SVM算法及其在乳腺X片微钙化点自动检测中的应用

万柏坤[1],王瑞平[1,2],朱 欣[1],綦宏志[1]
(1. 天津大学生物医学工程与科学仪器系,天津 300072;2. 清华大学生物医学工程系,北京 100080)

摘 要: 支持矢量机(SVM)是一种新的统计学习方法,其学习原则是使结构风险最小,而非经典学习方法所遵循经验风险最小原则.这使得SVM具有更强的泛化能力.并且,由于SVM求解的是凸二次优化问题,使之能保证所找到的极值解就是全局最优解.本文首次将SVM算法用于乳腺X影像微钙化点自动检测中,对临床实际病例的试用结果表明,同目前常用的基于经验风险最小的人工神经网络(ANN)分类方法相比,SVM具有更高的识别率,值得应用推广.

关键词: 支持矢量机; 结构风险最小; 经验风险最小; 微钙化点; 乳腺影像X片
中图分类号: TH776; TN911.73 **文献标识码:** A **文章编号:** 0372-2112 (2004) 04-0587-04

Principles of SVM and Its Application in Micro-calcifications Detection in Mammogram

WAN Bai-kun[1], WANG Rui-ping[1,2], ZHU Xin[1], QI Hong-zhi[1]
(1. *Dept of Biomedical Engineering, Tianjin University, Tianjin* 300072, *China*;
2. *Dept of Biomedical Engineering, Qinghua University, Beijing* 100080, *China*)

Abstract: Support vector machine (SVM) is a new statistical learning method. Compared with the classical machine learning methods, the learning discipline of SVM is to minimize the structural risk instead of empirical risk used in the learning discipline of classical methods, and SVM gives better generative performance. Because SVM algorithm is a convex quadratic optimization problem, the local optimal solution is certainly the global optimal one. In this paper, SVM algorithm is applied to detect the micro-calcifications in mammogram for the first time. The algorithm is tested with mammograms of clinical patients and results show that SVM method achieves a higher true positive in comparison with artificial neural network (ANN) based on the empirical risk minimization, and is valuable for application in clinical engineering.

Key words: support vector machine(SVM); structural risk minimization(SRM); empirical risk minimization(ERM); micro-calcification; mammogram

1 引言

乳腺X影像中微钙化点自动检测技术是指抽取有诊断价值的含微钙化点图像特征,并在此基础上进行特征优化,最后进行微钙化点的诊断分类.目前常用的分类方法有人工神经网络、模糊聚类、线性分类方法等.这些方法共同的理论基础是传统统计学,其研究的是样本数目趋于无穷大时的渐进理论.然而,在实际问题中,样本数往往是有限的,因此这些在理论上堪称优秀的分类方法在实际应用中表现却可能不尽人意[1].另外,这些方法数据处理能力差,片面强调克服训练错误,得到的可能是局部最优解,忽视了泛化性能的定量研究,产生的模型有时会产生过度拟合,或拟合程度较差现象.因此,寻找适合小样本的模式识别方法成为人们的研究目标[2].

支持矢量机(Support Vector Machine,SVM)是AT&Bell实验室V. Vapnik针对分类和回归问题(Classification and Regression),为适用于小样本学习问题而提出的通用学习算法[3].它根据VC(Vapnik-Chervonenkis)理论,基于结构风险最小(Structural Risk Minimization,SRM)原理[3],而非经验风险最小化(Empirical Risk Minimization,ERM)原理,从而能兼顾训练错误和泛化性能,开辟了机器学习算法的新天地.比目前常用的基于ERM原理的人工神经网络(Artificial Neural Network,ANN)的感知器学习算法性能更为优越.SVM用于模式识别问题,目的

收稿日期:2001-11-28;修回日期:2003-11-08
基金项目:天津市重点学科建设资金(No.津教委高[2000]-31)

是寻求泛化能力好的决策函数,即使是由有限训练样本得到的决策规则对独立的测试集仍能够得到小的误差.此外,SVM是求解凸二次优化问题,能够保证所找到的极值解就是全局最优解.目前,这种新的学习算法被建议用以替代多种传统的神经网络训练方法[3].本文以乳腺X影像中微钙化点自动检测为例,说明SVM算法的基本原理及其实用性能.

2 基于SVM的模式分类算法

经典统计学习方法遵循的是经验风险最小原则.而SVM的学习原则是使结构风险最小.对于微钙化点的自动检测而言,其检测结果需做出"是"或"非"的判断:"微钙化点"或"非微钙化点".因此,本文仅讨论将属于两个类型的训练向量进行分类的情况.下面以此为例简单介绍SVM原理和SVM算法.关于SVM的详细介绍可参考文献[4~6].

