第26卷第3期 长春师范学院学报(自然科学版) 2007年6月
Vol.26 No.3 Journal of Changchun Normal University(Natural Science) Jun.2007

支持向量机的算法研究

方 辉[1]，王 倩[2]
(1.渤海大学信息科学与工程学院，辽宁锦州 121013；
2.锦州市绿野环境工程开发中心，辽宁锦州 121000)

[摘 要] 支持向量机（support vector machine，SVM）是20世纪90年代发展起来的一种新型机器学习方法，是在统计学习理论基础上发展起来的一种新的数据挖掘方法，已广泛应用于模式识别与回归分析。并已成为国际机器学习界的研究热点。本文主要讨论其基本原理与SVM训练算法。

[关键词] 支持向量机；机器学习；分类
[中图分类号] TP274 [文献标识码] A [文章编号] 1008-178X(2007)03-0090-02

支持向量机是一种用来解决分类和回归问题的新的数据挖掘技术。由于SVM方法具有许多引人注目的优点和有前途的实验性能，越来越受重视，该技术已成为机器学习研究领域的热点，并已取得良好的效果，如手文本分类、手写识别、图像分类等。

支持向量机是在统计学习理论基础上发展起来的。1995年Vapnik和Chervonenkis提出了完整的统计学理论，并在此基础上发展了一种新的通用学习方法——SVM。

1 基本原理

支持向量机来源于分类问题，从本质上讲是一种前向神经网络，根据结构风险最小化准则，在使训练样本分类误差极小化的前提下，尽量提高分类器的泛化推广能力。

支持向量机的关键在于核函数，低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，我们就可以得到高维空间的分类函数。

2 支持向量机特点

利用最大间隔的思想降低分类器的VC维，实现结构风险最小化原则；利用Mercer核实现线性算法的非线性化；稀疏性，即少量样本（支持向量）的系数不为零，就推广性而言，较少的支持向量数在统计意义上对应好的推广能力，从计算角度看，支持向量减少了核形式判别式的计算量；算法设计成凸二次规划问题，避免了多解性。

3 SVM训练算法的研究

3.1 标准的SVM算法 标准的SVM算法的原始问题可以归结为如下的一个二次规划问题：

$$\min_{w,b} \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.} \quad y_i(<w, g(x_i)) + b) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0, i = 1, \cdots, l. \tag{1}$$

其中，$x_i \in R^n$ 为支持向量机的输入指标向量，$y_i \in \{-1, 1\}$ 为 x_i 所属类别，$i = 1, \cdots\cdots, l$。$K(x_i, x_j)$ 为核函数，它对应某特征空间Z中的内积，即 $K(x_i, x_j) = \langle g(x_i), g(x_j) \rangle$，变换 $g: x \to z$ 将样本从输入空间映射到特征空间。w为超平面的法向量，b为超平面的偏置，ξ_i 是松弛变量，C为惩罚因子。

对于支持向量机的训练并不是求解它的原始问题（1），而是求解它的对偶问题，其对偶问题描述如下：

$$\min_{\alpha} p(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

$$\text{s.t.} \quad y^T\alpha = 0,$$

$$0 \leq \alpha_i \leq C, \ i = 1, \cdots, l \tag{2}$$

[收稿日期] 2007-03-28

[作者简介] 方 辉（1980-），男，辽宁鞍山人，渤海大学信息科学与工程学院硕士研究生，从事数据挖掘研究。

其中：Hessian 矩阵 Q 是半正定的，$Q_{ij} = y_i y_j < g(x_i), g(x_j) > = y_i y_j K(x_i, x_j)$；$e = (1, 1, \cdots, 1)^T$；$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)^T$，$\alpha_i$ 是不等式约束 $y_i (w \cdot f(x_i) - b) \geq 1 - \xi_i$ 对应的 Lagrange 乘子。

训练支持向量机实际上就是求解如试（2）所示的一个二次规划问题。解决如此的优化问题在理论上并不困难，可以用牛顿法、拟牛顿法、梯度投影法、内点法等优化技术去解决，但原始 QP 方法在每次迭代时需要利用整个核矩阵去更新 *。而 Q 为一个 1 * 1 矩阵，并且不是稀疏阵，当数据量较大时，Q 占用的内存大大超过了目前计算机的内存容量。因此如何使 SVM 对大规模样本集的训练能力得以提高成为了现金 SVM 研究的重要问题。

