基于电子鼻传感器性能的互信息特征选择算法*

陶 洋,刘翔宇,梁志芳
(重庆邮电大学 通信与信息工程学院,重庆 400065)

摘 要:当前的互信息特征选择算法为提高泛化性能而未对专一应用领域进行优化,电子鼻传感器阵列优化作为一类特殊的特征选择问题,使用传统算法难以搜索出最优的特征子集。结合气体传感器阵列特殊的冗余性和特有的敏感性,提出了一种基于电子鼻传感器性能的互信息特征选择算法并对阵列进行优化,通过两种不同的电子鼻公开数据集验证了传感器特性对识别精度的影响,证明了所提出算法的有效性。

关键词:电子鼻;传感器阵列;互信息;特征选择
中图分类号:TN102 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.201007

中文引用格式:陶洋,刘翔宇,梁志芳. 基于电子鼻传感器性能的互信息特征选择算法[J]. 电子技术应用,2021,47(10):86-89.
英文引用格式:Tao Yang,Liu Xiangyu,Liang Zhifang. Mutual information feature selection algorithm based on electronic nose sensor performance[J]. Application of Electronic Technique,2021,47(10):86-89.

Mutual information feature selection algorithm based on electronic nose sensor performance

Tao Yang,Liu Xiangyu,Liang Zhifang
(School of Communication and Information Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China)

Abstract: In order to improve the generalization performance, most of the current mutual information feature selection algorithms have not optimized the specific application field, and electronic nose sensor array optimization is a special feature selection problem. It is difficult to search for the optimal feature subset using traditional algorithms. Combining the special redundancy and unique sensitivity of the gas sensor array, this paper proposes a mutual information feature selection optimization algorithm based on the performance of the electronic nose sensor. The influence of the sensor characteristics on the recognition accuracy is verified through two different electronic nose public data sets, which proved the effectiveness of the proposed algorithm.

Key words: electronic nose; sensor array; mutual information; feature selection

0 引言

相较于视觉的发达,人类的嗅觉并不出色。因此机器嗅觉可以在多个领域替代人工[1],实现对气体的检测与分析,例如环境质量监测[2]、食品安全[3]、医疗卫生等[4],对电子鼻系统的研究具有重大的价值。

电子鼻传感器阵列的优化是一类特殊的特征选择问题[5],主要表现在两个方面:

(1)电子鼻系统中的传感器普遍具有广谱效应[5],因此传感器之间的冗余有别于传统特征之间的冗余,在冗余度相同的情况下前者更倾向于较大的冗余分布,即筛选出较少重叠的特征;

(2)与传统特征选择不同,电子鼻传感器阵列更倾向于筛选出高敏感特征,即传感器对不同气体的响应有更大的幅度差。

综上所述,本文提出一种电子鼻传感器性能的互信息特征选择算法(Sensor Performance Mutal Information,SPMI),结合传感器特性进行特征子集的筛选,相较于现有算法获得了更优的识别精度。该算法的主要创新点有:

(1)针对候选特征与已选特征之间的冗余度设计权重函数,降低联合冗余信息离散程度小的特征权值,使得筛选出的特征之间相互冗余的数量降低;

(2)设计基于方差的特征敏感性评价函数,使得筛选的特征对目标响应具有更高的辨识度。

* 基金项目:国家重点研发计划项目(2019YFB2102001);重庆市基础研究与前沿探索项目(cstc2018jcyjAX0549);重庆市教育委员会科学技术研究项目(KJQN201800617)

1 现有算法分析

互信息特征选择算法根据评价函数 $J(X_m)$ 对特征 X_m 打分[6]，并根据需求的特征数量确定迭代次数，每次迭代从候选特征集合 M 中获得最高分特征并将其加入已选特征集合 N 中。

