

引用格式:陶洋,刘翔宇,梁志芳. 基于互信息特征选择的电子鼻传感器阵列优化算法[J]. 微电子学与计算机,2021,38(7):36-41. [TAO Y, LIU X Y, LIANG Z F. Electronic nose sensor array optimization algorithm based on Mutal Information Feature Selection[J]. Microelectronics & Computer,2021,38(7):36-41.]

基于互信息特征选择的电子鼻传感器阵列优化算法

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: 电子鼻传感器阵列目前面临传感器数量过多、数据维数过高等问题,导致电子鼻系统体积臃肿、模式识别算法精度差. 基于互信息的特征选择算法筛选出电子鼻传感阵列中传感器的最优子集,可以在降低电子鼻系统体积的同时,获得更好的模式识别精度. 目前传统的互信息特征选择算法没有考虑到气体与传感器的特性,应用于电子鼻系统获取的特征子集识别精度较低. 针对上述问题,提出一种基于电子鼻系统性能权重的互信息特征选择算法,该算法能够衡量特征的区别性、冗余性和敏感性并筛选出更适合电子鼻阵列系统的传感器子集. 将提出的方法应用于流量调制传感器阵列数据集和伤口细菌电子鼻数据集集中进行传感器筛选,获得的特征子集在相同的模式识别算法中达到了更高的精度.

关键词: 电子鼻;传感器;互信息;特征选择;性能权重;最优子集

中图分类号: TP212

文献标识码: A

文章编号: 1000-7180(2021)07-0036-06

Electronic nose sensor array optimization algorithm based on Mutal Information Feature Selection

TAO Yang, LIU Xiangyu, LIANG Zhifang

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Electronic nose sensor arrays are currently facing the problems of too many sensors and high data dimensions, which leads to bloated electronic nose systems and poor accuracy of pattern recognition algorithms. By selecting the optimal subset of sensors in the electronic nose sensing array through a feature selection algorithm based on mutual information, it is possible to reduce the volume of the electronic nose system and obtain better pattern recognition accuracy. At present, the traditional mutual information feature selection algorithm does not consider the characteristics of the gas and the sensor, and the recognition accuracy of the feature subset obtained by the electronic nose system is low. In this paper, a mutual information feature selection algorithm based on the performance weight of the electronic nose system is proposed. This algorithm can measure the features of distinguishability, redundancy and sensitivity to filter out a subset of sensors that are more suitable for the electronic nose array system. The method proposed is applied to the flow modulation sensor array data set and wound bacteria electronic nose data set for sensor screening, and the obtained feature subset achieves higher accuracy in the same pattern recognition algorithm.

Key words: electronic nose; sensor; mutal information; feature selection; performace weight; optimal subset

收稿日期: 2020-09-02; 修回日期: 2020-11-11

基金项目: 国家重点研发计划资助(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 引言

电子鼻是通过模拟哺乳动物的嗅觉器官实现气体识别的仿生系统,其传感器阵列模拟生物的嗅觉细胞,对气体信息进行采集;其模式识别算法模拟中枢神经,实现气体信息的分类与识别.电子鼻系统的应用领域非常广阔,如环境监测、食品安全、医疗卫生等,具有广阔的应用前景与重要的研究意义.

本文提出一种基于电子鼻系统性能权重的互信息特征选择算法(Performance Weight Mutal Information, PWMI),针对电子鼻系统中传感器的性能设计特征评价函数,获取更符合电子鼻系统特性的传感器子集,得到比传统特征选择算法更高的识别精度. PWMI 算法的主要创新点如下:

(1)针对复合气体间存在相同成分的情况,引入对特征的区分性能评价函数,使得能够区分相似标签的特征获得更高的分数,筛选出更符合电子鼻系统特性的传感器;

(2)针对候选特征与已选特征设计基于联合互信息的冗余性评价函数,尽可能筛选出相护独立的特征;

(3)设计基于方差的特征敏感性评价函数,用于筛选高敏感性的特征,剔除可能受到噪声影响的特征.

