第33卷 第17期 计 算 机 工 程 2007年9月
Vol.33 No.17 Computer Engineering September 2007
·博士论文· 文章编号：1000—3428(2007)17—046—03 文献标识码：A 中图分类号：TP273+.5

基于支持向量机的肿瘤形状特征分类

王 彬[1,3]，孙 蕾[2]
（1. 西安理工大学计算机学院，西安 710048；2. 西安电子科技大学经济管理学院，西安 710071；
3. 西安电子科技大学计算机学院，西安 710071）

摘 要：当两类中的样本数量差别较大时，支持向量机的分类能力将会下降。该文提出了一种支持向量机新算法——DFP-PSVM，将有约束条件的二次规划问题转换为无约束二次规划问题，并通过优化计算来实现。为了克服传统的蛇形算法不能收敛于边缘凹陷处以及初始化过于敏感的缺点，采用基于可变形模型的梯度矢量流方法，提取了乳腺X光片中的肿瘤区域，分析了3个基于边缘的价矩。将其他肿瘤形状特征作为DFP-PSVM分类算法的特征输入，进行恶性肿瘤和良性肿瘤的计算机辅助诊断。实验表明，在小样本、两类样本数量“严重不均衡”的情况下，该算法有着较强的分类能力。

关键词：可变形模型；梯度矢量流；肿瘤；形状特征；支撑向量机

Classification of Tumor Shape Feature Based on Support Vector Machine

WANG Bin[1,3], SUN Lei[2]
(1. School of Computer, Xi'an University of Technology, Xi'an 710048; 2. School of Economic and Management, Xidian Univ., Xi'an 710071;
3. School of Computer, Xidian Univ., Xi'an 710071)

【Abstract】 When two-class problem samples are very unbalanced, SVM has a poor performance. A novel SVM algorithm, DFP-SVM is presented to solve the problem implemented by transfering the quadratic program with limited condition into quadratic program without constraining condition. Optimal computation is conducted to achieve exciting results. In order to overcome the problems of traditional snake associated with poor convergence to boundary concavities and sensitive initialization, gradient vector flow based on deformable models is presented to segment tumor region. And three new moments based on boundary are also developed. The novel classifier applies the three moments and other shape features to classify the tumor into the malignant or the benign. Computational results indicate that the modified algorithm has a strong capability of classification for the unbalanced data of small set of samples related to two-class problems.

【Key words】 deformable model; gradient vector flow; tumor; shape feature; support vector machine

乳腺X光片已经被广泛地用于乳腺癌的早期诊断中，由于乳腺X线照片的对比度低、肿瘤组织不同，诊断结果往往有较高的假阳性或假阴性。而活组织切片检查会带来身体上的创伤，费用较高。计算机用于处理乳腺X光片的目的在于：

(1)从X光片中提取更多用肉眼看不到的特征；

(2)提高图像的质量，增强图像的特征，使医生更容易阅读；

(3)识别图像中的重要信息并进行量化，以便深入分析；

(4)能够在相对短的时间内分析大批量包含重要信息的片子，以减少放射科医生的负担，提高准确率。

应用先进的计算机技术，研究自动、计算机辅助的基于图像的肿瘤诊断是必要、实用的[1]。由于医学图像的分类需要很高的精确度，因此目前还没有一个广泛用于医学图像的分类器。笔者采用梯度矢量流算法对肿块进行分割，提取了5个基于形状的特征，并构建数据库，采用DFP-PSVM算法进行肿瘤辅助诊断，在小样本情况下，对“分布极不平衡”数据的分类准确率可提高到94%。

1 梯度矢量流

医学图像在采样过程中经常被污染，带有很多噪声。而且医学图像的格式多样，感兴趣区域的形状也多种多样。因此，用传统的分割方法如“边缘提取算子”和“阈值法”分割感兴趣的区域是很困难的[2]。为了解决这些问题，可变形模型(deformable models)已经成为医学图像分割最为活跃的研究领域之一，而且还可以得到轮廓边缘的数学描述。

传统的蛇形算法主要存在2个缺点：

(1)要求主动轮廓线的初始位置要靠近目标轮廓，此方法对初始位置很敏感；

(2)不能收敛于凹陷处。

梯度矢量场的方法通过矢量扩散方程扩大边缘梯度的映射范围，很好地解决了这些问题[2]。

梯度向量流定义为：使“能量方程最小”的静态外部力量场，即

$$V(x,y)=[u(x,y),v(x,y)]$$
$$\varepsilon=\iint(\mu(u_x^2+u_y^2+v_x^2+v_y^2)+|\nabla f|^2|v-\nabla f|^2)dxdy \quad (1)$$
$$E_{ext}=-|\nabla G(x,y)*I(x,y)|^2,\ f=-\nabla E_{ext}$$

