支持向量机理论及算法研究综述*

汪海燕，黎建辉，杨风雷
（中国科学院计算机网络信息中心 科学数据中心，北京 100190）

摘 要： 介绍了 SVM 的理论基础和它的多种主要算法及这些算法的利弊与发展现状，并介绍了 SVM 在现实生活中的应用原理及应用现状。最后分析了 SVM 在发展中的不足之处，指出了其研究方向及前景，并提出在分布式支持向量机这个方向上可以进行更深层次的研究。

关键词： 支持向量机；统计学习理论；训练算法；模糊支持向量机；多分类支持向量机；模式识别
中图分类号： TP301 **文献标志码：** A **文章编号：** 1001-3695(2014)05-1281-06
doi:10.3969/j.issn.1001-3695.2014.05.001

Overview of support vector machine analysis and algorithm

WANG Hai-yan, LI Jian-hui, YANG Feng-lei
(*Science Data Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: This paper introduced the theoretical basis of support vector machine, in addition, it described some algorithms about SVM and analysed their advantages and disadvantages and development status. Then it introduced the application principle of SVM in the real life and its application status. Finally, it analysed the deficiency in the development of SVM and pointed out the research direction and prospects, and it put forward distributed support vector machine which as a direction could be make a deeper research.

Key words: support vector machine(SVM); statistical learning theory(SLT); training algorithm; fuzzy support vector machines; multi-class support vector machines; pattern recognition

支持向量机(SVM)是 Cortes 等人于 1995 年提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能推广应用到函数拟合等其他机器学习问题中。

支持向量机[1]是建立在统计学习理论的 VC(Vapnik-Chervonenkis)维理论和结构风险最小原理基础上的，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中，以期获得最好的推广能力。支持向量机具有较强的理论基础，它能保证找到的极值解是全局最优解而非局部最小值，这也就决定了 SVM 方法对未知样本有较好的泛化能力，正因为这些优点，SVM 能良好地应用到模式识别、概率密度函数估计、时间序列预测、回归估计等领域，也被广泛应用到模式识别中的手写数字识别[2]、文本分类[3]、图像分类与识别[4]等众多领域中。

1 支持向量机理论

1.1 VC 维理论和结构风险最小原理

支持向量机是基于统计学习理论(SLT)的新型机器学习方法。机器学习主要研究的是计算机如何模拟或实现人类的学习能力，以获取新的知识和技能，重新组织已有的知识结构，使之不断改善自身的性能。机器学习的实现方法主要有以下三种：统计预测方法、经验非线性方法、统计学习理论。

SLT 是一种专门研究小样本情况下机器学习规律的理论[5]，该理论针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，而且追求在现有有限信息条件下能够得到最优结果。

SVM 是建立在 SLT 的 VC 维理论和结构风险最小原理基础上的。关于 VC 维理论[6]的定义是：对一个指示函数集，如果存在 h 个样本能够被一个函数集中的函数按所有可能的 2^h 种形式分开，则称这个函数集能够把 h 个样本打散，函数集的 VC 维就是它能打散的最大样本数目 h。VC 维本质上可以理解为问题的复杂程度，VC 维数越高，则该函数集的机器学习越复杂。关于结构风险最小原理，SLT 引入了泛化误差界的概念，该理论指出，机器学习的实际误差是由经验风险和置信风险两部分组成。

泛化误差界的公式如下：

$$R(w) \leqslant \mathrm{remp}(w) + \varphi(n/h) \tag{1}$$

其中：$R(w)$ 是实际风险，$\mathrm{remp}(w)$ 是经验风险，$\varphi(n/h)$ 是置信风险。置信风险与两个量相关：a) 样本数量，样本数量越大，机器学习结果越有可能正确，置信风险也就越小；b) 分类函数的 VC 维，VC 维的维数越大，泛化能力越差，置信风险就会越大。统计学习的目标就是从寻求经验风险最小化转变为寻求经验风险与置信风险的和最小，即结构风险最小。SVM 正是这样一种使结构风险最小的算法。

1.2 SVM 理论

SVM 理论的初衷是寻求一种处理两类数据分类问题的方法。SVM 旨在寻找一个超平面，使得训练样本集中不同类别

收稿日期：2013-08-13；修回日期：2013-10-18 基金项目：国家自然科学基金重点资助项目(91224006)；国家“十二五”科技支撑计划资助项目(2012BAK17B01-1)

作者简介：汪海燕(1984-)，女，河南许昌人，助理工程师，硕士，主要研究方向为 Web 数据挖掘、机器学习、文本分类(why@cnic.cn)；黎建辉(1973-)，男，研究员，博士，主要研究方向为海量数据管理、处理与挖掘分析的理论、方法及关键技术；杨风雷(1973-)，男，副研究员，博士，主要研究方向为 Web 数据挖掘、云计算和大数据处理、机器学习.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的点正好落在超平面的两侧，同时还要求超平面两侧的空白区域达到最大。对于二维两类线性可分数据，支持向量机理论上是能够实现最优分类的，推广到高维空间，最优分类线就叫做最优超平面。对于二维两类数据分类来说，给定的训练样本 $D_i = (x_i, y_i)$，$i = 1, \cdots, l$，$y_i \in \{+1, -1\}$，其中 x_i 为输入样本，y_i 为两类的类别值，其超平面为 $wx + b = 0$，样本点到超平面的间隔为 $\delta_i = \frac{1}{\| w \|} | g(x_i) |$。为使训练样本能正确分开，又要保证间隔最大，这个两类分类问题被转换成了一个带约束的最小值问题：

$$\min \frac{1}{2} \| w \|^2 \tag{2}$$

$$\text{subject to} \quad y_i [(wx_i) + b] - 1 \geqslant 0 \quad i = 1, \cdots, l (l \text{ 是样本数}) \tag{3}$$