2 现有互信息特征选择算法分析

在信息论中,互信息用来度量随机变量之间的独立程度^[1].基于互信息的特征选择算法将特征和特征所对应的标签视为随机变量,通过互信息衡量随机变量的相关程度.算法的流程如下:

首先在候选特征集合 M 中使用评价函数 $J(X_m)$ 对特征 X_m 打分,按照分数的高低从集合 M 中选择需求数量的特征加入已选特征集合 N 中,并将选中特征从集合 M 中删除,即可获取特征子集 N .其中标签变量为 L .

Battiti 等人于 1994 年提出 MIFS^[2] 算法,其评价函数如式(1)所示:

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

式中, $I(X_m; L)$ 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.可以发现 MIFS 算法更倾向于选择高标签相关低特征冗余的特征.

2000 年 Yang 等人提出了 JMI^[3] 算法如式(2),

该算法使用条件互信息消除冗余特征中无效信息:

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

Peng 等人于 2005 年提出了 $m\text{RMR}$ ^[3] 算法:

$$J_{m\text{RMR}}(X_m) = I(X_m; L) - \frac{1}{|N|} \sum_{X_n \in N} I(X_m; X_n) \quad (3)$$

该算法在 MIFS 的基础上修改平衡参数为已选特征数量的倒数,更具体的计算出候选特征和已选特征的集中趋势.

2018、2019 年 Gao 等人相继提出了 CFR^[5] 算法、MRMD^[6] 算法,分别为式(4)、式(5)所示.进一步提高特征之间冗余性的有效计算:

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

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

经过对现有的互信息特征选择算法分析,可以发现上述算法均在提高特征与标签相关性,降低特征之间的冗余性,而未考虑到电子鼻系统中传感器相关的特性.因此本文提出基于电子鼻系统性能权重的互信息特征选择算法.

3 基于性能权重的互信息特征选择算法

3.1 算法整体模型

本文针对传统的算法应用于电子鼻系统存在局限性问题,提出基于性能权重的互信息特征选择算法 PWMI.根据电子鼻系统特性,设计基于传感器特性的评价函数:

$$\begin{aligned} J(X_m) &= J_{\text{dsg}}(X_m) - J_{\text{rdu}}(X_m) + J_{\text{sen}}(X_m) \\ &= \frac{1}{|K|} \sum_{L_k \in K} \lambda_k \{I(X_m; L_m | L_k) - I(X_m; L_m; L_k)\} - \frac{1}{N} \sum_{X_n \in N} I(X_m; X_n; L_m) \\ &\quad + \alpha E(X_m) \end{aligned} \quad (6)$$

公式(6)中,对特征性能的评价函数由区分性评价函数 $J_{\text{dsg}}(X_m)$ 、冗余性评价函数 $J_{\text{rdu}}(X_m)$ 、敏感性评价函数 $J_{\text{sen}}(X_m)$ 构成.各子评价函数的推导和分析将在接下来的两节中进行详细分析.

3.2 特征区分性

在电子鼻的实际应用中,识别的目标气体往往并不是单一的成分,而是含有多种成分构成的复杂

气体,那么不同气体之间不可避免会存在相同的气体成分. 如果对气体化学成分进行分析并度量其中的冗余信息,无疑是非常困难的. 因此算法从传感器对气体的区分性能入手,通过获取特征与对应和非对应特征标签的互信息构造评价函数,表示特征对不同标签分布的区分性能.

将特征对应标签全排列,可构造出包含所有可能的非对应标签向量 L_k 的分布集合 K ,其中部分向量和特征对应的标签向量 L_m 具有一定的相似性,且不同的向量相似性不同. 为了使得与 L_m 相似性更高的 L_k 得到更大的权重,将 L_k 中与 L_m 相同的标签数量与标签总数量的比值作为相似因子 λ_k ,用于特征区分性的权值. 非对应标签分布集合 K 以及相似因子集合 λ 如下:

$$K = \{L_1, L_2, \dots, L_k\} \quad (7)$$

$$\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_k\} \quad (8)$$

对于特征 X_m 而言,其区分性能优良的情况最直观的为该变量与对应的标签变量互信息最大于与其他标签的互信息:

