

基于统计学习理论的支持向量机的分类方法

杨 斌, 路 游

(中国石油大学 计算机系, 北京 102249)

摘 要: 支持向量机是一种新型机器学习方法, 由于其出色的学习性能, 该技术已成为机器学习领域新的研究热点。介绍用于分类的支持向量机的统计学习理论基础, 在此基础上提出了支持向量机的分类算法, 讨论了支持向量机存在的问题, 对用于分类的支持向量机的应用前景进行了展望。

关键词: 支持向量机; 统计学习理论; 结构风险最小化; 数据分类

中图分类号: TP182

文献标识码: A

文章编号: 1673-629X(2006)11-0056-03

Classification Method of Support Vector Machine Based on Statistical Learning Theory

YANG Bin, LU You

(Department of Computer Science and Technology, China University of Petroleum, Beijing 102249, China)

Abstract: Support vector machines are a kind of novel machine learning method, which have become the hotspot of machine learning because of their excellent performance. In this paper, the elements of statistical learning theory for support vector machines used in classification and algorithms are introduced. The main issues of support vector machine are discussed, and the application foreground of support vector machine is prospected.

Key words: support vector machine; statistical learning theory; structural risk minimization; data classification

0 引 言

基于数据的机器学习是现代智能技术中的重要方面, 研究从观测数据(样本)出发寻找规律, 利用这些规律对未来数据或无法观测的数据进行预测。传统的机器学习的神经网络方法基于经验风险最小化原则(Empirical Risk Minimization, 简称 ERM), 泛化能力较差, 其网络结构选择存在过学习和局部极小点等问题, 目前无法克服^[1]。支持向量机(SVM)是 Vapnik 等人提出的一类新型机器学习方法^[2,3], 是以统计学习理论为基础的, 因而具有严格的理论和数学基础, 与神经网络的学习方法相比, 支持向量机是基于结构风险最小化(Structural Risk Minimization)原则, 保证了学习机器具有良好的泛化能力, 由于支持向量算法最终可转化为凸优化问题, 保证了算法的全局最优性, 避免了神经网络无法解决的局部最小问题^[4]。由于其出色的学习性能, 该技术已经成为机器学习界的研究热点, 随着研究的深入, SVM 已推广到多类分类问题中, 并展现了良好的学习和泛化性能。文中介绍用于分类的

支持向量机的理论基础, 其次提出了支持向量机的分类算法, 并分析了目前支持向量机存在的一些问题, 对其应用前景进行了展望。

1 支持向量机的统计学习理论基础

机器学习的目的是根据给定的训练样本求对某系统的输入输出之间依赖关系的估计, 使它能够对未知数据做出尽可能准确的估计。机器学习问题可以形式化地表示为: 输入变量与输出变量之间存在某种未知依赖关系, 即存在一个未知的联合概率 $p(x, y)$, 机器学习根据 n 个独立同分布观测样本:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \quad (1)$$

从给定的函数集 $F(x, w)$ 中选择具有最佳权值向量 w 的函数对依赖关系进行估计, 使实际响应的“最佳”逼近。函数逼近的质量常用损失函数或者偏差函数 $L(y, F(x, w))$ 表示。损失函数 $L(y, F(x, w))$ 的期望值定义为风险泛函:

$$R(w) = \int L(y, F(x, w)) dp(x, y) \quad (2)$$

式中, $p(x, y)$ 是输入向量 x 和期望向量 y 的联合概率分布。

学习的目的就是使风险函数最小。式(2)定义的期望风险函数最小化必须依赖关于联合概率 $p(x, y)$ 的信息。但是, 在实际的机器学习问题中, 只能利用样本式(1)的

收稿日期: 2006-03-14

基金项目: 中国石油大学科技创新基金资助(05C088)

作者简介: 杨 斌(1981-), 男, 安徽滁州人, 硕士研究生, 研究方向为机器学习、智能信息处理; 路 游, 副教授, 博士, 研究方向为机器学习。

信息,因此期望风险函数无法直接计算和最小化。根据概率论中大数定理的思想,人们自然想到用算术平均代替式(2)中的数学期望,于是定义

$$R_{\text{emp}}(w) = \frac{1}{N} \sum_{i=1}^N L(y_i; F(x_i, w)) \quad (3)$$

来逼近式(2)定义的期望风险函数。由于 $R_{\text{emp}}(w)$ 是用已知的训练样本(即经验数据)定义的,因此称作经验风险^[5](Empirical Risk Minimization,简称 ERM)。

但实际上得到的样本数是有限的,在样本数目有限的情况下,不能保证有好的预测效果,因此,需要一种能够指导人们在小样本情况下建立有效的学习和推广性理论。

根据统计学习理论,在二分类的情况下,经验风险和实际风险之间以概率 $1 - \eta$ 存在如下关系:

$$R(w) \leq R_{\text{emp}}(w) + \sqrt{\frac{h(\ln(2n/h) + 1) - \ln(\eta/4)}{n}} \quad (4)$$

