多气体的SVM数据融合定性识别方法

黄为勇[1,2],任子晖[1],童敏明[1]
HUANG Wei-yong[1,2],REN Zi-hui[1],TONG Min-ming[1]

1.中国矿业大学 信息与电气工程学院,江苏 徐州 221008
2.徐州工程学院 信电工程学院,江苏 徐州 221008
1.School of Information and Electrical Engineering,China University of Mining and Technology,Xuzhou,Jiangsu 221008,China
2.School of Information and Electronic Engineering,Xuzhou Institute of Technology,Xuzhou,Jiangsu 221008,China
E-mail:hwy@xzit.edu.cn

HUANG Wei-yong,REN Zi-hui,TONG Min-ming.Multi-gases qualitative identification using SVM and data fusion. Computer Engineering and Applications,2009,45(9):241-243.

Abstract: The traditional method for qualitative identification of multi-gases based on neural networks has the problems of over-fitting and poor generalization ability.In order to solve the drawbacks,this paper proposes a new method based on support vector machine (SVM) and multi-sensor data fusion,which uses multi-class classifier to fuse data of sensor array composed of several gas sensors,temperature sensor and humidity sensor,effectively eliminates the influence of ambient temperature and humidity on gas sensors,and reaches 100% qualitative identification rate.The experimental results show that the method is effective.
Key words: support vector machine(SVM);sensor array;data fusion;multi-gases qualitative identification

摘 要:针对基于神经网络的多气体定性识别方法中存在的过学习和泛化能力差的问题,提出了一种基于支持向量机(SVM)与多传感器数据融合的多气体定性识别方法。该方法采用结构化风险最小化准则的多类分类支持向量机对由多个气体传感器、温度和湿度传感器组成的传感阵列的数据进行融合,克服了传统方法的缺陷,消除了环境温度与湿度等因素的影响,实现了100%的定性识别率,实验结果证明了该方法的有效性。
关键词:支持向量机(SVM);传感器阵列;数据融合;多气体定性识别
DOI:10.3778/j.issn.1002-8331.2009.09.070 **文章编号:**1002-8331(2009)09-0241-03 **文献标识码:**A **中图分类号:**TP212.6

1 引言

气体传感器普遍存在交叉敏感,稳定性和选择性差,响应特性极易受到温度、湿度等环境因素影响的缺陷,采用单一气体传感器无法实现对多种气体进行准确的定性识别(即对待检测的气体的种类做出正确评价)。自1986年Miller R,Lange E[1]等人提出了利用传感器组来解决多种气体的检测问题以来,各国学者都在致力于利用传感器阵列实现多种气体的检测问题。随着神经网络技术的发展,人们逐渐采用神经网络对由多个敏感程度不同的气体传感器组成传感器阵列的输出信号进行融合,实现对多种气体的定性识别[2-4]。由于该方法采用经验风险最小化(ERM)准则,用十分复杂的模型去拟合有限的样本,存在过拟合问题,从而导致学习机器泛化能力的下降,气体的识别能力受到较大的影响。事实上,这种用ERM准则代替期望风险最小化的学习目的并没有充分的理论依据,只是一种直观上合理的想当然的做法[5]。

支持向量机(Support Vector Machine,简称SVM)是一种建立在VC维理论和结构风险最小化原则基础上的一种新型机器学习方法,在最小化经验风险的同时,最小化置信区间的上界,从而获得更强的泛化能力,有效地解决实际应用中样本不足的缺陷,以及神经网络的过学习、局部极小值和泛化能力差的问题[6-7]。

本文提出了一种基于支持向量机和多传感器数据融合的多气体定性识别方法。该方法采用多个气体传感器、湿度和温度传感器构成传感器阵列,应用基于支持向量机的数据融合方法对传感器阵列数据进行融合,消除温度、湿度等环境因素对传感器特性的影响,实现了小样本情况下多气体的自动定性识别,实验结果证明了该方法的有效性。

