量范围内的真实测量情况。

1 混合气体定性识别

1.1 数据预处理与特征图构建

本文采用数据集的传感器阵列具有四种不同的 MOS 传感器,型号分别为 TGS2600, TGS2602, TGS2610, TGS2620,每种类型 4 个单元。故每次采样可获得 16 个传感器测量值(数据排列方式为 TGS2602; TGS2602; TGS2600; TGS2600; TGS2610; TGS2610; TGS2620; TGS2620; TGS2602; TGS2602; TGS2600; TGS2600; TGS2610; TGS2610; TGS2620; TGS2620),实验采用传感器测量的稳态值,包含甲烷 CH_4 与乙烯 C_2H_4 95 种不同的浓度配比和 9 072 条可用数据。由于不同传感器输出分布不同,预处理时分别对每个维度进行归一化处理,假设第 i 维数据为 $\boldsymbol{X}_i=[x_i^0,x_i^1,\cdots,x_i^{9\,072}]^{\mathrm{T}}$, 第 i 维的最大值为 $x_i^{\max}$, 最小值为 $x_i^{\min}$, 归一化公式为:

$$x'^j_i=\frac{x_i^j-x_i^{\min}}{x_i^{\max}-x_i^{\min}} \tag{1}$$

CNN 常被用于解决图像识别、物体检测的任务，而输入的一般为二维或三维图像数据，故需要将归一化后的数据转化为二维数据。同时，由于单次采样受随机误差影响较大，为减小单次测量的干扰并增强检测鲁棒性，选取当混合气体处于同一浓度配比的 16 个相邻时刻构建特征图。取同一浓度配比 16 组连续数据 $\boldsymbol{x}'^{t},\cdots,\boldsymbol{x}'^{t+15}$ 构成大小为 16 × 16 的特征图 $\boldsymbol{X}$，公式为：

$$\boldsymbol{X}=\begin{bmatrix} x'^{t}_{0} & \cdots & x'^{t}_{15} \\ \cdots & \cdots & \cdots \\ x'^{t+15}_{0} & \cdots & x'^{t+15}_{15} \end{bmatrix} \tag{2}$$

1.2　GRCNN 神经网络结构

考虑到分类数较少，浅层卷积神经网络结构简单，训练和前向运算效率高，本文设计一个浅层的卷积神经网络 GRCNN 实现对不同气体定性识别和所处的浓度区间识别，CH_4 和 C_2H_4 混合气体检测结构如图 1 所示，根据气体种类，输入特征图大小为 16×16×1，输出为两个 4×1 的向量，分别表征两种气体所处的浓度区间，设第 i 种气体测量浓度最大值为 $c_i^{\max}$，则输出向量[1,0,0,0]，[0,1,0,0]，[0,0,1,0]，[0,0,0,1]分别表示气体所处浓度区间为 0，$\left(0,\frac{c_i^{\max}}{3}\right]$，$\left(\frac{c_i^{\max}}{3},\frac{2c_i^{\max}}{3}\right]$，$\left(\frac{2c_i^{\max}}{3},c_i^{\max}\right]$。

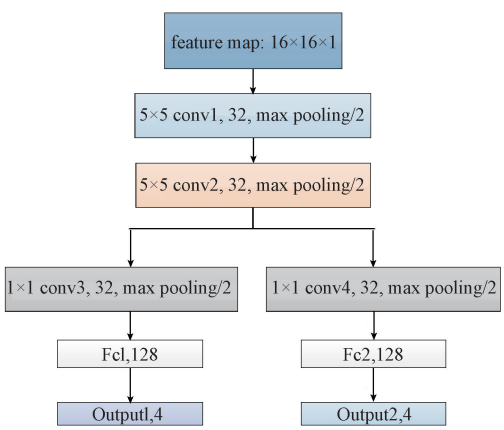


图 1　GRCNN 网络结构

Fig. 1　Block diagram of GRCNN

GRCNN 分为特征提取网络和分类网络，特征提取网络由卷积层 conv1+最大池化层，卷积层 conv2+最大池化层组成，conv1 和 conv2 均含 32 个大小为 5×5 的卷积核，最大池化层核大小为 2×2 且步长为 2，用于特征图的降维和提高模型泛化能力。最终得到大小为 4×4×32 的特征图；分类网络为两个分支网络，一定程度上可减小分类数目，防止模型过拟合。将得到特征图作为后两分支网络的输入，分支网络由卷积层、最大池化层、全连接层组成，用于对每种气体所处浓度区间识别。