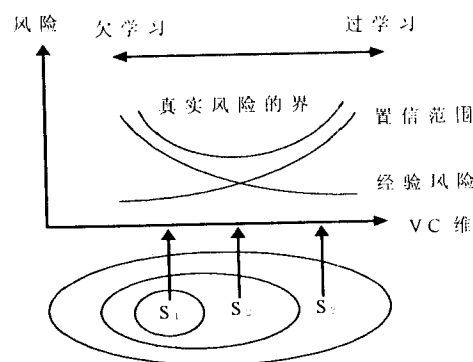
其中, h 是 VC 维, n 是样本数。

上式表明,学习机的实际风险由经验风险和置信区间两部分组成,它和学习集的 VC 维和训练样本数有关。可以简单地表示为:

$$R(w) = R_{\text{emp}}(w) + O(h/n) \quad (5)$$

上式表明,在有限训练样本的情况下,即使 $R_{\text{emp}}(w)$ 较小,也不能保证真实风险 $R(w)$ 取最小值。因此,希望找一种新的方法使 $R(w)$ 最小。

ERM 准则只强调经验风险最小(训练误差),没有最小化置信范围值,因此基于 ERM 准则的学习方法的学习能力强,但泛化能力较差,导致出现过学习现象,例如神经网络。最大化泛化能力不仅需要最小化经验风险,而且应最小化置信范围值。基于此思想,统计学习理论提出一种新的策略,即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,取得实际经验风险最小,这种思想称作结构风险最小化或有序风险最小化^[6](Structural Risk Minimization,简称 SRM)准则,如图 1 所示。



函数集子集: $S_1 \subset S_2 \subset S_3$, VC 维: $h_1 \leq h_2 \leq h_3$

图 1 结构风险最小化

SVM 是结构风险最小化思想的具体实现,它不像神经网络等传统方法那样以训练误差最小化作为优化目标,

而是以训练误差作为优化问题的约束条件,以置信范围值最小化作为优化目标。

2 用于分类的支持向量机

2.1 SVM 的基本思想

定义最优线性超平面,并把寻找最优线性超平面的算法归结为求解一个凸规划问题。进而基于 Mercer 核展开定理,通过非线性映射 φ ,把样本空间映射到一个高维乃至无穷维的特征空间(Hilbert 空间),使在样本空间中可以应用线性学习机的方法解决样本空间中的高度非线性分类问题。简单地说就是升维和线性化^[7]。

2.2 线性可分的最优分类面

支持向量机是从线性可分情况下的最优分类面发展而来的,也是统计学习理论中最实用的部分,考虑如图 2 的一个用某特征空间的超平面对给定训练数据集作二分类的问题。给定一组训练样本集 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$,其中 $x_i \in R^N$ 为 N 维向量, $y_i \in \{-1, 1\}$ 在线性可分的情况下,在特征空间中可以构造多个分割平面(如: H_1, H_2, \dots),这个超平面被定义为:

$$(w \cdot x) + b = 0 \quad (6)$$

同时,这个分类面能将两类 $(1, -1)$ 无误差地完全分开,即满足:

$$\begin{aligned} (w \cdot x_i) + b &\geq 1, \text{ for all } x_i \in 1 \\ (w \cdot x_i) + b &\leq -1, \text{ for all } x_i \in -1 \end{aligned} \quad (7)$$

在所有的分类面内,要寻找的是最优超平面,这个最优超平面是指满足两类的分类空隙 dist 最大,即每类距离超平面最近的样本到超平面的距离之和最大。这个距离被称为边(Margin),可以证明:

$$\text{dist} = \frac{2}{\|w\|} \quad (8)$$

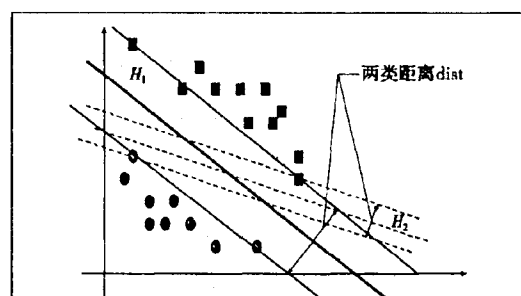


图 2 线性可分情况下的最优分类面

根据以上分析,求解最优超平面就相当于在式(7)的约束条件下,求式(8)的最大值,这样建立线性支持向量机的问题转化为求解如下的一个二次凸规划问题:

$$\begin{cases} \min \frac{1}{2} \|w\|^2 \\ \text{s.t. } y_i((w \cdot x_i) + b) \geq 1 \end{cases} \quad (9)$$

该约束优化问题可以用 Lagrange 方法求解,得到最优超平面决策函数为:

$$M(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i (x \cdot x_i) + b^*\right) \quad (10)$$

