Journal of Computer Applications
计算机应用,2012,32(3):881-884
ISSN 1001-9081
CODEN JYIIDU
2012-03-01
http://www.joca.cn

文章编号:1001-9081(2012)03-0881-04
doi:10.3724/SP.J.1087.2012.00881

支持向量数据描述在烟叶异物检测中的应用

黄仕建[1,2*]

(1.长江师范学院 物理学与电子工程学院,重庆 408100;
2.光电技术及系统教育部重点实验室(重庆大学),重庆 400044)
(*通信作者电子邮箱 huangshij520@126.com)

摘　要:针对烟叶异物检测中很难全面收集异物样本数据的问题,提出一种基于支持向量数据描述方法(SVDD)的烟叶异物检测方法。该方法只需要烟叶样本数据,就可建立单值分类器。首先,提取烟叶与几种典型异物的RGB分量与HSV分量;然后,选取烟叶的HV分量作为特征向量,训练SVDD分类器,实现烟叶异物的分类识别;最后,通过接受者操作特性(ROC)曲线对比了SVDD与其他3种方法的分类效果。实验结果表明,采用HV分量降低了数据维数,提高了计算效率;SVDD方法具有很好的分类效果和计算效率,能很好地区分烟叶与异物。

关键词:支持向量数据描述;异物检测;烟叶样本;HV分量;分类识别

中图分类号:TP274.3　　**文献标志码**:A

Application of support vector data description to detection of foreign bodies in tobacco

HUANG Shi-jian[1,2*]

(1. *School of Physics and Electron Engineering, Yangtze Normal University, Chongqing* 408100, *China*;
2. *Key Laboratory of Optoelectronic Technology and Systems (Chongqing University), Ministry of Education, Chongqing* 400044, *China*)

Abstract: It is difficult to fully collect foreign body sample in detecting foreign bodies from tobacco. A detection method based on Support Vector Data Description (SVDD) was proposed. Thus a one-class classifier can be developed by using tobacco samples soly. RGB and HSV of tobacco and several typical foreign bodies were firstly extracted; then the HV component was used as eigenvector. A developed SVDD classifier was applied to distinguish foreign bodies from tobacco by inputting the HV eigenvector. Finally through the Receiver Operating Characteristic (ROC) curve, the SVDD classifier was compared with three other methods in classification effect. The experimental results show that by adopting feature extraction with HV component, the data dimension was reduced and a higher computation efficiency was achieved. The SVDD classifier has a stronger classification ability and higher efficiency, which could distinguish foreign bodies from tobacco better.

Key words: Support Vector Data Description (SVDD); foreign body detection; tobacco sample; HV component; classification

0　引言

国内大多数烟草生产线上都采用金属探测仪检测和剔除金属异物,并辅之以人工剔除其他异物的方法[1-2],这些方法的异物剔除率受人为因素影响较大。因此研究烟叶异物自动检测系统极为重要,自动检测的关键技术多采用机器视觉技术与模式识别技术相结合[3],其中分类方法的选择是否恰当,直接影响最后的检测结果。但近年来国内外针对烟叶异物的分类技术却鲜有报道,而常用的可以借鉴的分类方法有贝叶斯分类算法、支持向量机等传统方法和支持向量数据描述[4-6](Support Vector Data Description, SVDD)等单值分类算法。但贝叶斯分类等传统分类方法在对训练样本不平衡领域如烟叶异物检测中异物种类繁杂甚至无法预知的情况时表现得力不从心,本文提出将支持向量数据描述方法运用到烟叶异物检测技术中,该方法已在很多训练样本不平衡的领域(如机械故障诊断[7-8]、语音识别[9]、图像识别等领域)得到了成功应用[10-12]。运用该方法只需用烟叶的HV分量数据训练单值分类器,就可实现分类,可以解决难以提取异物训练样本的问题。

1　颜色空间

RGB颜色空间的基本原理是采用红(R)、绿(G)、蓝(B)3个颜色分量来表示所有的颜色。HSV颜色空间模型是孟塞尔色彩空间的简化形式,直接采用彩色特性意义的3个分量:色度(H)、饱和度(S)、亮度(V)来描述颜色,更符合人对颜色的描述习惯。