1.3　GRCNN 网络损失函数与训练

网络采用分支结构减少网络全连接层参数个数和防止由于浓度区间内样本数较少造成的过拟合问题。假设训练样本模型输出经过 sigmoid 函数后的概率输出为 $\boldsymbol{y}_0^i=[y_0^0,y_0^1,y_0^2,y_0^3]^{\mathrm{T}}$，$\boldsymbol{y}_1^i=[y_1^0,y_1^1,y_1^2,y_1^3]^{\mathrm{T}}$，真值为 $\boldsymbol{t}_0^i=[t_0^0,t_0^1,t_0^2,t_0^3]^{\mathrm{T}}$，$\boldsymbol{t}_1^i=[t_1^0,t_1^1,t_1^2,t_1^3]^{\mathrm{T}}$。模型分别采用两个交叉熵损失函数 l_0^i,l_1^i 表示第 i 个样本预测结果损失，公式如下：

$$l_0^i=-\sum_{c=0}^{3}t_0^c\log y_0^c \tag{3}$$

$$l_1^i=-\sum_{c=0}^{3}t_1^c\log y_1^c \tag{4}$$

GRCNN 使用 ReLU 作为激活函数，损失函数为交叉熵损失与 L2 正则化之和。算法 batch 值大小设为 b；设特征提取网络部分参数表示为 ω_i；两个分类网络部分参数分别为 ω_j 和 ω_k。则训练时损失函数如下：

$$L_0=\frac{\sum_{i=0}^{b}l_0^i}{b}+\lambda\left(\sum_i\|\omega_i\|_2^2+\sum_j\|\omega_j\|_2^2\right) \tag{5}$$

$$L_1=\frac{\sum_{i=0}^{b}l_1^i}{b}+\lambda\left(\sum_i\|\omega_i\|_2^2+\sum_k\|\omega_k\|_2^2\right) \tag{6}$$

模型损失函数分为 L_0 与 L_1 两部分，神经网络采用 Adam(adaptive moment estimation)优化算法分别根据损失函数 L_0,L_1 优化特征提取网络和各自分类网络参数。

2　混合气体定量识别

2.1　KPCA 特征提取算法

在对数据完全无知的状态下，PCA 算法可能会丢失数据信息；虽然通过降维可减少特征个数，消除变量之间的关系，但无法解决非线性依赖问题[19]。设 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$ 表示观测样本。其中，每个样本 $\boldsymbol{x}_i$ 为 K 维列向量，N 表示观测样本总数，通过核函数 $\phi(\boldsymbol{X})$ 将原始数据 $\boldsymbol{X}$ 中向量 $\boldsymbol{x}_i$ 映射到高维(D 维)空间，称为特征空间，记作 $\boldsymbol{Z}$。

$$\phi(\boldsymbol{x}_i):\boldsymbol{R}^K\rightarrow\boldsymbol{R}^D,D>K \tag{7}$$

当核函数 $\phi(\boldsymbol{X})$ 已经中心化，即：

$$\sum_{i=1}^{M}\phi(\boldsymbol{x}_i)=0 \tag{8}$$

则在特征空间 $\boldsymbol{Z}$ 中协方差方程为：

$$\boldsymbol{C}_Z=\frac{1}{N}\phi(\boldsymbol{X})\phi(\boldsymbol{X})^{\mathrm{T}}=\frac{1}{N}\sum_{i=1}^{N}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}} \tag{9}$$

式中：$\boldsymbol{C}_Z$ 是 $D\times D$ 矩阵。求解协方差矩阵特征值，特征值求解方程为：

$$\phi(\boldsymbol{X})\phi(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{p}=\sum_{i=1}^{N}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{p}=\lambda\boldsymbol{p} \tag{10}$$