2 原理与算法

2.1 多传感器数据融合的多气体定性识别原理

目前很难找到只对某种气体敏感的传感器材料,单个气体传感器对不同气体敏感响应可能会有变化,但不具备自动识别

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.50534050);江苏省高校自然科学研究计划项目(No.06KJD460174)。
作者简介:黄为勇(1963-),男,博士生,副教授,主要研究领域为计算机测控与传感器技术、智能信息处理;任子晖(1962-),男,教授,博士生导师,主要研究领域为计算机测控技术;童敏明(1956-),男,教授,博士生导师,主要研究领域为传感器检测技术。
收稿日期:2008-01-28 **修回日期:**2008-04-15

气体种类和数量的能力,单个气体传感器功能十分有限。在多气体识别中,单一传感器不能保证识别的准确性和可靠性。由于多个传感器可同时提供待识别气体的冗余信息和互补信息,采用多个气体传感器构成的传感器阵列是气体定性识别的一个切实可行的方法[8]。

多传感器数据融合通过一定的技术融合手段,协调使用多个传感器信息,把不同传感器所提供的局部不完整信息和相关信息加以综合,消除多传感器之间可能存在的冗余和矛盾,并加以互补,降低其不确定性,获得对待识别气体的一致性认识。

另一方面,气体传感器特性极易受到温度和湿度等环境因素的影响,因此在气体识别系统中增加一个温度和一个湿度传感器。把环境温度和湿度作为变量同时测量,并与气体传感器的信息进行融合,从而获得对识别环境的一致性描述和正确的理解,使定性识别系统具有容错性,提高识别系统的正确性和鲁棒性。

多传感器系统中,各传感器所提供的信息都具有一定程度的不确定性,对这些不确定信息的融合过程实际上是一个不确定的推理过程[9]。依推理方法的不同,多传感器信息融合主要有贝叶斯估计、卡尔曼滤波、模糊推理和神经网络等方法,目前主要采用神经网络方法。鉴于神经网络方法的缺陷及支持向量机优秀的分类和泛化性能,应用多类分类支持向量机对由气体传感器、温度和湿度传感器所组成的传感器阵列的输出信号进行融合,挖掘传感器阵列所决定的多维空间所蕴涵的系统信息,消除温度和湿度等环境因素对传感器输出的影响,实现对不同气体的定性识别,达到气体的定性识别的目的。系统原理结构如图 1 所示。

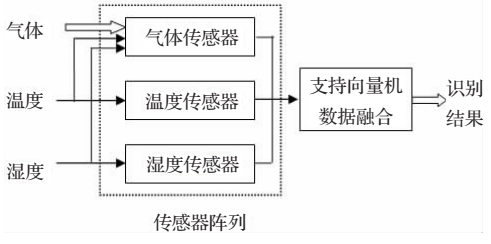

图 1　多气体定量识别原理结构图

基于支持向量机的多传感器数据融合方法是根据系统要求和融合形式,确定支持向量机的参数,通过系统学习确定支持向量机的模型,对各传感器数据进行融合,使系统具有较强的容错性和鲁棒性,实现对多种气体进行定性识别。

2.2　支持向量机分类原理

SVM 分类算法是从线性可分情况下的最优分类面(Optimal Hyperplane)提出的。所谓最优分类面就是要求分类面不但能将两类样本点无错误地分开,而且要使两类的分类空隙最大。设线性可分样本集(x_i,y_i),$i=1,2,\cdots,n$。d 维空间中线性判别函数的一般形式为 $g(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}$,分类面方程是 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0$,将判别函数进行归一化,使两类所有样本都满足$|g(x)|\geqslant 1$,此时离分类面最近的样本的$|g(x)|=1$,而要求分类面对所有样本都能正确分类,就是要求它满足:

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b})-1\geqslant 0 \quad i=1,2,\cdots,n \tag{1}$$

上式中使等号成立的那些样本叫做支持向量(Support Vectors)。两类样本的分类空隙(Margin)的间隔大小为 $2/\|\boldsymbol{w}\|$。因此,最优分类面问题可以表示成如下的约束优化问题,即在条件(1)的约束下,求函数

$$\phi(\boldsymbol{w})=\frac{1}{2}\|\boldsymbol{w}\|^2=\frac{1}{2}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})$$

