文章编号: 1001-0920(2003)03-0358-03

支持向量机和最小二乘支持向量机的比较及应用研究

阎威武, 邵惠鹤
(上海交通大学 自动化系, 上海 200030)

摘 要: 介绍和比较了支持向量机分类器和最小二乘支持向量机分类器的算法。并将支持向量机分类器和最小二乘支持向量机分类器应用于心脏病诊断, 取得了较高的准确率。所用数据来自 UCI benchmark 数据集。实验结果表明, 支持向量机和最小二乘支持向量机在医疗诊断中有很大的应用潜力。

关键词: 支持向量机; 分类器; 诊断
中图分类号: TP181 文献标识码: A

Application of support vector machines and least squares support vector machines to heart disease diagnoses

YAN Wei-wu, SHAO Hui-he
(Department of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Nonlinear classifiers algorithms of standard support vector machines (SVM) and least squares support vector machines (LS SVM) are discussed and compared. Then standard SVM nonlinear classifiers and LS SVM nonlinear classifiers are applied to heart disease diagnoses based on UCI benchmark data set. Comparing with other result, high accuracy rate is obtained in the prediction. Application of SVM and LS SVM to disease diagnoses indicates that SVM and LS SVM have potential application in medical.

Key words: Support vector machine; Classifiers; Diagnoses

1 引 言

统计学习理论是由 Vapnik[1] 建立的一种专门研究小样本情况下机器学习规律的理论, 支持向量机是在这一理论基础上发展而来的一种新的通用学习方法。支持向量机通过结构风险最小化原理来提高泛化能力, 它较好地解决了小样本、非线性、高维数、局部极小点等实际问题, 已在模式识别、信号处理、函数逼近等领域得到了应用[2, 3]。最小二乘支持向量机是支持向量机的一种扩展, 为了区别, 本文将 Vapnik 支持向量机称为标准支持向量机。并在比较标准支持向量机非线性分类器[3]和最小二乘支持向量机非线性分类器[4]的同时, 将它们用于心脏病的诊断。

2 标准支持向量机和最小二乘支持向量机的非线性分类器

支持向量机主要是基于如下思想: 首先选择一非线性映射 $\Psi(\cdot)$ 把 n 维样本向量 $(x_1, y_1), \cdots, (x_l, y_l) \in R^n \times \{+1, -1\}$ 从原空间 R^n 映射到特征空间, 在此高维特征空间中构造最优线性决策函数 $y(x) = \mathrm{sgn}[w \cdot \Psi(x) + b]$。在构造最优决策函数时, 利用了结构风险最小化原则, 同时引入了间隔的

收稿日期: 2001-10-09; 修回日期: 2002-01-04。
基金项目: 国家 973 重点基础研究发展基金资助项目(G1998030415)。
作者简介: 阎威武(1972—), 男, 甘肃陇西人, 博士生, 从事机器学习、数据挖掘等研究; 邵惠鹤(1936—), 男, 浙江宁波人, 教授, 博士生导师, 从事工业过程控制、智能控制等研究。

概念。并巧妙地利用原空间的核函数取代了高维特征空间的点积运算, 避免了复杂计算。

标准支持向量机和最小二乘支持向量机在利用结构风险原则时, 在优化目标中选取了不同的损失函数, 它们分别为误差 ξ_i(允许错分的松弛变量)和误差 ξ_i 的二范数。

对标准支持向量机, 优化问题为

$$\min J(w,\xi) = \frac{1}{2}w \cdot w + c\sum_{i=1}^{l}\xi_i$$

$$\text{s.t } y_i[(\Psi(x_i)\cdot w + b)] \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1,2,\cdots,l \tag{1}$$