其中 λ 和 $\boldsymbol{p}$ 分别表示特征值和特征向量，由于未定义映射 $\phi(\boldsymbol{x})$，所以上式无法直接求解。又因为特征向量 $\boldsymbol{p}$ 可由 $\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_N)$ 线性表示，由此存在 N 维列向量 $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$，满足式(11)。

$$\boldsymbol{p}=\sum_{i=1}^{N}\alpha_i\phi(\boldsymbol{x}_i)=\phi(\boldsymbol{X})\boldsymbol{\alpha} \tag{11}$$

将式(11)代入式(10)得：

$$\phi(\boldsymbol{X})\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})\boldsymbol{\alpha}=\lambda\phi(\boldsymbol{X})\boldsymbol{\alpha} \tag{12}$$

式(12)两边都左乘 $\phi(\boldsymbol{X})^{\mathrm{T}}$，得：

$$\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})\boldsymbol{\alpha}=\lambda\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})\boldsymbol{\alpha} \tag{13}$$

定义核矩阵 $\boldsymbol{K}=\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})$，$\boldsymbol{K}$ 为 $N\times N$ 的半正定对称矩阵，特征值问题转化为下式的非零特征值求解问题。

$$\boldsymbol{K}\boldsymbol{\alpha}=\lambda\boldsymbol{\alpha} \tag{14}$$

根据核技巧求解核矩阵 K 特征值 $\lambda_1,\lambda_2,\cdots,\lambda_N$，其中 $\lambda_1>\lambda_2>\cdots>\lambda_N$，对应特征向量为 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_N$。通过求解累计贡献率保留前 t 个特征向量，累计贡献率求解如式(15)所示。

$$r_{CCR}=\frac{\sum_{i=1}^{t}\lambda_i}{\sum_{i=1}^{N}\lambda_i}\times 100\% \tag{15}$$

假设测试集为含有 L 个样本的 $K\times L$ 矩阵 $\boldsymbol{X}_{\mathrm{test}}=[\boldsymbol{x}_{\mathrm{test}1},\boldsymbol{x}_{\mathrm{test}2},\cdots,\boldsymbol{x}_{\mathrm{test}L}]$，则 $\phi(\boldsymbol{X}_{\mathrm{test}})=[\phi(\boldsymbol{x}_{\mathrm{test}1}),\phi(\boldsymbol{x}_{\mathrm{test}2}),\cdots,\phi(\boldsymbol{x}_{\mathrm{test}L})]$，对其进行中心化后得到 $\tilde{\phi}(\boldsymbol{X}_{\mathrm{test}})$。将 $\tilde{\phi}(\boldsymbol{X}_{\mathrm{test}})$ 投影到高维空间特征向量 $\boldsymbol{p}$ 计算主成分：

$$\boldsymbol{t}_{\mathrm{test}}=\tilde{\phi}(\boldsymbol{X}_{\mathrm{test}})^{\mathrm{T}}\boldsymbol{p}=\tilde{\phi}(\boldsymbol{X}_{\mathrm{test}})^{\mathrm{T}}\phi(\boldsymbol{X})\boldsymbol{\alpha}=\tilde{\boldsymbol{K}}_{\mathrm{test}}\boldsymbol{\alpha} \tag{16}$$

其中 $\tilde{\boldsymbol{K}}_{\mathrm{test}}=\tilde{\phi}(\boldsymbol{X}_{\mathrm{test}})^{\mathrm{T}}\phi(\boldsymbol{X})$ 计算公式为：

$$\tilde{\boldsymbol{K}}_{\mathrm{test}}=\boldsymbol{K}_{\mathrm{test}}-\boldsymbol{K}_{\mathrm{test}}\cdot\boldsymbol{1}_N-\boldsymbol{1}_{NL}\cdot\boldsymbol{K}+\boldsymbol{1}_{NL}^{\mathrm{T}}\cdot\boldsymbol{K}\cdot\boldsymbol{1}_N \tag{17}$$

式中：$\boldsymbol{K}_{\mathrm{test}}=\phi(\boldsymbol{X}_{\mathrm{test}})^{\mathrm{T}}\phi(\boldsymbol{X})$，$\boldsymbol{1}_N$ 是一个 $N\times N$ 的矩阵，$\boldsymbol{1}_{NL}$ 为 $N\times L$ 矩阵，两矩阵每个元素都为 $1/N$。