2.1 VC维

VC维是一种定量反映函数集或学习机器的复杂性或者说是学习能力的概念.模式识别方法中VC维的直观定义是:对一个指示函数集,如果存在 h 个样本能够被函数集中的函数按所有可能的 2^h 种形式分开,则称函数集能够把 h 个样本打散(shattering).函数集的VC维是指它所能打散的最大样本数目 h.若对任意数目的样本都有函数能将它们打散,则函数集的VC维是无穷大.VC维越大则学习机器越复杂.目前尚没有通用的关于任意函数集VC维计算的理论,只对一些特殊的函数集知道其VC维,例如在 n 维实数空间中线性分类器和线性实函数的VC维是 $n+1$.对于一些比较复杂的,其VC维除了与函数集有关外,还受学习算法等的影响,其确定非常困难.

2.2 SVM原理

假定在 n 维空间有一 l 点的特征集 D,分属于两类 $+1$ 和 -1,即:

$$D=\{(X_i,y_i)\mid i\in\{1,2,\cdots,l\},X_i\in R^n,y_i\in\{+1,-1\}\} \tag{1}$$

二元分类器将寻找函数 f 将特征点从数据空间映射到类空间:

$$f:R^n\to\{+1,-1\}$$
$$X_i\to y_i \tag{2}$$

假定有一分类器,它的任务是学习 $X_i\to y_i$ 的映射.该分类器实际上是在一组函数 $\{f(X,v)\}$ 中求一个最优的函数 $f(X,v^*)$,使得最优分类器对测试数据出现错误分类的期望值最小,即:

$$\min R(v)=\int_{\frac{1}{2}}|y-f(X,v^*)|dP(X,y) \tag{3}$$

其中,$P(X,y)$为(X,y)的联合概率分布,$R(v)$为期望风险.由于分布 $P(X,y)$未知,实际上 $R(v)$无法计算,一般的统计学习方法根据经验风险最小化原则找到一个近似,即经验风险 $R_{emp}(v)$,定义为在训练集上的平均错误率的测量:

$$R_{emp}(v)=\frac{1}{2l}\sum_{i=1}^{l}|y_i-f(X_i,v)| \tag{4}$$

然而,经验风险的最小值未必收敛于期望风险的最小值.Vapnik已经证明当且仅当 $R(v)$依概率一致收敛于(样本数目趋于无穷)$R_{emp}(v)$,并且当且仅当假设空间$\{f(X,v):v\in\Lambda\}$的VC维是有限的时候,经验风险最小化原理才成立[7].Vapnik深入研究后得出如下结论:经验风险 $R_{emp}(v)$和期望风险 $R(v)$之间以$(1-\eta)$的概率保证:

$$R(v)\le R_{emp}(v)+\sqrt{\frac{h(\ln\frac{2l}{h}+1)-\ln\frac{\eta}{4}}{l}} \tag{5}$$

其中 η 为经验风险 $R_{emp}(v)$ 的概率,h 称为假设空间$\{f(X,v):v\in\Lambda\}$的VC维,l 为样本数,而$\sqrt{\frac{h(\ln(2l/h)+1)-\ln(\eta/4)}{l}}$称为VC置信度.显然,式(4)中经验风险 $R_{emp}(w)$最小并不能保证期望风险 $R(w)$最小.为了克服这一缺陷,Vapnik提出了结构风险最小原理:为了达到期望风险最小,应设法使式(5)两边同时最小,即VC维 h 和经验风险 $R_{emp}(v)$同时最小.实现策略就是把函数集构造为一个函数子集序列,使各个子集按照VC维的大小(亦即VC置信度的大小)排列,在每个子集中寻找最小经验风险,在子集间折衷考虑经验风险和置信范围,取得实际风险的最小.SRM原则使分类器在训练集和测试集上具有较好的总体性能.

2.3 SVM分类算法

SVM是一种基于结构风险最小化的分类器,通过解二次规划问题,寻找将数据分为两类的最佳超平面.