相对应的动态可变形轮廓线方程为

$$X_t(s,t)=\alpha X''(s)-\beta X''''(s)+V \quad (2)$$

从式(1)可以看出，当$|\nabla f|$很小时，能量主要由向量场的二阶偏导数平方和决定，产生一个缓慢变化的场。而当$|\nabla f|$很大时，被积函数主要由后一项决定，在$v=\nabla f$时能量达到最小，

作者简介：王 彬(1971－)，男，博士研究生、讲师，主研方向：网络通信与安全，数据挖掘，图形图像处理；孙 蕾，博士、副教授
收稿日期：2007-04-23 **E-mail：**wb@xaut.edu.cn

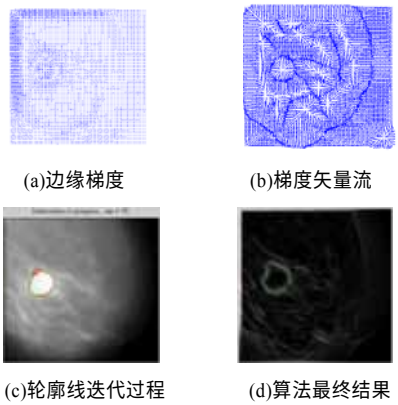
此时得到了所期望的结果，即梯度矢量流近似边缘梯度。图像的噪声越大，权重参数 μ 的取值越大。应用变分方法，求解以下欧拉方程，得到的梯度矢量流解为

$$\mu\nabla^2 u-(u-f_x)(f_x^2+f_y^2)=0$$
$$\mu\nabla^2 v-(v-f_y)(f_x^2+f_y^2)=0 \tag{3}$$

其中，∇^2是拉普拉斯算子。从式(2)可知，在图像灰度变化很小或没有变化的区域，$f(x,y)$的梯度等于零，u 和 v 的值可以通过离散和迭代来得到。

$$\mu_t(x,y,t)=\mu\nabla^2 u(x,y,t)-(u(x,y,t)-f_x(x,y,t))$$
$$(f_x^2(x,y,t)+f_y^2(x,y,t))$$
$$v_t(x,y,t)=\mu\nabla^2 v(x,y,t)-(v(x,y,t)-f_y(x,y,t)) \tag{4}$$
$$(f_x^2(x,y,t)+f_y^2(x,y,t))$$

迭代过程和计算结果见图 1。



(a)边缘梯度　(b)梯度矢量流

(c)轮廓线迭代过程　(d)算法最终结果

图 1　迭代过程和结果

2　形状特征提取

良性肿瘤和恶性肿瘤在形状上有着明显的差别。良性肿瘤的形状规则，边缘较光滑；恶性肿瘤形状不规则，边缘模糊粗糙。因此，应用最具表征力的肿瘤形状信息是计算机辅助诊断的一个大胆尝试和提高诊断精度的好方法。

2.1　似圆度(compactness)

“似圆度”是形状分析中最常用的一个方法，将“似圆度”值归一化，并随着形状的复杂性和凹凸度的增加而增大，值为零时形状是圆，公式为

$$C=1-\frac{4\pi a^2}{p^2}$$

其中，p 和 a 分别是周长和面积。

2.2　基于矩的形状特征

利用目标边缘可以计算出各种阶矩。笔者使用了 3 个新的阶矩：

$$F^1=\frac{[\frac{1}{N}\sum_{i=1}^{N}[z(i)-m_1]^2]^{1/2}}{\frac{1}{N}\sum_{i=1}^{N}z(i)} \qquad F^2=\frac{[\frac{1}{N}\sum_{i=1}^{N}[z(i)-m_1]^3]^{1/3}}{\frac{1}{N}\sum_{i=1}^{N}z(i)}$$

$$F^3=\frac{[\frac{1}{N}\sum_{i=1}^{N}[z(i)-m_1]^2]^{1/4}}{\frac{1}{N}\sum_{i=1}^{N}z(i)} \qquad m_1=\frac{1}{N}\sum_{i=1}^{N}z(i)$$

