



(12)发明专利申请

(10)申请公布号 CN 106295708 A

(43)申请公布日 2017.01.04

(21)申请号 201610686502.3

(22)申请日 2016.08.19

(71)申请人 重庆大学

地址 400044 重庆市沙坪坝区沙正街174号

(72)发明人 刘涛 李东琦 崔兴瑞 陈艳兵
武萌雅

(74)专利代理机构 北京国坤专利代理事务所
(普通合伙) 11491

代理人 姜彦

(51)Int.Cl.

G06K 9/62(2006.01)

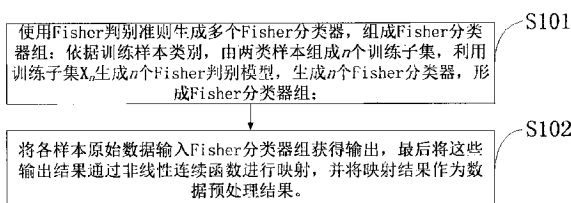
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于Fisher分类器组的连续型数据预处理方法

(57)摘要

本发明公开了一种基于Fisher分类器组的连续型数据预处理方法,使用Fisher判别准则生成多个Fisher分类器,组成Fisher分类器组;将各样本原始数据输入Fisher分类器组获得输出,最后将这些输出结果通过非线性连续函数进行映射,并将映射结果作为数据预处理结果。用Fisher分类器组的输出获得一定的样本分类冗余信息,随后在“非线性连续函数”映射过程中减小奇异值和野值对后续分类器的影响,最终可在不改变分类器性能的情况下,提高分类器的准确度;本发明提出的方法正确识别率为96.83%。



1.一种基于Fisher分类器组的连续型数据预处理方法,其特征在于,该基于Fisher分类器组的连续型数据预处理方法为:

使用Fisher判别准则生成n个Fisher分类器,组成Fisher分类器组:依据训练样本类别,由两类样本组成n个训练子集,利用训练子集 X_n 生成n个Fisher判别模型,生成n个Fisher分类器,形成Fisher分类器组;

将各样本原始数据输入Fisher分类器组获得输出,最后将输出结果通过非线性连续函数进行映射,并将映射结果作为数据预处理结果。

2.如权利要求1所述的基于Fisher分类器组的连续型数据预处理方法,其特征在于,该基于Fisher分类器组的连续型数据预处理方法包括以下步骤:

步骤一、训练子集选择与生成:通过获得若干组观察数据与所属类别的信息作为算法模型建立的依据,每一条信息称为一个训练样本,若干训练样本组成训练集;若训练样本有k类, $k \geq 2$;则依据训练样本类别,由两类样本组成 $\frac{k(k-1)}{2}$ 个训练子集,训练子集 X_n 表示为:

$$X_n = \{\{x_i\}, \{x_j\}\};$$

其中, $n \in \{1, 2, \dots, \frac{k(k-1)}{2}\}$, $i, j \in \{1, 2, \dots, n\}$ 且 $i \neq j$, $\{x_i\}$ 和 $\{x_j\}$ 分别表示训练集中第i和第j类样本的集合;

步骤二、Fisher分类器组生成:

利用训练子集 X_n 生成Fisher判别模型 $y_n = f_n(x)$,步骤如下:

1)求 X_n 中i, j两类样本的均值 $\bar{x}_i$ 和 $\bar{x}_j$;

2)求类内散度矩阵 S_{wn} :

$$S_{wn} = \sum_{x \in \{x_i\}} (x - \bar{x}_i)(x - \bar{x}_i)^T + \sum_{x \in \{x_j\}} (x - \bar{x}_j)(x - \bar{x}_j)^T;$$

其中 $(x - \bar{x}_i)^T$ 是 $(x - \bar{x}_i)$ 的转置矩阵;

3)求类间散度矩阵 S_{bn} :

$$S_{bn} = [\bar{x}_i - \bar{x}_j];$$

4)求投影方向 W_n :

$$W_n = S_{wn}^{-1} \cdot S_{bn};$$

5)求Fisher判别阈值 w_{0n} :

$$w_{0n} = W_n \cdot (\bar{x}_i + \bar{x}_j);$$

则得训练子集 X_n 对应的判别模型: $y_n = f_n(x) = W_n \cdot x - w_{0n}$;

