

文章编号:1001-9081(2007)06-1485-03

基于新型特征提取的寄生虫卵图像识别研究

罗泽举¹, 宋丽红², 伍小明³, 詹希美⁴

(1. 重庆工商大学 计算机科学与信息工程学院, 重庆 400067;

2. 重庆工商大学 实验实习中心, 重庆 400067;

3. 中山大学 数学与计算科学学院, 广东 广州 510275; 4. 中山大学 中山医学院, 广东 广州 510080)

(luozeju@126.com)

摘要:讨论了用支持向量机进行多分类的若干学习策略,提出了一种新型图像特征提取方法,以此来实现对鞭虫等九种寄生虫卵图像自动识别和分类,平均识别率优于传统神经网络,达到了 93.9%,为寄生虫卵图像识别提供了一种新方法。

关键词:支持向量机;新型特征提取;寄生虫卵识别

中图分类号:TP391.4 **文献标识码:**A

Study on recognition for parasite ovum images based on new method of feature extraction

LUO Ze-ju¹, SONG Li-hong², WU Xiao-ming³, ZHAN Xi-mei⁴

(1. School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067, China;

2. Center of Experiment and Practice, Chongqing Technology and Business University, Chongqing 400067, China;

3. School of Mathematics and Computational Sciences, Sun Yat-sen University, Guangzhou Guangdong 510275, China;

4. School of Pre-clinical Medicine, Sun Yat-sen University, Guangzhou Guangdong 510080, China)

Abstract: Several learning strategies about multi-class classification by using Support Vector Machine (SVM) being discussed in this paper, we proposed a new kind of method for images feature extraction, so as to realize recognition and classification automatically for the nine kinds of parasites ovum images such as Trichuris trichiura, etc. The average recognition rate is superior to that of traditional neural network and reaches 93.9%, providing a new method for parasites ovum's images recognition.

Key words: Support Vector Machines (SVM); new feature extraction; recognition for parasite ovum

支持向量机(Support Vector Machines, SVM)学习策略基于严格的统计学习理论^[1-3],是目前机器学习研究的热点领域,利用全局最优解,识别率高,运算时间少,它的多分类应用为准确诊断寄生虫以及相关疾病提供了切实可行的良好方法。

1 用 SVM 进行多分类问题的算法复杂度分析

设样本集为 $(x_1, y_1), \dots, (x_k, y_k)$ ($x_i \in R^n, y_i \in \{+1, -1\}$),假设将样本集分为 N 类,为了讨论上的方便,假设每类样本个数均为 n 个,即有 $k = Nn$ 。常见的针对多分类问题有一对一策略、一对其余策略和层次策略。由于 SVM 的主要计算体现在 Hessian 矩阵上,每个 Hessian 矩阵的元素是 $y_i y_j k(x_i, x_j)$,将每个 $y_i y_j k(x_i, x_j)$ 作为一个计算单元,下面讨论其算法的复杂度。

1) 一对一策略。有 $C_N^2 = N(N-1)/2$ 个支持向量机,最后分类时采取“投票”的方式决定分类结果,看哪个类的贡献大;两类样本其工作集大小是 $2n$,Hessian 矩阵元素的计算单元是 $(2n)^2$,因而总的计算单元是 $T_1 = 4n^2 C_N^2 = 2n^2 N(N-1)$;

2) 一对其余策略。一对其余策略建立 N 个支持向量机,每个支持向量机负责区分本类和非本类数据,最后分类结果

由有最大间隔的那个支持向量机决定。每个支持向量机样本工作集大小是 Nn ,Hessian 矩阵元素的计算单元是 $(Nn)^2$,因而总的计算单元是 $T_2 = (Nn)^2 N = n^2 N^3$;

3) 层次策略。层次策略源于二叉树搜索思想和一对其余策略,各组不允许重复,先将一类和其他类分开,再以同样的方法往下训练,直到各类被正确训练出,如图 1。

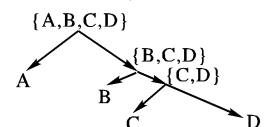


图 1 二叉树层次策略的 SVM

则有下面的定理:

定理 1 对于多分类问题,设样本集为 $(x_1, y_1), \dots, (x_k, y_k)$ ($x_i \in R^n, y_i \in \{+1, -1\}$),假设将样本集分为 N 类,

并设每类样本个数均为 n 个,即有 $k = Nn$,若以 Hessian 矩阵元素个数为算法的复杂度。则:

1) 一对一策略的复杂度为 $2n^2 N(N-1)$;

2) 一对其余策略的复杂度为 $n^2 N^3$;