RGB颜色空间有不均匀和不直观的缺点,HSV颜色空间的三分量相对独立,易通过设定不同权值将其融合在一起,具有计算量小等优点。

2　支持向量数据描述算法简述

SVDD的基本思想是把要描述的对象作为一个整体。假定一个目标集(Target)包含有n个需要描述的目标对象$\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in R^d; i = 1,2,\cdots,n\}$,构成单值分类器的$n$个学习样本。试图找到一个体积最小的超球体,使全部(或尽可能多)的$\boldsymbol{x}_i$都包含在该超球体内,而非目标样本(Outliers)就位于超球体外,为了增强分类的鲁棒性,引入松弛变量ξ_i。最小化超球体的体积是一个二次规划问题,即应满足:

收稿日期:2011-09-29;**修回日期**:2011-11-17。
作者简介:黄仕建(1983-),男,四川眉山人,助教,博士研究生,主要研究方向:模式识别、信息获取与处理。

$$\min f(R,\boldsymbol{a},\xi) = R^2 + C\sum_i^n \xi_i;\ i = 1,2,\cdots,n \tag{1}$$

约束条件：

$$\left\| \boldsymbol{x}_i - \boldsymbol{a} \right\|^2 \leqslant R^2 + \xi_i;\xi_i \geqslant 0$$

其中：$\boldsymbol{a}$ 为超球体球心；R 为超球体半径；C 为错分样本的惩罚系数，以实现在错分样本的比例和算法复杂程度之间的折中。上述问题可以转化为 Lagrange 极值问题：

$$L(R,\boldsymbol{a},\xi,\alpha,\gamma) = R^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i[R^2 + \xi_i - (\boldsymbol{x}_i \cdot \boldsymbol{x}_i - 2\boldsymbol{a} \cdot \boldsymbol{x}_i + \boldsymbol{a} \cdot \boldsymbol{a})] - \sum_{i=1}^n \gamma_i \xi_i \tag{2}$$

其中：$\alpha_i \geqslant 0$、$\gamma_i \geqslant 0$ 为 Lagrange 系数。对于每一个 $\boldsymbol{x}_i$，都有一个对应的 α_i 和 γ_i，经过变换，且用核函数代替内积，上述 Lagrange 优化目标函数可写为：

$$L(R,\boldsymbol{a},\xi,\alpha,\gamma) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_i) - \sum_{i=1,j=1}^n \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{3}$$

对其求最小值得出 α_i 的最优解 α_i^*。在实际计算中，多数的 α_i 将为 0，少部分 $\alpha_i \geqslant 0$，其不为 0 的 α_i 对应的样本称之为支持向量，只有这少部分的支持向量才决定了 $\boldsymbol{a}$ 和 R 的值，其他非支持向量因其对应的 $\alpha_i = 0$，在计算中将被忽略。因此这种方法的计算效率较高。半径 R 可由任一支持向量 $\boldsymbol{x}_k$ 按式(4) 求出：

$$R^2 = K(\boldsymbol{x}_k,\boldsymbol{x}_k) - 2\sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_k) + \sum_{i=1,j=1}^n \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{4}$$

对于一个新样本 z，判断它是否属于目标样本，有如下的判别函数：如果式(5) 成立，则 z 属于目标样本；否则为非目标样本。

$$R_z^2 = \| z - \boldsymbol{a} \|^2 = K(z,z) - 2\sum_{i=1}^n \alpha_i K(z,\boldsymbol{x}_i) + \sum_{i=1,j=1}^n \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \leqslant R^2 \tag{5}$$

3 实验研究

从人主观观察的角度看，烟叶异物在颜色、形状、大小、轻重、材质等方面都存在差异。但从机器视觉系统的角度来看，被检测物的轻重材质等特征难以获取，而形状、大小等参数又不具备明显的规律性，这使得颜色成了烟叶异物检测中最为重要的特征参量。

