文章编号:1673-9469(2009)02-0082-04

基于主成分分析和人工神经网络的酒类辨识

王英臣
(河北工程大学 科信学院,河北 邯郸 056038)

摘要:介绍了人工嗅觉系统对不同酒类样本的定性识别,尝试利用主成分分析(PCA)和人工神经网络(ANN)中改进的BP算法、改进的RBF算法(最近邻RBF与k均值RBF相结合选取中心的算法)和k均值RBF算法,实现对酒类的定性识别。实验结果表明:结合PCA的ANN方式为模式识别、分类提供了快速准确的辨识方法。

关键词:人工嗅觉系统;主成分分析;人工神经网络。
中图分类号:F224 **文献标识码**:A

The identification of wine based on PCA and ANN

WANG Ying-chen
(Kexin College, Hebei University of Engineering, Handan 056038, China)

Abstract: The qualitative identification of different wine through artificial electronic nose are introduced. PCA and ANN(including improved BP algorithm and RBF algorithm combining nearest neighbor - clustering algorithm and K - means clustering algorithm) are adopted to realize the identification. The test result indicated that the rapid and exact identification measure of ANN combining PCA is provided to the pattern identification.

Key words: artificial electronic nose; PCA; ANN

目前酒类辨识成为一个人工嗅觉系统研究的方向。酒是多种化学成份的混合物,这些化学物质可分为酸、酯、醛、醇等类型。决定酒质量的成份往往含量很低,但种类繁多。因此辨识酒类比辨识单一气体更具难度。H.V.Shurmer, J.W.Gardner[1-4]采用12种金属传感器对5种醇类进行了分类,取得较好的结果。M.Penza, G.Cassano通过ANN对意大利红酒进行辨识, María Luz Rodríguez - Méndez, Alvaro A. Arrieta, Vicente Parra等人通过3组传感器阵列的融合对红酒多模式辨识。

本文分别用BP和RBF对原始数据和主成分分析后的数据进行处理,均获得满意的分类效果。从而证明结合主成分分析的人工神经网络为模式辨识提供了一种有效方法,同时证明本文提出的改进的RBF算法比k均值RBF算法优越。

1 辨识方法

1.1 主成分分析(PCA)

主成分分析是利用降维的思想,把多指标转化为少数几个综合指标的多元统计分析方法。数据矩阵,包含指定的实验数据。主成分 PC_k 的表达形式如下:

$$PC_k = a_{1k}X_{1j} + a_{2k}X_{2j} + a_{3k}X_{3j} \quad (1)$$

其中 k 为主成分序号,n 为数据的维数,a_{ik} 为系数。在PCA学习中,对数据原始矩阵标准化后数据的相关矩阵[5]进行分析,得到主成分。主成分分析的主要特点是通过数据坐标轴的变化得到新的图示-得分图和负荷图。得分图经常被用来研究数据聚类、分类。负荷图被用来表示传感器和每个主成分的相对重要性和它们的相关性。

1.2 模式识别算法

人工神经网络(ANN)具有模拟人的部分形象

收稿日期:2008-10-06
作者简介:王英臣(1980-),男,河北邯郸人,硕士研究生,从事电子信息与检测技术方向的研究。

的能力,已经广泛应用于模式识别中,主要算法有BP、RBF 等。

反向传播学习算法:对反向传播(BP - Back Propagation)学习算法大量的研究产生了很多形式的修正和推广。文中为避免出现局部最小,算法采用了限幅函数法和加入 γ_1^{p1} 因子的方法。

径向基函数神经网络:径向基函数神经网络(RBF - Radial Basis Function),在选取 RBF 基函数中心时采用了最近邻聚类学习算法和 k - 均值聚类算法[6]相结合的方法。最近邻聚类学习算法是一种在线自适应聚类学习算法,不需要事先确定隐单元的个数,完成聚类所得到的 RBF 网络是最优的,并且此算法可以在线学习。

根据此算法构成的 RBF 网络是结合了最近邻聚类算法和 k - 均值聚类算法。半径 r 的大小决定了动态自适应 RBF 网络的复杂程度。r 越小,所得到的聚类数目就越多,计算量也越大。但由于 r 是一个一维参数,通常可以通过实验和误差信息找到一个适当的 r,这比同时确定隐单元的个数和一个合适的范数要方便的多。由于每一个输入 - 输出数据对都可能产生一个新的聚类,因此这种动态自适应 RBF 网络,实际上同时在进行参数和结构两个过程的自适应调整。