本文也把式(2)和(3)的问题叫做二次规划(quadratic programming，QP)，由于它的可行域是一个凸集，也可以叫做凸二次规划。在线性不可分的情况下，需要在条件式(3)中加入一个松弛变量 $\xi_i \geqslant 0$、惩罚因子 C，则上面所述凸二次规划问题就变成

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i \tag{4}$$

$$\text{subject to} \quad y_i [(wx_i) + b] \geqslant 1 - \xi_i$$
$$i = 1, \cdots, l (l \text{ 是样本数}); \xi_i \geqslant 0 \tag{5}$$

其中：$C > 0$ 为一个常数，其大小决定了对错分样本惩罚的程度。

下面是上述二次规划问题的求解问题，为了求解，引入了 Lagrange 函数：

$$L = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{l} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{l} \alpha_i \tag{6}$$

其中：$\alpha_i > 0$，为 Lagrange 系数，求解上述问题后得出的最优分类函数为

$$f(x) = \text{sgn}\{ [\sum_{j=1}^{l} \alpha_j^* y_j (x_j \cdot x_i)] + b^* \} \tag{7}$$

对于线性不可分的情况，SVM 的主要思想是将输入向量映射到一个高维的特征向量空间，并在该特征空间中构造最优分类面。将 x 作非线性映射 Φ: $R^n \rightarrow H$，H 为高维特征空间，则有

$$x \rightarrow \Phi(x) = (\Phi_1(x), \Phi_2(x), \cdots, \Phi_l(x))^T \tag{8}$$

求解则可得到最优分类函数为

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i y_i \Phi(x_i) \cdot \Phi(x) + b) \tag{9}$$

2 SVM 算法

SVM 研究的主要问题就是凸二次规划的求解，传统的 SVM 算法在计算上存在许多问题，包括训练算法速度慢、算法复杂而难以实现、测试阶段运算量大、抗击噪声及孤立点能力差等。所以在 SVM 算法的研究中，如何提高训练速度、减少训练时间、建立实用的学习算法是亟待解决的问题。

为了优化凸二次规划求解问题，已有很多学者提出了相应算法，典型的算法有选块算法(chunking algorithm)、分解算法(decomposition algorithm)、序列最小优化算法(sequential minimal optimization，SMO)，另外还有模糊 SVM 学习算法、采用光滑化技巧的学习算法、基于标准 SVM 的变形算法；近年来比较热门的还有多分类支持向量机(multi-class SVM)等。

2.1 选块算法

选块算法最早是由 Boser 等人[7]提出来的。它的主要思想是去掉核矩阵中 Lagrange 乘子为零的样本对应的行和列，这样得到的结果与用之前矩阵计算得到的结果相同，这就使得大型复杂的二次规划问题转换为一系列小的子问题。选块算法的核心是预测样本集中哪些样本的 Lagrange 乘子是零，为零的被丢弃，那些非零(支持向量)的则需要保留下来。然而实际的支持向量是未知的，于是引入了 KKT(Karush-Kuhn-Tucker)条件进行逐步迭代，最终得到全局最优解。选块算法将矩阵规模由训练样本数目的平方减少到非零 Lagrange 乘子的样本数的平方，在很大程度上降低了训练过程对存储容量的要求，因此能够大大提高训练速度。然而这也只是在一定程度上解决了大数据集的 QP 问题，选块算法的最终目的是找出所有的支持向量(support vector)，因而需要存储相应的核函数矩阵。所以选块算法在本质上还是受 support vector 数目的影响，并未从根本上解决内存不足问题。

2.2 分解算法

分解算法最早是由 Osuna 等人[8]提出的，它的主要思想也是把大型的 QP 问题分解成一系列小的子问题，但是它与选块算法又是不同的，它是活动集(active set)方法的一个变形。它的算法过程是：在每次迭代中，把 Lagrange 乘子 α_i 分为工作集和非工作集，非工作集在本次迭代中保持不变；工作集为自由变量，在本次迭代中对这些自由变量进行优化，这个过程一直进行下去，直到满足停止条件为止。这个算法的关键在于选择最优工作集来提高该算法的收敛性和收敛速度，但在实际的过程中工作集的选取是随机的，所以限制了其收敛速度。

对于上述算法中的不足，Joachims 对 Osuna 的方法进行了一系列的改进，提出了选择工作集及其他的一些策略与技术，并实现了这一算法，最终形成了软件 SVM^{light}。在选择工作集方面，其思想是从全部变量中挑选出 q 个变量，这些变量应满足下述条件：不等于零且与目标函数可行的最速下降方向对应，这个可以通过求解一个简单的线性优化问题解决。

SVM^{light} 算法还引入了 shrinking 技术、caching 技术、梯度增量修正技术等。它对大规模问题，尤其是支持向量比例较小或者多数支持向量在边界上的情况效果明显。