2.2 GBDT 浓度识别

GBDT 被广泛运用于分类、回归等问题，属于 Boosting 算法族，集成学习通过构建和结合多个学习器来完成任务。GBDT 每一轮迭代拟合残差学习一个 CART 树作为弱学习器，通过弱学习器组成的基函数的线性组合，不断减小训练中出现的残差。GBDT 算法流程如图 2 所示，模型可描述为：

$$f_M(\boldsymbol{x})=\sum_{t=1}^{M}T(\boldsymbol{x};\gamma_t) \tag{18}$$

其中模型共训练 M 轮，每轮产生一个弱学习器 $T(\boldsymbol{x};\gamma_t)$，设 $\boldsymbol{D}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$ 为训练样本，前一轮迭代得到的学习器是 $f_{t-1}(\boldsymbol{x})$，损失函数是 $L(y,f_{t-1}(\boldsymbol{x}))$，损失函数使用平方损失函数，即：

$$L(y_i,f_t(\boldsymbol{x}_i))=(y_i-f_{t-1}(\boldsymbol{x}_i)-T(\boldsymbol{x}_i;\gamma_t))^2 \tag{19}$$

利用损失函数最小化求解 γ_t，即：

$$\gamma_t=\mathrm{argmin}\sum_{i=1}^{n}L(y_i,f_{t-1}(\boldsymbol{x}_i)-T(\boldsymbol{x};\gamma_t)) \tag{20}$$

令残差 $r_{i,t}=y_i-f_{t-1}(\boldsymbol{x}_i)$，则：

$$\gamma_t=\mathrm{argmin}\sum_{i=1}^{n}(r_{i,t}-T(x_i;\gamma_t))^2 \tag{21}$$

最终输出强学习器 $f_M(\boldsymbol{x})$，GBDT 算法训练步骤如下。

1）初始化弱学习器：

$$f_0(\boldsymbol{x})=\mathrm{argmin}\sum_{i=1}^{n}L(y_i,\gamma_0) \tag{22}$$

2）迭代训练，其中 $t=1,2,\cdots,M$。

(1)对于样本 $i=1,2,\cdots,n$，计算残差：

$$r_{i,t}=y_i-f_{t-1}(\boldsymbol{x}_i) \tag{23}$$

(2)根据残差训练回归树参数 γ_t。

(3)模型更新：

$$f_t(\boldsymbol{x})=f_{t-1}(\boldsymbol{x})+T(\boldsymbol{x};\gamma_t) \tag{24}$$

3）返回强学习器 $f_M(\boldsymbol{x})$。

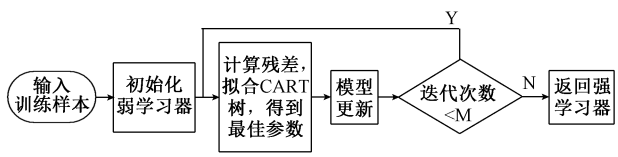


图 2 GBDT 算法流程

Fig. 2 Flow chart of GBDT

3 混合气体检测算法

本文提出的混合气体检测分为气体所处浓度区间识别和气体浓度定量检测两部分。浓度区间分类方法训练算法为：选取数据集中 95 种不同浓度配比二元混合气体的传感器矩阵数据，为减少单一数据干扰，每种浓度配比至少有 32 组数据；将 62 种浓度配比数据作为训练集，33 种浓度配比数据作为测试集，训练集与测试集比例约等于 2∶1。根据每种气体所处浓度区间设定对应的数据标签，同一浓度配比数据组合成特征图；使用训练集组合成的特征图进行浓度区间识别训练并使用测试集特征图进行测试。