$$I(X_m; L_m) > I(X_m; L_k) \quad (9)$$

很显然,设计评价函数最直接的方式就是将这两者做差,差值越大则特征的区分性越强,如:

$$J_{\text{dsg}}(X_m) = I(X_m; L_m) - I(X_m; L_k) \quad (10)$$

单一的标签分布的区分性能并不能很好的衡量该特征的实际区分性能,因此需要将特征与集合 K 中所有的标签分布的互信息差值求均值. 在累加的过程中,特征、对应标签和非对应标签分布的联合互信息会被多次累加,所以在每次求差值的时候减去此联合分布,即可获得新的评价函数:

$$J_{\text{dsg}}(X_m) = \frac{1}{|K|} \sum_{L_k \in K} \{I(X_m; L_m) - I(X_m; L_k) - I(X_m; L_m; L_k)\} \quad (11)$$

然而,当特征与对应标签的相关性低于非对应标签的相关性时,特征的评价函数可能会出现负数,并因此影响到后续特征的冗余性、灵敏性的评价. 出现负互信息的主要原因是因为公式(11)中减去了无效的互信息. 重新分析特征与非对应标签的互信息可知,此互信息是由三者的联合互信息和条件互信息构成,即:

$$I(X_m; L_k) = I(X_m; L_m; L_k) + I(X_m; L_k | L_m) \quad (12)$$

其中特征对应标签的条件互信息 $I(X_m; L_k | L_m)$ 并不会影响特征与对应标签互信息的大小,且当 $I(X_m; L_k | L_m) > I(X_m; L_m | L_k)$ 时,就会使得区

分性评价函数为负,为了去除无效互信息 $I(X_m; L_k | L_m)$,通过联合互信息 $I(X_m; L_m; L_k)$ 代替特征与非对应标签的互信息 $I(X_m; L_k)$,可得新的区分性评价函数:

$$\begin{aligned} J_{\text{dsg}}(X_m) &= \frac{1}{|K|} \sum_{L_k \in K} \{I(X_m; L_m) - I(X_m; L_m; L_k) \\ &\quad - I(X_m; L_m; L_k)\} \\ &= \frac{1}{|K|} \sum_{L_k \in K} \{I(X_m; L_m | L_k) - I(X_m; L_m; L_k)\} \end{aligned} \quad (13)$$

最后将每一个标签 L_k 对应的相似因子 λ_k 作为权值带入公式,得到最终的区分性评价函数:

$$J_{\text{dsg}}(X_m) = \frac{1}{|K|} \sum_{L_k \in K} \lambda_k \{I(X_m; L_m | L_k) - I(X_m; L_m; L_k)\} \quad (14)$$

3.3 特征冗余性

器件的广谱响应特性能给模式识别算法带来丰富的标签信息的同时,也会使得不同的传感器对同一目标气体产生响应,即不同的特征变量间存在冗余信息.

据前文所述,特征 X_m 与标签 L_m 的全部互信息可表示为 $I(X_m; L_m)$,当另外的特征 X 对 L_m 同样有互信息时,原互信息降低为 $I(X_m; L_m | X)$. 因此通过总互信息减去有效互信息得到冗余互信息:

$$\begin{aligned} J_{\text{rdi}}(X_m) &= I(X_m; L_m) - I(X_m; L_m | X) \\ &= I(X_m; L_m; X) \end{aligned} \quad (15)$$

由于算法的目的是尽可能筛选出独立的特征子集 N ,故算法只关心集合 N 中特征之间的独立性,由此可得:

$$J_{\text{rdi}}(X_m) = \frac{1}{|N|} \sum_{X_n \in N} I(X_m; X_n; L_m) \quad (16)$$

3.4 特征敏感性

在理想的特征选择模型中,特征变量变化的幅度并不会对模式识别的结果造成影响. 然而在电子鼻系统的实际应用中,对目标气体响应幅度小、灵敏度低的传感器更容易受空气中非目标气体、电路热效应等噪声的影响,因此算法针对特征的方差设计敏感性评价函数.