2.3 序列最小优化算法

SMO 算法是分解算法的一种特殊情况，它最早是由 Platt[9]提出来的，这个算法的主要特点是其工作集中只有两个样本，减少到两个的原因是等式线性约束的存在要求至少有两个 Lagrange 乘子发生变化。由于只有两个变量，迭代中这样的 QP 子问题都能用解析方法得到最优解，从而避开了复杂的求解优化过程，也提高了算法的收敛速度，又不需要大的矩阵存储空间。SMO 对于工作集的选择采用的是启发式策略，具体的办法可以参考文献[10]。SMO 算法对线性核的情况最有效，而对于非线性问题则达不到满意的结果，原因是在线性情形下，每次最小优化后的重置都是简单运算，但是在非线性情形下，误差的重置必须对全部支持向量逐个地计算核函数，而核函数的计算比较复杂，也就需要占有较多的时间。

之后 Platt 又对 SMO 算法进行了一些改进，如借鉴 SVM^{light} 对 SMO 算法进行优化，引入了 shrinking 技术和 kernel cache 思想来提高收敛速度。但是由于 shrinking 技术一般都是在优化迭代即将结束时才使用，并且 shrinking 中不完全包含集合的最优非边界乘子，重构目标函数的梯度需要花费很大的代价，所以基于 shrinking 技术优化还是有其局限性的[11]。

由于 Platt 启发式策略存在不确定性，人们将重心转移到可行性方向方法上。Keerthi 等人[12]在后来的研究中提出了利用双阈值优化条件来确定工作集的优化算法，并提出了基于 violating pair 和 τ-violating pair 的 GSMO(generalized SMO)的概

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

念,证明了 GSMO 在任何迭代停止标准和停止容忍范围内可以在有限步骤停止并收敛。文献[13]中的两种优化改进都是 GSMO 的特殊情况。

在核函数占主要地位的 SMO 算法中,合适的缓存替换策略在算法的性能改进中扮演了非常重要的角色,但是大多数 SMO 算法中工作集的选择如 Platt 的启发式、Keerthi 的改进和可行方向方法,目的都是使目标函数尽可能多地下降,最快找到目标函数的最小值。文献[14]利用边界支持向量数据变化平缓、缓存更新带来的计算小于重复计算 BSV 核函数的特征,提出了一种针对非线性可分样本的高效缓存策略,并把它推广到 SMO 之外的分解算法中。文献[15]在可行性方向策略的基础上考虑 kernel cache 的利用效率,给出了一种收益代价平衡的工作集选择方法。文献[16]提出了 SMO 工作集选择算法的通用框架模型,即选取与最大违反对成一定函数关系的乘子为工作集,并提出了基于未定的 Q 矩阵的分解算法。文献[17]提出了一种基于目标函数二阶展开 SMO 算法的工作集选择法。文献[18]提出了基于微粒群优化的 SMO 算法的双层优化原理,并通过仿真进行了应用研究,验证了其有效性。文献[19]对 SMO 算法进行改进的办法是在选取工作集时,选取优化步长最大的违反 KKT 条件的样本和其配对样本,并且对求解过程进行简化,从而使训练过程速度更快。文献[20]从变量优化和变量选择等多个方面对 SMO 算法进行改进,并且基于改进后的算法进行了仿真实验,取得了满意的效果。

2.4　模糊支持向量机

在进行数据挖掘的时候,数据集中会存在很多的噪声,也称为模糊信息,这些信息在用 SVM 算法进行预测的时候,使得其性能不能得到好的发挥,甚至是无能为力,在这种情况下 Lin 等人提出了 FSVM[21]。该算法将模糊数学与 SVM 相结合,主要用于处理训练样本中的噪声数据[22],其核心思想是用异常数据检测方法对训练数据集中的数据进行检测,检测出来异常数据并赋予它很小的隶属度,这些异常数据也就是噪声或孤立点;而对支持向量赋予较大的隶属度,这样做的目的就是把噪声或孤立点从有效样本中分离出去。

模糊支持向量机能够提高分类精度,又能解决异常数据造成的过学习问题,但是在实际的应用中,也存在一些不足之处,如异常数据可能会有很多,或者这些异常数据服从某种分布,在这种情况下如果还按上述算法并分离出这些异常数据,就会造成信息的丢失,从而使支持向量机的泛化能力受到影响。另外,模糊支持向量机还存在核函数计算量大、所需内存大、训练时间长等问题,所以如何优化模糊 SVM 的训练速度也是至关重要的。

针对上述问题,有很多学者对模糊支持向量机进行了改进,文献[23]针对模糊支持向量机普遍存在训练时间过长的难题,提出了使用截集 C-均值聚类的方法对训练样本进行聚类处理,以聚类中心作为新的样本进行训练,并用数值进行了实验,实验结果证明,与一般的 FSVM 相比,有效提高了分类速度和精度。文献[24]针对 Lin 等人提出的基于类中心距离的模糊隶属度设计方法不能有效区分噪声或孤立点,而且可能降低支持向量的隶属度等不足,通过引入一个半径控制因子,提出了一种改进的隶属度函数设计方法,这种方法在不增加时间复杂度的情况下能有效提高 FSVM 的分类精度。文献[25]针对传统 FSVM 对非球形分布数据不合理的现象,使用类内超平面代替类中心,提出了基于样本到超平面距离的新隶属度函数设计方法,该方法克服了传统方法的不足,降低隶属度函数对样本集几个形状的依赖,提高模糊支持向量机的泛化能力。

2.5　最小二乘支持向量机

LS-SVM 最早是由 Suykens 等人[26]于 1999 年提出的。LS-SVM 将不等式约束换成了等式约束,将凸二次规划问题转变成了一个线性方程组,从而能够明确得到解的表达式,LS-SVM 由以下优化问题刻画:

$$\min \quad \frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \tag{10}$$

$$\text{subject to} \quad y_i(\langle w,\Phi(x_i)\rangle+b)=1-\xi_i \quad i=1,\cdots,l;\xi_i\geqslant 0 \tag{11}$$