浓度定量检测算法如下：将上述训练集和测试集根据浓度分为单一气体和混合气体训练集与测试集；以混合气体 C_2H_4 和 CH_4 为例，针对单一气体(C_2H_4 或 CH_4)和混合气体(C_2H_4 和 CH_4)的训练集构造核矩阵 $\boldsymbol{K}_1(\cdot)$、$\boldsymbol{K}_2(\cdot)$、$\boldsymbol{K}_3(\cdot)$、$\boldsymbol{K}_4(\cdot)$ 通过 KPCA 提取训练样本特征；针对单一气体和混合气体 4 种情况分别训练 4 种 GBDT 模型 $f_{M1}(\boldsymbol{x})$ $f_{M2}(\boldsymbol{x})$ $f_{M3}(\boldsymbol{x})$ $f_{M4}(\boldsymbol{x})$ 分别用于单一气体下 C_2H_4、CH_4 和混合气体下 C_2H_4、CH_4 的浓度检测。

混合气体定性定量检测算法如图 3 所示。先选取测试集中同浓度配比数据构成特征图，经过 GRCNN 后得

到 C_2H_4 与 CH_4 各自所处浓度区间。根据所处浓度区间不难得到当前气体混合情况。若检测到所有气体所处浓度范围均为 0，结束检测，否则选择对应的核矩阵进行特征提取并选择对应的 GBDT 模型进行浓度预测。

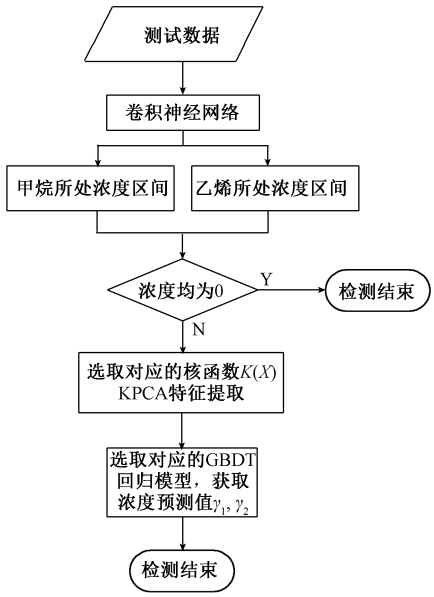


图 3　混合气体检测算法流程

Fig. 3　Flow chart of mixed gas detection algorithm

4　实验样本与结果

4.1　实验样本组成

实验样本由 95 种不同浓度组合的传感器阵列数据组成。C_2H_4 的浓度范围为 $0\sim300\times10^{-6}$，CH_4 的浓度范围为 $0\sim20\times10^{-6}$。其中每种浓度采样频率为 10 Hz，每次采样连续采样 3.2 s，每个浓度组合至少得到 32×16 大小的样本。训练集包含 62 种浓度组合，数据大小为 1 984×16；测试集包含 33 种浓度组合，数据大小为 1 056×16。

4.2　气体浓度区间判别实验结果分析

GRCNN 神经网络算法使用 ReLU 作为激活函数，损失函数为交叉熵损失与 L2 正则化之和，用于防止过拟合且使模型更稳定。优化算法 batch 值大小设为 $b=100$，初始学习率为 0.001。作为对比的 BP 神经网络具有 16 个输入神经元，10 个隐含层神经元，4 个输出层神经元；损失函数为均方误差与 L2 正则化之和；优化算法为随机梯度下降法；优化算法 batch 设置为 100，学习率为 0.015。PCA 算法选取主成分累计贡献率大于 95% 的主成分个数，将数据由原来的 16 维减小到 12 维后通过 BP 神经网络进行定性识别。

如表 1 与图 4 所示，GRCNN 的气体成分四分类定性识别平均准确率达到 98.7%，比 3 层的 BP 神经网络高出 12%，比 PCA 的分类方法高出了 6.8%，说明了相对于 BP 神经网络容易陷入局部最小值且收敛速度慢的缺点（图 5），本文提出的网络结构能够较好提取多维信号中的特征信息且识别精度较高、收敛速度快，有效的提高了测试样本识别准确率。

表 1　BP 神经网络和 GRCNN 分类准确率

Table 1　Classification accuracy of BP neural network and GRCNN

气体类别	样本数目	测试样本识别准确率/%		
		BP	PCA+BP	GRCNN
C_2H_4	384	79.3	89	95.8
CH_4	224	100	71.4	100
混合气体	704	86.5	100	100
平均		86.7	91.9	98.7