对于线性可分的问题,要寻求参数(w,b),使

$$\begin{cases}w\cdot X_i+b\ge 1, & y_i=1\\ w\cdot X_i+b\le -1, & y_i=-1\end{cases} \tag{6}$$

超平面空间由所有的指示函数集(7)构成:

$$f_{w,b}=\operatorname{sign}(w\cdot X+b) \tag{7}$$

为减少分类平面的重复,对(w,b)进行如下约束:

$$\min_{i=1,\cdots,l}|w\cdot X_i+b|=1 \tag{8}$$

满足式(8)的超平面称为典型超平面.点 X 到(w,b)所确定超平面的距离为

$$d(X,w,b)=\frac{|w\cdot X_i+b|}{\|w\|} \tag{9}$$

式中$\|\cdot\|$运算指求其模值.根据约束条件(8),典型超平面到最近点的距离为$1/\|w\|$,使分类超平面到最近点的距离最大实际上就是对分类器泛化能力的控制,则求解最佳(w,b)参数的问题便可归结为二次规划问题:

$$\min_{w,b}\frac{1}{2}\|w\|^2\ s.t.\ y_i(w\cdot X_i+b)\ge 1\quad i=1,\cdots,l \tag{10}$$

优化的结果:

$$w^*=\sum_{i=1}^{l}\lambda_i^*y_iX_i \tag{11}$$

$$b^*=y_i-w^*\cdot X_i \tag{12}$$

其中,$\lambda_i,i=1,2,\cdots,l$ 为Lagrange优化方法在 $y_i(\omega\cdot X_i+b)\ge 1$ 约束条件下构造的Lagrange多项式的非负向量.满足式(12)约束条件的向量(X_i^*,y_i^*)被称为支持矢量(Support Vectors,SV),对应于分类边界上的点.这些点决定了分类的边界和分类器的性能.最佳学习模型参数为(λ^*,w^*,b^*).分类的决策函数可以记为:

$$f(X)=\mathrm{sign}\left(\sum_{i=1}^{l}y_i\lambda_i^*(X\cdot X_i)+b^*\right) \tag{13}$$

对于非线性分类的情况,通过非线性变换转化为某个高维空间中的线性问题,在变换空间求最优分类面.首先需要对输入向量进行映射:

$$X\to\phi(X)=(\alpha_1\phi_1(X),\alpha_2\phi_2(X),\cdots,\alpha_n\phi_n(X),\cdots) \tag{14}$$

经过映射后的 SVM 决策函数为

$$f(X)=\mathrm{sign}(\phi(X)\cdot w^*+b^*)$$

$$=\mathrm{sign}\left(\sum_{i=1}^{l}y_i\lambda_i^*\phi(X)\cdot\phi(X_i)+b^*\right) \tag{15}$$

在式(15)中,定义核函数 $k(X,y)$:

$$k(X,y)=\phi(X)\cdot\phi(y) \tag{16}$$

核函数 $k(X,y)$代入式(15),得到非线性分类的 SVM 决策函数:

$$f(X)=\mathrm{sign}\left(\sum_{i=1}^{l}y_i\lambda_i^*k(X,X_i)+b^*\right) \tag{17}$$

常用的核函数有多项式、高斯径向基函数、指数型径向基函数、多层感知器样条函数、张量积核函数等.本文用的是多项式核函数.

2.4 基于 SVM 分类算法的微钙化点检测

为了检验 SVM 在乳腺影像中微钙化点自动检测的效果,本文选择了天津医科大学附属肿瘤医院 10 名患者的 414 个微钙化点.真实微钙化点像素由富有经验的乳腺科专家确认.其中 200 个微钙化嫌疑样点(66 微钙化点,134 非微钙化点)为训练例,214 个微钙化嫌疑样点(70 例微钙化点,144 非微钙化点)为测试例.SVM 算法的输入向量为微钙化点的特征向量.本文取经小波去噪后由尺度 4 和尺度 5 重建图像的归一化像素值、对比度、方差、面积五个特征参数组成特征矢量.模式识别过程分为机器学习和测试,以及模式的在线识别三个步骤.

首先,需对微钙化点进行机器学习.学习目的是找到优化参数(w,b),再根据式(17)对待检测的乳腺 X 片进行微钙化点检测.选择输入向量映射核的类型,并计算核函数 $k(X,y)$;通过二次规划计算出(λ^*,w^*,b^*),根据测试的结果,对所选择的核进行调整,将最佳的学习模型(λ^*,w^*,b^*)存入数据库中,以便在线检测时调用(如图 1 所示).检测方法如图 2 所示,其中包括图像预处理,特征提取,最后调用 SVM 模型,检测微钙化点等流程.图像预处理包括数字化乳腺影像的规格化处理和感兴趣区域的提取[8].