在烟叶异物检测中最重要的是识别出烟叶与非烟叶(即异物)，如再进一步识别出异物的具体类型费时费力且没有必要，因此考虑用单值分类方法对烟叶异物进行分类识别，把烟叶识为目标样本，将所有的异物都识为非目标样本。用烟叶的颜色特征数据训练 SVDD 单值分类器，即建立一个超球体紧紧包围住烟叶数据，再将烟叶异物混杂的特征数据输入训练好的分类器进行分类识别，把落在超球体内的样本判别为烟叶，把落在超球体外面的样本判别为异物。

在异物中选取橙纸、黑橡胶、灰纸箱、红纸、黄海绵、黄皮带、绿纸这几种典型异物作为非目标样本分析。在相同摄像条件下拍摄烟叶与异物图像，并对烟叶和异物各抽取一定数量的样本点，提取出相应的 RGB 分量，以及 HSV 分量。表 1 和表 2 分别为对烟叶异物的 RGB 与 HSV 各分量的均值方差统计。

表 1 烟叶和异物 RGB 各分量均值与方差统计

样本类别	均值			方差		
	R	G	B	R	G	B
烟叶	129.5182	49.3926	26.4015	17.0904	15.3265	13.8100
橙纸	242.9645	147.6333	49.1395	6.4347	6.7288	13.3758
黑橡胶	27.7864	18.9185	22.2494	5.0896	5.7976	12.8682
灰纸箱	133.3106	97.1953	55.5460	6.7092	7.4395	13.2898
红纸	181.1205	36.2183	37.0451	6.2400	6.5084	13.0252
黄海绵	254.0429	251.8454	128.9912	2.8066	5.0813	14.9703
黄皮带	126.1132	59.5856	25.3158	9.1526	8.4775	12.3641
绿纸	43.5210	110.2044	46.7324	5.1499	6.2135	13.5238

表 2 烟叶和异物 HSV 各分量均值与方差统计

样本类别	均值			方差		
	H	S	V	H	S	V
烟叶	0.0403	0.6294	0.3041	0.0174	0.0501	0.0689
橙纸	0.0845	0.7978	0.9528	0.0073	0.0546	0.0252
黑橡胶	0.4486	0.4796	0.1220	0.3884	0.1551	0.0351
灰纸箱	0.0880	0.5832	0.5228	0.0224	0.0980	0.0263
红纸	0.4328	0.8235	0.7103	0.4850	0.0385	0.0245
黄海绵	0.1639	0.4930	0.9976	0.0060	0.0582	0.0089
黄皮带	0.0582	0.7994	0.4946	0.0545	0.0965	0.0359
绿纸	0.3413	0.6361	0.4322	0.0318	0.0491	0.0244

在 HSV 三分量中色度 H 非常适合用来描述颜色特征，其识别效果一般较好，亮度 V 与色度 H 具有较高的独立性。从表 2 可见烟叶异物的 S 分量分布较广，不利于作为模式识别的特征参量。为了尽可能地提高烟叶异物分类识别的时效性，放弃了饱和度 S 分量，把色度 H 和亮度 V 结合起来组成二维的色度空间对烟叶与异物进行分类识别。

要想使分类器达到预定的目标，必须对其进行适当的训练。选取 100 组烟叶的 RGB 数据和 HV 数据作为训练样本

训练 SVDD 分类器。

为使分类器的训练结果和测试结果能够直观地展示出来,引入参数 $\lambda_k = R_k - R$。其中:R_k 表示第 k 个样本到超球体中心的距离,按式(5)求出;R 表示超球体的半径,按式(4)求出。若 $\lambda_k \leqslant 0$,则表示该样本属于烟叶;反之则属于异物。图 1 ~2分别为利用烟叶 RGB 数据与 HV 数据训练出的超球体分类器结果图,表 3 给出了两个超球体分类器的数据对比。