定理 2 对于多分类问题,设样本集为 $(x_1, y_1), \dots, (x_k, y_k)$ ($x_i \in R^n, y_i \in \{+1, -1\}$),假设将样本集分为 N 类,并设每类样本个数均为 n 个,即有 $k = Nn$,若以 Hessian 矩阵元素个数为算法的复杂度,则层次策略的复杂度为:

收稿日期:2006-12-07;修订日期:2007-02-12

作者简介:罗泽举(1965-),男,重庆人,副教授,博士,主要研究方向:机器学习与模式识别、生物信息学; 宋丽红(1969-),女,四川越西人,实验师,主要研究方向:机器学习、计算机应用; 伍小明(1948-),女,副教授,主要研究方向:信息系统与计算机网络、图像处理与模式识别; 詹希美(1945-),男,教授,博士生导师,主要研究方向:寄生虫学。

$$n^2 \left[\frac{N(N+1)(2N+1)}{6} - 1 \right]$$

证明: 如图1, 基于层次策略的分类问题要建立 $N-1$ 个支持向量机。其中第一层计算单元是 $(nN)^2$, 第二层计算单元是 $(nN-n)^2$, 如此递推, 第 $N-1$ 层计算单元是 $[nN - (nN - 2n)]^2 = (2n)^2$; 故总的计算单元是:

$$\begin{aligned} T_3 &= (nN)^2 + (nN-n)^2 + \dots + (2n)^2 \\ &= n^2 [N^2 + (N-1)^2 + \dots + 2^2] \\ &= n^2 \left[\frac{N(N+1)(2N+1)}{6} - 1 \right] \end{aligned}$$

证毕。

定理3 对于多分类问题, 设样本集为 $(x_1, y_1), \dots, (x_k, y_k)$ ($x_i \in R^n, y_i \in \{+1, -1\}$), 假设将样本集分为 N 类, 并设每类样本个数均为 n 个, 即有 $k = Nn$, 若以 Hessian 矩阵元素个数为算法的复杂度。一对一策略的复杂度为 T_1 , 一对其余策略的复杂度为 T_2 , 层次策略的复杂度为 T_3 , 则有:

$$T_2 \geq T_3 \geq T_1$$

可见, 一对一策略的算法复杂度是上述三种算法中最小的, 即计算时间是最少的。因此, 以下的实验均采用一对一策略进行分类。

2 寄生虫卵图像识别

2.1 国内外寄生虫卵图像识别的现状

虽然国内外有关研究人员都在尝试利用计算机进行寄生虫病原体的自动识别研究^[4], 但利用计算机对寄生虫卵图像进行自动识别仍有不少困难, 主要体现在以下几方面:

1) 寄生虫的种类多, 使得在图像预处理时很难找到能适合所有虫卵的方法, 寄生虫卵的形态颜色各异, 使得选取区分各种虫卵的特征很困难;

2) 由于图像拍摄装置的差异、拍摄环境的不同, 即使是同一种虫卵, 拍摄出来的图像在背景和虫卵本身的颜色等方面也可能存在差异, 这也会影响识别效果;

3) 寄生虫卵本身在不同的时期也会有不同的形态, 有的甚至相差很大, 如蛔虫卵在未受精未脱蛋白膜时期和已受精已脱蛋白膜时期就明显不同。

在图像预处理上, 一般都是采取图像平滑化处理的手段来消除或抑制噪声, 使用的是在空域或频域上的较传统的算法, 如邻域平均法、低通滤波法等, 另外还采取了一些灰度级调整、锐化等措施来进行图像增强; 在图像分割上, 大多是采用灰度阈值分割法, 把虫卵图像转化为二值图像来实现目标与背景的分割; 在特征提取上, 提取的主要是虫卵的形态特征和颜色特征, 形态特征如虫卵面积、周长、长宽比、圆形成度等, 这些特征的计算一般依靠分割后的虫卵的轮廓而进行。颜色特征如各种颜色信号的均值、方差、色度变化、偏度、峰度等。

上述这些尝试虽取得一定效果, 但存在以下一些局限性, 如: 某些系统需要交互式的用手工进行一些参数测量, 离真正意义上的自动化还有一段距离; 识别方法特定于某一细分类下的寄生虫的虫卵; 由于以往的设备限制, 识别分类方法只针对于灰度虫卵图像; 在特征抽取中仅考虑形态特征; 在图像分割时仅以灰度为依据, 忽略了虫卵图像的彩色信息等等^[4,5]。

2.2 新型图像预处理和特征提取

针对上述图像预处理和特征提取的缺陷, 我们的实验采用以下新方法。

2.2.1 图像预处理

1) 对显微镜下拍摄出来的原始数字图像利用金字塔压

缩算法进行压缩, 以便于下一步处理;

2) 依据图像的 B 信号的特点, 将虫卵从压缩后的图像里剪切出来。这样不仅图像文件所占的空间大大减小, 而且去除了大量的噪声;