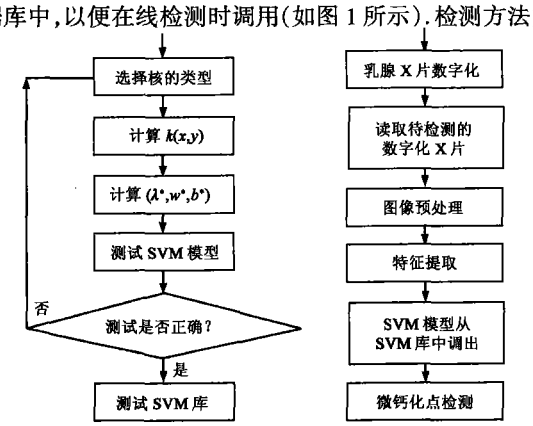


图 1　SVM 模型训练程序框图

图 2　微钙化点的在线识别

3 实验结果

本文分别用 ANN 和 SVM 算法对 214 个微钙化点进行检测,结果如图 3 和图 4 所示(图中纵坐标表示分类结果输出值:+1.0 为微钙化点,-1.0 为非微钙化点).由图可以看出,ANN 算法的误检率为 10/214(图 3),而 SVM 算法则使之降为 1/214(图 4).显然,SVM 具有更准确的检出率.

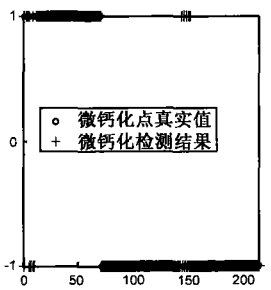


图 3　ANN 对 214 个微钙化点的分类结果

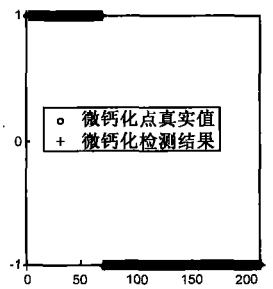


图 4　SVM 对 214 个微钙化点的分类结果

4 讨论

本文首次应用 SVM 算法实现了针对乳腺 X 线影像中微钙化点自动检测的分类器.按照 SVM 思想,这个分类器经过训练之后,蕴涵了训练集合中对分类起作用的那些样本的信息.微钙化点检测的核心问题就是找到一个分类器,用来对一幅待测的乳腺 X 胶片找出微钙化点.实际上,本文所设计的这个分类器具有很好的普适性,只要改变训练数据,通过训练,一般而言,该分类器能对任何可分类的模式进行分类.本文在微钙化点检测中,还通过对数据进行了一些预先处理,以及精心选取训练样本,使其更具有代表性,从而使分类器具有更好的泛化能力,这对达到更好的分类效果是十分必要的.总之,对 SVM 这种新的统计学习分类方法,值得进一步研究并推广其应用.

参考文献:

[1] 张学工.关于统计学习理论与支持矢量机[J].自动化学报,2000,26(1):32-42.

[2] Vladimir N Vapnik. An overview of statistical learning theory[J]. IEEE Trans. On Neural Network, 1999, 10(5): 988-999.

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

[4] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995: 1-15.

[5] S R Gunn. Support Vector Machines for Classification and Regression [R]. Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.

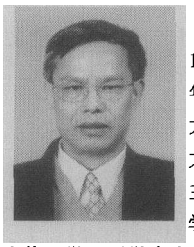
[6] M O Stitson, J A E Weston, A Gammerman, et al. Theory of Support

Vector Machines. Technical Report[R], CSD-TR-96-17. Department of Computer Science Egham, Surrey TW20 0EX, England, Royal Holloway University of London, December 31, 1996.

[7] V Vapnik, A Y Chervoknenkis. On the uniform convergence of relative frequencies events to their probabilities[J]. Theory of Probable and Its Application, 1971, 16(2): 263 - 280.

[8] 王瑞平, 万柏坤, 朱欣. 乳腺癌早期诊断的计算机处理研究[J]. 天津大学学报, 2002, 35(4): 497 - 500.

作者简介:



万柏坤 男, 1945 年 12 月出生于江西南昌, 1968 年毕业于中国科技大学核电子学专业, 1982 年获中科院理学硕士学位, 先后在日本东京工业大学和东京医科齿科大学作访问学者, 现为天津大学生物医学工程系教授、博士生导师、系主任, 主要从事生物医学信息检测与处理, 任中国电子学会医学图像处理与分析专业委员会委员、中国生物医学工程学会生物信息与控制分会委员、教育部生物医学工程教学指导委员会委员、美国纽约科学院会员(ID # 469859-7).

王瑞平 女, 1974 年 9 月出生于内蒙巴蒙, 清华大学生物医学工程系博士后, 于 1997 年、2000 年在天津医科大学获得"生物医学工程"工学学士学位和医学硕士学位, 2003 年在天津大学获得"生物医学工程"工学博士学位, 研究方向为生物医学图像处理和模式识别.