广西科学院学报
Journal of Guangxi Academy of Sciences

2008,24(4):287~290
Vol. 24,No. 4 November 2008

基于支持向量机的红细胞彩色图像分割算法*
Segmentation of Blood Cells Image Based on Support Vector Machines

黄文明,邓珍荣,计冬华
HUANG Wen-ming,DENG Zhen-rong,JI Dong-hua

(桂林电子科技大学计算机与控制学院,广西桂林 541004)
(Computer and Control College, Guilin University of Electronic Technology, Guilin,Guangxi, 541004,China)

摘要:针对细胞图像分割中红细胞目标提取和重叠红细胞分割两个难点,提出一种基于支持向量机(SVM)的红细胞彩色图像分割算法,并通过实验对算法进行验证。该算法利用SVM对原始图像进行红细胞提取,把原始细胞分割成红细胞和背景两类目标区域,然后使用改进距离标记的分水岭算法对红细胞区域进行重叠分割。算法选择线性不可分的SVM模型和核函数RBF($C=1,\xi_i=0.2$)时能够较好的分割红细胞彩色图像。

关键词:红细胞图像 支持向量机 分割
中图法分类号:TP391 **文献标识码**:A **文章编号**:1002-7378(2008)04-0287-04

Abstract: Aiming at the difficulties of cells image segmentation which involve the blood image target extraction and segmentation of clustering blood cells, a segmentation algorithm based on the Support Vector Machines (SVM) for blood cells image is presented. The algorithm extracted the blood cells from the original image by SVM. The image was divided into of blood cells area and background. Then the improved distance marked watershed method was used to separate the blood cells from the overlapping areas. The choice of non-linear dividable SVM types and kernel RBF function at ($C=1,\xi_i=0.2$)results in satisfactory blood cells image segmentation.

Key words:blood cell image,Support Vector Machines, separation

细胞的临床分类识别对各种血液病的诊断有极其重要的地位。红细胞是血液中最多的一种细胞。常见的红细胞异常主要表现在红细胞的大小、形态、染色性,血红蛋白量及分布状况以及包涵体等几个方面。准确地对红细胞进行分割识别并提取各自的参数,从而识别红细胞的异常形态,对临床诊断也有着重要价值[1]。红细胞的分割识别工作包括红细胞的目标提取和重叠红细胞间的分割两个方面。目标提取是细胞图像分类识别的基础,目前比较常用的目标提取方法可分为监督/半监督型和非监督型两种[2]。在红细胞图像中经常出现细胞重叠现象,这也是医学细胞图象处理与分析领域的一个重点和难点,目前的研究已提出一些处理重叠细胞图象分割的方法[3~6],如数字形态学分割算法,寻找凹点算法和分水岭算法,各有优缺点,但是都未能很好地分割重叠细胞图像。

本文针对传统方法的缺陷和细胞图像的特点,提出一种基于支持向量机(SVM)的红细胞彩色图像分割算法,该算法基于SVM[7~10]和距离标记的分水岭算法[5,6,11]来实现红细胞彩色图像的分类识别,并通过实验对算法的核函数、支持向量类型及相关参数进行比较和分析,寻求实现最佳的分割效果。

1 支持向量机分类

1.1 SVM原理[7~10]

SVM是在小样本统计学习理论基础上发展起来一种新的机器学习方法,它以结构风险最小化准则为理论基础,通过适当地选择函数子集及该子集

收稿日期:2008-10-12
作者简介:黄文明(1963-),男,副教授,主要从事图像图形处理、网络计算和软件工程等研究。
* 2006年国家级火炬计划项目(2006GH041397)资助。

中的判别函数，根据有限样本信息在模型的复杂性和学习能力之间寻求最佳折衷，使学习机器的实际风险达到最小，保证通过有限训练样本得到小误差分类器。SVM 在解决小样本、非线性及高维分类等方面具有很大的优越性[4]，其基本思想是：把在输入空间中的线性不可分的数据集，通过内积核函数，非线性的映射到高维特征空间后；变为线性可分的数据集，随后在高维特征空间建立一个不但能将两类正确分开，而且使分类间隔最大的最优分类面。图1是SVM思想在二维空间中的原理图，其中H为最优分类面，H1、H2分别为过各类样本中离分类线最近的、且平行于分类线的直线，H1、H2之间的距离叫做分类间隔d。