对最小二乘支持向量机, 优化问题为

$$\min J(w,\xi) = \frac{1}{2}w \cdot w + c\sum_{i=1}^{l}\xi_i^2$$

$$\text{s.t. } y_i[(\Psi(x_i)\cdot w + b)] = 1 - \xi_i$$

$$i = 1,2,\cdots,l \tag{2}$$

用拉格朗日法求解上述优化问题, 标准支持向量机优化问题转化为下面的二次规划

$$\max W(a) = -\frac{1}{2}\sum_{i,j=1}^{l}a_i y_i y_j K(x_i,x_j)a_j + \sum_{i=1}^{l}a_i$$

$$\text{s.t } \sum_{i=1}^{l}a_i y_i = 0$$

$$0 \leq a_i \leq c, \quad i = 1,2,\cdots,l \tag{3}$$

最小二乘支持向量机优化问题转化为求解线性方程

$$\begin{bmatrix} 0 & y_1 & \cdots & y_l \\ y_1 & y_1y_1K(x_1,x_1)+1/c & \cdots & y_1y_lK(x_1,x_l) \\ \vdots & \vdots & \ddots & \vdots \\ y_l & y_ly_1K(x_l,x_1) & \cdots & y_ly_lK(x_l,x_l)+1/c \end{bmatrix} \times \begin{bmatrix} b \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{4}$$

3　基于支持向量机的心脏病诊断

心脏病诊断系统是根据病理检测结果来诊断病人的心脏状况。本文将标准支持向量机分类器和最小二乘支持向量机分类器应用于心脏病诊断。所用的数据样本可从 UCI 机器学习问题库得到[4]。该数据库有303个样本, 病理检测有75项, 心脏病状况分为 5 类(Value 0 1 2 3 4)。为简单起见, 在实际诊断时, 只利用病理检测中的 13 项: Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal(详细说明略)。心脏病状况分为两类: 有(Value 1, 2, 3, 4) 和无(Value 0)。这样, 每个数据样本包括 13 个属性, 所有数据样本被分为两类。

3.1　数据预处理

首先对属性值进行归一化

$$\bar{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{5}$$

归一化后的属性值 $\bar{x}_i \in [0,1]$。"有"用 + 1 表示, "无"用 - 1 表示, 这样问题就抽象为属性集 A 到分类集 C 的映射, $A \times C \rightarrow [0,1] \times \{+1,-1\}$。

3.2　核函数

本研究中核函数选取径向基函数

$$K(x,x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right) \tag{6}$$

其中: $|x - x_i| = \sqrt{\sum_{k=1}^{n}(x^k - x_i^k)^2}$, σ 为核宽度。

3.3　结果及讨论

从数据库中, 选取 200 个样本作为训练集, 剩余的样本作为测试集。所有的实验在 1 台 Pentium 350 MHz 内存 128 MB 的计算机上进行, 语言使用 Matlab 6.0。

训练和测试的结果见表 1, 与其他方法所得结果的比较见表 2。标准支持向量机分类器和最小二乘支持向量机分类器的测试仿真分别如图 1 和图 2 所示, 图中圆代表分类器的分类结果, 叉代表样本的实际值。如果圆和叉重叠, 则分类是准确的; 如果圆和叉不重叠, 则分类是错误的。在图 1 和图 2 中, 上面的子图为测试集的测试结果, 中间和下面的子图为训练集的测试结果。

表 1　训练和测试的结果

		准确率 /%	运行时间 /s
SVM	训练集	90.8 ± 1.27	120
	测试集	83.5 ± 1.92	
LS SVM	训练集	90.6 ± 2.11	80
	测试集	81.5 ± 2.29	

表 2　与其他方法的比较

方　法	准确率 /%
NTgroeth[5]	77
C4[5]	74.5
Classit[5]	78.9
SVM	83.5
LS SVM	81.5

从表 1 可见, 标准支持向量机分类器比最小二乘支持向量机分类器具有更高的准确率, 而最小二乘支持向量机分类器收敛速度更快。标准支持向量机求解一个凸二次规划, 所得的解是唯一的最优解, 但当数据量较大时, 求解过程所需计算资源很大。最小二乘支持向量机求解一个线性方程, 虽然不能保