根据 KKT 条件,式(10)(11)的问题等价于如下线性系统:

$$\begin{cases} w=\sum_{i=1}^{l}\alpha_i\Phi(x_i) \\ \sum_{i=1}^{l}\alpha_i=0 \\ \alpha_i=C\xi_i,\forall i \\ \xi_i=y_i-(\langle w,\Phi(x_i)\rangle+b),\forall i \end{cases} \tag{12}$$

简化式(12),消掉 w 与 ξ_i,可得到关于 α_i 和 b 的方程组:

$$\begin{pmatrix} \tilde{K} & e \\ e^{\mathrm{T}} & 0 \end{pmatrix}\cdot\begin{pmatrix} \alpha \\ b \end{pmatrix}=\begin{pmatrix} y \\ 0 \end{pmatrix} \tag{13}$$

其中: $e\in R^{(l)}$ 为元素为 1 的向量,$\tilde{K}=(\tilde{k}(x_i,x_j))_{l\times l}$ 为正定矩阵,其中 $\tilde{k}(x_i,x_j)=k(x_i,x_j)+\frac{1}{C}\delta_{ij}$。这样,LS-SVM 的问题就转换为求式(14)的解的问题了。它的判决函数为

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{l}\alpha_i k(x_i,x)+b] \tag{14}$$

LS-SVM 与标准的 SVM 相比,它用等式约束条件代替不等式约束条件,使求解二次规划问题转换为求解一个等式方程组,从而大大减少了训练时间,但这样也丧失了标准 SVM 的稀疏性。LS-SVM 算法中几乎所有的样本都是支持向量,这在处理大样本数据时就很难得到较好的分类效果。文献[27]采用了增量的学习方法,用 LS-SVM 算法对大样本数据进行学习训练,有效地处理了大样本数据的问题。文献[28]提出了一种基于统计分析的 LS-SVM 稀疏化算法(PCA-LSSVM)来对 LS-SVM 算法进行改进,并用它进行了一些测试,结果表明 PCA-LLSVM 由于采用了统计方法,所以稀疏性高于标准 SVM,并保持了良好的预测性能。文献[29]对目标函数含二次损失函数、样本特征空间分布形状不规则的情况提出了混合参数优化算法,用待优化参数重构 LS-SVM 的目标函数,通过自适应遗传算法、交叉验证来优化目标函数、选择最优的核和其他参数,实验结果表明,经过优化的 LS-SVM 建立起来的预测模型有较高的训练、泛化精度。

2.6　拉格朗日支持向量机

LSVM 算法是由 Mangasarion 等人[30]于 2001 年提出来的,它的算法思想是对对偶问题进行分析,先得出对偶问题的解,再由对偶问题的解求出原问题的解。它主要是处理线性分类问题的,可以由以下优化问题刻画:

$$\min \quad \frac{1}{2}(\|w\|^2+b^2)+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \tag{15}$$

$$\text{subject to} \quad y_i(\langle w,x_i\rangle+b)\geqslant 1-\xi_i \quad i=1,\cdots,l;\xi_i\geqslant 0 \tag{16}$$

与 LS-SVM 算法相比,目标函数中增加了 $b^2/2$,这相当于在每个样本中增加了一个常数特征 1,然后求经过原点的最优超平面。这对高维空间来说是比较弱的条件,这样目标函数就是强凸的,其对偶问题中的约束条件消失了,其对偶问题是:

$$\max_{\alpha} \quad e^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}Q\alpha \tag{17}$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$\text{subject to} \quad \alpha \geqslant 0 \tag{18}$$

其中: e 是全为 1 的列向量, $Q = \frac{I}{C} + HH^{\mathrm{T}}$, I 是单位方阵, H 是 $l \times (d+1)$ 矩阵, 其第 i 行是 $[y_i x_i, -y_i]$, d 是样本维数。由 KKT 条件, 上述对偶问题等价于: 用此定理 $0 \leqslant a \perp b \geqslant 0 \Leftrightarrow a = (a - \beta b)_+$, $\beta > 0$, 则可得到 LSVM 算法的迭代公式为

$$\alpha^{i+1} = Q^{-1}(e + ((Q\alpha^i - e) - \beta\alpha^i)_+) \quad i = 0, 1, \cdots \tag{19}$$

LSVM 算法在线性分类的应用中, 对于大样本集的分类处理相比其他算法如 SMO 收敛速度要快很多, 也相对简易, 但是在推广到非线性问题中时就只能处理中等规模的样本集。

2.7 多分类支持向量机

SVM 的研究初期主要是针对两类问题的分类, 但是在现实生活的应用中却普遍是多分类问题, 怎样将 SVM 有效地应用在多分类问题中一直也是许多学者近年来研究的热点与难点。多分类支持向量机是将 SVM 从两分类算法推广到多分类的算法[31~34]。目前多分类支持向量机算法的思路主要有两种:

a) 在经典 SVM 的基础上对其目标函数进行优化, 构造多分类模型, 进而实现多分类问题。这种方法为一次性求解法, 由于其计算复杂度比较高, 在实际应用中效率低, 所以并不常用。

b) 将多分类问题归结为多个两分类问题, 这样也就将一个复杂的问题转换为若干个简单的问题, 正因为这个优点, 这一思想下的算法研究得到了极大的发展。常用的算法主要有一对多[35,36]、一对一[37,38]、导向无环图[39,40]、二叉树[41]四种。文献[42]对这四种算法分别进行了详细的描述, 在训练复杂度、测试复杂度和分类准确率方面作了理论分析, 并利用数据对分析结果进行了验证。分析得出, 导向无环图的分类性能最优, 一对一的分类性能次之, 二叉树的分类性能较差, 一对多的分类性能最差; 在训练和测试耗时方面, 二叉树耗时最短, 导向无环图次之, 一对一和一对多的耗时相对较长。