这 3 个价矩与Gupta[3]等提出的价矩均有以下 2 个优点：

(1)无量纲，易于比较和组合。

(2)F^3值越大，目标形状边缘越粗糙，能够很好地表征乳腺肿瘤边缘的粗糙度。

2.3　傅里叶描述子

采用傅里叶描述子的优点是将 2-D 的问题转化为 1-D 问题，即

$$X(n)=\frac{1}{N}\sum_{k=0}^{N-1}X_k e^{-jk2\pi/N}$$

$$NFD(k)=R+I_j \qquad \|NFD\|=(R^2+I^2)^{1/2}$$

本程序所用的傅里叶描述子为

$$FD=\frac{1}{n}\sum_{k=1}^{n}\|NFD(k)\| \quad n=6$$

2.4　弦长

弦长的表达式为

$$M_c=\frac{1}{K}\sum_{i=1}^{K}L_i$$

其中，K为边缘上点有数量；L_i为边缘上点和点之间的距离。

2.5　半径

半径的表达式为

$$R=\frac{1}{n\sum_{i=0}^{n}|C-V_i|}$$

其中，C为肿瘤的的中心(x_{center}, y_{center})；V_i为目标边缘上的点(x_i, y_i)；$|C-V_i|=\sqrt{(x_i-x_{center})^2+(y_i-y_{center})^2}$。

3　模式识别(分类)

3.1　PSVM

支持向量机的基本思想是：通过非线性变换，将输入空间变换到一个高维空间，在这个高维空间中求得最优线性分类面，而这种非线性变换是通过定义适当的内积函数来实现的。在实际运用中，发现在样本量有限且两类样本数量悬殊的情况下，算法的分类能力比较低。笔者提出了 DFP-PSVM 算法来提高 SVM 对不均衡数据的分类能力。这里只对本次实验所采用的非线性算法进行简要的描述，而略去了对支持向量机的详细描述。

Vapnik 给出的标准线性 SVM(其核函数为线性核)为：

$$Min\, ve'y+\tfrac{1}{2}\omega'\omega$$

其中，v 是大于 0 的参数。

$$\text{s.t}\quad D(A\omega-e\gamma)+y\geq e \text{ and } y\geq 0 \tag{5}$$

相应的线性分类器为

$$x'\omega-\gamma\begin{cases}>0 & then\quad x\in A^+\\ =0 & then\quad x\in A+\ or\ A^-\\ <0 & then\quad x\in A^-\end{cases} \tag{6}$$

将式(5)做进一步的变化：将y的 1 范数改为 2 范数以去除约束条件 $y\geq 0$；添加γ^2，这样通过优化分类面的方向和位置来确定最优分类面，而不是像标准的支持向量机那样只优化其方向。新的支持向量机如下：

$$Min\,\frac{v}{2}\|y\|^2+\tfrac{1}{2}(\omega'\omega+\gamma^2)$$
$$\text{s.t}\quad D(A\omega-e\gamma)+y=e \tag{7}$$

通过拉格朗日乘子法，对 ω,γ,μ,y 分别求梯度，并令其为零而得到 KKT 条件为

$$\omega=A'D\mu \quad \gamma=-e'D\mu \quad y=\frac{\mu}{v}$$
$$D(A\omega-e\gamma)+y-e=0 \tag{8}$$

为了得到非线性分类器，将 $\omega=A'D\mu$ 代入式(7)的限制条件中，并用非线性核函数 $K(A,A')$ 代替 AA'，得到式(9)：

$$Min\frac{v}{2}\|y\|^2+\frac{1}{2}(\mu'\mu+\gamma^2)$$
$$\text{s.t}\quad D(K(A,A')D\mu-e\gamma)+y=e \tag{9}$$

以下用 K 代替 $K(A,A')$，得到如下的拉格朗日函数(υ 为拉格朗日乘子)：

$$L(\mu,\gamma,\upsilon,y)=\frac{v}{2}\|y\|^2+\frac{1}{2}\left\|\begin{bmatrix}\mu\\ \gamma\end{bmatrix}\right\|^2-\upsilon'(D(KD\mu-e\gamma)+y-e) \tag{10}$$