的最小值。利用 Lagrange 优化方法将上述最优问题化为如下凸二次规划的对偶问题:

$$\begin{cases}\max\sum_{i=1}^{n}\boldsymbol{a}_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}_i\boldsymbol{y}_j(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\\ \text{s.t}\quad \boldsymbol{a}_i\geqslant 0,i=1,\cdots,n\\ \sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{y}_i=0\end{cases}$$

这是一个不等式约束下二次函数机制问题,存在唯一最优解。若 $\boldsymbol{\alpha}_i^*$ 为最优解,则有

$$\boldsymbol{w}^*=\sum_{i=1}^{n}\boldsymbol{a}_i^*\boldsymbol{y}_i\boldsymbol{x}_i$$

式中,$\boldsymbol{\alpha}_i^*$ 不为零的样本,即为支持向量。解上述问题后得到的最优分类函数是:

$$f(x)=\mathrm{sgn}((\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}^*)=\mathrm{sgn}(\sum_{i=1}^{n}\boldsymbol{a}_i^*\boldsymbol{y}_i\boldsymbol{x}_i^*\boldsymbol{x}+\boldsymbol{b}^*)$$

在线性不可分的情况下,可以引入松弛变量 $\boldsymbol{\xi}_i$,$(\boldsymbol{\xi}_i\geqslant 0,i=1,2,\cdots n)$,使超平面 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0$ 满足:

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b})\geqslant 1-\boldsymbol{\xi}_i$$

当 $0<\xi_i<1$ 时样本点 $\boldsymbol{x}_i$ 仍旧被正确分类,而当 $\xi_i\geqslant 1$ 时样本点 $\boldsymbol{x}_i$ 被错分。为此,引入以下目标函数:

$$\psi(w,\xi)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{n}\boldsymbol{\xi}_i$$

其中 C 是一个正常数,称为惩罚因子,它起着控制对错分样本惩罚程度的作用,实现错分样本的比例与算法复杂度之间的折中。C 值越大,表示主要把重点放在减少分类错误上,C 值越小,主要把重点放在分离超平面上,避免过学习问题。在实际应用中,可通过交叉验证方法得到合适的 C 值。

若在原始空间中线性不可分,则可以通过非线性变换 Φ 将输入空间变换到一个高维空间,然后在这个新空间中求取最优线性分类面,而这种非线性变换是通过定义适当的核函数(内积函数)实现的。令:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)\rangle$$

用核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 代替最优分类平面中的点积 $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$,就相当于把原特征空间变换到了某一新的特征空间。此时优化函数变为:

$$Q(\alpha)=\sum_{i=1}^{n}\boldsymbol{a}_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}_i\boldsymbol{y}_jK(\boldsymbol{x}_i\boldsymbol{x}_j)$$

而相应的判别函数式则为:

$$f(x)=\mathrm{sgn}[(\boldsymbol{w}^*)^{\mathrm{T}}\Phi(\boldsymbol{x})+\boldsymbol{b}^*]=\mathrm{sgn}(\sum_{i=1}^{n}\boldsymbol{a}_i^*\boldsymbol{y}_iK(\boldsymbol{x}_i\boldsymbol{x})+\boldsymbol{b}^*) \tag{2}$$

其中 $\boldsymbol{x}_i$ 为支持向量,$\boldsymbol{x}$ 为未知向量,这就是支持向量机。

根据 Hilbert-Schmidt 定理,只要一种核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 满足 Mercer 条件,它就可作为这里的内积使用,采用不同的内积(核)函数将导致不同的支持向量机算法。常用的核函数为径向基核函数:

$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\{-\frac{\|x-\boldsymbol{x}_i\|^2}{\sigma^2 2}\}$,其中宽度系统 σ 可由实验确定。