1994 年，Battiti 等人提出 MIFS 算法[7]，其评价函数如式(1)所示，该算法将特征与类的互信息作为相关性，将候选特征与已选特征之间的互信息作为冗余性，通过系数 β 平衡两者之间的权重：

$$J_{\mathrm{MIFS}}(X_m)=I(X_m;L)-\beta\sum_{X_n\in N}I(X_m;X_n) \tag{1}$$

其中，$I(X_m;L)$ 为特征与标签的相关性；$I(X_m;X_n)$ 表示特征之间的相关性；β 作为平衡系数，权衡两者之间的权重。

2000 年，Yang 等人提出了 JMI 算法[8]，该算法在特征与类的相关性基础上，通过条件互信息进一步消除其中的无效信息：

$$J_{\mathrm{JMI}}(X_m)=I(X_m;L)-\frac{1}{|N|}\sum_{X_n\in N}I(X_m;L)-I(X_m;L|X_n) \tag{2}$$

2005 年，Peng 等人提出了 mRMR 算法[9]，该算法确定特征之间的冗余性的系数为已选特征数量的倒数，计算出冗余程度的集中趋势：

$$J_{\mathrm{mRMR}}(X_m)=I(X_m;L)-\frac{1}{|N|}\sum_{X_n\in N}I(X_m;X_n) \tag{3}$$

2018、2019 年 Gao 等人相继提出了 CFR 算法[10]、MRMD 算法[11]，分别如式(4)、式(5)所示：

$$J_{\mathrm{CFR}}(X_m)=\sum_{X_n\in N}I(X_m;L|X_n)-I(X_m;X_n;L) \tag{4}$$

$$J_{\mathrm{MRMD}}(X_m)=I(X_m;L)-\frac{1}{|N|}\sum_{X_n\in N}I(X_m;X_n)-I(X_m;L|X_n) \tag{5}$$

通过上述分析，可知现有算法并未根据电子鼻传感器阵列特性进行优化，使得其在电子鼻传感器数据集上未能筛选出最优特征子集。据此本文提出基于传感器性能的互信息特征选择算法。

2 传感器性能的互信息特征选择算法

2.1 评价函数

互信息特征选择算法主要通过评价函数作为准则筛选特征，SPMI 算法的评价函数分为三部分：特征相关性 $J_{\mathrm{rev}}(X_m)$、特征冗余性 $J_{\mathrm{rdu}}(X_m)$ 以及特征敏感性 $J_{\mathrm{sen}}(X_m)$：

$$J_{\mathrm{SPMI}}(X_m)=J_{\mathrm{rev}}(X_m)-J_{\mathrm{rdu}}(X_m)+J_{\mathrm{sen}}(X_m) \tag{6}$$

$$J_{\mathrm{SPMI}}(X_m)=I(X_m;L)-\frac{1}{|N|}\sum_{X_n\in N}\frac{1}{1+\sigma}[I(X_m;X_n;L)-I(X_m;L|X_n)]+\alpha E(X_m) \tag{7}$$

可以发现此评价函数采用特征与类的互信息衡量特征的相关性：

$$J_{\mathrm{rev}}(X_m)=I(X_m;L) \tag{8}$$

2.2 特征冗余性

候选特征与已选特征的互信息 $I(X_m;X_n)$ 可以简单地看作特征的冗余性。然而由于标签信息的存在，此互信息可根据是否存在标签信息划分为两个部分：

$$I(X_m;X_n)=I(X_m;X_n;L)+I(X_m;X_n|L) \tag{9}$$

其中，$I(X_m;X_n|L)$ 中不带任何标签信息，被称为类外冗余。类外冗余因不带标签信息可不被考虑在冗余范围内，因此，可以缩小冗余信息为类内冗余 $I(X_m;X_n;L)$。同时减去此特征与标签特有的条件互信息 $I(X_m;L|X_n)$，以最大化相关性[12]。可得到冗余互信息评价函数为：

$$J_{\mathrm{rdu}}(X_m)=\frac{1}{|N|}\sum_{X_n\in N}I(X_m;X_n;L_m)-I(X_m;L_m|X_n) \tag{10}$$