此外, 很多学者还对上述算法进行了改进, 文献[43]提出了一种新的基于几何分布二叉树支持向量机多分类算法, 该方法是考虑到二叉树分类向量机分类的效果与二叉树的结构密切相关, 若要获得更好的效果和更高的效率, 就要使得二叉树高度尽量小而两个子类尽量易分, 实验表明这种新的方法具有较高的分类准确率和效率。文献[44]针对一对一多分类算法因在分类时存在不可分区域而影响了其分类效果的问题, 提出了一种一对一与基于紧密度判决相结合的多分类方法。这种方法采用基于紧密度决策解决不可分区, 依据样本到类中心之间的距离和基于 KNN(K-nearest neighbor)的样本分布情况相结合的方式构建判别函数来确定类别归属, 并用数据进行了测试, 其结果表明, 该方法能有效地解决不可分区域问题, 并且表现出比其他算法更好的性能。

3 SVM 应用

支持向量机方法在理论基础上有较强的优势, 它能够保证找到的极值解就是全局最优解而非局部最小值, 这也就决定了 SVM 方法对未知样本有较好的泛化能力。正因为这些优点, 使得 SVM 在应用方面得到了很多领域相关学者的广泛重视, 在回归估计、概率密度函数估计、模式识别等领域均有其应用成果, 其中模式识别是 SVM 方法的主要应用领域。在模式识别方面, SVM 方法主要应用于手写数字识别、语音识别、人脸检测与识别、文本分类等方面。

3.1 手写数字识别

手写数字识别(handwritten digit recognition)是模式识别学科的传统研究领域, 是光学字符识别技术的一个分支, 它主要研究的问题是如何让计算机自动辨认手写在纸张上的阿拉伯数字。SVM 方法应用在现实世界的第一个例子就是手写数字识别问题, Boser 等人用多种方法对美国邮政手写数字库识别进行实验[45], 结果表明, 人工识别错误率是 2.5%, 5 层神经网络错误率是 5.1%, 三种 SVM 方法的错误率分别是 4.0%、4.1%和 4.2%。这个实验的成功使得 SVM 方法成为数字识别领域的新工具。

SVM 方法对手写数字识别的研究主要集中在预处理、特征提取和分类器等三个方面。文献[46]就是针对预处理技术中二值化准确度的问题进行了研究, 对预处理中笔画丢失与断开和小内孔问题的解决提供了依据。文献[47]针对自由手写体因其书写风格差异大、上下文无关及识别准确度要求高等原因导致其识别难度大的问题, 提出了一种基于组合结构特征的自由手写体数字识别新算法, 通过扩展的字符结构特征识别算法自动、鲁棒地提取手写体数字字符端点、分叉点、横线等多种结构特征, 并组合应用这些结构特征构造决策树完成手写体字符的自动识别。文献[48]针对目前手写数字识别精度不高的问题, 提出了基于手写数字图像的空间、旋转、层次和结构特性的特征提取方法, 这种新的方法是把手写数字的统计和结构特征结合起来, 以特征提取方法为基础, 利用 LIBSVM 算法对手写数字特征进行了训练和识别。

3.2 语音识别

语音识别技术是人机接口应用的前沿技术之一, 目的是使计算机听懂人类的语言, 能够实现人类与计算机语言上的互通信息。

在语音识别技术中, 大多数在有背景噪声或训练和测试环境不同的情况下效果大打折扣, 所以某种程度上来说它们只适合于识别“干净”的语音。SVM 方法由于其能较好地解决小样本、非线性、高维数和局部极小点等实际问题的优点, 它比隐马尔可夫模型(hidden Markov models, HMM)、人工神经网络等方法具有更好的泛化能力和分类精确性, 更适合于语音识别。

国内早期将 SVM 方法用于语音识别[49,50]是在汉语数字和孤立词的语音识别方面, 实验的效果与 HMM 的基本相同, 但是在这个实验中, SVM 方法所用的核函数是最基本的, 并没有对它们进行描述。再者, 这些实验也是在“干净”即非噪声干扰的环境中完成的。文献[51]介绍了一种混合系统, 它将支持向量机与段长分布隐马尔可夫模型结合起来, 并用于普通话数字语音识别实验中, 效果虽有所改善, 但并不理想, 并且算法复杂。文献[52]提出了一种使用 SVM 在 HMM 系统基础上进行二次识别来提高易混淆语音识别率的方法, 这个方法通过引入置信度估计环节来提高系统的效率和性能。通过实验, 这种方法在对识别速度影响较小的情况下, 可使识别率比采用 SVM/HMM 混合结构模型的识别率有明显提高。文献[53]是针对语音识别系统在噪声环境中识别率差的问题, 提出了一种基于生境共享机制的并行结构人工鱼群算法优化 SVM 参数的方法, 并用实验证明了其有效性。