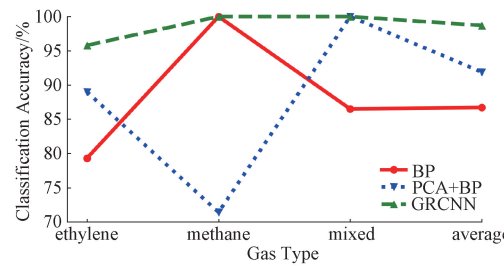
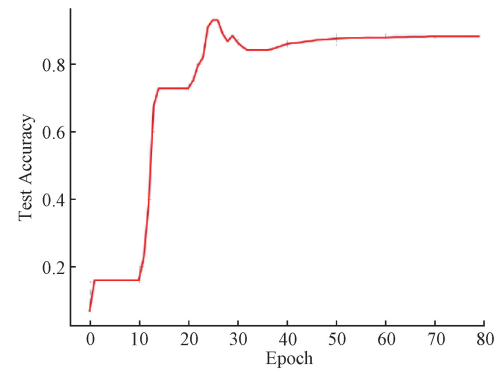


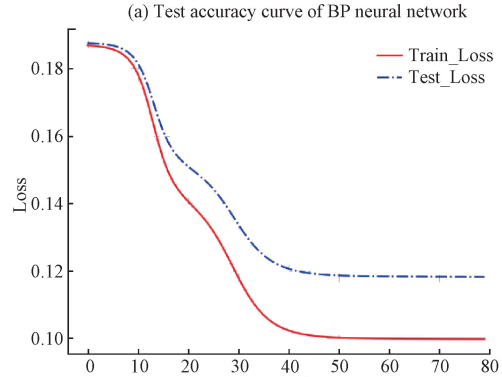
图 4　分类准确率折线

Fig. 4　Line chart of classification accuracy



(a) BP神经网络测试精度曲线

(a) Test accuracy curve of BP neural network



(b) BP神经网络训练损失与测试损失曲线

(b) Training and test loss curve of BP neural network

图 5　BP 神经网络训练过程

Fig. 5　Diagram of training process of BP neural network

如表2所示,分别使用两个分支网络对两种气体所处浓度区间进行识别,模型对 C_2H_4 所处浓度区间判别准确率达到了97%,对 CH_4 浓度所处区间判别准确率达到了95.4%,测试集中两种气体所处浓度区间16分类的准确率大于92.4%,均优于传统算法。训练过程准确率变化如图6所示。

表2　GRCNN气体浓度区间识别准确率

Table 2　Identification accuracy of gas concentration interval of GRCNN

气体类别	样本数目	C_2H_4 浓度区间识别准确率/%	CH_4 浓度区间识别准确率/%
C_2H_4	384	91.7	100
CH_4	224	100	71.4
混合气体	704	100	100
平均		97.7	95.4

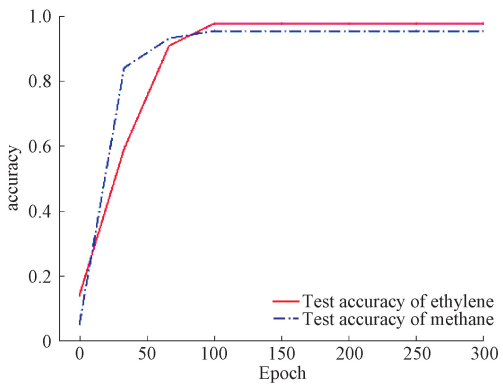


图6　GRCNN气体浓度区间判别测试精度曲线

Fig. 6　Test accuracy curve of GRCNN

4.3　混合气体浓度定量实验结果分析

根据前面两种气体浓度区间分类的结果对单一气体和混合气体使用不同的KPCA与GBDT模型进行回归,本文中的KPCA算法核函数采用高斯径向基核函数,其中核系数 $\sigma=5$。根据多次试验,主成分贡献率为98%时,模型相对平均误差最小,数据维数由16维增加到20维。GBDT模型采用平方损失函数,学习率设置为0.1,经过交叉验证,C_2H_4 回归模型中的CART回归树最大深度设为3,CH_4 回归模型中的CART树最大深度设为2。作为对比的SVR算法使用高斯径向基核函数,核函数系数设置为输入特征个数的倒数,惩罚因子设为1。