在不同的环境下,传感器阵列受到干扰的影响大小是不同的,例如在实验室环境中传感器阵列受到的影响远小于户外、风洞等环境. 为了调节不同环境下特征敏感性占总评价函数的比重,需要设置惩罚因子 α 作为特征方差的系数. 在低环境干扰下降低特征敏感性的权重,反之则提高:

$$J_{\text{sen}}(X_m) = \alpha E(X_m)$$

(17)

3.5 算法流程与优势分析

本文提出的算法综合了电子鼻系统的特性,综合式(14)、式(16)和式(17)即可得到式(6)的评价函数.为了筛选出最优特征子集,算法采用逐步选取而非直接选取.即通过迭代逐步获取候选特征集合中每次得分最高的特征.

算法的详细流程如表 1 所示.

表 1 PWMI 算法流程

算法	PWMI 算法
输入	候选特征集 M 、标签变量 L ;
计算	最优特征子集 N ;
初始化	根据需要的特征数量确定迭代次数 n ;全排列标签 L 构造集合 K ;同时构造对应的相似因子 λ ;构造空集 N ;设置惩罚因子 α .
迭代开始	当迭代次数小于 n 时:
	1:开始遍历候选特征集 M ;
	2:取出特征 X_m ,通过评价函数公式 6 获取该特征的分
	数 R_m ;
	3:遍历完成;
	4:取出分数集合 R 中获得最高分的特征 $X_{\max}$;
输出	5:将特征 $X_{\max}$ 加入集合 N ,从候选特征集 M 中删除特征 $X_{\max}$;
	6:返回迭代开始部分;
	结束

综上所述,应用基于互信息的特征选择算法于电子鼻系统中,PWMI 算法具有以下优势:

- (1)设计基于电子鼻传感器特性的特征评价函数,充分衡量了特征的区分性、冗余性和敏感性.
- (2)算法通过逐步选取得分最高特征,更精确的计算特征子集之间的互信息.

4 实验验证与分析

4.1 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^[1].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^[7].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,同样将传感器的最大响应值作为候选特征.

4.2 实验设置

本文在上述两种电子鼻传感器阵列的数据集中,通过所提出算法和其他算法筛选出占传感器总数 30%至 60%的传感器子集,对筛选出的传感器子集统一使用支持向量机(Suport Vektor Machine, SVM)算法进行模式识别,对不同算法筛选出来子集达到的识别精度做对比.

算法中特征敏感性的惩罚因子是为了降低环境对传感器阵列的影响,以期获得更好的识别精度,因此在实验设置阶段需要探寻惩罚因子对最终识别精度的影响.图 1 和图 2 为在不同惩罚因子下,筛选出不同数量的特征子集获得的识别精度的三维网格图.

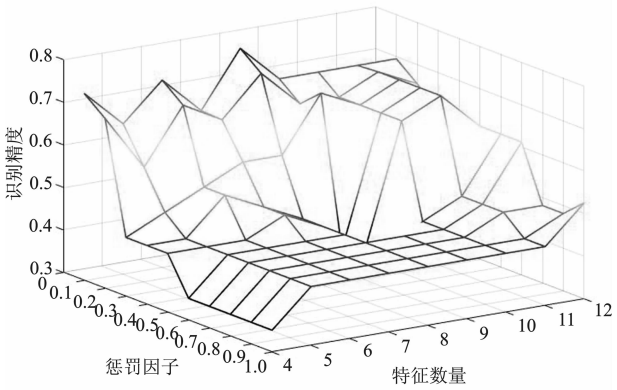


图 1 流量调制数据集惩罚因子对识别精度的影响

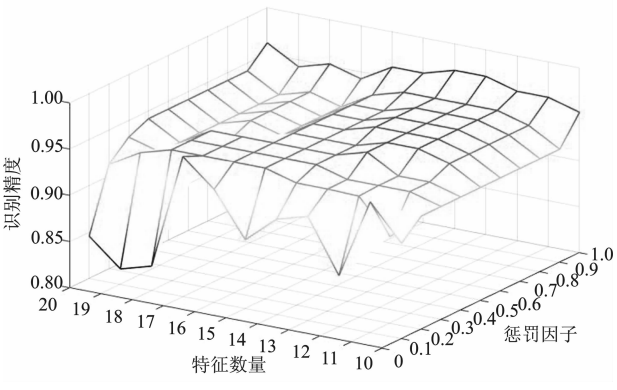


图 2 伤口细菌数据集惩罚因子对识别精度的影响

由图可知,选择合适的惩罚因子对识别精度有较大的影响,而一定范围内特征数量对精度的影响有限;在流量调制数据集中,惩罚因子在[0.1,0.5]区间中,各特征数量的识别精度显著下降,在其他区间中精度变化起伏不大;在伤口细菌数据集中,惩罚因子在[0.1,0.3]区间外时精度变化趋

