

基于多类支持向量机的棉花异性纤维分类方法

杨文柱, 卢素魁, 王思乐

(河北大学 数学与计算机学院, 河北 保定 071002)

(wenzhuyang@163.com)

摘要:提出一种基于多类支持向量机的棉花异性纤维分类方法,以期解决棉花异性纤维的在线分类难题。该方法首先对异性纤维目标图像进行颜色、形状和纹理特征提取,形成用于精确描述异性纤维目标的特征向量;然后分别构建3种不同体系结构的多类支持向量机用于棉花异性纤维的分类;最后采用交叉验证法对所构建的3种多类支持向量机进行测试。测试结果表明,基于有向无环图的一对一多类支持向量机在分类精度和分类速度上更适合用于棉花异性纤维在线分类。

关键词:异性纤维;在线分类;特征向量;多类支持向量机;留一交叉验证

中图分类号: TP391.41 **文献标志码:** A

Classification of cotton foreign fibers based on multi-class support vector machine

YANG Wen-zhu, LU Su-kui, WANG Si-le

(College of Mathematics and Computer Science, Hebei University, Baoding Hebei 071002, China)

Abstract: This paper proposed a new classification method based on Multi-class Support Vector Machine (MSVM) which aimed at solving the problems in online classification of cotton foreign fibers. Firstly the features of color, shape and texture of the foreign fiber objects were extracted to create the feature vectors. Secondly three kinds of multi-class support vector machines were constructed for foreign fiber classification. These three MSVMs were tested with the obtained feature vectors using leave-one-out cross validation. The experimental results show that the one-against-one directed acyclic graph MSVM is the fastest one and is fitter for online classification of foreign fibers.

Key words: foreign fiber; online classification; feature vector; multi-class support vector machine; leave-one-out cross validation

0 引言

在棉花异性纤维含量在线计量系统中,高质量的图像采集^[1]、高效的图像分割^[2]、准确的目标分类、合理的计量模型是实现异性纤维含量精确计量的关键。本文仅对棉花异性纤维目标分类方法进行研究。

分类是运用某种决策标准将待分类数据集中的每个元素分配到类别的某个有限集合的过程^[3]。在各类别总体概率分布已知并且类别数目固定的情况下,使用线性或分段线性分类器可以获得较满意的分类效果^[4]。当各类别的总体概率分布未知或可变时,可使用神经网络^[5]、模糊逻辑^[6]等分类方法。上述方法皆为监督分类方法。当类别数未知或可变时,则需要采用非监督分类方法。本文研究对象为有限确定种类的棉花异性纤维,因此属于监督分类。

传统模式识别方法及神经网络等都是建立大样本基础上的,当样本数目较少时很难获得满意的分类结果。在统计学学习理论上发展起来的支持向量机(Support Vector Machine, SVM)是一种新的模式识别方法,对解决小样本、非线性及高维模式识别问题有很好的效果^[4]。SVM利用结构风险最小化代替传统模式识别中的经验风险最小化,很好地解决了小样本学习问题,并通过核函数将非线性空间问题映

射到线性空间来降低算法复杂度。标准支持向量机最初是为解决两类别分类而设计的,目前实现多类支持向量机(Multi-class SVM, MSVM)的方法主要有一对多方法和一对一方方法^[7]。支持向量机已经在禽蛋无损检测^[8]、人脸识别^[9]、比萨饼质量检测^[3]、步态识别^[10]、车牌识别^[11]、医学超声图像分类^[7]等方面进行了成功应用。

为解决棉花异性纤维在线分类中存在的分类正确率低、适应性差问题,本文设计了3种不同结构的MSVM,并利用获取的棉花异性纤维样本数据对所构建的3种MSVM进行了分类效果测试,旨在找到一种分类正确率高、适应性强的分类方法,以期解决棉花异性纤维在线分类难题。

1 棉花异性纤维样本集的构建

1.1 棉花异性纤维目标的形成

使用棉花异性纤维检测实验平台进行图像采集,得到各种典型异性纤维的实时图像。典型的异性纤维图像如图1所示。

从采集的大量棉花异性纤维图像中选择含有不同种类异性纤维的图像79幅作为样本,其中红色布条12幅,红色丙纶丝12幅,麻绳12幅,黑色塑料布14幅,鸡毛12幅,头发17幅。这些图像经过手工处理,去掉了图像中的棉花叶、棉花籽

