基于支持向量机的红细胞彩色图像分割①

A Red Blood Cells Color Image Segmentation Algorithm Based on Support Vector Machines

黄建灯　陈庆全　(桂林电子科技大学 信息科技学院 广西 桂林 541004)

摘　要： 对红细胞进行精确的分割并统计各类参数在临床医学中有着重要的意义，但现有经典算法很难实现自动分割并获得较高准确率。红细胞的分割主要有两个难点：一是红细胞的目标提取；二是重叠红细胞的分割。本文首先利用 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，然后对红细胞区域进行重叠分割（本文中使用改进距离标记的分水岭算法），最终得到各个红细胞的统计数据。为获得最佳的分割效果，本文通过实验对核函数、支持向量类型及相关参数进行了详细捷、准确、全自动的实现红细胞的分割识别。

关键词： 目标

1　引言

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，也一直是生物医学工程研究中一个十分活跃的领域。红细胞，又称为红血球，是血液中最多的一种细胞，红细胞的异常形态，对临床诊断也有着重要价值。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。由此可见，能准确的对红细胞进行分割识别并提取各自的参数，对临床医学有着重大的意义。但是，在实际中红细胞间的相互重叠现象的普遍，这使得血细胞的分类和识别又具有一定的难度。

红细胞的分割识别工作主要可分为两个方面：红细胞的目标提取和重叠红细胞间的分割。本文主要利用支持向量机（SVM）和距离标记的分水岭算法来实现血细胞的分类识别。首先以像数点为单位，利用支持向量机对样本图象中的所有像数点进行分类，分出红细胞像数点和背景区域像数点，由此来完成样本图象的目标提取；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；最后统计各个红细的特征参数。具体算法流程如下图所示：

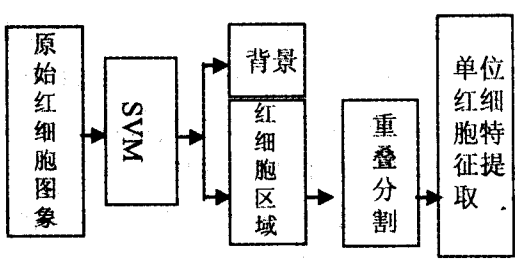

图 1　本文算法流程图

2　支持向量机 SVM

2.1 SVM 基本理论

支持向量机(support vector machine,SVM)是 Vapnik 等人根据统计学习理论提出的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中的判别函数，使学习机器的实际风险达到最小，保证了通过有限训练样本得到的小误差分类器，它在解决小样本、非线性及高维分类等方面具有很大的优越性。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

① 收稿时间:2008-07-04

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 2 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。

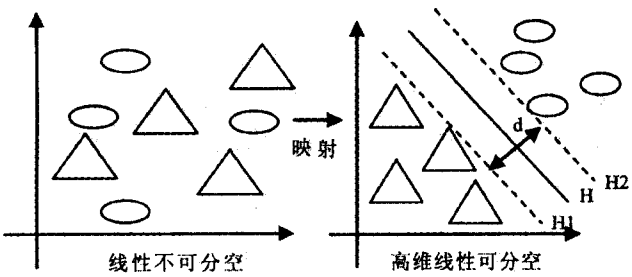

图 2　SVM 原理图

2.2 线性与非线性分类问题

支持向量基能很好的解决线性问题域及非线性问题域的分类问题，其解决这两类问题的数学理论求解如下：

给定训练样本集：$D=\{(x_i, y_i), i=1,2,3...l\}$，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

$$y_i\left[\langle w,x\rangle+b\right]\geq 1 \qquad i=1，\cdots，l \tag{1}$$

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

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

$$\text{s.t.}\ \ y_i\left[\langle w,x\rangle+b\right]\geq 1 \quad i=1，\cdots，l \tag{2}$$

最优解可以通过求解拉格朗日函数的鞍点得到，有

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

其中，a 为拉格朗日乘子。依据经典的拉格朗日对偶理论，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简单。其对偶问题的形式为：

$$\max -\frac{1}{2}\sum_{i=1}^{l}a_i a_j y_i y_j\langle x_i,x_j\rangle+\sum_{k=1}^{l}a_k$$

$$\text{s.t.}\ \sum_{i=1}^{l}a_i y_j=0 \quad 0\leq a_i \quad i=1，\cdots，l \tag{4}$$