在多类问题中,需要组合多个支持向量机进行分别,目前多类分类支持向量机方法主要有“1-v-r”和“1-v-1”两种方法。

表1 学习样本数据

样本标号	传感器响应(经放大调理后)					被测气体样本
	SensorCO	SensorCO$_2$	SensorNO	SensorT	SensorH	
1	0.051 8	0.001 2	0.004 5	0.292 8	1.366	CO 50 ppm
3	0.151 9	0.001 7	0.004 3	0.292 6	1.371	CO 150 ppm
5	0.251 3	0.001 6	0.004 6	0.293 0	1.359	CO 250 ppm
7	0.352 4	0.001 3	0.004 1	0.292 7	1.377	CO 350 ppm
9	0.450 9	0.001 5	0.004 0	0.289 7	1.349	CO 450 ppm
11	0.001 9	0.019 2	0.003 8	0.293 8	1.368	SO_2 5 ppm
13	0.001 8	0.055 9	0.003 1	0.292 8	1.369	SO_2 15 ppm
15	0.001 9	0.092 9	0.002 8	0.292 9	1.370	SO_2 25 ppm
17	0.001 9	0.130 6	0.002 3	0.294 0	1.367	SO_2 35 ppm
19	0.002 0	0.167 8	0.002 1	0.292 8	1.367	SO_2 45 ppm
21	0.001 8	0.001 6	0.050 7	0.293 8	1.378	NO 1 ppm
23	0.002 1	0.001 9	0.143 0	0.293 4	1.359	NO 3 ppm
25	0.001 9	0.001 3	0.236 3	0.294 1	1.361	NO 5 ppm
27	0.001 8	0.001 2	0.331 0	0.292 9	1.359	NO 7 ppm
29	0.002 0	0.001 5	0.424 3	0.293 5	1.377	NO 9 ppm

表2 定性识别结果

样本标号	传感器响应(经放大调理后)					识别类别	实际气体
	SensorCO	SensorCO$_2$	SensorNO	SensorT	SensorH		
2	0.102 2	0.001 5	0.004 6	0.293 5	1.375	1	CO
4	0.202 6	0.001 1	0.004 5	0.291 9	1.378	1	CO
6	0.300 9	0.001 5	0.003 9	0.300 1	1.380	1	CO
8	0.402 5	0.001 4	0.004 2	0.293 2	1.358	1	CO
10	0.502 8	0.001 5	0.004 2	0.292 0	1.369	1	CO
12	0.002 0	0.037 9	0.003 5	0.293 5	1.371	2	SO_2
14	0.002 2	0.074 8	0.002 9	0.293 4	1.372	2	SO_2
16	0.001 6	0.111 8	0.002 5	0.293 6	1.359	2	SO_2
18	0.002 1	0.149 5	0.002 2	0.293 5	1.358	2	SO_2
20	0.001 9	0.186 5	0.001 9	0.293 6	1.358	2	SO_2
22	0.001 7	0.001 5	0.097 1	0.293 6	1.379	3	NO
24	0.002 0	0.001 1	0.189 6	0.293 4	1.349	3	NO
26	0.002 0	0.001 2	0.285 1	0.293 4	1.374	3	NO
28	0.002 1	0.001 4	0.377 5	0.293 1	1.380	3	NO
30	0.001 8	0.001 6	0.470 9	0.293 8	1.368	3	NO

对于 K 类分类,“1-v-r”方法需构建 K 个 SVM 两类分类器,每一个分类器分别将某一类的样本从其他类的样本中鉴别出来。第 i 类的 SVM 训练时,样本中属于第 i 类的样本标为正的一类,其他样本标为负的一类。而“1-v-1”方法只是为 K 类训练样本构造所有可能的两类分类器,每类仅仅在 K 类的两类训练样本上训练,这样一共需 $K(k-1)/2$ 个 SVM 分类器。这两种方法各有利弊,可根据场合灵活选用。