收稿日期: 2011-05-10; **修回日期:** 2011-06-17。 **基金项目:** 国家自然科学基金资助项目(30971693); 现代精细农业系统集成研究教育部重点实验室开放基金资助项目; 河北大学校内基金资助项目。

作者简介: 杨文柱(1968-),男,河北保定人,副教授,博士,主要研究方向:机器视觉、人工智能; 卢素魁(1959-),男,河北保定人,副教授,硕士,主要研究方向:机器学习; 王思乐(1971-),男,河北保定人,讲师,硕士研究生,主要研究方向:机器视觉。

屑等伪异性纤维,以简化图像处理和目标分类过程。

利用文献[2]中的图像处理方法对上述图像进行处理,共得到 356 个异性纤维目标,其中红色布条 55 个,红色丙纶丝 27 个,麻绳 118 个,黑色塑料布 77 个,鸡毛 49 个,头发 30 个。部分典型目标如图 2 所示。

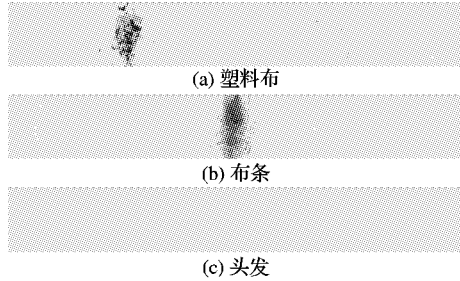


图 1 棉花异性纤维原始彩色图像

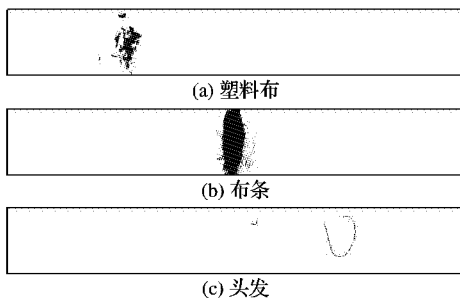


图 2 图像处理后的棉花异性纤维目标

1.2 特征提取

提取每个异性纤维目标的颜色、形状和纹理特征,组成表示该目标的特征向量。其中,颜色特征是根据分割得到的异性纤维目标的坐标在原始 RGB 图像的对应位置提取的,包括红色均值、绿色均值、蓝色均值、红绿蓝总均值以及红绿蓝三个分量的标准差等 5 个特征。形状特征也是直接在分割得到的异性纤维目标上提取的,包括形状因子、外观比、扩展比例、充实度、偏心率、球状性和欧拉数等 7 个特征。纹理特征则是根据分割得到的异性纤维目标的外接矩形,在灰度图像上的对应区域提取的,包括基于灰度直方图的平均亮度、平均对比度、平滑度、三阶矩、一致性、熵,以及基于灰度共生矩阵的角二阶矩、熵和对比度等 9 个特征。这些特征按顺序组成 21 维特征向量。

2 多类支持向量机的构建

标准 SVM 只能实现对两类问题的分类,若要实现多类分类,就需要将多个标准 SVM 通过某种方法组织在一起,形成 MSVM。目前构建 MSVM 的方法主要有一对多方法和一对一方法。

2.1 基于决策树的一对多 MSVM

构建一个能实现 K 分类的基于决策树的一对多多类支持向量机 (One-Against-All Decision-Tree Based MSVM, OAA-DTB MSVM),首先需要构建 $K-1$ 个两分类 SVM,然后再组织成决策树。构造的第 i 个两分类 SVM 用于实现对第 i 类与剩余其他类之间的分类,其决策函数描述如下:

$$f^i(X) = \sum_{n=1}^N y_n^i \alpha_n^i k(X_n, X) + b^i$$

式中 N 为训练样本集中所有类别样本的总个数; $y_n^i \in \{+1, -1\}$, 当样本 X_n 属于第 i 类时,其类别标识 $y_n^i = +1$, 否则 $y_n^i = -1$; α_n^i 为拉格朗日系数; $k(\cdot)$ 为核函数; b^i 为用于区分第 i 类和其他类的分类阈值; X 为待分类目标的特征向量。

