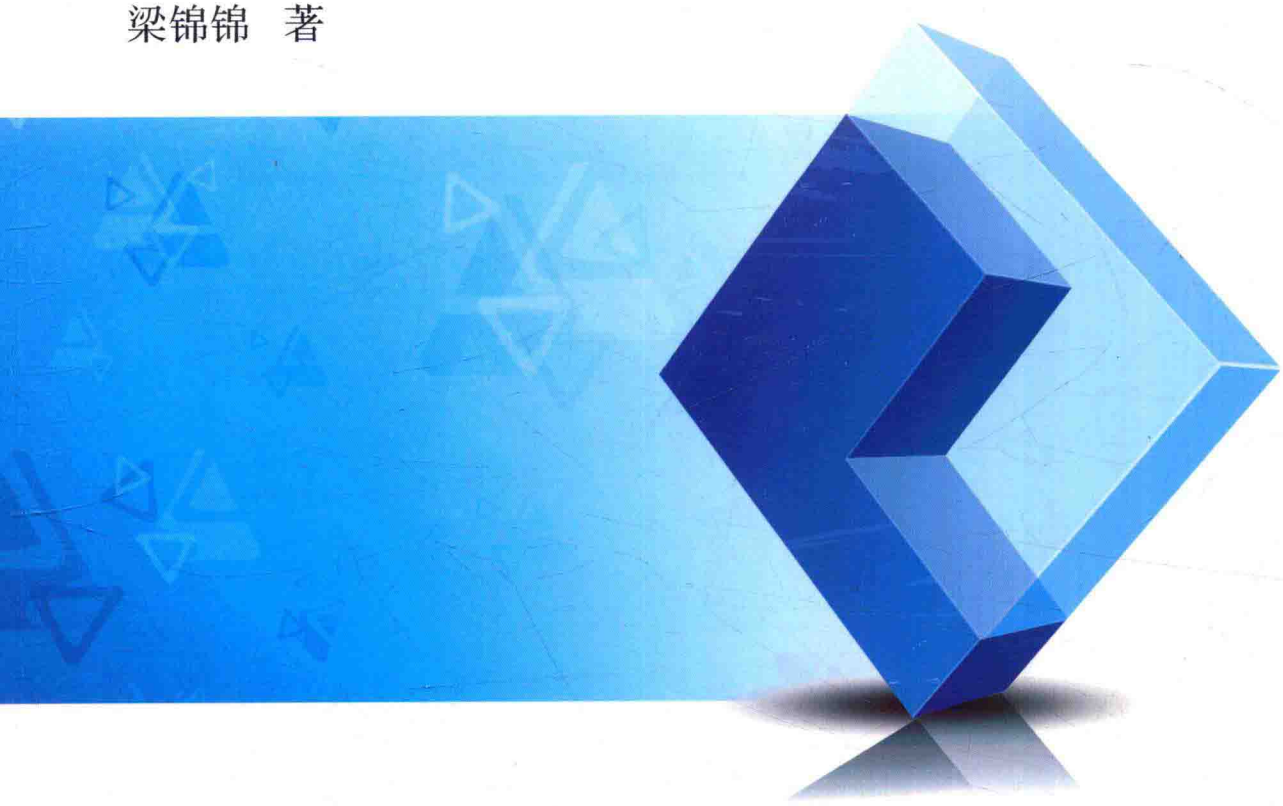
支持向量机算法及在大规模样本集的应用

梁锦锦 著



中国石化出版社
//WWW.SINOPEC-PRESS.COM

支持向量机算法及在大规模样本集的应用

梁锦锦 著

中国石化出版社

内 容 提 要

本书是关于支持向量机理论及算法的专著。全书共分为八章，第1章介绍了数据挖掘算法的发展历程、支持向量机的研究现状；第2章介绍了统计学习理论和最优化理论中的重要概念和定理；第3章在最优分类超平面基础之上，详细阐述了三类情形下的支持向量机模型；第4章至第6章依次介绍了最小二乘支持向量机、支持向量域描述、光滑支持向量机；第7章将支持向量机与其他算法集成，展示在大规模样本集上的分类性能；第8章展望支持向量机未来的发展趋势。