3.3 人脸检测与识别

人脸检测的基本思想[54]是用知识的或统计的方法对人脸建模, 比较所有可能的待检测区域与人脸模型的匹配度, 从而得到可能存在人脸的区域。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Osuna 等人[53]最早将 SVM 方法用在人脸检测技术中，方法是训练非线性的 SVM 分类器将人脸与非人脸进行分类，并取得了良好的效果。但是这个方法仍有其不足之处：a) SVM 方法本身的训练就需要大量的存储空间，而人脸检测需要收集大量的非人脸样本，这样就会影响分类器的性能；b) 非线性 SVM 分类器需要大量的支持向量，这就导致分类器的速度很慢。为了解决这些问题，文献[56]提出了一种基于层次型支持向量机的正面直立人脸检测方法，这个新型的分类器是由一个线性 SVM 组合和一个非线性 SVM 组成，由前者在保证检测率的情况下快速排除图像中绝大部分非人脸区域，后者再对人脸候选区域进行进一步的确认。实验证明，这个方法不仅有较高的检测率和较低的误检率，而且具有较小的计算量。人脸检测的技术一般都能很好地解决正面端正情况下的检测问题，但是对于姿态变化下的检测技术还没有得到很好的解决。文献[57]针对上述问题提出了基于 SVM 的人脸姿态判定算法，将人脸姿态划分为六类，从一个多姿态人脸库中手工标定出 1 800幅人脸图像作为训练样本集，分别训练基于支持向量分类(support vector classification，SVC)和基于支持向量回归(support vector regression，SVR)两种分类器；另外标定出 300 幅人脸图像作为测试样本，实验结果表明，SVM 方法对于解决姿态判定问题是很有效的。文献[58]提出了一种多姿态人脸特征定位方法，这种方法是让支持向量机先经过大量多姿态五官样本的训练，然后再用它在搜索区域中辨别候选的眼、鼻及嘴区域，实现多姿态人脸五官的精确定位。实验证明，这种方法有良好的鲁棒性和精确性。

3.4　文本分类

文本分类(text categorization 或 text classification)就是根据给定文本的内容，将其判别为事先确定的若干个文本类别中的某一类或某几类的过程。

最先将 SVM 方法用于文本分类中的是 Joachims[59]和 Dumais[60]，他们在不同的语料库中进行了反复的实验，实验结果表明，SVM 方法在文本分析中的应用比其他方法如贝叶斯、决策树更有效，也有较好的泛化能力，并且克服了高维表示中的困难。这一研究结果引起了很多相关学者的注意，他们对 SVM 方法应用在文本分类中进行了很多研究。文献[61]提出了 SVM 主动学习策略，并将其应用在文本分类中。文献[62]针对文本过滤过程中训练样本少的问题，研究了一种交互式 SVM。文献[63]提出了训练直推式 SVM 的方法，以解决在混合文本训练集上训练 SVM 的问题。文献[64]在训练直推式 SVM 的基础上提出了基于 SVM 的渐进直推式分类学习算法，并取得了良好的效果。文献[65]针对大规模文档的高效分类提出了一种基于加权近似支持向量机文本分类算法，这种方法是在近似支持向量机的基础上增加权值从而得到了改进。实验结果证明，这种算法的分类质量和速度都有所改进。文献[66]在多项式核函数支持向量机的基础上，将条件正定核混合多项式改进为一种混合型核函数，并用实验证明了其优越性。

4　结束语

SVM 以统计学习理论为基础，有极其严格的理论依据，建立在 VC 维理论和结构风险最小原理基础上，同时又引入了核函数，使其算法可以向高维空间映射，但又避免了复杂的计算，并有效克服了位数灾难的问题。由于上述这些比较显著的优点，它也被应用在了很多的领域并取得了好的成果。虽然 SVM 理论和算法经过这么多年的研究与应用，也都有了很大程度上的发展与进步，但是在一些问题上，如训练速度、核函数的选取、计算存储容量等方面还需要发展与完善，进一步的研究方向包括：

a) 在自身算法的完善方面。SVM 的性能在很大程度上还是依赖于核函数的选取，所以在核函数的选取上还需要有进一步的发展；SVM 的性能还表现在训练效率和泛化能力上，所以怎样提高这两个问题也是需要进一步研究的重点。

b) 在算法结合其他学科的改进方面。上述已经说到 SVM 的训练速度与精度也是其性能的一种重要标准，怎样在大规模样本集的情况下提高其训练速度和训练精度也是一个需要进一步研究和改进的方面。虽然自 SVM 提出以来也出现了很多新型的算法，如本文介绍的 FSVM、LS-SVM、LSVM，还有一些其他算法，如 SOR(successive overrelaxation)、SSVM(smooth support vector machines)、ASVM(adaptive support vector machines)、PSVM(parallel support vector machines)、TWSVMS(twin support vector machines)、GSVM(granular support vector machines)等。这些算法都能针对某些方面提高其性能，如收敛速度、泛化能力，但是这些新型算法也都各有其缺点，如 FSVM 训练速度慢、LSVM 在非线性问题中不能处理大样本数据问题等。所以在算法的改进方面有待进一步的完善。近年来，有研究者提出了一些分布式支持向量机，这种算法的发展虽然还不是很成熟，但是它面对大规模的训练样本集有很好的优势，因而这方面的改进与完善也将会是一个比较好的研究方向。

c) 在应用方面。虽然 SVM 在理论上有比较突出的优势，在算法上也一直有所改进并有很多新型算法的出现，并应用在模式识别、回归估计等领域。但是它在这些领域的应用方面相比理论和算法的研究还是很滞后，很多实验的研究报道也仅局限于仿真与对比实验，所以在应用方面还需要进一步加强应用到生活中的众多领域。

参考文献：

[1] CRISTIANINI N，TAYLOR J S. 支持向量机导论[M]. 李国正，王猛，曾华军，译. 北京：电子工业出版社，2004.