2.3 基于支持向量机多传感器数据融合的多气体定性识别算法

设 d 维的传感器阵列的输出信号为 x_i,气体的类别为 y_i,待分类的气体种类为 m,通过实验建立样本集(x_i,y_i),i=1,2,…n,n 为样本容量。一般情况下,气体类型模式一般不会太多,可选择“1-v-r”多类分类方法。考虑到适当改变径向基函数的参数可以逼近其他形式的核函数,故采用径向基函数作为核函数。

根据上述支持向量机多类分类原理,多气体定性识别算法可描述为:

步骤 1 数据准备,调整 y_i,若气体属于 q 类,y_{qi}=1,否则 y_{qi}=-1。

步骤 2 建立支持向量机多类分类器把训练样本通过 Φ 映射到高维特征空间,选择适当的径向基函数参数和惩罚参数 C,利用训练样本(x_i,y_i)求解二次优化就问题,以获得(a,b)及其对应的支持向量,从而得到识别模型。重复步骤 2 m 次,得到 m 个识别模型。

步骤 3 利用获得的识别模型,根据气体类型输入模式,判断其类型。若第 q 个识别模型的输出为 1,则为第 q 类气体;若第 q 个识别模型的输出为-1,则被测气体不是第 q 类气体。对每一输入,只有一个识别模型的输出为 1,否则识别模型要重新训练。

3 实验与结果

根据文献[10],采用静态配气法对一氧化碳、二氧化硫和一氧化氮三种气体分别配制了 10 个不同的浓度。使用由三个气体传感器(SensorCO、SensorSO_2 和 SensorNO),一个温度传感器(SensorT)和一个湿度传感器(SensorH)的传感器形成的传感器阵列,所有传感器信号经信号调理电路和多功能数据采集卡进行采集,并送入计算机。

实验测量了不同情况下气体传感器阵列的响应,共获得 30 组数据。将实验测得的数据分为两部分,其中样本标号为单数的 15 组为学习样本(训练样本),样本标号为偶数的 15 组数据用于测试(系统验证)。学习样本数据如表 1 所示。

将 CO、SO_2 和 NO 类别分别定义为 1、2、3 类。由表 1 中的每一行中 5 个传感器响应值组成 $\boldsymbol{x}_i$ 向量和气体的类别 $\boldsymbol{y}_i$ 向量组成训练目标向量构成训练集$(\boldsymbol{x}_i,\boldsymbol{y}_i)$,$(i=1,2,\cdots 15)$,其中 $\boldsymbol{x}_i \in R5$,$\boldsymbol{y}_i \in \{1,2,3\}$。实验时首先用训练集样本进行训练,并利用训练后模型对标号为偶数的测试样本进行测试。

支持向量机识别模型的建立、训练和预测算法均在 MATLAB6.5 环境下编程实现。通过交叉验证进行径向基函数的宽度系数 σ 和惩罚系数 C 的选择。实验表明,在径向基函数的宽度系统 σ=1 时,C 值大于 23 时,即可实现 100%的定性识别准确率,结果如表 2 所示。

对比实验表明,采用目前常用的基于神经网络的数据融合方法,其识别结果皆有错误,而且在融合前必须对传感器数据进行特定的预处理。而该方法无需对传感器输出信号进行任何预处理即可得到 100%的定性识别率。通过现场测试,该方法可在 0.1 秒内完成(HP 7650,AMD Sempron 3400+1.8 G,512 M),完全可以满足在线识别的要求。

4 结束语

提出的基于支持向量机和多传感器数据融合的多气体定性识别方法具有充分理论依据的支持,克服了神经网络方法存在的过学习、泛化能力差的缺陷,消除了温度、湿度等环境因素对传感器特性的影响,鲁棒性好,识别能力强,尤其适合小样本的学习,是实现多气体定性识别的一种有效方法。

参考文献:

[1] Miller R,Lange E.Multidimensional sensor for gas analysis[J].Sensor and Actuators,1986(9):39-48.

[2] Hong Hyung-Ki.Gas identification using micro gas sensor array and neural-network pattern recognition[J].Sensor and Actuators B,1996(33):68-71.

(下转 248 页)