3) 将剪切后的虫卵图像进行旋转, 使每个虫卵的长径垂直于水平线。这一步已经实现了自动化, 不必对每个虫卵图像计算都人工计算旋转的角度。

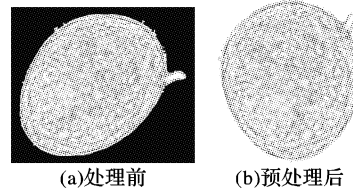


图2 血吸虫卵预处理前和预处理后的图像

2.2.2 特征提取

1) 对旋转后的图像进行尺度为 2 和 3 的 Haar 小波分解, 从水平和对角两个方向提取分解结构的高频系数^[5]。

2) 计算以上提取的高频系数的标准差和均值绝对偏差作为特征。这样就有: 尺度为 2 的小波分解水平方向上的高频系数的标准差; 尺度为 2 的小波分解水平方向上的高频系数的均值绝对偏差; 尺度为 2 的小波分解对角方向上的高频系数的标准差; 尺度为 2 的小波分解对角方向上的高频系数的均值绝对偏差; 尺度为 3 的小波分解水平方向上的高频系数的标准差; 尺度为 3 的小波分解水平方向上的高频系数的均值绝对偏差; 尺度为 3 的小波分解对角方向上的高频系数的标准差; 尺度为 3 的小波分解对角方向上的高频系数的均值绝对偏差; 共 8 个特征。

3) 取图像的香农熵、对数能量熵、范数熵和能量作为特征, 共 4 个特征。

对以上 2, 3 两步提取的 12 个特征组成总共 12 维的特征向量, 输入到设计好的分类器中, 进行学习训练和识别分类。

3 实验结果

本文采用的识别流程如图 3 所示。

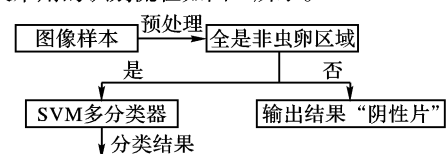


图3 图像样本的识别流程

上述识别流程先将不是虫卵的图片剔除, 再进入 SVM 多分类器进行分类识别, 多分类器采用一对一策略进行分类。

九种寄生虫卵包括鞭虫、布氏姜片虫、钩虫、华支睾吸虫、蛔虫、蛲虫、绦虫、卫氏并殖吸虫、血吸虫。其中蛔虫包括未受精未脱蛋白膜和已受精已脱蛋白膜两种形态, 另外, 九种寄生虫中由于蛔虫有两种形态, 将它们合为一类, 因此总共有九类数据, 以下分别以 1, 2, 3, 4, 5, 6, 7, 8, 9 表示。每类数据取 30 个样本来进行训练, 另外用 30 个样本来做测试; 核函数方面, 选取线性核、二阶多项式、径向基函数及 B 样条函数, 所得实验结果如表 1、表 2 所示, 平均识别率达到 93.9%, 比神经网络高出 2.6% (表 1 中 NN 是神经网络 Neural Network 的缩写)。

从实验结果可知, 支持向量机的识别率明显优于神经网络。线性核的识别率最好, 达 98.1%, 多项式次之, 径向基函数一般, B 样条最差; 从所耗时间来看, B 样条所需时间几乎是线性核的 6 倍; 支持向量方面, 线性核、二阶多项式属正常; 而 B

样条却几乎没有支持向量,表明数据远离间隔面;而径向基函数几乎全部训练样本都成了支持向量,这说明如果实验数据已经在原空间线性可分,随着模型复杂性提高,映射后新的特征空间却反而将两类数据混在一起了。上述事实表明,模型的选取要适合所给的实验数据,并不是模型越复杂越好。

表1 九类寄生虫卵的平均识别率

类别	平均识别率(%)	
	SVM	NN
1	96	91
2	95	93
3	93	88
4	93	89
5	93	92
6	95	95
7	92	90
8	93	91
9	95	93

表2 未受精未脱蛋白膜蛔虫类和鞭虫类训练时

方法	支持向量/%	耗时/s	平均识别率/%
线性核	2.6	5.3	98.1
多项式核	3.4	5.4	96.3
径向基核	97.4	5.6	95.3
B样条核	0.0	31.5	86.5
线性神经网络	-	4.3	92.1
BP神经网络	-	6.3	91.5
径向基神经网络	-	8.8	87.2

(上接第1484页)