2 实验及结果分析

2.1 实验系统

实验采用了目前流行的半导体传感器组成传感器阵列,组成人工嗅觉系统前端的感应部分。传感器阵列包含了费加罗公司的 4 种传感器,NEMETO 公司的 2 种传感器并另加入了霍尼韦尔温度传感器和湿度传感器,一共 8 个传感器。传感器阵列的输出信号,经过模数转换器(ADAM - 4017)转化为数字信号。因为 ADAM - 4017 采用 RS485 协议,要连接到计算机串口上并传输数据就必须经过 RS485 - RS232 转换器,这样就可以在计算机串口上得到传感器阵列的输出信号。实验样本为一种白酒(太白酒五十度),一种葡萄酒(威龙天然红葡萄酒)和一种啤酒(蓝马澳麦)。

实验中对每种酒取 20 组样本,取 100% 原液,每隔 5% 浓度配成 20 种不同浓度的样本,3 种酒共 60 组样本。实验时每组样本测量 120s,对串口每 2s 采样一次,共得 60 个测量值,计算其平均值作为样本数据值。最终的数据每种酒取 12 个样本作为训练样本,8 个样本作为测试样本。训练样本集是 36 个样本,测试样本集是 24 个样本。

2.2 原始数据的主成分分析

传感器阵列输出的信号,组成实验样本矩阵 ,对 进行主成分分析。表 1 给出了特征值和主成分的贡献率和累积贡献率。第一主成分占到所有信息量的 96.601%,可见它的重要性,前两个主成分累积占到所有信息量的 99.877%。由特征值比较法确定了主成分适合的个数为 2,即 2 个主成分即可取代原来 8 个变量来进行模式识别。

表 1 PCA 分析后各主成分特征值和贡献率

Tab.1 Eigenvalue and percentage contribution by the PCA analysis performed

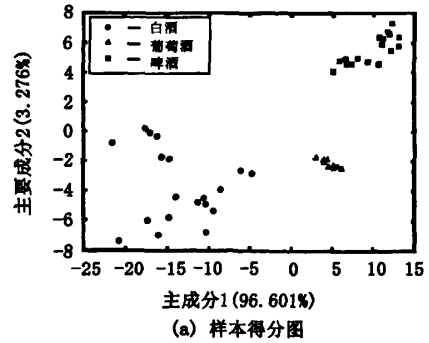
主成分	特征值	贡献率(%)	累积贡献率(%)
PC1	7.7281	96.601	96.601
PC2	0.2621	3.276	99.877
PC3	0.0060	0.075	99.952
PC4	0.0020	0.025	99.977
PC5	0.0009	0.011	99.988
PC6	0.0006	0.007	99.995
PC7	0.0003	0.003	99.998
PC8	0.0001	0.002	100

图 1 是 3 种样本在 PC1 - PC2 得分图和负荷图,从图(a)中可以看出,白酒和啤酒这 2 种样本可以明显地分开。负荷图中相似的负荷代表冗余度,因为在图(b)中没有相似的负荷,所以在原始数据中没有冗余的信息,也就是传感器阵列中没有冗余信息的传感器。

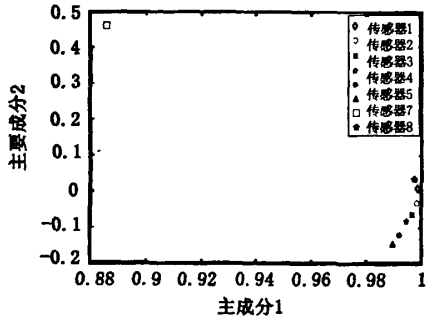
2.3 数据的 ANN 分析

在网络结构上,分别以原始数据(8 个传感器预处理[3]后信号)和 2 个主成分作为输入,对于 BP 算法,确定了对于原始数据为输入单元的网络结构为 8 - 12 - 3,2 个主成分作为输入单元的网络结构为 2 - 12 - 3。k 均值 RBF 中隐层单元数为 31,而改进的 RBF 中隐层单元个数为自适应调整,其他的均于 BP 相同。输出单元为 3 个,设定其输出为[1 0 0]为白酒,[0 1 0]为葡萄酒,[0 0 1]为啤酒。