从 KKT 条件得出 μ,γ,y 及拉格朗日乘子 υ，即

$$\mu = DK'D\upsilon,\ \gamma = -e'D\upsilon,\ \ y = \frac{\upsilon}{\nu}$$

$$D(KD\mu - e\gamma) + y = e \tag{11}$$

$$\upsilon = (\frac{I}{\nu} + D(KK' + ee')D^{-1}e = (\frac{I}{\nu} + GG')^{-1}e \tag{12}$$

$$G = D[K \ \ -e] \tag{13}$$

非线性 SVM 分类面可以从线性分类面的公式中推导出来，即

$$x'A'D\mu - \gamma = 0$$

用 $K(x', A')$ 代替上式中的内积 $x'A'$，并将式(11)中的 μ 和 γ 代入，得到如下非线性 SVM 分类超平面，即

$$\begin{aligned} &K(x', A')D\mu - \gamma \\ &= K(x', A')DDK(A, A')'D\upsilon + e'D\upsilon \\ &= (K(x', A')K(A, A')' + e')D\upsilon = 0 \end{aligned} \tag{14}$$

相应的非线性 SVM 分类器为

$$(K(x', A')K(A, A)' + e')D\upsilon \begin{cases} > 0 & x \in A^+ \\ < 0 & x \in A^- \\ = 0 & x \in A^+ \ or \ A^- \end{cases} \tag{15}$$

非线性 SVM 分类器算法步骤如下：

step1 选择一个核函数 K(A,A')。

step2 用式(13)定义 G。

step3 用式(12)计算拉格朗日乘子 υ。

Step4 应用式(14)得到非线性 SVM 分类超平面。

Step5 应用式(15)对新的样本进行分类。

本项目采用上述算法进行乳腺肿瘤良性/恶性分类。当良性肿瘤样本大超过恶性肿瘤样本时，此算法的分类精度明显下降，反过来亦如此。然而由于各种主客观原因，在有限样本的实际问题中，两类样本数量悬殊的情况经常发生。因此，DFP-SVM 算法能很好地解决此问题。

3.2 DFP-PSVM 算法

DFP-PSVM 算法通过将约束性数学规划问题转换为无约束性规划问题来实现的。式(1)中的两个约束条件可以用 $y = (e - D(A\omega - e\gamma))_+$ 替换，得到无约束最小化问题，即

$$\min \frac{\nu}{2}\left\|(e - D(A\omega - e\gamma))_+\right\|^2 + \frac{1}{2}\left\|\begin{bmatrix}\omega \\ \gamma\end{bmatrix}\right\|^2 \tag{16}$$

其中，$(e - D(A\omega - e\gamma))_+ = 0$ 表示($e - D(A\omega - e\gamma)$)的负成分为零，即

$$(e - D(A\omega - e\gamma))_+ = \max\{o, (e - D(A\omega - e\gamma)\}$$

设 $\begin{bmatrix}\varpi \\ \bar{\gamma}\end{bmatrix}$ 是式(16)的唯一解，用 $\lambda\varpi$ 代换 ω 就把此问题转换成了一个简单的二维空间的凸规划问题，即

$$\min f(\lambda, \gamma) = \frac{\nu}{2}\left\|(e - D(\lambda A\overline{\omega} - e\gamma))_+\right\|^2 + \frac{1}{2}\left\|\begin{bmatrix}\lambda\overline{\omega} \\ \gamma\end{bmatrix}\right\|^2 \tag{17}$$

以上问题可用无约束极小问题的有效算法 DFP 算法快速求解，定义

$$d(\lambda, \gamma) = e - D(\lambda A\overline{\omega} - e\gamma)$$

DFP-SVM 算法所需要的偏导数为

$$\nabla f(\lambda, \gamma) = \begin{bmatrix} -\nu\overline{\omega}'A'D(d(\lambda, \gamma))_+ + \left\|\overline{\omega}\right\|^2\lambda \\ \nu e'D(d(\lambda, \gamma))_+ + \gamma \end{bmatrix}$$

该算法的具体步骤如下：

(1)用第 1 节的算法计算出 $\overline{\omega}$ 和 $\bar{\gamma}$ ；

(2)令λ^0=1，γ^0= $\bar{\gamma}$ ，H_0=I, j=0；

(3)求g_0= $\nabla f(\lambda^j, \gamma^j)$ ，若 $\nabla f(\lambda^j, \gamma^j) = 0$ 则停机，否则p_0=-g_0；

(4)用 $\frac{d(f(\begin{bmatrix}\lambda_j \\ \gamma^j\end{bmatrix} + \alpha\ p_j)}{d\alpha} = 0$ 确定α_j；

(5)$\begin{bmatrix}\lambda^{j+1} \\ \gamma^{j+1}\end{bmatrix} = \begin{bmatrix}\lambda^j \\ \gamma^j\end{bmatrix} + \alpha_j p_j$ ；