本书理论体系完整，材料取舍得当，内容全面新颖，充分反映了近年来支持向量机理论的研究动态，适合从事支持向量机理论研究和数据挖掘专业的技术人员和教师参考，也可以作为相关学科专业的研究生教材。

图书在版编目（CIP）数据

支持向量机算法及在大规模样本集的应用 / 梁锦锦著 .
—北京 ：中国石化出版社，2017.6
ISBN 978-7-5114-4505-6

Ⅰ. ①支… Ⅱ. ①梁… Ⅲ. ①向量计算机-算法理论
Ⅳ. ①TE301.6

中国版本图书馆 CIP 数据核字(2017)第 142329 号

中国石化出版社出版发行
地址：北京市朝阳区吉市口路 9 号
邮编：100020　电话：(010)59964500
发行部电话：(010)59964526
http://www.sinopec-press.com
E-mail：press@sinopec.com
北京柏力行彩印有限公司印刷
全国各地新华书店经销
*
787×1092 毫米 16 开本 8.5 印张 208 千字
2017 年 8 月第 1 版　2017 年 8 月第 1 次印刷
定价：28.00 元

前　言

支持向量机(Support Vector Machine，SVM)是一种新兴的模式识别方法，由Vapnik等在统计学习理论和最优化理论的基础上提出，具有全局最优、分类精度高等优点，在解决小样本、非线性、高维问题中表现出独特的优势。历经十几年的发展，支持向量机已成为国内外专家、学者的研究热点，并在数据挖掘、人工智能、医学、经济、社会等各个领域得到了广泛应用。

多年来，作者一直从事支持向量机算法的研究，特别是在光滑算法和集成算法方面取得了一些有价值的研究成果。本书是作者在总结攻读博士学位期间研究工作的基础上，结合工作以来所完成的有关支持向量机的预抽取策略及光滑技术方面的科研项目，并吸收了国内外同行的最新研究成果，提炼和整理而成的。本书系统介绍了支持向量机理论的基本概念、研究现状和发展趋势，重点讨论了最新研究成果中的变形支持向量机算法，如最小二乘支持向量机、支持向量域描述等，同时引入聚类算法、最近邻算法、粒子群优化算法来设计集成算法，进而解决大规模样本集的学习问题。本书选择人工数据集和基准数据集，经由数值试验来验证算法的分类性能。

本书可供应用数学、计算机、信息科学等领域的高年级本科生、研究生、工程技术人员和科技工作者阅读参考。本书具有如下特点：

(1) 着重从理论角度出发，突出可读性和实用性；

(2) 内容由浅入深，易于理解；

(3) 算法均给出翔实的步骤，并分析相关实验结果。

本书各章的主要内容包括：第1章，介绍机器学习理论的发展历程，尤其是支持向量机的理论研究、算法研究和应用研究。第2章，阐述统计学习理论和最优化理论，根据最大间隔分类超平面理论依次导出线性可分、线性不可分和非线性可分的支持向量机模型，并分析支持向量的特性。第3章，阐述了标准支持向量机算法的相关理论，对训练样本进行标准化处理消除不同量纲的影响，再设计近邻域比率修剪算法来剔除异常点。第4章，研究了最小二乘支持向量机算法，设计原空间求解模型来提高分类精度，修改目标函数构造具有稀疏性的模型来减少支持向量数目；这两种模型在非线性空间中均对核函数进行了乔列斯基分解。第5章，研究了支持向量域描述算法，根据样本自身特性设

置两种预抽取策略，一种是自中心距离比值策略，另一种是信赖度量策略，以提取到的样本代替全部样本参与最终训练，加快了算法的分类速度。第 6 章，研究了光滑对角加权支持向量机，在目标函数中采用松弛向量的二次损失函数，利用最优解条件将松弛向量表述为加函数形式之后应用熵函数进行近似，在隐空间中采用这一技术导出光滑支持向量机模型，取消了核函数对称正定这一限制，并运用粒子群优化算法确定隐空间支持向量机的参数，设计实例从分类精度、训练时间、迭代次数三个方面来展示算法的分类性能。第 7 章，运用同心超球面分割技术减少每次训练 SVM 的样本规模，并利用支持向量与全体训练样本等价的特性，提取支持向量集参与最终训练，展示算法在分类精度和分类时间上的优势；将支持向量域描述(Support Vector Machine，SVDD)与最近邻、聚类等其他算法集成，选取大规模样本集，展示集成算法的分类精度、分类时间以及随核参数的变化趋势；第 8 章，展望了支持向量机未来的发展趋势。