3.2 算法的改进　分解算法最早是由 Osuna 提出的，后来又经 Joachims t. 等人对其进行改进。Platt 于 1998 年提出了 SMO（sequential minimal optimaization）算法，该算法可以说是 Osuna 分解算法的一个特例，工作集中只有 2 个乘子，其优点是针对 2 个乘子的二次规划问题可以有解析解的形式，避免了每次迭代中调用标准的优化算法。从而使原问题通过分析的方法加以快速解决，大大提高了运算速度。

为了进一步提高 SMO 算法的运行效率和收敛速度，研究人员对此做出了巨大的努力，提出了很多改进办法。文献［1］使用了双阈值方法对原有的 SMO 算法在最优性条件及工作集选择做出了改进，克服了 SMO 算法采用单一阈值所带来的可能使算法进入混乱状态和使算法变得不够高效的缺点。在工作集选择方面他提出了两种改进策略，第一种工作集选择算法类似于 SMO 算法中的工作集选择；第二种工作集选择算法就是最大违反对法。实验表明这种算法明显比原来的 SMO 算法更加高效，同时在理论上对该算法的收敛性 Keerthi 也给出了说明。文献［6］中说明了最大违反对法是利用了目标函数的一阶近似信息，为了选择更好的训练集，它提出了一种利用二阶近似信息来选取工作集的快速训练算法，并且在理论上证明了该算法是线性收敛的，实验表明该算法比最大违反对法收敛更加快速。文献［7］给出了 SMO 算法的一种可行方向解释，在综合考虑与工作集相关的目标函数的下降量和计算代价的情况下，提出了一种收益代价平衡的工作集选择法，提高了缓存的效率，试验证明该方法特别适用于样本较多、支持向量较多、非有界支持向量较多的情况。

4 结束语

作为一种尚未成熟的技术，支持向量机仍有很多局限和不足，最重要的就是核函数及参数的构造和选择问题，核函数的选择影响着分类器的性能，如何根据待解决问题的先验知识和实际样本数据选择和构造合适的核函数、确定核函数的参数等问题都缺乏相应的理论指导。其次，支持向量机的训练速度极大地受到训练规模的影响。因此，支持向量机队多类问题的处理仍有待进一步研究和改善。

［参 考 文 献］

[1]S.S.Keerthi,S.K.Shevade,C.Bhattachayya et al.Improvements to Platt's SMO Algorithm for SVM Classifier Design[J].Neural Computation,2001.

[2]E.Osuna.An Improved Training Algorithm for Support Vector Machines[J].InProc.IEEE Neural Networks in Signal Processing'97,1997.

[3]T.Joachims.Making Large-scale Support Vector Machine Learning Practical[A].Advances in Kernel Methods - Support Vector Learning,MIT Press,1998.

[4]John C.Platt.Fast Training of Support Vector Machines Using Sequential Minimal Optimization[A].Advances in Kernel Method - Support Vector Learning.MIT press,1999.

[5]邓乃扬,田英杰.数据挖掘中的新方法－支持向量机[M].北京:科学出版社,2004.

[6]Fan Rong-En,Chen Pai-Hsuen,Lin Chih-Jen.Working Set Selection Using Second Order Information for Training Support Vector Machines[J].Journal of Machine Learning Research,2005.

[7]李建民,张钹,林福宗.序贯最小优化的改进算法[J].软件学报,2003.

Research of Algorithm on Support Vector Machine

FANG Hui[1], WANG Qian[2]

(1.Department of Information Science and Technology,Bohai University,Jinzhou 121013 China;
2.Jinzhou Greenfield Environmental Development Center,Jinzhou 121000,China)

Abstract: Support vector machine is one new machine learning method which developped in 1990s and it is a novel data mining technique based on statistical learning theory and has been used in pattern classification and regression estimation widely.It has become the focus in international machine learning research.This article mainly discusses its basic principle and the SVM training algorithm.

Key words: support vector machine;machine learning;class