6)按照步骤1)至步骤5)的方法求出每个训练子集对应的Fisher判别模型,生成 $\frac{k(k-1)}{2}$ 个Fisher分类器,形成Fisher分类器组,则分类器组输出 $\tilde{y}$ 表示为:

$$\tilde{y} = \{y_1, y_2, \dots, y_n\} = \left\{ f_1(x), \dots, f_{\frac{k(k-1)}{2}}(x) \right\};$$

步骤三、非线性连续型函数映射方法包括:

利用非线性连续函数对Fisher分类器组输出 $\tilde{\mathbf{y}}$ 进行映射,令 $\mathbf{y}_n^* = \varphi(\mathbf{y}_n)$ 为第n个Fisher分类器输出的非线性映射且:

$$\mathbf{y}_n^* = \varphi(\mathbf{y}_n) = e^{-\left(\frac{\mathbf{y}_n}{2a \cdot w_{0n}}\right)^2};$$

其中a是为增强算法泛化性能而引入的松弛变量, $a > 0$;若Fisher分类器组由k个分类器组成,则 $\mathbf{Y} = \{\mathbf{y}_1^*, \mathbf{y}_2^*, \dots, \mathbf{y}_k^*\}$ 为数据预处理结果。

3.如权利要求1所述的基于Fisher分类器组的连续型数据预处理方法,其特征在于,使用人工嗅觉系统对测试物质,进行分类识别,人工嗅觉系统由32个气敏传感器组成,每种测试物质进行12次采集,每次采集过程中清洁空气即基线采集时间为3分钟,被测物进样时间为3分钟,清洗时间为4分钟;

每次采集完成后,将第i个传感器的响应记为 Δr_i :

$$\Delta r_i = r_i^{gas} - r_i^{base};$$

其中 r_i^{gas} 为被测物进样阶段第i个传感器响应的平均值, r_i^{base} 为基线采集阶段第i个传感器响应的平均值,则每次采集获得32维样本,对于每种物质的样本,取其中3个作为训练样本,剩余为测试样本,松弛变量取 $a = 5.76$ 。

一种基于Fisher分类器组的连续型数据预处理方法

技术领域

[0001] 本发明属于数据处理技术领域,尤其涉及一种基于Fisher分类器组的连续型数据预处理方法。

背景技术

[0002] 随着人工智能技术的不断发展,模拟人类嗅觉感知过程的人工嗅觉系统应运而生,该系统由“气体传感器阵列”和“模式识别”两大部分构成。其中,“气体传感器阵列”由多种具有广谱响应的气体传感器组成,对各种气体产生不同的“响应图谱”;“模式识别”部分则利用合适的数学模型对“响应图谱”进行判别,为保证识别结果的准确性与鲁棒性,通常需要在图谱数据进入“模式识别”方法之前进行“数据预处理”。

[0003] 当前的人工嗅觉系统“数据预处理”方法主要包括主成分分析(PCA)、独立分量分析(ICA)、时频变换、归一化处理等。PCA以多维信号相关性作为判断准则,在降维过程中尽可能多地保留有用信息;ICA则以信号独立性为判断依据,对信号进行分解与重组,但存在信号分解过程中会损失原始信号中的幅度信息;时频变换则适用于有效信号与干扰信号频率特性明显相异的情况;归一化处理着重于解决信号幅度差异过大对模式识别方法造成的影响。

[0004] 然而,以上数据预处理方法的处理原则并不以最优分类为目标,即预处理过程中可能将有益于分类的信息剔除而保留某些对分类无用的信息,进而导致:对于同一分类器,往往数据经过预处理后的分类准确率弱于预处理前。因此需要一种以分类准确率为导向的数据预处理方法,保证在不增加分类器复杂度的情况下,获得较好的分类准确率。

发明内容

[0005] 本发明的目的在于提供一种基于Fisher分类器组的连续型数据预处理方法,旨在解决背景技术中提及的问题。

[0006] 本发明提供一种基于Fisher分类器的人工嗅觉系统数据预处理方法,该基于Fisher分类器的人工嗅觉系统数据预处理方法为:

[0007] 使用Fisher判别准则生成多个Fisher分类器,组成Fisher分类器组:依据训练样本类别,由两类样本组成n个训练子集,利用训练子集 X_n 生成n个Fisher判别模型,生成n个Fisher分类器,形成Fisher分类器组;