图 2 和图 3 为上述 2 种结构构成的 ANN 算法的误差和训练次数。算法是在 CPU 为 P42.4、内存 512 M 以 VC + + 实现。在 BP 和 RBF 两种算法中,最终误差均设为 0.01,BP 算法中学习率是调整的,初始时输出层和隐层学习率是 0.8,隐层和输



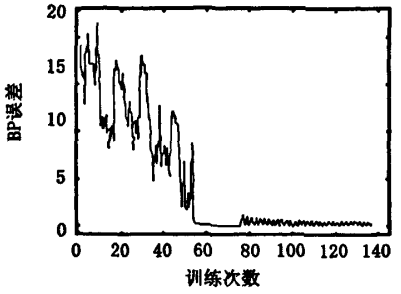
(a) 样本得分图



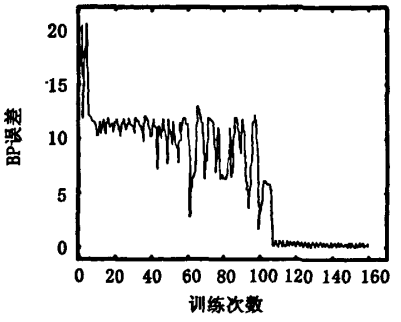
(b) 样本负荷图

图1 得分图和负荷图

Fig. 1 Scores plot and loads plot



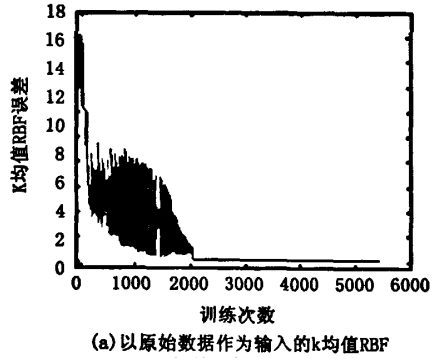
(a) 以原始数据作为输入



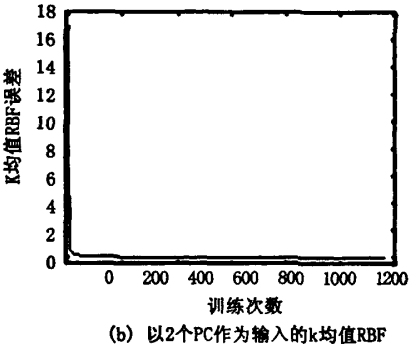
(b) 以2个PC作为输入

图2 BP算法的误差和训练次数

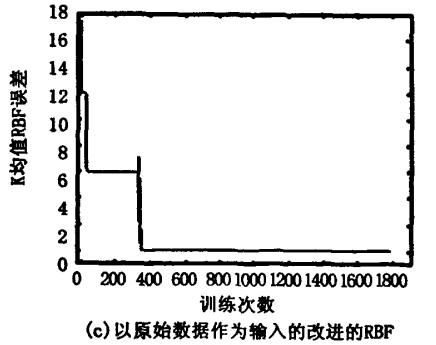
Fig. 2 Train step and err of BP algorithm