(6)$\delta_{j+1} = \begin{bmatrix}\lambda^{j+1} \\ \gamma^{j+1}\end{bmatrix} - \begin{bmatrix}\lambda^j \\ \gamma^j\end{bmatrix} = \alpha_j p_j$

如果 $\left\|\delta_{j+1}\right\| \le \varepsilon$ 则停机，$\begin{bmatrix}\lambda^* \\ \gamma^*\end{bmatrix} = \begin{bmatrix}\lambda^{j+1} \\ \gamma^{j+1}\end{bmatrix}$，$\begin{bmatrix}\omega \\ \gamma\end{bmatrix} = \begin{bmatrix}\lambda^*\varpi \\ \gamma^*\end{bmatrix}$ ；

(7)否则，$g_{j+1} = \nabla f(\lambda^{j+1}, \gamma^{j+1})$, $\Delta g_j = g_{j+1} - g_j$ ；

(8)利用DFP公式求H_{j+1}，即

$$H_{j+1} = H_j + \frac{\delta_{j+1}\delta_{j+1}^T}{\delta_{j+1}^T\Delta g_{j+1}} - \frac{H_j\Delta g_{j+1}\Delta g_{j+1}^T H_j}{\Delta g_{j+1}^T H_j \Delta g_{j+1}}$$ ；

(9) $p_{j+1} = -H_{j+1}g_{j+1}$ $j = j+1$；

(10)返回步骤(4)；

(11)应用所得到的最优解 $\begin{bmatrix}\omega \\ \gamma\end{bmatrix}$ 进行分类。

而对于非线性问题，同样可将式(8)变换成无约束条件的数学规划问题(式(18))，再用 DFP-SVM 算法求解。

$$Min\frac{\nu}{2}\left\|e - D(K(A, A')D\mu - e\gamma)\right\|^2 + \frac{1}{2}\left\|\begin{bmatrix}\mu \\ \gamma\end{bmatrix}\right\| \tag{18}$$

4 实验结果与分析

采用 MIAS 库中的乳腺肿瘤 X 光图像，其中良性样本 85 个、恶性样本 15 个。首先对所有样本进行去噪、增强等预处理，然后提取纹理和肿瘤形状特征，在这些特征构成的特征库上进行分类。应用第 1 节中的算法(只有 83%的分类准确率)，在 15 个恶性样本中只有 1 个样本被正确分类，在 85 个良性样本中有 82 个被正确分类。而用 DFP-SVM 方法的分类准确率提高到 94%，其中 84 个良性样本和 10 个恶性样本被正确分类。实验证明，在小样本情况下，对分布极不平衡的数据，DFP-SVM 算法能减弱多样本类对少样本类的影响，从而具有较高的分类能力。

5 结束语

在进行基于肿瘤 X 光片的数据挖掘之前，对X光片进行处理，然后提取数字化肿瘤病灶的特征，应用数据挖掘算法进行计算辅助分析和决策。本文针对乳腺肿瘤的特点，提出应用基于可变形模型的方法分割肿瘤区域，提取了 5 个基于形状的特征，在这些特征构建的数据库上，采用一种 DFP-PSVM算法进行肿瘤的辅助诊断，取得了良好的效果。基于图像样本的分类精度的提高与图像的预处理、特征提取和选择也有着密切的关系[3,4]。特征选择将是今后研究的重点，获取最佳特征组合是提高分类精度和效率的重要途径，如纹理、形状和统计特征的筛选和有机组合。

参考文献

1 田　捷, 包尚联, 周明全. 医学影像处理与分析[M]. 北京: 电子工业出版社, 2003.

2 Cohen L D, Cohen I. Finite-element Methods for Active Contour Models and Balloons for 2-D and 3-D Images[J]. IEEE Trans. on Pattern and Machine Intell., 1993, 15(11): 1131-1147.

3 Xu C, Prince J L. Snakes, Shapes and Gradient Vector Flow[J]. IEEE Transactions on Image Processing, 1998, 7(3): 359-369.

4 Gupta L, Srinath M D. Contour Sequence Moments for Classification of Closed Planar Shapes[J]. Pattern Recognition, 1987, 20(3): 267-272.

5 Ranagayyan R M. Measures of Acutance and Shape for Classification of Breast Tumors[J]. IEEE Transaction on Medical Image, 1997, 16(6): 700-810.