[0008] 将各样本原始数据输入Fisher分类器组获得输出,最后将这些输出结果通过非线性连续函数进行映射,并将映射结果作为数据预处理结果。

[0009] 该基于Fisher分类器组的连续型数据预处理方法包括以下步骤:

[0010] 步骤一、训练子集选择与生成:通过获得若干组观察数据与所属类别的信息作为算法模型建立的依据,每一条信息称为一个训练样本,若干训练样本组成训练集;若训练样本有k类, $k \geq 2$;则依据训练样本类别,由两类样本组成

$\frac{k(k-1)}{2}$ 个训练子集,训练子集 X_n 表示

为:

[0011] $X_n = \{\{x_i\}, \{x_j\}\};$

[0012] 其中, $n \in \left\{1, 2, \dots, \frac{k(k-1)}{2}\right\}, i, j \in \{1, 2, \dots, n\}$ 且 $i \neq j, \{x_i\}$ 和 $\{x_j\}$ 分别表示训练集中

第i和第j类样本的集合;

[0013] 步骤二、Fisher分类器组生成:

[0014] 利用训练子集 X_n 生成Fisher判别模型 $y_n = f_n(x)$, 步骤如下:

[0015] 1)求 X_n 中 i, j 两类样本的均值 $\bar{x}_i$ 和 $\bar{x}_j$;

[0016] 2)求类内散度矩阵 S_{wn} :

[0017] $S_{wn} = \sum_{x \in \{x_i\}} (x - \bar{x}_i)(x - \bar{x}_i)^T + \sum_{x \in \{x_j\}} (x - \bar{x}_j)(x - \bar{x}_j)^T;$

[0018] 其中 $(x - \bar{x}_i)^T$ 是 $(x - \bar{x}_i)$ 的转置矩阵;

[0019] 3)求类间散度矩阵 S_{bn} :

[0020] $S_{bn} = [\bar{x}_i - \bar{x}_j];$

[0021] 4)求投影方向 W_n :

[0022] $W_n = S_{wn}^{-1} \cdot S_{bn};$

[0023] 5)求Fisher判别阈值 w_{0n} :

[0024] $w_{0n} = W_n \cdot (\bar{x}_i + \bar{x}_j);$

[0025] 则得训练子集 X_n 对应的判别模型: $y_n = f_n(x) = W_n \cdot x - w_{0n};$

[0026] 6)按照步骤1)至步骤5)的方法求出每个训练子集对应的Fisher判别模型,生成 $\frac{k(k-1)}{2}$ 个Fisher分类器,形成Fisher分类器组,则分类器组输出 $\tilde{y}$ 可表示为:

[0027] $\tilde{y} = \{y_1, y_2, \dots, y_n\} = \left\{f_1(x), \dots, f_{\frac{k(k-1)}{2}}(x)\right\};$

[0028] 步骤三、非线性连续型函数映射方法包括:

[0029] 利用非线性连续函数对Fisher分类器组输出 $\tilde{y}$ 进行映射,令 $y_n^* = \varphi(y_n)$ 为第n个Fisher分类器输出的非线性映射且:

[0030] $y_n^* = \varphi(y_n) = e^{-\left(\frac{y_n}{2a \cdot w_{0n}}\right)^2};$

[0031] 其中 $a(a > 0)$ 是为增强算法泛化性能而引入的松弛变量;若Fisher分类器组由k个分类器组成,则 $Y = \{y_1^*, y_2^*, \dots, y_k^*\}$ 为数据预处理结果。

[0032] 进一步,使用人工嗅觉系统对测试物质,进行分类识别,人工嗅觉系统由32个气敏传感器组成,每种测试物质进行12次采集,每次采集过程中清洁空气即基线采集时间为3分钟,被测物进样时间为3分钟,清洗时间为4分钟;

[0033] 每次采集完成后,将第i个传感器的响应记为 Δr_i :

[0034] $\Delta r_i = r_i^{gas} - r_i^{base};$

[0035] 其中 r_i^{gas} 为被测物进样阶段第i个传感器响应的平均值, r_i^{base} 为基线采集阶段第i个传感器响应的平均值,则每次采集可获得一个32维的样本,对于每种物质的样本,取其中3个作为训练样本,剩余为测试样本,松弛变量取 $a=5.76$ 。

[0036] 本发明所涉方法优势在于:

[0037] 1、相较于其它数据预处理方法,利用Fisher分类器组将样本的类别信息作为先验知识,使预处理后的数据更易被正确分类;