然而当多个特征的冗余度相同时，传感器阵列更倾向于较大的冗余分布，因此可以根据均值 μ 和标准差 σ 表征特征的类内冗余分布：

$$\sigma=\sqrt{\frac{1}{|N|}\sum_{X_n\in N}[I(X_m;X_n;L)-\mu]^2} \tag{11}$$

$$\mu=\frac{1}{|N|}\sum_{X_n\in N}I(X_m;X_n;L) \tag{12}$$

类内冗余标准差 σ 越大，此特征的冗余离散程度越高，相同冗余度下与已选特征冗余的数量越小。由于冗余性评价函数与整体评价函数呈现负相关，因此将此标准差的倒数作为权重函数 W，并通过常数项扩大函数定义域：

$$W(X_m)=\frac{1}{1+\sigma} \tag{13}$$

最后将函数 $W(X_m)$ 作为特征冗余度的权值，获得特征冗余评价函数：

$$J_{\mathrm{rdu}}(X_m)=\frac{1}{|N|}\sum_{X_n\in N}\frac{1}{1+\sigma}[I(X_m;X_n;L_m)-I(X_m;L_m|X_n)] \tag{14}$$

2.3 特征敏感性

电子鼻传感器阵列所识别的目标气体之间往往具有相关性，例如混合气体的识别中，不同目标气体之间可能只有浓度的差异。这就要求传感器特征具有足够的敏感性区分不同的目标气体。

SPMI 算法根据特征变量的方差设计特征的敏感性评价函数。特征的方差越大，特征分量与均值的差异也就越高，传感器特征对不同目标气体的辨识能力也就越强。同时设置系数 α 平衡敏感性占总评价函数的比重：

$$J_{\mathrm{sen}}(X_m)=\alpha E(X_m) \tag{15}$$

2.4 算法流程与优势分析

SPMI 算法通过逐步迭代获取候选特征集合中每次得分最高的特征，算法的详细流程如下：

输入：候选特征集 M、标签向量 L；

输出：最优特征子集 N；

//初始化

(1)根据需要的特征数量确定迭代次数 n；

(2)构造空集 N、设置敏感性系数 α；

//迭代过程

(1)开始遍历候选特征集 M；

(2)取出特征 X,通过评价函数式(6)获取该特征的分数 R;
(3)完成遍历;
(4)取出分数集合 R 中获得最高分的特征 X;
(5)将此特征加入集合 N,从候选特征集 M 中删除特征 X;
(6)返回迭代开始部分;
//结束
输出特征集合 N。

综上所述,SPMI 算法结合传感器阵列特性,做出以下优化:

(1)基于类内冗余度获得最大相关性,并设计冗余度的标准差为权重函数,以筛选出冗余离散度更高的特征;

(2)基于特征方差设计敏感性,以筛选出对不同目标气体更敏感的特征。

3 实验验证与分析

3.1 实验数据集介绍

本文使用了两个数据集验证所提出算法性能:

(1)加州大学欧文分校(University of California Irvine,UCI)机器学习库中收录的流量调制下气体传感器阵列数据集[13]。其中包含了从 16 个金属氧化物传感器在气流调制条件下获取的 58 个时间序列内的响应。调制的气流为丙酮、乙醇以及二者的气态混合物,实验将传感器时间序列内稳态最大响应值作为候选特征。

(2)重庆大学生物感知与智能信息处理实验室采集的伤口细菌电子鼻公开数据集[14]。实验采用了 34 个化学传感器获取对大肠杆菌培养液、金黄色葡萄球菌培养液、铜绿假单胞菌培养液以及任意两种混合培养液的响应,同样将传感器稳态最大响应值作为候选特征。

3.2 实验设置

实验将在上述数据集中运行泛化性能良好的现有算法(MIM、JMI、mRMR、MIFS、CFR、MRMD)以及所提出算法(SPMI)进行对比,获得各对比算法筛选出特征子集,并在分类算法(支持向量机(Suport Vertor Machine,SVM))下获得特征子集的识别精度[15]。

经过多次实验,确定在流量调制数据集下特征敏感性的系数 $\alpha=1$,伤口细菌数据集下系数 $\alpha=0.3$,能达到最优效果。支持向量机的核函数采用径向基函数(Radial Basis Function,RBF)能获得更好的分类精度。