本书涉及的研究内容和成果，得到了国家自然科学基金“基于拓扑控制的无线传感器网络性能优化”(60674108)，陕西省教育厅专项研究计划“基于预抽取策略和优化方法的支持向量机算法研究”(2010JK773)和西安石油大学“支持向量机的光滑算法研究”(134010109)资助。

希望本书能够给有志于在数据挖掘及支持向量机领域，开展研究工作的广大学者以新的启示和收获，对促进支持向量机理论的研究及在相关领域中的应用发挥积极作用。当前世界正处于信息快速发展时代，知识的更替日新月异；由于作者理论水平有限，以及所做研究工作的局限性，书中难免存在不妥之处，恳请广大读者批评指正。

目　录

1 绪论

支持向量机(Support Vector Machine，SVM)是数据挖掘领域的一种新型有监督模式识别方法，对未见样本具有良好的预测能力，并在模式识别、函数拟合和密度估计等领域得到了广泛的应用。SVM 具有深厚的统计学习理论和最优化理论基础：设计结构风险最小化原则，在最大化分类间隔和最小化分类误差之间取得折衷。SVM 较好地解决了小样本、高维数及非线性等实际问题，具有拟合精度高、选择参数少、推广能力强和全局最优等特点。

本章将依次介绍机器学习理论、学习问题的分类和支持向量机算法。

1.1 机器学习理论

人类智慧的一个很重要方面就是学习能力，表现为从过去的数据和以往的知识中学习并获取规律的能力；通过归纳学习，得到对客观世界的认识和规律。运用得到的规律不仅可以解释已知的实例，而且能够对未知的现象做出正确的预测和判断，我们称这种对未知现象的预测能力为推广能力。

20 世纪 90 年代以来，随着信息技术和数据库技术的迅猛发展，人们可以非常方便地获取和存储大量的数据。面对大量的数据，传统的数据分析工具(如管理信息系统)只能进行一些表层的处理(如查询、统计学)，而不能获得数据之间的内在关系和隐含的信息。为了摆脱“数据丰富、知识贫乏”的困境，迫切需要一种技术和工具，能够智能地将数据自动转化为有用信息和知识。

用计算机来模拟上述学习能力的问题被称为“基于数据的机器学习问题”，其任务就是：设计某种方法和模型。通过对已知数据的学习，找到数据内在的相互依赖关系，从而对未知数据进行预测或判断，使机器具有良好的推广能力。

关于机器学习的研究，最早可以追溯到 20 世纪 50 年代，其研究和发展历程可分为如下四个阶段：

1) Rosenblatt 感知器(20 世纪 60 年代)

1962 年 Rosenblatt 提出第一个学习机器的模型，称为感知器。它在机器学习领域中有着里程碑的意义，标志着人们对学习过程进行数学研究的真正开始。Rosenblatt 将感知器的模型表示为一个计算程序，并通过试验说明了这个模型具有一定的泛化能力。Novikoff 证明了感知器的第一个定理：若学习样本能够以间隔 ρ 被分割开，则感知器至多需要 $[R^2/\rho^2]$ 次修改，就可以将学习样本分隔开。该定理在创建学习理论中起着十分重要的作用。

2) 学习理论的创立(20 世纪 60~70 年代)

在这段时间里，统计学习理论得到了很大的发展，提出了 VC(V. Vapnik and A. Chervonenkis)维和 VC 熵的概念。利用这些概念发现了泛函空间的大数定律，并通过研究

它与学习过程的联系得到了关于收敛速率的非渐近界的主要结论；同期还提出了用于解决不适定问题的正则化理论，这对学习理论的发展产生了深远的影响。此外，作为统计学和信息论中最伟大的思想之一，算法复杂度的思想，也在这一时期萌芽。