图3和图4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,PWMI、CFR、MRMD算法筛选出来特征的数量增加的同时,识别的精度也在增加.说明此类算法筛选的特征之间低冗余,因此随着特征数量的增加,蕴含的数据信息越丰富,识别的精度也就越高.当特征子集达到一定数量时,蕴含的数据信息趋于饱和,则精度不再增加.而对于MIM、JMI等算法,筛选出的特征之间存在较大的冗余,影响了识别的精度.在伤口细菌的数据集中也是如此,三种能够筛选出低冗余特征子集的算法能获得较高的精度,剩余的算法的精度较低.

表2和表3列出了在两个数据集中不同算法筛选出特征子集的识别精度值.可以发现消除特征间冗余性能较强的算法中,PWMI算法通过衡量了特征对标签的区分性能,并通过敏感性评价函数过滤部分受到噪声影响的特征,筛选出的特征子集获得了更好的识别精度.

5 结束语

基于互信息的特征选择算法可以很好的衡量电子鼻传感器之间的冗余信息,为了将基于互信息的特征选择算法应用于电子鼻系统中,本文提出了基于性能权重的互信息特征选择算法(Performance Weight Mutal Information, PWMI).针对电子鼻系统特性,衡量特征对标签的区分性、特征与特征之间的冗余性、特征自身的敏感性三个方面,筛选出最优特征子集进行模式识别.实验表明,在电子鼻的相关数据集中,算法能够筛选出获得更高精度的传感器特征子集,有效的优化了基于互信息的特征选择算法在电子鼻传感器上的应用.由于传感器工作的复杂环境,算法还可以衡量更多相关的性能以获取更优的传感器子集,今后的工作将改进PWMI算法,将传感器对识别精度可能造成影响的性能如稳定性纳入计算范围,进一步提高算法筛选出特征子集的识别精度.

参考文献:

[1] ZIYATDINOV A, FONOLLOSA J, FERNÁNDEZ L,

et al. Bioinspired early detection through gas flow modulation in chemo-sensory systems[J]. *Sensors and Actuators B: Chemical*, 2015(206): 538-547. DOI: 10.1016/j.snb.2014.09.001.

- [2] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. *IEEE Transactions on Neural Networks*, 1994, 5(4): 537-550. DOI: 10.1109/72.298224.
- [3] YANG H H, MOODY J. Data visualization and feature selection: new algorithms for nongaussian data [C]//*Proceedings of the 12th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 1999: 687-693.
- [4] PENG H C, LONG F H, 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. DOI: 10.1109/TPAMI.2005.159.
- [5] GAO W F, HU L, ZHANG P, et al. Feature selection considering the composition of feature relevancy[J]. *Pattern Recognition Letters*, 2018(112): 70-74. DOI: 10.1016/j.patrec.2018.06.005.
- [6] GAO W F, HU L, ZHANG P. Feature redundancy term variation for mutual information-based feature selection[J]. *Applied Intelligence*, 2020(50): 1272-1288. DOI: 10.1007/s10489-019-01597-z.
- [7] SUN H, TIAN F C, LIANG Z F, 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. DOI: 10.1109/TIE.2017.2694353.

作者简介:

陶 洋 男,(1964-),博士后,教授,博士生导师.研究方向为机器学习、模式识别等.

刘翔宇 男,(1996-),硕士研究生.研究方向为机器嗅觉、模式识别.

梁志芳(通讯作者) 女,(1989-),博士,讲师.研究方向为机器嗅觉、模式识别、迁移学习等.

E-mail: liangzf@cqupt.edu.cn.