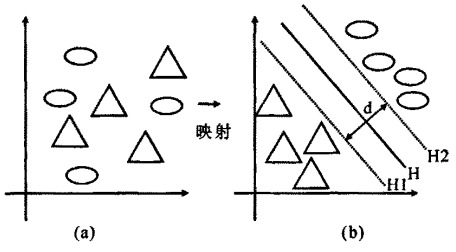


图1 SVM原理

(a)线性不可分空间；(b)高维线性可分空间

1.2 线性与非线性SVM分类问题

SVM分类函数形式上类似于一个神经网络，输出是中间节点的线性组合，每个中间节点对应一个支持向量与输入样本的内积。考虑一个两类训练样本集的分类问题 $D=\{(x_i,y_i),i=1,2,3,\cdots,l\}$，其中 $x_i\in R^n$，$y_i\in\{+1,-1\}$存在超平面$(w,x)+b=0$，使得训练样本集完全正确分开，同时满足距离超平面最近的两类点间隔最大，为样本集被超平面最优划分。归一化超平面方程，使得对所有样本集满足的约束条件为

$$y_i[\langle w,x\rangle+b]\geqslant 1,i=1,\ldots,l。 \quad (1)$$

此时分类间隔为$\frac{2}{\|w\|}$，最大间隔等价于使$\|w\|^2$最小。最大分类间隔是SVM的核心思想之一，它实际上是对学习机推广能力的控制，间隔越大，学习机泛化能力越强。因此，寻找最优分类超平面问题，可以转化二次规划问题，即

$$\min \frac{1}{2}\|w\|^2,$$

$$\text{s.t.}\quad y_i[\langle w,x\rangle+b]\geqslant 1,i=1,\ldots,l。 \quad (2)$$

最优解可以通过求解拉格朗日函数的鞍点[7]得到，有

$$\phi(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}a_i(y_i[\langle w,x_i\rangle+b]-1), \quad (3)$$

其中，a为拉格朗日乘子。依据经典的拉格朗日对偶理论[7]，可以将原问题式(3)变换为对偶问题，这将使得求解最优问题变得更加简单。其对偶问题的形式为

$$\max_a -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle+\sum_{k=1}^{l}a_k,$$

$$\text{s.t.}\ \sum_{i=1}^{l}a_iy_j=0,0\leqslant a_i,i=1,\ldots,1, \quad (4)$$

其中，a_i为Lagrange系数，$a_j>0$的样本称之为支持向量(SV)[7]，由此得到的支持向量机(即判决函数)为

$$f(x)=\text{sgn}\Big(\sum_{i=1}^{l}a_i^*y_i(x\cdot x_i)+b^*\Big)。 \quad (5)$$

在高维特征空间中，如果训练样本集线性不可分，或事先不知道它是否线性可分，将允许存在一定数量的误分类样本，折衷考虑最少错分样本和最大分类间隔，得到广义最优超平面。求解广义最优超平面的对偶问题与线性可分情况几乎完全相同，最优决策函数的形式与(5)式一样，只是约束条件变为$0\leqslant a_i\leqslant C$，其中，惩罚参数$C$。

由于对偶形式中只出现两向量的内积运算，Vapnik等提出采用满足Mercer条件的核函数$K(x_i,x_j)$来代替内积运算，实现非线性软间隔分类[8]。常用的核函数包括线性核(LINEAR)和非线性核多项式核(POLY)、径向基核(RBF)等，其核形式的最优判别函数为

$$f(x)=\text{sgn}\Big(\sum_{i=1}^{l}a_i^*y_iK(x\cdot x_i)+b^*\Big)。$$

2 基于SVM的红细胞彩色图像分割算法

将分割问题转化为分类问题求解是SVM算法的特点，由于分割的目的在于将目标区(细胞)从背景区域中分离出来，其实质也是一个分类问题。SVM这类性能优良的分类器引入，为红细胞图像分割提供了新的思路。基于支持向量机的红细胞彩色图像分割算法首先以像数点为单位，利用支持向量机对样本图像中的所有像数点进行分类[8]，分出红细胞像数点和背景区域像素点，由此来完成样本图像的目标提取，而所有红细胞像数点的集会就是我们所提取的目标。对于提取出的红细胞，再利用基于极限腐蚀距离标记的分水岭算法，进行重叠分割[3]。具体算法流程如图2所示。