求解得到的最优解为 $a^*=\left(a_1^*,\cdots,a_l^*\right)^T$。这样，计算得到

$$w^*=\sum_{i=1}^{l}a_i y_i x_i$$

$$b^*=-\frac{1}{2}\langle w^*,x_m+x_n\rangle \tag{5}$$

其中，m 和 n 是两类中任意的支持向量(SV)。依据 KKT 互补条件，其中只有少量最靠近超平面样本点的 a_i 值不为零，Vapnik 等人称之为 SV。由式(5)可以看，超平面只是由训练样本集中的一个很小的子集(SV 集)决定，最终决策函数为：

$$f(x)=\text{sgn}(\sum_{i=1}^{1}a_i^* yi(x,x_i)+b_i \tag{6}$$

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，在式（1）中引入非负松弛变量 $\xi_i\geq 0$，i=1,2,3…l，则变为：

$$y_i\left[\langle w,x_i+b\rangle\right]\geq 1-\xi_i \qquad i=1,2,3,\cdots l \tag{7}$$

将目标函数改为 $\phi(w,\xi)=\frac{1}{2}\|w\|^2-C\sum_{i=1}^{l}\xi_i$ 最小，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。其中，惩罚参数 C 作为综合这两个目标的权重。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，只是约束条件变为 $0\leq a_i\leq C$，最优决策函数的形式与式(6)一样。

由于对偶形式中只出现两向量的内积运算，Vapnik 等人提出采用满足 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。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：

$$f(x)=\text{sgn}(\sum_{i=1}^{l}a_i^* y_i K(x,x_i)+b^*) \tag{8}$$

其中b^*是分类阈值,可由任一向量求得。

由上可知，线性不可分数据集只需确定b^*及核函数$K\left(x_i,x_j\right)$经过式（8）计算就可得最优的分类结果集。本文的核函数使用径向基函数$K(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{\sigma^2})$。其中，$x_i$是支持向量。

正因为支持向量基有小样本、非线性及高维分类等方面具大的优越性，对于分类红细胞图像中的细胞和背景具有良好的应用场景，本文正是利用该方法作为图像分割的第一步。

3 支持向量机的红细胞图像分割算法的实现

3.1 基于支持向量机的目标提取

将分割问题转化为分类问题求解是本文算法的特点,由于分割的目的在于将目标区(红细胞)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。本文采用基于支持向量机的目标提取算法，即把红细胞区域提取问题转化为细胞像数点分类问题，利用 SVM 把整个图像的像数点分成两类，一类为红细胞像数点，另一类为背景区域像数点，而所有红细胞像数点的集合就是我们所提取的目标。

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（即它处于空间中的位置，或者说它周围像数点信息）。本文的红细胞提取算法中，也考虑了这两个特征。所以，本文 SVM 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。实验证明，其比单纯的以 RGB 作为特征分量具有更好的分割效果。

本文所采用的 SVM 算法步骤为：

(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。

此算法只需要训练一次，生成支持向量分类器，就可对同类图片直接进行预测.通过上述方法，经少量样本训练以后，可实现完全自动目标提取，且准确率较高。

3.2 重叠分割算法的研究

在血细胞图象中经常出现细胞重叠现象，这通常会严重影响后续的统计分析和分类识别。于是，对于重叠细胞分割算法的研究，也就成为了医学细胞图象处理与分析领域的一个重点和难点。目前比较常用的重叠分割算法主要有三种：第一种是数字形态学分割算法；第二种是寻找凹点算法；第三种是分水岭算法，他们各有其优缺点。

数学形态学分割算法是利用二值腐蚀和膨胀运算对重叠细胞的二值图像作处理，该分割方法原理简单，运算速度快，但由腐蚀和膨胀并不是一对可逆运算，其分割效果有一的定误差。

寻找凹点算法的原理是首先找到二值图像中的凹点，然后将凹点进行匹配找到分离点，然后将重叠细胞分离。但由于图像的本身是数字的，决定了其边缘并不是连续的，并且往往有许多的干扰，因此很难准确的识别出属于细胞分割点的凹点。因此，凹点法还很难应用到实际中。