根据 Vapnik^[8] 等的分析,判定分类面函数的 VC 维存在如下的定理:假设训练样本完全包含在一个最大直径为 D_{\max} 的球内,不同类别样本之间的最小边际距离是 M_{\min} ,则分类面函数的 VC 维 h 满足

$$h \leq D_{\max}^2 / M_{\min}^2 + 1 \quad (11)$$

可见,SVM通过最大化边际距离 M_{\min} ,实现对 VC 维大小的控制,降低模型复杂度,从而体现 SRM 原理。

2.3 线性不可分的广义最优分类面

考虑到可能存在一些样本不能被超平面正确分类,即对线性不可分情况,可以引入松弛变量 $\xi_i \geq 0, i = 1, 2, \dots, l$,得到新的凸规划问题:

$$\begin{cases} \min & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t.} & y_i((w \cdot x_i) + b) \geq 1 - \xi_i \quad (i = 1, \dots, l) \\ & \xi_i \geq 0 \quad (i = 1, \dots, l) \end{cases} \quad (12)$$

求解问题(12)与求解问题(9)本质上是一样的。得到的最优超平面决策函数仍然为:

$$M(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i (x \cdot x_i) + b^*\right)$$

对于多类线性分类问题的一种解决办法是把它转化为多个二类线性分类问题解决。 K 类分类问题可以转化为 K 个二类划分问题。其中每个二类划分都是判断样本点属于第 i 类或不属于第 i 类。

2.4 高维空间的最优分类面

对于空间 L 内非线性分类问题,可以通过一非线性变换 $\Phi(x)$,将数据 x 从原空间 L 映射到一个高维特征空间 H ,再在空间 H 建立最优分类面。这时的分类函数是:

$$M(x) = \text{sgn}((w^* \cdot \Phi(x)) + b^*) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i (\Phi(x_i) \cdot \Phi(x)) + b^*\right) \quad (13)$$

这里只是用 $\Phi(x)$ 和 $\Phi(x_i)$ 代替了 x 和 x_i ,因此计算过程相同。根据 Mercer 定理知由点积定义的核必是 Mercer 核: $K(x, y) = (\Phi(x) \cdot \Phi(y))$,则上式可以化简为:

$$M(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i K(x_i, x) + b^*\right) \quad (14)$$

这种核函数的变换处理,为支持向量机提供了极大的灵活性,使其有了更广泛的应用范围。常见的核函数类

型有:多项式核函数、径向基函数 RBF、样条核函数。

3 结 论

支持向量机是基于统计学习理论的新的机器学习方法,具有严格的理论基础,能够较好地解决小样本、非线性、高维数和局部最小点等问题,在许多问题上它有着其他统计学习方法难以比拟的优越性,支持向量机在模式识别(字符识别、文本自动分类、人脸检测、头的姿态识别)、函数逼近、时间序列预测、故障识别和预测、信息安全、电力系统及电力电子等方面都有很好的应用前景,因此成为 20 世纪 90 年代末发展最快的研究方向之一。文中深入推导了用于解决分类问题的 SVM 方法,与其它方法相比,支持向量机具有泛化性强、效率高等特点。但由于支持向量机是一种尚未成熟的新技术,它目前仍有很多局限,其最大的局限就是核函数的选择和参数的确定,虽然目前已有一些研究者对使用先验知识选择核进行了研究,但对于特定问题选择最佳的核仍是一个难以解决的问题;另一方面,支持向量机的训练速度极大地受到训练集规模的影响;此外,支持向量机对多类问题的处理能力仍有待进一步研究和改善。

参考文献:

- [1] Vapnik V. The nature of statistical learning theory[M]. New York: Springer Verlag, 1995.
- [2] Cortes C, Vapnik V. Support Vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.
- [3] Cortes C, Vapnik V. Support vector networks[J]. Machine learning, 1995, 20(1): 273-297.
- [4] Haykin S. 神经网络原理[M]. 叶世伟,史忠植,译. 北京:机械工业出版社,2004.
- [5] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
- [6] 边肇祺,张学工. 模式识别(第2版)[M]. 北京:清华大学出版社,2000: 284-303.
- [7] 陈永义. 处理非线性问题分类和回归问题的一种新方法——支持向量机方法简介[J]. 应用气象学报, 2004(2): 345-354.
- [8] Vapnik V. Statistical learning theory[M]. New York: John Wiley & Sons, 1998.

(上接第 55 页)

开发的 VoIP 系统就是利用这种机制成功地完成了 NAT 穿透。经过反复的测试,此平台性能稳定,适用了各种中小营运商的需要。

参考文献:

- [1] RFC3261. SIP: Session Initiation Protocol[S]. 2002.
- [2] RFC1631. NAT: The IP Network Address Translator[S].

2002.

- [3] Collins D. VoIP 技术与应用[M]. 舒华英,李 勇,等译. 北京:人民邮电出版社,2003.
- [4] Camarillo G. SIP 揭密[M]. 白建军,彭 晖,田 敏,等译. 北京:人民邮电出版社,2003.
- [5] RFC3266. Support for IPv6 in Session Description Protocol [S]. 2002.