如图7所示,在不使用KPCA算法进行特征提取时,GBDT模型平均相对误差为4.8%,明显好于支持向量回归算法(15.0%)和弹性回归算法(26.4%);如图8所示,在使用KPCA算法进行特征提取后,GBDT模型平均相对误差为4.1%,显著低于SVR与Elastic Net模型。

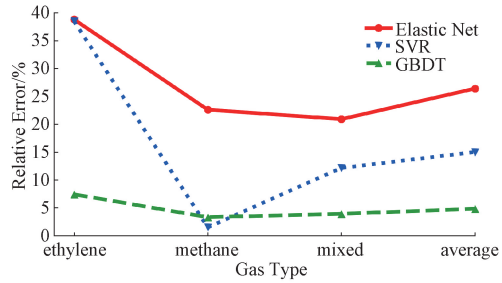


图7　浓度回归相对误差折线

Fig. 7　Line chart of relative error of concentration regression

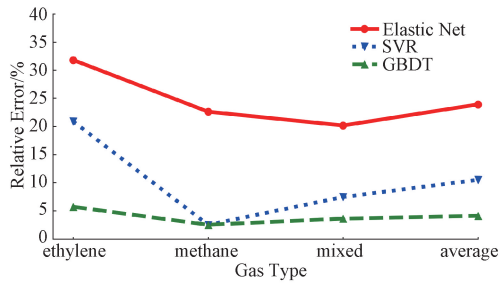


图8　特征提取后浓度回归相对误差折线

Fig. 8　Line chart of relative error of concentration regression after feature extraction

5　结　论

针对混合气体检测中准确度较低的问题,本文提出了一种基于机器学习的新型混合气体定性识别和浓度定量检测的算法。提出基于卷积神经网络的气体所处浓度区间识别方法,不仅实现了对混合气体的定性识别,而且实现对气体所处浓度区间的识别,混合气体定性测试识别率达到了98.6%以上,对气体所处浓度区间识别准确率在95.4%以上,具有一定的产品化价值;在浓度检测方面,提出了KPCA结合GBDT的浓度检测算法,对单一气体和混合气体训练不同的模型,根据前面定性识别的结果,使用不同的模型进行浓度检测,测试平均相对误差小于4.1%,相对于弹性网络回归、支持向量机回归,所提出的KPCA结合GBDT浓度定量检测算法平均相对误差分别降低了19.8%、6.4%,表明了方法的有效性。

参考文献

[1] MIGUEL P, LAURA E-G. A 21st century technique for food control: Electronic noses[J]. Analytica Chimica Acta, 2009, 638(1):1-15.

[2] 窦仁超. 气体传感器在国外航天器上的应用[J]. 仪器仪表学报,2016,17(5):1187-1200.
DOU R CH. Application of gas sensors on foreign

spacecraft[J]. Chinese Journal of Scientific Instrument, 2016,37(5):1187-1200.

[3] 杜利农,柴春祥,郭美娟. 电子鼻在水产品品质检测中的应用研究进展[J]. 电子测量技术,2014,37(5):80-84.
DU L N, CHAI CH X, GUO M J. The application of electronic nose in quality detection of aquatic product[J]. Electronic Measurement Technology, 2014, 37(5):80-84.

[4] SANKARAN S, KHOT LR, PANIGRAHI S. Biology and applications of olfactory sensing system: A review[J]. Sensors & Actuators B Chemical, 2012, 171-172(8): 1-17.

[5] ZHANG L, TIAN F C, DANG L J, et al. A novel background interferences elimination method in electronic nose using pattern recognition[J]. Sensors & Actuators A Physical, 2013, 201(10):254-263.

[6] ZHANG L, TIAN F C. Performance study of multilayer perceptrons in a low-cost electronic nose[J]. IEEE Transactions on Instrumentation & Measurement, 2014, 63(7):1670-1679.