常用的核函数主要有线性核函数、多项式核函数、径向基

核函数和 Sigmoid 核函数等。鉴于高斯径向基核函数具有的良好学习能力和适应性,本文亦采用了高斯径向基核函数,定义为

$$k(X_n, X) = \exp\left\{-\frac{\|X_n - X\|^2}{\sigma^2}\right\}$$

式中 σ 为训练样本的方差。

第 i 个两分类 SVM 的输出由决策函数 $f^i(X)$ 值的符号决定,即:若 $f^i(X) > 0$ 则输出 $+1$,表示 X 属于第 i 类;否则输出 -1 ,表示 X 属于剩余的其他类。

在训练第 i 个两分类 SVM 时,令第 i 类训练样本的类别标识为 $+1$,其他类训练样本的类别标识为 -1 。通过训练可以获得每个两分类 SVM 决策函数中的拉格朗日系数 α_n^i 和分类阈值 b^i 。

将上述 $K-1$ 个两分类 SVM 组织成决策树,就可以得到对 K 类目标进行分类的 MSVM。用于实现对 6 类棉花异性纤维分类的基于决策树的一对多 MSVM 的逻辑结构如图 3 所示。

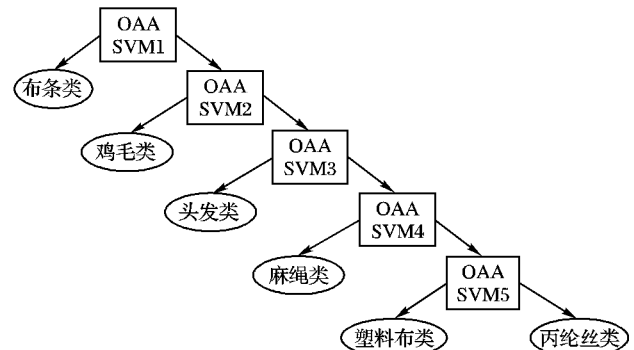


图 3 基于决策树的一对多 MSVM

2.2 基于投票的一对一 MSVM

为构建一个能实现 K 分类的一对一 MSVM,需要构建 $K(K-1)/2$ 个两分类 SVM,每个两分类 SVM 用于实现其中一类跟另一类的分类。设 C_{ij}^i 为某两分类 SVM,用于实现对第 i 类与第 j 类的分类,则其决策函数可描述为:

$$f^{ij}(X) = \sum_{n=1}^M y_n^{ij} \alpha_n^{ij} k(X_n^i, X) + b^{ij}$$

其中: M 为第 i 类与第 j 类训练样本的总个数; $y_n^{ij} \in \{+1, -1\}$, 当样本 X_n^i 属于第 i 类时,其类别标识 $y_n^{ij} = +1$, 当属于第 j 类时, $y_n^{ij} = -1$; α_n^{ij} 为拉格朗日系数; $k(\cdot)$ 为核函数; b^{ij} 为用于区分第 i 类和第 j 类的分类阈值; X 为待分类目标的特征向量。

C_{ij}^i 的输出由决策函数 $f^{ij}(X)$ 值的符号决定,即:若 $f^{ij}(X) > 0$ 则输出 $+1$,表示 X 属于第 i 类;否则输出 -1 ,表示 X 属于第 j 类。

在训练 C_{ij}^i 时,令第 i 类训练样本的类别标识为 $+1$,第 j 类训练样本的类别标识为 -1 ,通过训练可以获得 C_{ij}^i 决策函数中的拉格朗日系数 α_n^{ij} 和分类阈值 b^{ij} 。

基于投票的一对多类支持向量机 (One-Against-One Voting Based MSVM, OAO-VB MSVM) 是将待分类目标 X 在上述的 $K(K-1)/2$ 个两分类 SVM 上分别进行分类,并设置计数器对分类结果进行统计,最后将 X 判定为得票最多的类。

2.3 基于有向无环图的一对一 MSVM

基于有向无环图的一对多类支持向量机 (One-Against-One Directed Acyclic Graph MSVM, OAO-DAG MSVM) 同样需要事先构建 $K(K-1)/2$ 个两分类 SVM,且其构建和训练方式与 2.2 节所述的两分类 SVM 完全一样。唯一不同之处在于对 C_{ij}^i 输出的解释。在 OAO-DAG MSVM 中,若 C_{ij}^i 输出 $+1$,表