3）人工神经网络

1986 年，Lecun、Rumelhart、Hiton 和 Williams 独立地提出了构造感知器所有神经元的向量系数的后向传播算法(Back Propagation，BP)。此方法采用连续的 Sigmoid 函数修改了感知器的 McCulloch-Pitts 神经元模型，使得对于任意固定的输入都存在对应于所有神经元的所有系数的梯度。利用计算出的梯度，人们可以应用任何基于梯度的方法来构造对预期函数的无限逼近。此后，感知器也被称为神经网络。神经网络对于逼近实数值、离散值或向量值的目标函数提供了一种健壮性很强的方法，并在很多实际问题中取得了惊人的成功。由于缺乏统一的数学理论，人工神经网络容易产生过拟合现象，虽然对训练数据表现良好，但是网络的泛化能力较低。

4）统计学习理论

统计学习理论(Statistical Learning Theory，SLT)是一种专门研究小样本情况下机器学习规律的理论，它对小样本统计问题建立了一套新的理论体系。在这种体系下的统计推理规则不仅考虑了对渐进性能的要求，而且追求在现有的有限信息下的最优结果。Vapnik 等从 20 世纪 60 年代开始致力于这方面的研究。到 20 世纪 90 年代中期，随 SLT 的不断发展和成熟，也由于 ANN 等学习算法在理论上缺乏实质性的进展，SLT 开始受到越来越广泛的重视。

SLT 对有限样本情况下模式识别的一些根本性问题进行了系统地研究，它能将很多现有的方法纳入其中，在很大程度上解决了模型选择和过学习问题、非线性和维数灾难问题、局部最小值问题等。SLT 的一个核心概念就是 VC 维，它是描述函数集(或学习问题)的复杂性(或学习能力)的一个重要指标。在 VC 维的基础上发展出一系列关于统计学习的一致性(Consistency)、收敛速度(Convergence Speed)、泛化性能(Generalization Performance)等重要结论。

1.1.1 学习问题的一般模型

学习问题是依据经验数据选取依赖关系的问题。图 1.1 以基于实例的学习问题为例，给出学习问题的一般模型。

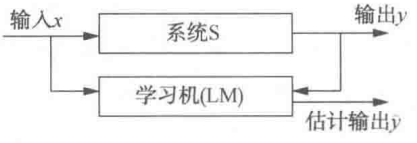


图 1.1 学习问题的一般模型

模型由 3 个部分组成：

(1) 输入数据(或称实例)，它由发生器按照某一未知的概率分布函数 $F(x)$，独立同分布地产生，表现形式为向量 $x \in X$。

(2) 系统(或称目标算子、训练算子、训练器),返回与输入向量 $x \in X$ 所对应的输出值 y。

(3) 学习机器(Learning Machine, LM),它设置某一规则作为相应的训练器,对发生器所产生的某一特定向量(不妨记为 x_i),预测相应的输出 y_i。学习机器的目标是构造一个适当的训练器,用它来逼近训练算子。

假设训练集 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$ 由 l 个独立同分布地观测点对组成,且服从联合概率分布函数 $F(x, y)=F(x)F(y \mid x)$。为了构造一个分类器,来逼近待求的未知算子,学习机器从如下目标中选择一个作为追求目标:

(1) 模仿训练器的算子:试图构造一个算子,对于一个给定的发生器 G,该算子可以对训练器的输出提供最佳的预测结果。

(2) 辨识训练器算子:试图构造一个非常接近于训练器算子的算子。

"构造一个算子"形式上的意义是学习机器可以构造一个实现某一固定函数集的机器,学习过程的意义是从函数集中选取一个适当函数。这两种追求目标预示着处理学习问题的两种不同方法。

1.1.2 经验风险最小化原则

传统的机器学习方法采用经验风险最小化(Empirical Risk Minimization, ERM)原则:根据观测样本损失的算数平均来定义经验风险,并根据经验风险来逼近给定的期望风险,通过最小化经验风险求得学习机器的参数。根据经验风险最小化理论的关键定理,下面两个事实等价:

(1) 经验风险最小化原则是一致的。

(2) 特定的经验过程收敛。

经验风险最小化原则指出,应该不惜一切代价来最小化经验风险。根据风险上界的相关结论,通过最小化经验风险值可以控制基于固定数量经验数据最小化风险泛函的过程;这是因为风险上界随着经验风险值的减少而减少。经验风险最小化原则对于规模较大的样本集能够取得好的结果,但却没有可靠的理论依据。