[0038] 2、利用多个简单的二分类器进行数据映射,为分类器在多分类场景下提供更多有用信息;

[0039] 3、“非线性连续函数”映射可减小连续空间中奇异值和野值对后续分类器的影响;综上,本发明所涉方法可提高分类器在多分类识别中的正确率。

[0040] 现有技术中,数据归一化后,不采用任何预处理方法正确识别率为92.06%,使用PCA方法并取第1和第2主成分作为预处理结果正确识别率为50.79%;本发明提出的方法正确识别率为96.83%。

附图说明

[0041] 图1是本发明实施例提供的基于Fisher分类器组的连续型数据预处理方法流程图。

具体实施方式

[0042] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0043] 下面结合附图对本发明的应用原理作详细的描述。

[0044] 如图1所示:本发明实施例的基于Fisher分类器组的连续型数据预处理方法包括以下步骤:

[0045] S101:使用Fisher判别准则生成多个Fisher分类器,组成Fisher分类器组:依据训练样本类别,由两类样本组成n个训练子集,利用训练子集 X_n 生成n个Fisher判别模型,生成n个Fisher分类器,形成Fisher分类器组;

[0046] S102:将各样本原始数据输入Fisher分类器组获得输出,最后将这些输出结果通过非线性连续函数进行映射,并将映射结果作为数据预处理结果。

[0047] 该基于Fisher分类器组的连续型数据预处理方法具体包括以下步骤:

[0048] 步骤一、训练子集选择与生成:通过获得若干组观察数据与所属类别的信息作为算法模型建立的依据,每一条信息称为一个训练样本,若干训练样本组成训练集;若训练样本有k类, $k \geq 2$;则依据训练样本类别,由两类样本组成 $\frac{k(k-1)}{2}$ 个训练子集,训练子集 X_n 表示为:

[0049] $X_n = \{\{x_i\}, \{x_j\}\};$

[0050] 其中, $n \in \left\{1, 2, \dots, \frac{k(k-1)}{2}\right\}, i, j \in \{1, 2, \dots, n\}$ 且 $i \neq j, \{x_i\}$ 和 $\{x_j\}$ 分别表示训练集中

第i和第j类样本的集合；

[0051] 步骤二、Fisher分类器组生成：

[0052] 利用训练子集 X_n 生成Fisher判别模型 $y_n=f_n(x)$ ，步骤如下：

[0053] 1)求 X_n 中i,j两类样本的均值 $\bar{\mathbf{x}}_i$ 和 $\bar{\mathbf{x}}_j$ ；

[0054] 2)求类内散度矩阵 S_{wn} ：

$$[0055] \quad \mathbf{S}_{wn} = \sum_{\mathbf{x} \in \{\mathbf{x}_i\}} (\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)^T + \sum_{\mathbf{x} \in \{\mathbf{x}_j\}} (\mathbf{x} - \bar{\mathbf{x}}_j)(\mathbf{x} - \bar{\mathbf{x}}_j)^T ;$$

[0056] 其中 $(\mathbf{x} - \bar{\mathbf{x}}_i)^T$ 是 $(\mathbf{x} - \bar{\mathbf{x}}_i)$ 的转置矩阵；

[0057] 3)求类间散度矩阵 S_{bn} ：

$$[0058] \quad \mathbf{S}_{bn} = [\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j] ;$$

[0059] 4)求投影方向 $\mathbf{W}_n$ ：

$$[0060] \quad \mathbf{W}_n = \mathbf{S}_{wn}^{-1} \cdot \mathbf{S}_{bn} ;$$

[0061] 5)求Fisher判别阈值 w_{0n} ：

$$[0062] \quad w_{0n} = \mathbf{W}_n \cdot (\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j) ;$$

[0063] 则得训练子集 X_n 对应的判别模型： $y_n=f_n(x)=\mathbf{W}_n \cdot \mathbf{x}-w_{0n}$ ；

[0064] 6)按照步骤1)至步骤5)的方法求出每个训练子集对应的Fisher判别模型，生成 $\frac{k(k-1)}{2}$ 个Fisher分类器，形成Fisher分类器组，则分类器组输出 $\tilde{\mathbf{y}}$ 可表示为：

$$[0065] \quad \tilde{\mathbf{y}} = \{y_1, y_2, \dots, y_n\} = \left\{ f_1(\mathbf{x}), \dots, f_{\frac{k(k-1)}{2}}(\mathbf{x}) \right\} ;$$