示待识别目标 X 不属于第 j 类,是否属于第 i 类还不确定;而若输出 -1 ,则表示 X 不属于第 i 类,是否属于第 j 类也不确定。只有经过整个有向无环图到达叶子节点,才能最终确定 X 的具体类别。

用于实现对 6 类棉花异性纤维分类的 OAO-DAG MSVM 的逻辑结构如图 4 所示。

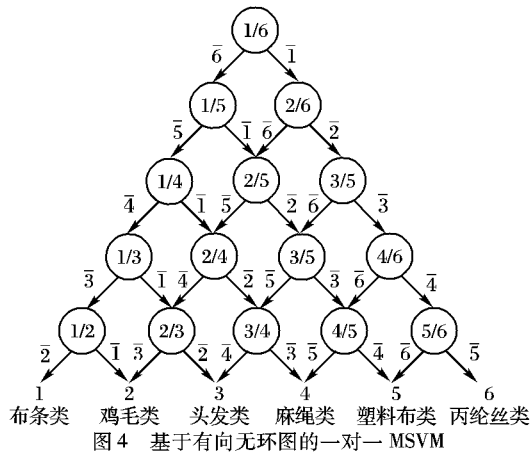


图 4 基于有向无环图的一对一 MSVM

3 实验结果与讨论

三种 MSVM 利用 Matlab 7.0 编程实现,其中的标准 SVM 使用了 Steve Gunn 编写的支持向量机 Matlab 工具箱。三种 MSVM 在 PC 上用构建的棉花异性纤维样本数据集对其进行测试。PC 的处理器为主频 2.20 GHz 的 Intel Core 2 Duo CPU,内存为 1 GB。

由于样本数量相对较少,为保证实验结果的可靠性,实验方案采取留一交叉验证法 (leave-one-out cross validation)。在测试某 MSVM 对第 i 类异性纤维的分类正确率时,按顺序每次取第 i 类中的 1 个样本作为测试样本,其余作为训练样本,直到所有样本皆测试 1 次为止,最后统计分类正确率。三种 MSVM 对 6 种棉花异性纤维的分类结果如表 1~3 所示(其中,表 1 的平均分类正确率为 79.25%,平均分类时间为 24.8 ms,表 2 的平均分类正确率为 93.57%,平均分类时间为 38.2 ms,表 3 的平均分类正确率为 92.34%,平均分类时间为 12.0 ms)。

表 1 基于决策树的一对多 MASM 分类结果

异性纤维样本	分类结果						异性纤维总数	分类正确率/%	分类时间/ms
	塑料布	布条	麻绳	头发	丙纶丝	鸡毛			
塑料布	69	1	2	0	0	5	77	89.61	35.7
布条	0	53	0	0	2	0	55	96.36	9.1
麻绳	1	1	109	0	1	6	118	92.37	28.7
头发	0	0	1	26	0	3	30	86.67	21.9
丙纶丝	0	1	13	0	10	3	27	37.04	33.0
鸡毛	3	0	7	0	3	36	49	73.47	20.4

表 2 基于投票的一对一 MASM 分类结果

异性纤维样本	分类结果						异性纤维总数	分类正确率/%	分类时间/ms
	塑料布	布条	麻绳	头发	丙纶丝	鸡毛			
塑料布	73	0	0	0	0	4	77	94.81	37.7
布条	0	53	0	0	2	0	55	96.36	37.5
麻绳	1	0	112	0	0	5	118	94.92	38.3
头发	1	0	1	28	0	0	30	93.33	39.6
丙纶丝	0	1	0	0	26	0	27	96.30	37.6
鸡毛	4	0	3	0	0	42	49	85.71	38.6

表 3 基于有向无环图的一对一 MASM 分类结果

异性纤维样本	分类结果						异性纤维总数	分类正确率/%	分类时间/ms
	塑料布	布条	麻绳	头发	丙纶丝	鸡毛			
塑料布	72	0	0	1	0	4	77	93.51	13.6
布条	0	53	0	0	2	0	55	96.36	12.5
麻绳	1	0	112	0	0	5	118	94.92	13.2
头发	1	0	1	28	0	0	30	93.33	10.4
丙纶丝	0	1	0	0	26	0	27	96.30	10.4
鸡毛	4	0	6	0	0	39	49	79.59	11.8