[2] 曾水玲，徐蔚鸿. 基于支持向量机的手写体数字识别[J]. 计算机与数字工程，2006，34(10)：104-106.

[3] 刘晓亮，丁世飞. SVM 用于文本分类的适用性[J]. 计算机工程与科学，2010，32(6)：106-108.

[4] 张瑜慧，胡学龙，陈琳. 基于支持向量机的图像分类[J]. 扬州大学学报，2007，10(2)：42-46.

[5] VAPNIK V N. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2000.

[6] 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.

[7] BOSER B E，GUYOU I M，VAPNIK V N. A training algorithm for optimal margin classifiers[C]//Proc of the 5th Annual Workshop on Computational Learning Theory. New York：ACM Press，1992：144-152.

[8] OSUNA E，FRENUD R，GIROSI F. An improved training algorithm for support vector machines[C]//Proc of IEEE Workshop on Neural Networks for Signal Processing. 1997：276-285.

[9] PLATT J. Fast training of support vector machines using sequential minimal optimization[C]//Advances in Kernel Methods-Support Vector Learning. Cambridge，MA：MIT Press，1998.

[10] PLATT J. Fast training of support vector machines using sequential minimal optimization[M]//Advances in Kernel Methods-Support Vector Learning. Cambridge，MA：MIT Press，1998.

[11] 李向东，王进华. 支持向量机分解算法研究[J]. 计算机与数字工程，2007，35(5)：9-12.