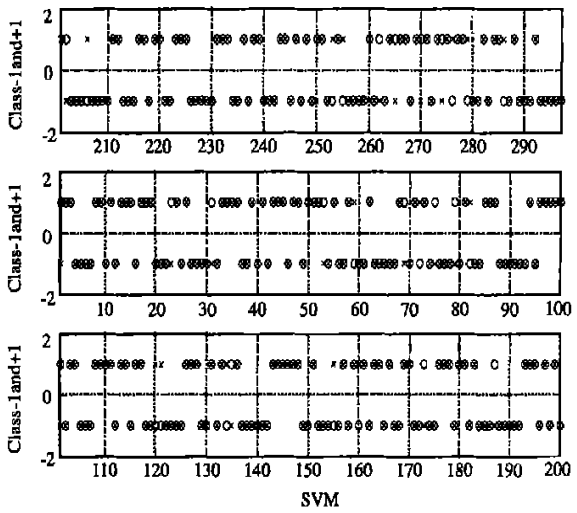

图 1 标准支持向量机分类器的测试仿真

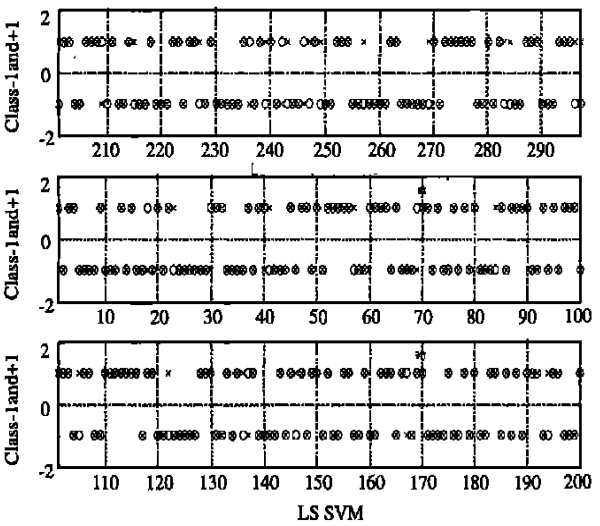

图 2 最小二乘支持向量机分类器的测试仿真

证其解为所求问题的全局最优解, 但线性方程的求解速度要比二次规划快, 且所需计算资源少。

通过与其他方法所得结果[5] 比较发现, 标准支持向量机分类器和最小二乘支持向量机分类器均具有较高的准确率。这说明支持向量机能较好地解决小样本、非线性等实际问题, 具有很强的泛化能力。二者在心脏病的诊断等医疗诊断中具有很大应用潜力。

4 结 论

本文比较了标准支持向量机分类器和最小二乘支持向量机分类器, 并将它们用于心脏病诊断。标准支持向量机分类器和最小二乘支持向量机分类器均具有较高的准确率。标准支持向量机求解一个凸二次规划, 其解是唯一的且为最优解, 这样不存在一般神经网络的局部极值问题。最小二乘支持向量机求解线性方程, 其解满足极值条件, 但不能保证是全局最优解, 最小二乘支持向量机具有更快的求解速度, 求解所需的计算资源较少。两者能较好地解决小样本、非线性等实际问题, 在医疗诊断中均具有很大应用潜力。

参考文献(References):

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

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

[3] Vapnik V N. *The Nature of Statistical Learning Theory*[M]. New York: Springer-Verlag, 1999.

[4] Probenl L P. A set of neural network benchmark problem and benchmark rules[R]. Germany: University Karlsruhe, 1994.

[5] Turney P D. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm[J]. *J of Artificial Intelligence Research*, 1995, 2: 369-409.

(上接第 357 页)

[7] Lehner P, Connor M F, Sak S, et al. Cognitive biases and time stress in team decision making[J]. *IEEE Trans on Systems, Man and Cybernetics, Part A*, 1997, 27(5): 698-703.

[8] Song A. Mathur A, Pattipati K R. Design of process parameters using robust design techniques and multiple criteria optimization[J]. *IEEE Trans on Systems, Man and Cybernetics*, 1995, 25(11): 1437-1446.

[9] Handley H A, Zaidi Z R, Levis A H. The use of simulation models in model driven experimentation[A]. *Proc of the* 1999 *Command & Control Research & Technology Symposium*[C]. Newport, 1999.