从测试结果可以看出,一对一 MSVM,无论采用投票决策还是有向无环图决策,其分类正确率都比一对多 MSVM 高。OAO-DTB MSVM 的分类正确率只有 79.25%,尤其是对丙纶丝的分类正确率仅为 37.04%,因此该分类器不能满足异性纤维在线计量对分类精度的要求。OAO-DAG MSVM 的平均分类正确率为 92.34%,要比 OAO-VB MSVM 的 93.57% 稍低一点,但都能满足异性纤维计量对分类精度的要求。二者的差别主要表现在对塑料布和鸡毛的分类上,前者的分类正确

率分别为 93.51% 和 79.59%,后者为 94.81% 和 85.71%。在分类速度上,OAO-DAG MSVM 完成一次分类平均需要 12.0 ms,相对于 OAO-VB MSVM 的 38.2 ms,快了二倍多。这是因为 OAO-VB MSVM 进行一次分类需要经过 15 个两分类 SVM 的判断,而 OAO-DAG MSVM 仅需要经过 5 个两分类 SVM 的判断。因此,在分类精度相差不大的情况下,OAO-DAG MSVM 要比 OAO-VB MSVM 的实时性更强,更适合用于在线分类系统。

(下转第 3452 页)

GB2312 字符集代码在字库芯片的地址码,然后将该地址码送给字库芯片提取该地址所对应的 GB2312 字符集,再将 GB2312 字符集代码通过字库芯片转化提取相对应的 16×16 点阵代码,通过 SPI 口将代码送给单片机,进而单片机将点阵代码送给点阵屏并控制点阵屏将数据显示出来。

工作时,当手机编辑短信发送数据,GPRS 模块将接收数据并通过串口发送给单片机,单片机识别到正确密码后,后面的数据就视为有效数据。有效数据开始位可自定义。比如为“#”,那么软件将认为“#”后是控制 LED 显示的亮度、速度、移动方向的数据。如果有效数据的开始位不为“#”,那么将默认此次数据为 LED 要显示的新内容,系统将把数据按顺序存入指定的 FLASH 地址范围内,待复位后新数据将显示于屏幕上。特别要指出的是,为了保证显示屏工作的可靠性,要在显示循环程序中设置好“喂狗”参数。

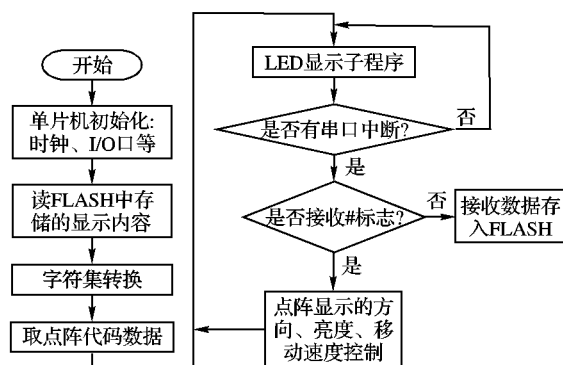


图7 系统软件流程

4 结语

系统基本达到了设计的任务要求,实现了通过手机 GPRS 远程更新显示屏显示内容的目的。本设计是由手机通

过短信编辑显示数据的传送,所以具有通用性。但是,有少数手机不支持 Unicode 字符集格式,就无法采用这个方案。

同时,由于短信通信是双向的,本系统还可实现设备的简单智能诊断,当 LED 显示系统出了问题时能实时通知管理者,实现简单功能的无人值守显示系统设备。

参考文献:

- [1] 韩斌杰. GPRS 原理及其网络优化[M]. 北京: 机械工业出版社, 2003.
- [2] 诸昌铃. LED 显示屏系统原理及工程技术[M]. 成都: 电子科技大学出版社, 2000.
- [3] 靳槐, 鄒芝权, 李骥, 等. 基于 51 系列单片机的 LED 显示屏开发技术[M]. 北京: 北京航空航天大学出版社, 2009.
- [4] C8051F410/1/2/3 混合信号 ISP FLASH 微控制器数据手册[EB/OL]. [2011-02-20]. <http://www.xhl.com.cn/upfile/Flash/2011/4/20110426161609.pdf>.
- [5] 李保凤, 郭新志. 手机短信编解码分析及其 C++ 程序实现[J]. 安阳学院学报, 2009(2): 65-67.
- [6] 孙英. 短信收发中的 PDU 编码分析[J]. 内蒙古科技与经济, 2007(1): 142-144.
- [7] 博讯科技. AT 命令集使用手册[K/EB]. [2011-02-20]. http://www.bocon.com.cn/Document/wireless/zigbee_atcmd_1.7.pdf.
- [8] 吴刚, 朱一. 短信服务 PDU 收发技术研究[J]. 装备制造技术, 2008(12): 75-77.
- [9] 潘琢金, 孙德龙, 夏秀峰. C8051F 单片机应用解析[M]. 北京: 北京航空航天大学出版社, 2002.
- [10] 童长飞. C8051F 系列单片机开发与 C 语言编程[M]. 北京: 北京航空航天大学出版社, 2005.
- [11] 谭浩强. C 程序设计[M]. 4 版. 北京: 清华大学出版社, 2010.
- [12] 华成英, 童诗白. 模拟电子技术基础[M]. 4 版. 北京: 高等教育出版社, 2010.
- [13] 康华光. 电子技术基础: 数字部分[M]. 5 版. 北京: 高等教育出版社, 2010.

(上接第 3448 页)

4 结语

通过提取异性纤维目标的颜色、形状和纹理特征,构成高效的特征向量。从采集的大量棉花异性纤维图像中选择有代表性的图像作为样本,通过提取图像分割产生的异性纤维目标的特征,形成样本数据集。

设计了 OAA-DTB MSVM、OAO-VB MSVM 和 OAO-DAG MSVM 3 种多类支持向量机。采用交叉验证法对 3 种 MSVM 用棉花异性纤维样本数据集进行了测试。实验结果表明, OAA-DTB MSVM 无法满足异性纤维在线计量对分类精度的要求,而两种 OAO MSVM 皆可满足分类精度要求,尤其是 OAO-DAG MSVM,更适合在棉花异性纤维在线分类系统上应用。

由于选择的异性纤维种类有限,且忽略了伪异性纤维造成的影响,因此该方法还需要进一步的实验验证。另外,对检测过程中出现的未知异性纤维种类如何进行自动识别也是需要进一步研究的内容。

参考文献:

- [1] YANG WENZHU, LI DAOLIANG, WEI XINHUA, et al. An automated visual inspection system for foreign fiber detection in lint [C]// Proceedings of the WRI Global Congress on Intelligent Systems. Washington, DC: IEEE Computer Society, 2009: 364-368.

- [2] 杨文柱, 李道亮, 魏新华, 等. 棉花异性纤维图像分割方法[J]. 农业机械学报, 2009, 40(3): 156-160, 171.
- [3] DU CHENG-JIN, SUN DA-WEN. Multi-classification of pizza using computer vision and support vector machine[J]. Journal of Food Engineering, 2008, 86(2): 234-242.
- [4] 边肇祺, 张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社, 2000.
- [5] OU G B, MURPHEY Y L. Multi-class pattern classification using neural networks[J]. Pattern Recognition, 2007, 40(1): 4-18.
- [6] AMO A, MONTERO J, BIGING G, et al. Fuzzy classification systems[J]. European Journal of Operational Research, 2004, 156(2): 495-507.
- [7] HORNG M H. Multi-class support vector machine for classification of the ultrasonic images of supraspinatus[J]. Expert Systems with Applications, 2009, 36(4): 8124-8133.
- [8] 何丽红, 刘金刚, 文友先. 基于粗糙集与支持向量机的禽蛋蛋壳无损检测[J]. 农业机械学报, 2009, 40(3): 167-171.
- [9] 彭中亚, 程国建. 基于独立成分分析和核向量机的人脸识别[J]. 计算机工程, 2010, 36(7): 193-194.
- [10] 路远. 基于模糊支持向量机的步态识别[J]. 计算机工程, 2009, 35(21): 189-191.
- [11] 张旭光, 罗以宁, 艾必刚, 等. 基于 SVM 和排列模型的白色车牌分类[J]. 计算机工程与应用, 2010, 46(29): 207-210.