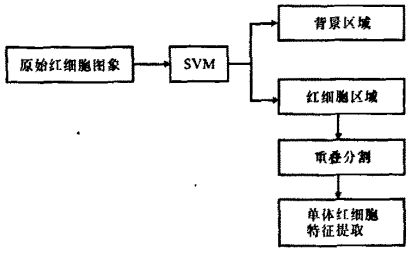


图 2 基于 SVM 的红细胞彩色图像分割算法流程

2.1 目标提取

通过寻找图像像素之间的特征的差别,即从像素点本身的特征和周围的环境(临近的像素点)出发,寻找差异,然后将各类像素点区分出来。支持向量机图像目标提取对于每一个像数点而言,它既具有颜色特征,即它的RGB值,也有它的空间特征。所以,算法的输入特征分量的包括:当前像数点的R、G、B,以当前像数点为中心的3*3方阵的平均AveR、AveG、AveB这六个特征分量。

基于SVM的红细胞彩色图像分割算法对目标提取的步骤为:(1)由操作者通过观察,选择若干正样本种子点和负样本种子点;(2)把样本种子点及其八领域像数点作为SVM训练样本,其R、G、B、AveR、AveG、AveB值作为样本的特征分量;(3)利用标记好的样本对SVM进行训练,生成支持向量分类器;(4)对待分割图像的所有像数点,使用训练好的支持向量分类器进行预测,得到提取出红细胞的二值图像。

2.2 距离标记的分水岭算法

在红细胞图像中经常出现细胞重叠现象,传统的图像分割算法各有局限性,不能取得令人满意的效果,这通常会严重影响后续的统计分析和分类识别。为此,我们采用基于改进距离标记的分水岭算法[4],即在目标被提取的基础上,首先用极限腐蚀的方法对目标区域进行距离标记,然后采用分水岭算法进行重叠分割。该算法从以下两个方面来避免过分割:一是对于小面积区域,即单体细胞区域不进行分水岭分割,对于可疑重叠区域才利用分水岭算法进行分割,这个主要基于面积因素;二是忽略图像的灰度信息,因为这是引起分水岭过分割的主要原因,而是利用像数点间的几何信息,即距离来替代像数点的灰度值,然后进行分水岭分割。

3 实验分析

基于SVM的红细胞彩色图像分割算法在Windows 2000 Professional系统下,使用VC++6.0编程实现,其中使用LIBSVM2.83程序库来实现SVM方法的红细胞提取。对于SVM,不同样本分布对应的SVM模型选择,不同的模型选择就有不同的提取效果。根据文献[6]知道红细胞样本线性不可分,因此,使用线性不可分的SVM(即C-SVM)进行训练。对于不同的核函数,其提取效果也不相同,我们选用交叉验证法来获得的不同核函数各自的参数选择,即把训练样本分为n部分,然后以其中一部分进行训练,以其余部分作预测,最终判定预测的准确性,从而不断调整参数,以达到最优效果。

图3为红细胞原图和正负样本的选取图,图4是在C-SVM模型下,使用不同核函数在各自最优参数下所获得的效果比较。

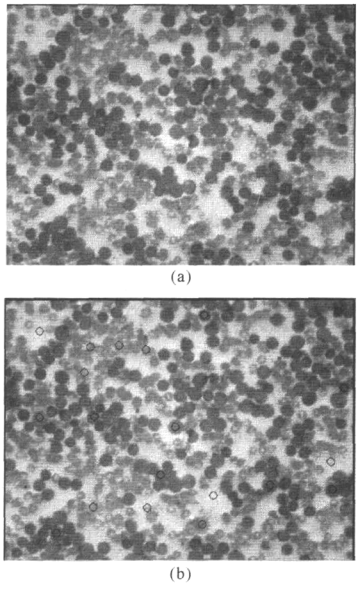
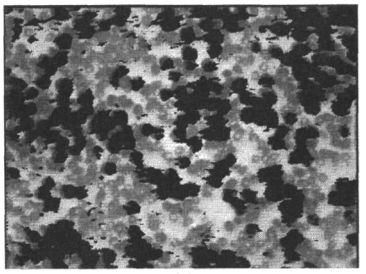
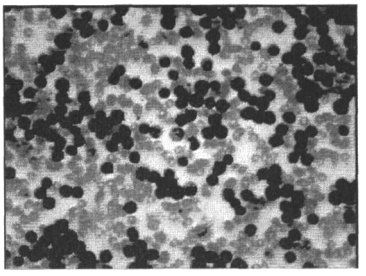