(1) 根据概率论中的大数定律,当样本数目趋于无穷大时,经验风险在概率意义下趋近于期望风险,但并不保证经验风险和期望风险在同一点取得最小值。

(2) 没有理论保证样本数无穷大条件下得到的学习机器在样本数有限时仍能具有好的分类结果。

1.1.3 模型复杂度和推广能力

ERM 原则通过最小化训练误差来实现最小化测试误差的目的,但一味地追求经验误差最小化并不总能获得好的预测效果;因为可以采用足够复杂的学习机器使得训练误差任意减小。最极端的情况是,学习机器"记住"所有的训练样本使得训练误差为零,但这种学习机器几乎不具备推广能力。某些情况下,训练误差过小反而导致推广能力的下降,也即真实风险的增加,这就是过学习问题。

有限样本情况下,学习机器的模型复杂度和泛化能力之间存在以下矛盾:

(1) 经验风险对学习机器的性能有一定的影响，但不起决定作用。执行 ERM 原则并不总能提高学习机器的泛化性能。

(2) 复杂度高的学习机器往往具有较低的经验风险。执行 ERM 原则将使学习机器变得越来越复杂。

(3) 学习机器的复杂度对其性能有较大影响。泛化性能好的学习机器应该具有与实际问题相对应的复杂度。

如何根据实际问题在学习机器的模型复杂度和泛化能力之间取得合理的折衷是一个非常重要的问题，统计学习理论(Statistical Learning Theory，SLT)的发展为此问题的解决提供了坚实的理论基础。

1.1.4 结构风险最小化原则

结构风险最小化原则(Structural Risk Minimization，SRM)是小规模样本集学习的归纳原则。给定经验数据，SRM 准则从给定的函数集中寻找一个函数，对经验数据的输入和输出给出最佳逼近，并给出刻画函数集的容量值之间的最优关系。

根据推广能力界的结论，我们知道风险上界由经验风险和置信范围两项之和构成。当训练样本集的规模较小时，经验数据上的风险值较小，往往并不能保证实际风险值也较小。因此，在最小化实际风险的过程中，必须同时考虑经验风险和置信范围。当固定观测数据集的数量时，经验风险取决于函数集中的某个特定函数；而置信范围取决于整个函数集的 VC 维，也即函数集被打散的程度。

结构风险最小化原则可以表述为：为给定函数集提供一个容许结构，然后在整个给定的结构元素上找到最小化风险的函数。首先，将函数集合结构化，建立 VC 维与各函数子结构之间的关系；其次，通过控制对函数结构的选择来达到控制 VC 维的目的。分析 SRM 准则的渐进性质，得到下列三个结论。

(1) 结构风险最小化原则是一致的；

(2) 结构风险最小化原则所选出函数具有最小保证风险值；

(3) 收敛速率的界。

1.2 支持向量机算法及研究现状

1992 年到 1995 年期间，Vapnik 等在 SLT 理论的基础上提出了一种新的机器学习方法——支持向量机。SVM 方法具有坚实的理论基础，被视作研究机器学习问题的一个基本框架。SVM 根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。SVM 可以有效地解决高维数据模型的构建问题，并具有泛化能力强、收敛到全局最优解、维数不敏感等优点；因而 SVM 被成功地应用于模式识别、回归估计和概率密度函数估计等领域。

相比于神经网络等传统的学习方法，SVM 具有许多优异的性质。概括地说，它的优点主要体现在以下几个方面：

(1) 针对小样本情况提出，其最优解是基于现有的有限样本信息得到，而并非基于样本

数目趋于无穷大时的样本信息得到。

(2) 算法设计为一个凸二次规划，保证了解的全局最优性，解决了传统学习方法无法避免的局部极小值问题。

(3) 采用结构风险最小化原则，同时考虑了经验风险和置信范围的最小化，保证了学习机器具有良好的泛化能力。

(4) 采用核函数技术，巧妙地解决了算法复杂度与输入向量维数密切相关的问题。通过一个非线性映射函数，将输入空间中的非线性分类问题映射到一个高维的特征空间，并在此高维空间构造线性判别函数。由于判别函数仅由样本间的内积决定，SVM 通过引入核函数而不需要知道非线性映射的具体形式和高维空间的维数。定义不同的核函数，SVM 可以实现许多其他现有的学习算法。