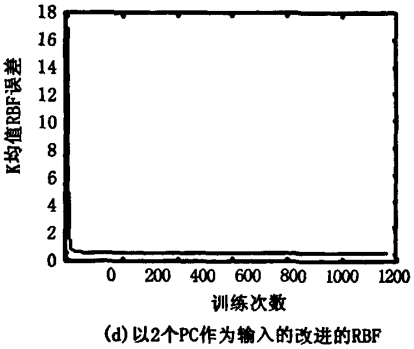
(a) 以原始数据作为输入的k均值RBF训练次数和误差



(b) 以2个PC作为输入的k均值RBF训练次数和误差



(c) 以原始数据作为输入的改进的RBF训练次和误差



(d) 以2个PC作为输入的改进的RBF训练次数和误差

图3 RBF算法的误差和训练次数

Fig. 3 Train step and err of RBF algorithm

表2 4种方式的分类正确率

Tab.2 Identification accurate rate of four types

序号	样本(个数)	预测结果					
		BP	PCA－BP	改进RBF	PCA－改进RBF	K均值RBF	PCA－K均值RBF
1	白　酒(8)	白　酒(8)	白　酒(8)	白　酒(8)	白　酒(8)	白　酒(8)	白　酒(8) 葡萄酒(6)
2	葡萄酒(8)	葡萄酒(8)	葡萄酒(7) 啤　酒(1)	葡萄酒(8)	葡萄酒(7) 白　酒(1)	葡萄酒(8)	白　酒(1) 未辨识(1)
3	啤　酒(8)	啤　酒(8)	啤　酒8)	啤　酒(8)	啤　酒(5) 未辨识(3)	啤　酒(4) 葡萄酒(4)	啤酒(6) 未辨识(2)
预测结果(%)		100	95.38	100	83.33	83.33	83.33

入层学习率是0.7,当误差小于2时,二者分别为0.3和0.2。改进RBF算法中,通过调r的大小,比较误差和训练次数,将隐层单元数确定为27个,最终的辨识结果见表2。

2.4 实验辨识结果

表2中第2列是测试样本种类,括号中是样本个数,其他各列为4种方式的辨识结果,括号中是辨识结果的样本个数。表2说明BP和改进RBF算法对原始数据的分类正确率为100%,而k均值RBF算法的分类正确率为83.33%。经过PCA分析后BP算法对于由PC1和PC2作为输入的样本的分类正确率为95.83%,而2种RBF算法分类正确率为83.33%。

由PC1和PC2构成的输入样本经过BP算法训练,可以很好的辨识测试样本,只有1个误判,即将1个葡萄酒样本误判为啤酒。经本文提出的改进RBF算法训练,由PC1和PC2构成的输入样本训练共有4个误判,1个葡萄酒样本误判为白酒,3个啤酒样本未辨识出结果。而对于k均值RBF算法,由PC1和PC2构成的输入样本训练共有4个误判,1个葡萄酒样本误判为白酒,1个葡萄酒样本未辨识出结果,2个啤酒样本未辨识出结果。

图2中,对于BP算法,达到相同的训练误差,由PC1和PC2构成的输入样本训练次数比由原始样本构成的输入样本的训练次数多,并产生一个误判。图3中,对于2种RBF算法,由PC1和PC2构成的输入样本训练次数比由原始样本构成的输入样本训练次数少,可以迅速减小误差,但是结果误判率却上升。由图(a)、(b)可知,k均值RBF算法比本文的改进RBF算法训练次数多2倍,并且前者的网络结构也更复杂,隐层单元数比后者的要多5个,表明了后者具有学习时间短,计算量小和网络性能优良的优点。

3 结论

1)通过传感器阵列对于3种酒辨识的实验,以8个传感器形成的8维变量作为原始数据,ANN的2种算法(BP、改进RBF)对它们的辨识率是100%,而k均值RBF的辨识率是83.33%。

2)在原始数据变量8维的基础上,经过PCA分析后,可以将维数减小到2维,将这2维变量作为输入,ANN的BP算法对于样本的分类正确率可以达到95.83%,RBF算法分类正确率可以达到83.33%。

参考文献:

[1] SHURMER H V, GARDNER J W. Intelligent vapour discrimination using a composite 12－element sensor array[J]. Sensors and Actuators, 1990, (1):256－260.

[2] SHURMER H V, GARDNER J W. The application of discrimination techniques to alcohols and tobaccos using tin oxide sensor[J]. Sensors and Actuators, 1989, 18:359－369.

[3] GARDNER J W, HINES E L, TANG H C. Detection of vapours and odours from a multisensor array using pattern recognition[J]. Sensors and Actuators B, 1991, (4): 109－115.

[4] SHURMER H V, GARDNER J W. Odours discrimination with an electronic nose[J]. Sensors and Actuators, 1992, (8):1－11.

[5] 范　杰. 主成分分析法的数值实现算法[J].河北工程大学学报(自然科学版), 2007, 24(4): 103－105..

[6] 李清政, 钟建伟. 基于神经网络法的配电网状态估计[J] .河北建筑科技学院学报,2006, 23(3):83－86.

(责任编辑　闫纯有)