图 3 红细胞和正负样本的选取
(a)原图;(b)正负样本的选取

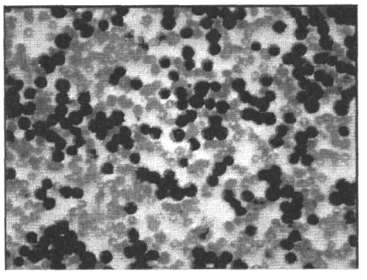
由图4的分割效果来看,在红细胞目标提取中,RBF核函数具有最好的提取效果,但是如果参数设置不够好的时候,会出现提取不完全现象;POLY核函数存在过提取的缺点;而LINEAR核函数提取的目标普遍存在过提取的缺点,而且过提取情况比较严重,而使用SIGMOID核函数虽也能得到很高的分割精度,在综合效果上略逊于RBF核函数。经过分析和比较,我们选择线性不可分的SVM模型和核函数RBF($C=1,\xi_i=0.2$)作为实验最终结果。



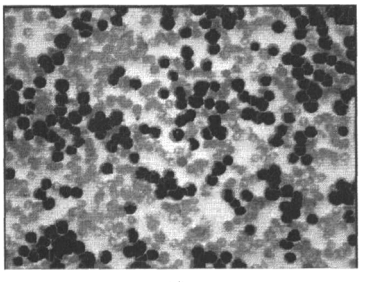
(a)



(b)



(c)



(d)

图4 不同核函数的效果比较及最终结果

(a)LINEAR:$C=1$,$\xi_i=0.2$;(b)POLY:DEGREE=2,$C=1$,$\xi_i=0.01$,Coef0=3;(c)SIGMOID:$C=500$, Coef0=2, G=0.000003;(d)RBF:$C=1$,$\xi_i=0.2$。

4 结束语

本文通过使用最新的具有小样本分类优势的SVM对红细胞图像进行分割，该算法首先利用SVM把红细胞从原始图像中提取出来，然后采用距离标记的分水岭算法，对红细胞区域进行重叠分割。实验结果显示，该算法具有较高的准确性、自动性和鲁棒性，具有较高的实际应用价值。

参考文献：

[1] 林开颜，吴军辉，徐立鸿. 彩色图像分割方法综述[J]. 中国图像图形学报，2005，10(1)：1-10.

[2] Otsu N. A threshold selection method from gray-level histogram[J]. IEEE Transactions on Systems, Man, and Cybernetics，1979，9(1)：62-66.

[3] 郭戈，平西建，胡敏. 分水岭算法在重叠细胞图象分割中的应用[J]. 微计算机信息，2005，21(8-3)：68-69.

[4] Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991，2(4)：583-598.

[5] 计冬华，黄文明，李春妍. 基于改进距离标记的彩色细胞图象分割[J]. 计算机应用，2007，27：436-439.

[6] Borgefors G. Distance transformations in digital images [J]. Computer, Graphics and Image Process，1986，34：344-371.

[7] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery，1998，2(2)：121-167

[8] 潘晨，闰相国，郑崇勋，等. 用于彩色图像分割的支持向量机的快速训练[J]. 模式识别与人工智能，2005，18(4)：392-398.

[9] 曾明，张建勋，王湘晖，等. 基于支持向量机的血液细胞核彩色图像分割[J]. 光电子·激光，2006，17(4)：479-483.

[10] 李美娟，王文伟，杨定楚. 等基于支持向量机的多光谱显微细胞图像分割[J]. 计算机工程与应用，2006(8)：37-43.

[11] 游迎荣，范影乐，庞全. 基于距离变换的粘连细胞分割方法[J]. 计算机工程与应用，2005，20：206-208.

（责任编辑：韦廷宗）