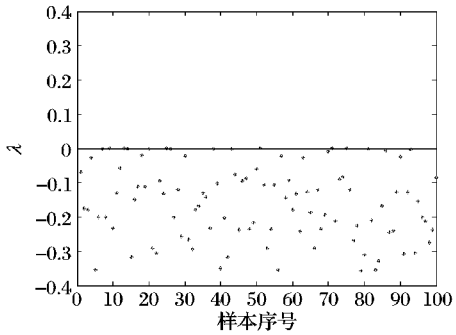

图 1　RGB 数据训练结果

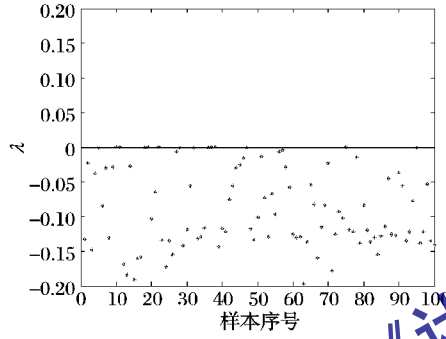

图 2　HV 数据训练结果

表 3　RGB 数据与 HV 数据训练对比

数据类型	$\bar{\lambda}$	超球体半径 R	训练用时/s
RGB	-0.1550	0.3624	1.8175
HV	-0.0869	0.2564	1.2743

从图 1 ~ 2 可以直观地看到所训练的两个超球体分类器都可以把全部烟叶包围在超球体内,但两个分类器还是存在差别,从表 3 可看出 HV 的烟叶样本 λ_k 的平均值 $\bar{\lambda}$ 绝对值更小、超球体半径更小(这意味着超球体把烟叶包裹得更紧凑,更利于把异物排除在超球体外),且训练用时更短。

为了检测两个分类器的分类性能,选取 45 组烟叶和 15 组异物(烟叶编号为 1 ~45,异物编号为 46 ~60)作为测试样本,并提取样本的 RGB 分量和 HV 分量作为测试样本集,分别对两个分类器进行测试。测试结果分别如图 3 ~4 所示,表 4对比了两个分类器的测试结果。

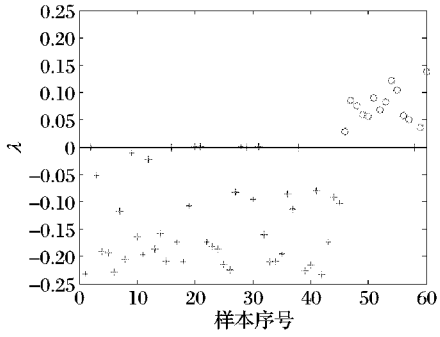

图 3　RGB 数据测试结果

从图 3 ~4 和表 4 可看出:两个 SVDD 分类器测试结果中,前者把 58 号异物误判为烟叶,而后者识别完全正确;后者的测试样本 λ_k 平均值 $\bar{\lambda}$ 之差更大,这说明用 HV 分量作为样本数据时,烟叶与异物在超球体分类器上的相对距离更远,这更有利于烟叶异物的分类识别;用 HV 数据训练的分类器,单个样本测试用时更少,提高了烟叶异物检测的效率。因此把烟叶异物的 HV 分量作为特征参量用于烟叶异物的分类识别更为合适。

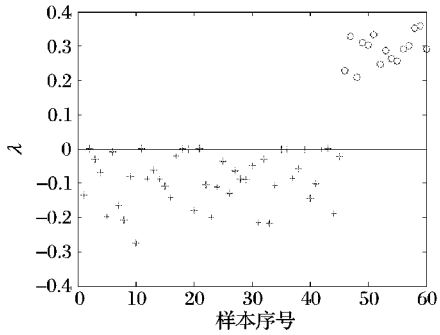

图 4　HV 数据测试结果

表 4　测试结果对比

数据类型	烟叶数量	异物数量	烟叶-异物 $\bar{\lambda}$ 差	单个样本测试用时/s
RGB	46	14	0.2017	0.0257
HV	45	15	0.3773	0.0163