分水岭算法[2-3]的原理，可以根据水面浸地形的过程来说明。但其对微弱边缘较敏感，能够检测出粘连和重叠细胞的边缘，同时，它也可检测出均匀区域中的低对比度变化区域，而且对噪声也比较敏感，所以其容易造成过分割。

本文采用的是基于距离标记的分水岭算法，它的原理是：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（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；二是忽略图象的灰度信息，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。由于篇幅原因，笔者在此不作详细介绍，具体可参考论文[4]。实验表明，该分割算法能有效分割开重叠细胞，较其他算法具有更高的分割准确率。

4　实验结果与讨论

本文算法在 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，即线性可分 SVM、线性软间隔 SVM 及各种非线性核函数形式 SVM 等，不同的模型选择就有不同的提取效果。从论文[5]中，我们知道，我们的样本线性不可分，因此，我们使用线性不可分的 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，而不同核函数各自的参数选择，笔者选用的是交叉验证法来获得的，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。下面是在 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：

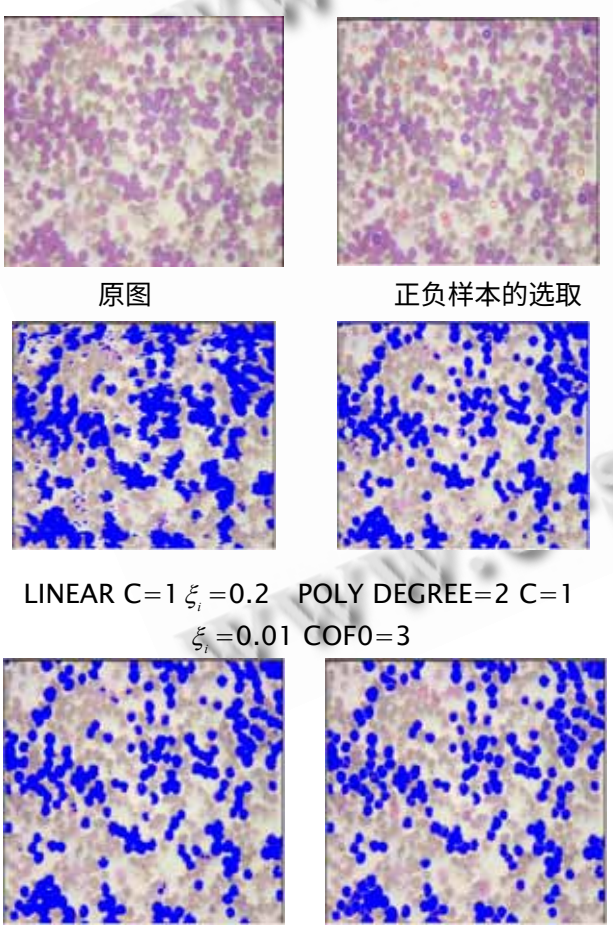

图 3 不同核函数的效果比较及最终结果图

由图 3 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（即使参数调的很好，还是避免不了）；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。另外，在实验中，笔者发现：对于非线性问题，使用非线性核做 SVM 分类，比用线性核做分类具有更好的抗干扰性。本文重叠分割算法与其它算法的具体比较可参考文献[4]。

5　结论

本文通过使用最新的具有小样本分类优势的支持向量机(SVM)对红细胞图象进行分割，该算法首先利用 SVM 把红细胞从原始图象中提取出来，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，最终得到每个红细胞的各类统计数据。实验结果显示，该算法较传统的目标提取和重叠分割算法具有更高的准确性，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.中国图象图形学报,2005,10(1):1–10.
2 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.微计算机信息,2005,21,(8-3):68–69.
3 Andr'e Bleau,Lean L J.Watershed-based segmentation and region merging. Computer Vision and Image Understanding,2000,77:317–326.
4 计冬华,黄文明,李春妍.基于改进距离标记的彩色细胞图像分割.计算机应用,2007,27(6):436–437,439.
5 曾明,张建勋,王湘晖,赵雅静,陈少杰.基于支持向量机的血液细胞核彩色图像分割.光电子·激光,2006,17(4):479–483.
6 李美娟,王文伟,杨定楚,王思贤.基于支持向量机的多光谱显微细胞图像分割.计算机工程与应用,2006,08：37–43.
7 潘晨,闰相国,郑崇勋,杨勇.用于彩色图像分割的支持向量机的快速训练.模式识别与人工智能,2005,18(4):392–398.