目前，对支持向量机的研究已经成为国际学术的一个研究热点，国内外专家在此领域也展开了广泛的研究。以 Schölkopf 为首的研究小组提出的一些关于 SVM 的成果代表了目前该领域的最高成就：如快速学习算法、大规模学习算法、半监督学习算法、主动学习算法等；以 Mangasarian 为首的团队在光滑 SSVM 的学习算法设计、收敛性分析等方面取得了许多研究成果；Tax 等提出的 SVDD 在很多领域也取得了广泛的应用。国内学者也对 SVM 开展了广泛的研究，但大多数研究内容和方法是对国外先进成果的跟踪和吸收，达到国际领先或者具有突破意义的理论和应用成果仍寥寥无几。

SVM 方法尚处在蓬勃发展的阶段，有很多方面尚不成熟、不完善，也有很多问题亟待解决。比如许多理论尚不能在实际算法中实现；有关 SVM 的某些理论解释也并非完美；对学习机器的 VC 维进行分析尚没有通用的方法；实际问题中对于核函数的选择和构造缺乏理论指导；在计算上存在算法速度慢、算法复杂而且难以实现以及检测阶段运算量大的问题；分类性能受噪音及孤立点的影响较大；多类分类等问题。掌握 SLT 和 SVM 的相关基础性知识，投入精力并展开卓有成效的研究有着重要的意义。

1.2.1 理论研究

近几年涌现的大量理论研究成果完善和发展了 SVM 理论，更为其应用研究奠定了坚实的基础；主要研究成果归结如下。

1) 模型选择

称求解 SVM 的二次规划中惩罚参数、核函数类别与核参数的选择问题为“模型选择”。模型选择是 SVM 中一个公开的问题。

核参数从某种程度上定义了一个高维特征空间，反映了数据的先验知识。核参数的选取过程也是模型比较的过程。如何构造与实际问题相适应的核函数一直是支持向量机研究的重要课题。Amari 提出利用实验数据修正已有的核函数，使之与问题相吻合。文献分别针对文本分类、图像处理等具体问题构造不同类型的核函数，获得的泛化性能优于径向基核或多项式核。

惩罚参数和核参数中不同参数的组合构成不同的模型，这些参数决定了 SVM 的泛化性能。简单、准确和有效地估计泛化性能是进行模型选择、参数优化以及核函数优化的基础。最简单的是借助设计者的经验进行手工调整，但所得结果缺乏可靠性。更加规范化的方法可

分为以下三类：

（1）确认集和交叉验证法。

确认集法首先定义训练集、确认集和测试集，然后采用不同的参数组合对训练集的数据进行训练，将训练得到的 SVM 在确认集上验证，选出误差最小的 SVM 对测试集进行测试。交叉验证法（Cross Validation，CV）将数据集随机分割成互不重叠的 k 个子集，应用 $k-1$ 个子集训练，剩下的一个子集测试；整个计算过程重复 k 次，故交叉验证的计算量较大。留一法（Leave One Out，LOO）是一种极端的交叉验证方法，每次仅保留一个样本用于测试，其余样本用于训练；该方法可以给出对期望泛化误差的几乎无偏估计。

（2）基于界的方法。

该类方法最小化不同的泛化误差界以求取最佳模型参数。Burgers 提出 VC 维界 $h \leqslant D^2 \| w \|^2 + 1$；当 $D^2 \| w \|^2 + 1 > l + 1$ 时，Burgers 取 $h = l + 1$。Vapnik 和 Chapelle 指出 $\frac{N_{L\infty}}{l} \leqslant \frac{S\max(D, 1/\sqrt{C}) \sum_{i=1}^{n} \alpha_i + m}{l}$，称 S 是支持向量跨距，并称该式的右边为跨距边界；Vapnik 以 $LOO_{Err} \leqslant \frac{1}{4l} D^2 \| w \|^2$ 作为半径间隔界，而 Chapelle 对该半径间隔界做出修改 $LOO_{Err} \leqslant \frac{1}{l}[D^2 \| w \|^2 + (D^2 C + 1) \sum_{i=1}^{l} \xi_i]$。另外还有 Wahba-Lin 方法，Jaakkola-Haussler 方法和 Opper-Winther 方法等。