将 SVDD 与另外 3 种分类方法包括自编码神经网络法(Autoenc)、K 中心法(K-center)和混合高斯模型法(Mog),进行分类性能比较。分别采用相同的训练集(即图 2 使用训练样本集)训练 4 种不同方法的分类器,并用相同的测试集(即图 4 使用测试样本集)对 4 个分类器进行测试。接受者操作特性曲线(Receiver Operating Characteristic,ROC)曲线描述了在不同参数控制下分类器的异物漏检率与烟叶识别率之间的关系,能很好地反映出分类器的分类性能,经计算各分类器的 ROC 曲线如图 5 所示;为了更准确定量地比较 4 个分类器的分类性能,计算每条 ROC 曲线下面积(Area Under the Curve,AUC),AUC 值分布在 0 ~1,其值越大表示分类器分类性能越好。4 种分类器的最终测试结果对比如表 5 所示。

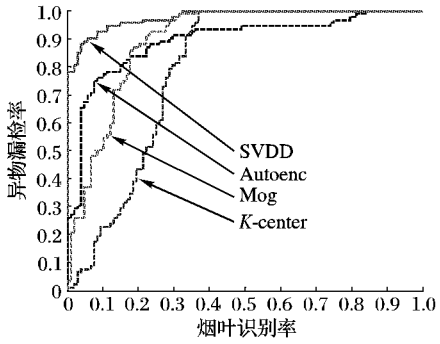

图 5　各分类器 ROC 曲线

表 5　4 种分类器测试结果对比

分类方法	烟叶数量	异物数量	AUC 值	单个样本测试用时/s
SVDD	45	15	0.9778	0.0163
Autoenc	44	16	0.8934	0.0357
K-center	49	11	0.7906	0.0187
Mog	46	14	0.8962	0.0233

从图 5 可看出 SVDD 的 ROC 曲线明显好于其他 3 种算法。从表 5 中可见,SVDD 的 AUC 值最大,且其测试结果也最准确,其他 3 种分类器都出现有误判的情况,其中 *K*-center 分类器最严重,出现了 4 个异物漏检的情况,这在实际烟叶异物检测系统中是不被允许的。在单个样本测试用时方面,SVDD 分类器略快于 *K*-center 分类器,而 Autoenc 分类器用时太长,降低了系统的实效性。实验证明,SVDD 方法在烟叶异物分类识别中的综合性能优于其他 3 种方法。

4 结语

把烟叶异物检测问题看作单值分类问题,运用支持向量数据描述方法对烟叶异物进行分类识别,解决了难以准确提取异物特征参量的难题。把 HV 颜色分量作为烟叶异物分类识别的特征参量,在保证分类效果的情况下降低特征数据的维数,提高了系统的时效性;与其他分类方法相比,基于 HV 颜色分量特征提取的 SVDD 方法分类效果好,计算效率高。如何将烟叶异物的颜色特征与纹理、形状等特征相结合,以及如何将其他信号处理方法与 SVDD 结合以进一步提高分类效果和计算效率,将是下一步研究的方向。

参考文献:

[1] 何浩,张乐年. 烟叶异物剔除系统中图像处理卡的研究[J]. 电气技术与自化,2010,39(6):156 - 158.

[2] 戴勇强. 成熟度与烟叶质量的关系及其在烟叶分级中的判断[J]. 现代农业科技,2011,3(6): 37 - 38.

[3] 阎瑞琼,韩力群,陈晋东. 计算机技术在烟叶检测与分级领域的应用[J]. 烟草科技,2001(3):13 - 15.

[4] TAX D M J, DUIN R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45 - 66.

[5] OSUNA E, FREUND R, GIROSI F. Support vector machines: Training and applications, # 1602[R]. Cambridge: MTT, 1997.

[6] TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11): 1191 - 1199.

[7] 李凌均,张周锁,何正嘉. 基于支持向量数据描述的机械故障诊断研究[J]. 西安交通大学学报,2003,37(9) :910 - 913.

[8] JOHNSON E A, LAM H F, KATAFYGIOTIS L S, *et al*. Phase I IASC-ASCE structural health monitoring benchmark problem using simulated data[J]. Journal of Engineering Mechanics, 2004, 130, (1):3 - 15.