[0066] 步骤三、非线性连续型函数映射方法包括：

[0067] 利用非线性连续函数对Fisher分类器组输出 $\tilde{\mathbf{y}}$ 进行映射，令 $\mathbf{y}_n^* = \varphi(\mathbf{y}_n)$ 为第n个Fisher分类器输出的非线性映射且：

$$[0068] \quad \mathbf{y}_n^* = \varphi(\mathbf{y}_n) = e^{-\left(\frac{y_n}{2a \cdot w_{0n}}\right)^2} ;$$

[0069] 其中a(a>0)是为增强算法泛化性能而引入的松弛变量；若Fisher分类器组由k个分类器组成，则 $\mathbf{Y} = \{y_1^*, y_2^*, \dots, y_k^*\}$ 为数据预处理结果。

[0070] 下面结合具体实施例对本发明的应用原理作进一步描述。

[0071] 实施例1：

[0072] 使用人工嗅觉系统对七种物质，包括：啤酒、白酒、葡萄酒、绿茶、红茶、乌龙茶和普洱茶，进行分类识别。人工嗅觉系统由32个气敏传感器组成，系统对传感器阵列响应的采样速率为1Hz，采样精度为16bit。

[0073] 每种测试物质进行12次采集，共计84次采集。每次采集过程中清洁空气(基线)采集时间为3分钟，被测物进样时间为3分钟，清洗时间为4分钟。

[0074] 每次采集完成后，将第i个传感器的响应记为 Δr_i ：

$$[0075] \quad \Delta r_i = r_i^{gas} - r_i^{base} ;$$

[0076] 其中 r_i^{gas} 为被测物进样阶段第i个传感器响应的平均值， r_i^{base} 为基线采集阶段第i个

传感器响应的平均值,则每次采集可获得一个32维的样本,样本共计84个。对于每种物质的样本,取其中3个作为训练样本,其余为测试样本。松弛变量取 $a=5.76$ 。

[0077] 数据预处理方式有3种:

[0078] ①不采用任何预处理方法;

[0079] ②使用PCA方法并取第1和第2主成分作为预处理结果;

[0080] ③本发明提出的方法;

[0081] 模式识别方法:k近邻法(k-NN)。

[0082]

数据预处理方法	识别正确率
无	92.06%
PCA方法	50.79%
本专利方法	96.83%

[0083] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,均应包含在本发明的保护范围之内。

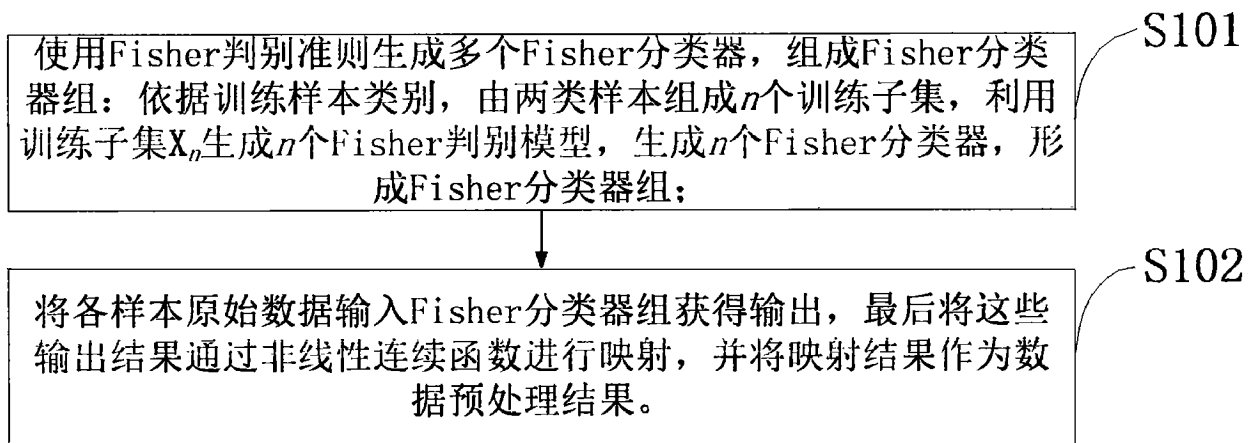


图1