（3）统计的方法。

通过定义输入空间上函数的先验，统计的方法能清晰地解释核的作用，从而为数据建模提供一个新的框架。Smola 指出 SVM 中不同核的应用可以解释为对函数空间定义的不同先验分布。Macky 提出了一种贝叶斯证据框架，进行网络结构参数的选择。Kwok 用证据框架讨论了支持向量机分类问题。

2）变形 SVM

随着对 SVM 研究的深入，许多研究学者提出一系列变形 SVM：如 Vapnik 等提出的可调罚参数 C-SVM 系列，Schölkopf 等提出的用于分类和回归问题的 V-SVM 系列等。Suykens 等提出最小二乘支持向量机（Least Square SVM，LSSVM），采用 L_2 范数衡量违反约束的误差总量，并将不等式约束改为等式约束。LSSVM 将学习问题转化为求解一个线性方程组的问题，可以引入最小二乘法求解。Roobaert 提出直接支持向量机（Direct SVM），采用启发式搜索方法在所有训练样本中搜索支持向量，并避免了求解二次规划的最优化过程。周伟达等学者研究了 SVM 的线性规划求解方法。

Mangasarian 等研究了基于标准 SVM 的扰动模型，或取消对目标函数凸性和核函数正定性的限制，或采用二次损失函数，或对目标函数加上分类超平面偏置的二次扰动；这些变形模型使相应对偶二次规划的最优性条件得到了简化。Mangasarian 等同时构造了一系列迭代型求解算法：如 GSVM，SOR，LSVM，RSVM 等。

3）支持向量域描述

支持向量域描述(Support Vector Domain Description，SVDD)由 Tax 等于 1999 年提出，目前有关 SVDD 的研究和应用的文献大量涌现。学者或侧重于应用 SVDD 建立分类器，进行分类、探测和排除噪点；或拓宽 SVDD 的应用领域，对遥感图像、人脸识别和拒识判别等问题设计相应的求解算法；或通过调节 SVDD 的核参数对样本点进行聚类，显示“核”的灵活性；或研究 SVDD 和其他学习算法的集成，提高分类性能。

4）多类分类

最初的 SVM 是针对两类分类问题设计的，如何有效地将其推广到多类分类问题，是当前 SVM 的一个重要研究方向。构造多类 SVM 分类器的方法主要分两类。一类直接在经典 SVM 理论的基础上，对新模型的目标函数进行优化将二分类 SVM 扩展成多分类 SVM。该类算法存在目标函数复杂、实现困难、计算复杂度高等问题。另一类通过组合多个二值 SVM 分类器来构造多类分类算法，如“一对一”(One-Against-One)、“一对多”(One-Against-Rest)、有向无环图支持向量机(DAGSVM)等。该类算法操作简单、易于实现，在实际中得到了广泛的应用。

1.2.2 训练算法

训练 SVM 最终归结为求解一个凸二次规划，而相应的求解算法被称为“训练算法”或“学习算法”。一些学者针对该二次规划提出的解决方法包括块算法、分解算法、并行学习算法等；一些学者从提升泛化能力的角度研究了 SVM 的集成学习算法；一些学者采用预处理的方法提取潜在支持向量等。

训练算法的研究内容归结如下：

1）块算法

Vapnik 提出的块算法(Chunking)基于这样的事实：去掉 Lagrange 乘子为零的学习样本不影响原 QP 问题的解。具体做法是：随机拿出一个子集训练，保留支持向量，用学习结果检验出剩余样本中违犯 KKT 条件的样本，将它们与本次训练的支持向量合并为新的训练集，然后重新训练直到所有样本都满足 KKT 条件。该方法降低了求解的存储和计算需求，但如果支持向量的数目本身就比较多，工作集规模将随迭代次数的增加而增大。已有学者证明了 Chunking 算法的收敛性。