[9] XIN D, WU Z H, ZHANG W F. Support vector domain description for speaker recognition [C]// Proceedings of the 2001 IEEE Signal Processing Society Workshop of Neural Networks for Signal Processing XI. Piscataway, NJ: IEEE Press, 2001:481 - 488.

[10] 杨敏,张焕国,傅建明,等. 基于支持向量数据描述的异常检测方法[J]. 计算机工程, 2005, 31(3):39 - 41.

[11] 赵学风,段晨东,刘义艳,等. 一种基于支持向量数据描述的 损伤诊断方法[J]. 系统仿真学报,2008,20(6):1570 - 1573.

[12] 赵旭文,缪志敏,周志杰,等. 基于 SVDD 的认知无线电网络仿冒主用户检测技术[J]. 信号处理,2010,26(7):974 - 979.

(上接第 863 页)

(Virtual Link, VL),配置最大帧长度(Maximum Frame Size, MFS)和最小帧间隔(Bandwidth Allocation Gap, BAG)等属性。

6)设置不同模块上 AFDX 接口卡收发数据帧。

7)编译并加载分区及分区中的应用程序以及 AFDX 网络的虚链路配置信息,运行分区及分区中的应用程序,启动测试。

8)在内核操作系统层使用“i”命令查看对应的应用分区是否创建。

9)使用“partitionShow”命令查看分区状态是否正常。

10)启动周期进程进行 AFDX 数据通信,验证数据收发内容是否一致。

测试结果表明,移植后的 BSP 能够正确地按照配置信息创建分区并启动分区驻留的应用程序,时间和空间分区机制工作正常,AFDX 驱动程序能够按照事先规划的虚链路进行通信,并通过虚分区与分区内驻留的应用程序进行端口通信,采样端口和队列端口均能正常工作,多个模块可以通过 AFDX 网络进行数据通信。经验证,经过移植后的 C2K 板卡支持 VxWorks 653 及 AFDX 接口,时间和空间隔离机制工作正常,分区之间及分区与外部接口通信正常,可以根据分区配置表建立分区、加载软件并执行相应的功能。

5 结语

通过移植和开发 C2K 的 BSP 和 AFDX 驱动程序,首次实现了高性价比的商用单板计算机 C2K 对 VxWorks 653 的支持,并提供了 AFDX 接口,形成了分布式的分区计算机系统,为 IMA 航空电子的各种应用提供了开发平台,可用来支持驻留应用的开发和调试,大幅度降低了开发和维护成本。

参考文献:

[1] RUSHBY J. Partitioning in avionics architectures: requirements, mechanisms, and assurance, DOT/FAA/AR-99/58[R]. California: SRI International, Computer Science Laboratory, 2000.

[2] Airlines Electronic Engineering Committee. Aavionics application software standard interface ARINC specification 653-1[S]. Annapolis, Maryland: Aeronautial Radio, Inc., 2003.

[3] Airlines Electronic Cngineering Committee. Aircraft data network part 7 Avionics Full Duplex Switched Ethernet (AFDX) network [M]. Annapolis, Maryland: Aeronautial Radio, Inc., 2005.

[4] 上海飞机设计研究院. C919 大型客机项目介绍[EB/OL]. [2011-11-13]. http://www.sadri.com.cn/cpzs1/c919_8.html.

[5] PRISAZNUK P J. ARINC 653 role in Integrated Moduler Avionics (IMA) [C]// DASC 2008: IEEE/AIAA 27th (ICIP), Digital Avionics Systems Conference. Piscataway, NJ: IEEE Press, 2008: 1.E.5-1 - 1.E.5-10.

[6] BSP migration guidelines for platform safety critical ARINC 6 5 3 [EB/CD]. Alamida, California: Windriver, Inc., 2007.

[7] VxWorks 653 programmer's guide, 2.2[EB/CD]. Alamida, California: Windriver, Inc., 2007.

[8] C2K compactPCI 6U single board computer user's guide[EB/CD]. Albuquerque, NM: SBS Technologies, 2006.

[9] VxWorks 653 configuration and build guide 2.2[EB/CD]. Alamida, California: Windriver, Inc., 2007.