3.3 实验结果

本次实验在流量调制数据集和伤口细菌数据集下获取的特征子集的识别精度趋势如图 1 和图 2 所示,精度的数值如表 1 和表 2 所示。

图 1 和图 2 展现了在两个数据集上应用不同的算法得到的传感器子集在进行模式识别精度的趋势变化。可以发现在流量调制数据集中,当特征数量在 30%~50%

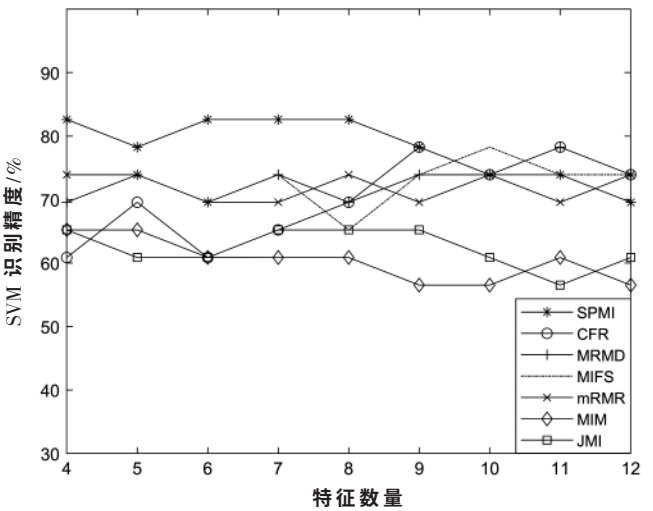

图 1 流量调制数据集特征子集识别精度折线图

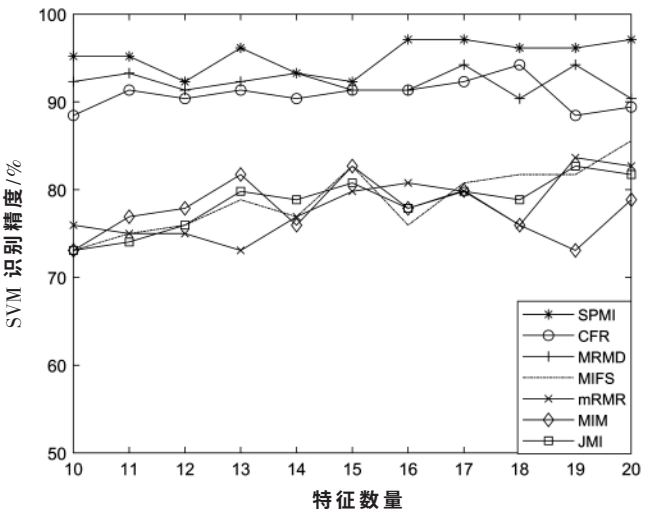

图 2 伤口细菌数据集特征子集识别精度折线图

表 1 流量调制数据集下各算法筛选传感器的识别精度 (%)

算法	数量								
	4	5	6	7	8	9	10	11	12
MIM	65.22	65.22	60.89	60.87	60.87	56.52	56.52	60.87	56.52
JMI	65.22	60.87	60.87	65.22	65.22	65.22	60.87	56.52	60.87
mRMR	73.91	73.91	69.57	69.57	73.91	69.57	73.91	69.57	73.91
MIFS	73.91	73.91	69.57	73.91	65.22	73.91	78.26	73.91	73.91
MRMD	69.57	73.91	69.57	73.91	69.57	73.91	73.91	78.26	73.91
CFR	60.87	69.57	60.87	65.22	69.57	78.22	73.91	78.26	73.91
SPMI	82.61	78.26	82.61	82.61	82.61	78.26	73.91	73.91	69.57

表 2 伤口细菌数据集下各算法筛选传感器的识别精度

(%)

算法	数量										
	10	11	12	13	14	15	16	17	18	19	20
MIM	73.08	76.92	77.88	81.73	75.96	82.69	77.88	79.91	75.96	73.08	78.85
JMI	73.08	74.04	75.96	79.81	78.85	80.77	77.88	79.81	78.85	82.69	81.73
mRMR	75.96	75.00	75.00	73.08	76.92	79.81	80.77	79.81	75.96	83.65	82.69
MIFS	73.08	75.00	75.96	78.85	76.92	82.69	75.96	80.77	81.73	81.73	85.58
MRMD	92.32	93.27	91.35	92.31	93.27	91.35	91.35	94.24	90.38	94.23	90.38
CFR	88.46	91.35	90.38	91.35	90.38	91.35	91.35	92.31	94.23	88.46	89.42
SPMI	95.19	95.19	92.30	96.15	93.26	92.30	97.12	97.12	96.15	96.15	97.12