2）工作集方法

工作集方法最早由 Osuna 提出，把训练样本集分解成工作集和非工作集，其中工作集的大小固定在算法可以容忍的限度内。每次迭代只对工作集中样本进行 QP 求解，而非工作集中所有样本的 Lagrange 乘子保持不变。迭代过程是将剩余样本中部分严重违反 KKT 条件的样本与工作集中的样本进行等量交换，即使支持向量的个数超过工作集的大小也不改变工作集的规模。

Joachims 提出 SVM^{light} 的 SVM 分解学习算法，采用类似 Zoutendijkd 可行方法的策略来确定工作集，提出 Shrinking 方法以估计出部分支持向量和非支持向量，从而有效地减小了 QP 问题的规模，最后采用 Kernel Cache 技巧以减少海赛矩阵中元素的计算次数。Lin 证明了

SVM^{light}的收敛性，Hsu 和 Lin 指出工作集的选择并不简单，任何好的策略可能都会遇到例外；并通过实验观测，简化了 Joachims 采用的方法，在复杂情况下提高了收敛速度。

3）SMO

Platt 等提出序列最小优化算法（Sequential Minimal Optimization，SMO），将工作集的规模缩减为两个样本，使得迭代过程中每一步子问题的最优解均可直接用解析的方法求出，并避开了复杂的数值求解优化问题的过程；同时 SMO 的内存需求与训练数据的大小呈线性关系，不需要额外的矩阵存储；这一算法的收敛性也在相关文献进行了证明。S. S. Keerhi 等改进了 SMO 算法，在判别最优条件时采用两个阈值代替一个阈值，从而使算法更合理、收敛更快。

4）增量或并行学习

增量学习的基本思想是：首先对部分训练样本进行 SVM 训练，当加入一个新样本时，在原最优解的基础上对 SVM 再进行训练，同时保留支持向量样本，抛弃非支持向量样本。并行学习的基本思想是：将待学习的样本集 S 分割成若干子集，并且分派到并行系统的若干个计算节点上进行单独地 SVM 训练；期间，各个节点通过互相交换支持向量以不断提高本地 SVM 的泛化性能，并最终收敛到全局最优解。很多学者并行化算法进行了研究；其中的典型代表包含 Dong 等采用近似的块对角核矩阵对大规模样本集进行分割，以及 Graf 等提出一种二叉级联结构的并行学习算法 Cascade SVM。

5）集成学习

近几年来，支持向量机的集成学习算法受到广泛关注，而成为众多学者的研究焦点。人们逐渐意识到，即便是对于稳定的 SVM 方法，特殊设计的集成学习算法也能够明显提升 SVM 的泛化性能。Valentini 等分析了 SVM 分类器的期望误差、偏差和方差随着模型参数的变化趋势，提出 LoBag 集成学习算法。Robert 等指出扰动样本的输入特征空间能够获得具有显著差异性的成员分类器，从而提升 SVM 的泛化性能。

6）其他一些算法

观测样本中占少数的支持向量决定了 SVM 的解，已有学者转向研究支持向量的预选取方法：如类中心距离比值方法、“卫向量”方法、TransRed 和 GetBorder 方法等。该类预处理方法依据训练样本集的几何分布信息或者密度信息来选取潜在支持向量集进行训练，从而在一定程度上缩小待解 QP 问题的规模，降低计算复杂度并加快 SVM 的训练速度。

观测样本中的噪声影响着 SVM 的分类性能。C. F. Lin 和 H. P. Huang 等提出模糊支持向量机 FSVM，根据不同样本对分类超平面的贡献设定相应的隶属度，使得噪声由于隶属度较小而减弱对分类的影响。隶属度函数的设计是 FSVM 的关键，这要求隶属度函数必须能够客观、准确地反映系统中样本存在的不确定性。为提高识别或分类的精度，一些学者研究了噪声的硬剔除方法，如采用核主成分分析进行除噪，采用 K 近邻（K Nearest Neighbor，KNN）删除噪声或野点等。

1.2.3 应用研究

SVM 具有拟合精度高、选择参数少、推广能力强、对维数不敏感和全局最优等特点，因而其在各个领域的应用研究尤为活跃。