[7] LIN S W, YING K C, CHEN S C, et al. Particle swarm optimization for parameter determination and feature selection of support vector machines[J]. Expert Systems with Applications, 2008, 35(4): 1817-1824.

[8] MARCO S, GUTIERREZ-GALVEZ A. Signal and data processing for machine olfaction and chemical sensing: A review[J]. IEEE Sensors Journal, 2012, 12(11): 3189-3214.

[9] LEOPOLD J H, BOS L D J, STERK P J, et al. Comparison of classification methods in breath analysis by electronic nose[J]. Journal of Breath Research, 2015, 9(4):046002.

[10] YAN J, GUO X Z, DUAN S K, et al. Electronic nose feature extraction methods: A review[J]. Sensors, 2015, 15(11):27804-27831.

[11] ZHANG L, TIAN F C, PEI G S. A novel sensor selection using pattern recognition in electronic nose[J]. Measurement, 2014(54): 31-39.

[12] 宋凯,王祁,林定选. 基于 ICA 的气体模式识别方法研究[J]. 仪表技术与传感器,2009(Z1):41-43.
SONG K, WANG Q, LIN D X. ICA based gas pattern recognition method[J]. Instrument Technique and Sensor, 2009(Z1):41-43.

[13] 余道洋,戚功美,瞿顶军,等. 基于 SVM 和 PCA 的痕量多组分气体检测方法[J]. 模式识别与人工智能,2015,28(8):720-727.
YU D Y,QI G M,QU D J,et al. Detection method of trace multi-component gases based on SVM and PCA[J]. PR & AI, 2015,28(8):720-727.

[14] DANG LJ, TIAN F C, ZHANG L, et al. A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose[J]. Sensors & Actuators A Physical, 2014, 207(1):67-74.

[15] 严玥,江赟,严实. 利用 RBF 网络的火电厂氮氧化物浓度检测方法[J]. 电子测量与仪器学报,2017,31(1):45-50.
YAN Y, JIANG Y, YAN SH. Detection method of NO_x concentration in coal fired power plant using RBF network[J]. Journal Of Electronic Measurement And Instrumentation, 2017,31(1):45-50.

[16] 许永辉,陈寅生,张铭. MOS 传感器阵列的二元混合气体检测方法研究[J]. 仪器仪表学报,2018,39(5):179-187.
XU Y H, CHEN Y SH, ZHANG M. Binary mixed gas detection method using MOS sensor array[J]. Chinese Journal of Scientific Instrument, 2018, 39(5): 179-187.

[17] 何怡刚,苏蓓蕾,李兵,等. 基于 AJAFSA-SVM 温度补偿算法的 SF_6 泄漏检测方法研究[J]. 电子测量与仪器学报,2018,32(8):42-49.
HE Y G, SU B L, LI B, et al. Research on SF_6 leakage detection based on AJAFSA-SVM temperature compensation algorithm[J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(8): 42-49.

[18] REN S, HE K, GIRSHICK R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[19] HOFMANN T, SCHOLKOPF B, SMOLA A J. Kernel methods in machine learning[J]. Annals of Statistics, 2008,36(3):1171-1220.

[20] SCHONLAU M. Boosted regression (boosting): An introductory tutorial and a stata plugin[J]. Stata Journal, 2005,5(3): 330-354.

作者简介

谭光韬,2017 年于东华大学获得学士学位,现为同济大学硕士研究生,主要研究方向为模式识别与气体传感器。
E-mail:gt_tan@ tongji. edu. cn

Tan Guangtao received his B. Sc. degree from Donghua University in 2017. Now

he is a M. Sc. candidate at Tongji University. His main research interests include pattern recognition and gas sensor.

张文文,现为同济大学博士研究生,主要研究方向为机器学习,智能信息处理与模式识别。

E-mail:zhangwenwen_1203@163.com

Zhang Wenwen is a Ph. D. candidate at Tongji University. His main research interests are machine learning, Intelligent information processing and pattern recognition.

王磊,现为同济大学教授,博士生导师,主要研究方向为传感器检测技术与测量系统。

E-mail:leiwang@tongji.edu.cn

Wang Lei is a professor and Ph. D. supervisor at Tongji University. His main research interests are sensor detection technology and measurement system.