的范围时，SPMI 算法筛选的特征子集获得了最好的识别精度，随着特征数量的继续上升该算法也保持了良好的效果；在伤口细菌数据集下，SPMI 算法相较于对比算法则始终保持了最高的识别精度。

表 1 和表 2 列出了在两个数据集中不同算法筛选出特征子集具体的识别精度值。可以发现，各算法在伤口细菌数据集下的识别精度普遍高于流量调制数据集；相同数据集同等特征数量下，SPMI 算法精度提升的最大值均能达到 20%以上。

4 结论

本文针对电子鼻系统特性提出一种基于传感器性能的互信息特征选择算法，并在电子鼻相关数据集中筛选特征子集验证识别精度。实验表明，SPMI 算法确实能够针对传感器特性进行有效优化，筛选出的传感器特征子集能够获得更高的识别精度，且相较于现有的互信息特征选择算法筛选出的子集有较大的提升。

现阶段电子鼻系统工作的环境较为复杂，所识别的目标气体常为混合气体而非单质，因此针对多标记的特征选择算法更为契合电子鼻系统。今后的工作将继续改进 SPMI 算法，使得算法能够充分考虑在多目标下传感器特征的性能，进一步扩大算法的应用范围。

参考文献

[1] SCOTT S M, JAMES D, ALI Z. Data analysis for electronic nose systems[J]. Microchimica Acta, 2006, 156(3-4): 183-207.

[2] TASTAN M, GÖKOZAN H. Real-time monitoring of indoor air quality with Internet of Things-based E-Nose[J]. Applied Sciences, 2019, 9(16): 3435.

[3] Li Zhenfeng, Wang Ning, VIGNEAULT C. Electronic nose and electronic tongue in food production and processing[J]. Stewart Postharvest Review, 2006, 2(4): 1-5.

[4] GARDNER J W, SHIN H W, HINES E L. An electronic nose system to diagnose illness[J]. Sensors and Actuators B: Chemical, 2000, 70(1-3): 19-24.

[5] WILSON D M, GARROD S, HOYT S, et al. Array optimization and preprocessing techniques for chemical sensing microsystems[J]. Sensors Update, 2002, 10(1): 77-106.

[6] 周生彬，黄叶金. 基于互信息的变量选择方法[J]. 统计与决策，2020，36(1)：20-23.

[7] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks, 1994, 5(4): 537-550.

[8] YANG H H, MOODY J. Data visualization and feature selection: new algorithms for nongaussian data[C]//Advances in Neural Information Processing Systems, 2000: 687-693.

[9] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[10] GAO W, HU L, ZHANG P, et al. Feature selection considering the composition of feature relevancy[J]. Pattern Recognition Letters, 2018, 112: 70-74.

[11] GAO W, HU L, ZHANG P. Feature redundancy term variation for mutual information-based feature selection[J]. Applied Intelligence, 2020, 50(4): 1272-1288.

[12] ZHOU H F, ZHANG Y, ZHANG Y J, et al. Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy[J]. Applied Intelligence, 2018, 49(3): 883-896.

[13] ZIYATDINOV A, FONOLLOSA J, FERNANDEZ L, et al. Bioinspired early detection through gas flow modulation in chemo-sensory systems[J]. Sensors and Actuators B: Chemical, 2015, 206: 538-547.

[14] SUN H, TIAN F, LIANG Z, et al. Sensor array optimization of electronic nose for detection of bacteria in wound infection[J]. IEEE Transactions on Industrial Electronics, 2017, 64(9): 7350-7358.

[15] QIU S, GAO L, WANG J. Classification and regression of ELM, LVQ and SVM for E-nose data of strawberry juice[J]. Journal of Food Engineering, 2015, 144: 77-85.

(收稿日期：2020-10-13)

作者简介：

陶洋(1964-)，男，博士后，博士生导师，教授，主要研究方向：机器学习、模式识别。

刘翔宇(1996-)，男，硕士研究生，主要研究方向：机器嗅觉、模式识别。

梁志芳(1989-)，通信作者，女，博士，讲师，主要研究方向：机器嗅觉、模式识别、迁移学习等，E-mail：liangzf@cqupt.edu.cn。

扫码下载电子文档

版权声明