倍(即 $3r$),这样的使得一些灰度分布不明显的图像检测结果有了一些提高。图10显示了经过改变模板半径得到的不同结果。

4 结语

本文的算法是首先得到人脸图像的谷区域图像,再得到眼睛的候选点,综合分离度信息和霍夫变换的结果来计算每一个眼睛对的代价函数,找到比较准确的眼睛位置。通过实验对比和数据统计,发现此算法鲁棒性比较好,速度也在毫秒数量级上。如果修改或简化一些判别准则,可以将速度再次提高,用于实时检测。但对戴眼镜的人脸区域会有不同程度的误差,这是由于眼镜的反光、边框等因素造成的,因此基于灰度信息的算法难以避免。要想取得更好的结果,需要对图像进一步作比较满意的预处理才能根本解决上述问题(比如考虑除去眼镜的算法)。

参考文献:

- [1] FREUND Y, SCHAPIRE RE. A decision-theoretic generalization of on-line learning [J]. Journal of Computer and System Sciences, 1997, (55): 119 - 139.
- [2] 刘直芳,游志胜,王运琼. 基于一种新的非线性彩色空间的人脸检测[J]. 数据采集与处理, 2004, 19(2): 407 - 409.
- [3] BRUNELLI R, POGGIO T. Face recognition: features versus templates[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1993, 15(10): 1042 - 1052.
- [4] 杨清凤,游志胜,张先玉. 基于豪斯多夫距离的快速多人脸检测算法[J]. 电子科技大学学报, 2004, 33(4): 160 - 166.
- [5] REISFELD D, YESHURUN Y. Preprocessing of faceimages: detec-

4 结语

本文提出了一种新的方法提取寄生虫卵的特征数据,在分析了支持向量机的多分类应用算法复杂度基础上,用一对一策略对多类寄生虫卵图像进行识别,显示出比传统神经网络更强的分类能力。需要注意的是,SVM和核函数的选取密切相关,本文中径向基核函数反而不如线性核,说明并不是模型越复杂越好,模型的选取一定要适合特征数据的数据结构,在本文中我们提取了图像样本的12维的特征向量,是对特征提取的新的尝试,取得了良好的识别效果。

参考文献:

- [1] VAPNIK VN. Statistical learning theory[M]. New York: Wiley, 1998. 100 - 128.
- [2] FU Y, SHEN R, LU H. Watermarking scheme based on support vector machine for colour images[J]. Electronics Letters, 2004, 40(16,5): 986 - 987.
- [3] GORDAN M, KOTROPOULOS C, PITAS I. Application of support vector machines classifiers to visual speech recognition[A]. Image Processing, 2002. Proceedings. 2002 International Conference on [C]. 2002, 3: 129 - 132.
- [4] 赵亚娥. 人体寄生虫虫卵图像的自动识别[J]. 中国体视学与图像分析, 1997, 2(3): 135 - 138.
- [5] PETROCK A, REISMAN S, WEINER S, et al. Wavelet analysis of masseter muscle activity resulting from stimulation of hypothalamic behavioral sites[A]. Bioengineering Conference. Proceedings of the IEEE 28th Annual Northeast[C]. 2002. 49 - 50.
- [6] 王忠,胡步发,严世榕. 一种改进的对称变换应用于人脸图像眼睛定位[J]. 计算机应用, 2004, 24(11): 119 - 121.
- [7] REINDERS MJT, KOCH RWC, GERBRADS JJ. Locating facial features in image sequences using neural networks[A]. Proceedings of Automatic Face and Gesture Recognition[C]. 1996. 230 - 235.
- [8] ZHOU ZH, GENG X. Projection functions for eye detection[J]. Pattern Recognition, 2004, 37(5): 1049 - 1056.
- [9] 肖柏旭,张丽静. 基于分流抑制机制的卷积神经网络人脸检测法[J]. 计算机应用, 2006, 26(9): 46 - 48.
- [10] VIOLA P, JONES MJ. Robust real-time object detection[EB/OL]. Technical Report Series, <http://crl.research.compaq.com>, February, 2001.
- [11] COOTES TF, EDWARDS GJ, TAYLOR CJ. Active appearance models[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2001, 23(6): 681 - 685.
- [12] SHAN Y, LIU Z, ZHANG Z. Model-based bundle adjustment with application to face modeling[J]. Proc. of. IEEE International Conference on Computer Vision, 2001, (2): 644 - 651.
- [13] LI Q, SUN LY, DUAN JY. Web page viewing behavior of users: an eye-tracking study[J]. Proceedings of ICSSM'05. 2005, (1): 244 - 249.
- [14] WANG JG, SUN E, VENKATESWARLU R. Eye gaze estimation from a single image of one eye[A]. Proceedings of the Ninth IEEE International Conference on Computer Vision[C]. 2003. 136 - 143.
- [15] KAWAGUCHI T, RIZON M. Iris detection using intensity and edge information[J]. Pattern Recognition, 2003, (36): 549 - 562.
- [16] KIMME C, BALLARD D, SKLANSKY J. Finding circles by an array of accumulators[J]. Communication ACM, 1975, 18(2): 120 - 122.