[12] KEERTHI S S，GIIBERT E G. Convergence of a generalized SMO al-

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

gorithm for SVM classifier design [J]. Machine Learning, 2002, 46(1-3): 351-360.
[13] KEERTHI S S, SHEVADE S, BHATTACHARYY C. Improvements to Platt's SMO algorithm for SVM classifier design [J]. Neural Computation, 2002, 13(3): 637-649.
[14] 孙剑, 郑南宁, 张志华. 一种训练支撑向量机的改进贯序最小优化算法 [J]. 软件学报, 2002, 13(10): 2007-2013.
[15] 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, 14(5): 919-925.
[16] 周晓剑, 马义中, 朱嘉钢. SMO 算法的简化及其在非正定核条件下的应用 [J]. 计算机研究与发展, 2010, 47(11): 1962-1969.
[17] FAN R E, CHEN P H, LIN C J. Working set selection using second order information for training SVM [J]. Journal of Machine Learning Research, 2005, 6(12): 1889-1918.
[18] 翟永杰, 王子杰, 黄宝海, 等. 基于 PSO 优化的 SMO 算法研究及应用 [J]. 华北电力大学学报, 2008, 35(1): 57-61.
[19] 王越, 吕奇峰, 王泉, 等. 一种改进的支持向量机序列最小优化算法 [J]. 重庆理工大学学报, 2013, 27(3): 76-79.
[20] 朱齐丹, 张智, 邢卓异. 支持向量机改进序列最小优化学习算法 [J]. 哈尔滨工程大学学报, 2007, 28(2): 183-188.
[21] LIN Chun-fu, WANG Sheng-de. Fuzzy support vector machines [J]. IEEE Trans on Neural Networks, 2002, 3(2): 464-471.
[22] LI Xue-hua, SHU Lan. Fuzzy theory based support vector machines classifier [C] //Proc of the 5th International Conference on Fuzzy Systems and Knowledge Discovery. 2008: 600-604.
[23] 邬啸, 魏延, 吴瑕. 改进的双隶属度模糊支持向量机 [J]. 重庆师范大学学报, 2011, 28(5): 49-52.
[24] 刘三阳, 杜喆. 一种改进的模糊支持向量机算法 [J]. 智能系统学报, 2007, 2(3): 30-33.
[25] 杜喆, 刘三阳, 齐小刚. 一种新隶属度函数的模糊支持向量机 [J]. 系统仿真学报, 2009, 21(7): 1901-1903.
[26] SUYKENS J, VANDEWALLE J. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293-300.
[27] 万辉, 魏延. 一种改进的最小二乘支持向量机算法 [J]. 重庆师范大学学报, 2010, 27(4): 69-72.
[28] 刘翠芬, 柏海滨. 改进的最小二乘支持向量机在入侵检测系统中的应用 [J]. 现代计算机, 2009(6): 39-42.
[29] 张伟, 胡昌华, 焦李成. 基于混合参数优化的 LLSVM 与时间序列预测 [J]. 电子测量与仪器学报, 2007, 21(5): 55-59.
[30] MANGASARIAN O L, MUSICANT R. Lagrangian support vector machines [J]. Journal of Machine Learning Research, 2001, 1(9): 161-177.
[31] DIETTERICH T G, BAKIRI G. Solving multi-class learning problems via error-correcting output codes [J]. Journal of Artificial Intelligent Research, 1995, 2(1): 263-286.
[32] WESTON J, WATKINS C. Multi-class support vector machines [M]. Brussels: [s. n.], 1999.
[33] 应伟, 王正鸥, 安金龙. 一种基于改进的支持向量机的多类文本分类方法 [J]. 计算机工程, 2006, 32(16): 74-76.
[34] 张苗, 张德贤. 多类支持向量机文本分类方法 [J]. 计算机技术与发展, 2008, 18(3): 139-141.
[35] POLAT K, GUNES S. A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems [J]. Expert Systems with Application, 2009, 36(2): 1587-1592.
[36] 郭磊, 陈进, 朱义. 小波支持向量机在滚动轴承故障诊断中的应用 [J]. 上海交通大学学报, 2009, 43(4): 678-682.
[37] 韩兰胜, 邹梦松, 刘其文. 多类支持向量机的病毒行为检测方法 [J]. 计算机应用, 2010, 30(1): 181-185.
[38] ARINDAM C, KAJAL D, DIPAK C. A comparative study of kernels for the multi-class support vector [C] //Proc of the 4th International Conference on Natural Computation. Washington DC: IEEE Computer Society, 2008: 3-7.
[39] MANIKANDAN J, VENKATARAMANI B. Study and evaluation of a multi-class SVM classifier using diminishing learning technique [J]. Neurocomputing, 2010, 73(10-12): 1676-1685.
[40] WIDODO A, YANG B S. Support vector machine in machine condition monitoring and fault diagnosis [J]. Mechanical Systems and Signal Processing, 2007, 21(6): 2560-2574.
[41] 吴德会. 基于多分类支持向量机的智能辅助质量诊断研究 [J]. 系统仿真学报, 2009, 21(6): 1689-1693.
[42] 薛宁静. 多类支持向量机分类器对比研究 [J]. 计算机工程与设计, 2011, 32(5): 1792-1795.
[43] 李雷, 房小萍, 张宁. 一种基于几何分布的新支持向量机多分类方法 [J]. 计算机技术与发展, 2012, 22(11): 172-175.
[44] 单玉刚, 王宏, 董爽. 改进的一对一支持向量机多分类算法 [J]. 计算机工程与设计, 2012, 33(5): 1837-1841.
[45] 边肇祺, 张学工. 模式识别 [M]. 北京: 清华大学出版社, 2000.
[46] 马立权, 李维, 蔡韩辉, 等. 手写数字识别中的预处理技术研究 [J]. 仪器仪表学报, 2001, 22(3): 263-265.
[47] 陈军胜. 组合结构特征的自由手写体数字识别算法研究 [J]. 计算机工程与应用, 2013, 49(5): 179-184.
[48] 双小川, 张克. 基于统计和结构特征的手写数字识别研究 [J]. 计算机工程与设计, 2012, 33(4): 1533-1537.
[49] 苏毅, 吴文虎, 郑方, 等. 基于支持向量机的语音识别研究 [C] // 第六届全国人机语音通讯学术会议论文集. 2001.
[50] XIE Xiang, JING Ming. Mandarin digits speech recognition using support vector machines [J]. Journal of Beijing Institute of Technology, 2005, 14(1): 9-12.
[51] LIU Jing-wei, WANG Zuo-ying, XIAO Xi. A hybrid SVM/DDBHMM decision fusion modeling for robust continuous digital speech recognition [J]. Pattern Recognition Letters, 2007, 28(8): 912-920.
[52] 王欢良, 韩纪庆, 李海峰, 等. 基于 HMM/SVM 两级结构的汉语易混淆语音识别 [J]. 模式识别与人工智能, 2006, 19(5): 578-584.
[53] 白静, 杨利红, 张雪英. 一种面向语音识别的抗噪 SVM 参数优化方法 [J]. 中南大学学报, 2013, 44(2): 604-611.
[54] CHELLAPPA R, WILSON C L, SIROHEY S. Human and machine recognition of faces: a survey [J]. Processing of the IEEE, 1995, 83(5): 705-740.
[55] OSUNA E, FRENUD R. Training support vector machines: an application to face detection [C] //Proc of Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 1997: 130-136.
[56] 马勇, 丁晓青. 基于层次型支持向量机的人脸检测 [J]. 清华大学学报: 自然科学版, 2003, 43(1): 35-38.
[57] 叶航军, 白雪生, 徐光祐. 基于支持向量机的人脸姿态判定 [J]. 清华大学学报: 自然科学版, 2003, 43(1): 67-70.
[58] 傅由甲, 相入喜, 黄鸿, 等. 基于支持向量机的多姿态人脸特征定位 [J]. 计算机工程, 2011, 37(17): 7-10.
[59] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features [C] //Proc of the 10th European Conference on Machine Learning. 1998: 137-142.
[60] DUMAINS S. Using SVMs for text categorization [J]. IEEE Intelligent Systems, 1998, 13(4): 21-23.
[61] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. Journal of Machines Learning Research, 2002, 2(1): 45-66.
[62] 卢增祥, 李衍达. 交互 SVM 学习算法及其在文本信息过滤中的应用 [J]. 清华大学学报: 自然科学版, 1999, 39(7): 93-97.
[63] JOACHIMS T. Transductive inference for text classification using support vector machines [C] //Proc of the 16th International Conference on Machine Learning. 1999: 200-209.
[64] 陈毅松, 汪国平, 董士海. 基于支持向量机的渐进直推式分类学习算法 [J]. 软件学报, 2003, 14(3): 451-460.
[65] 庄东, 陈英. 基于加权近似支持向量机的文本分类 [J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1787-1790.
[66] 赖苏, 熊忠阳, 江帆, 等. 利用改进的多项式核函数支持向量机进行文本分类 [J]. 重庆大学学报: 自然科学版, 2012, 35(S1): 46-51.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net