中国科学: 数学 2020 年 第 50 卷 第 9 期: 1233 ~ 1248
SCIENTIA SINICA Mathematica
论 文

支持向量机的关键问题和展望

献给越民义教授 100 华诞

邵元海[1], 刘黎明[2], 黄凌伟[1,3], 邓乃扬[4*]

1. 海南大学管理学院, 海口 570228;
2. 首都经济贸易大学统计学院, 北京 100083;
3. 海南大学经济学院, 海口 570228;
4. 中国农业大学理学院, 北京 100083
E-mail: shaoyuanhai@hainanu.edu.cn, llm5609@163.com, xhuanglw@163.com, dengnaiyang@cau.edu.cn

收稿日期: 2020-01-13; 接受日期: 2020-03-19; 网络出版日期: 2020-09-14; * 通信作者
国家自然科学基金 (批准号: 11871183, 61866010 和 11926349)、海南省自然科学基金 (批准号: 118QN181) 和北京市自然科学基金 (批准号: 9172003) 资助项目

摘要 作为机器学习的主要方法之一, 支持向量机不仅有坚实的统计学习理论基础, 而且在众多领域中表现出优秀的泛化性能, 因此受到了广泛关注. 然而近几年来, 相比于深度学习的蓬勃发展, 支持向量机的研究进展缓慢. 本文从支持向量机的本质出发, 探讨支持向量机的理论方法与深度学习等机器学习热点研究的交叉与融合, 提出一些新的思路. 具体地, 包括 3 个方面: 支持向量机的大间隔原则及其带来的低密度性、核映射的高维划分技巧及其统计学习理论, 以及支持向量机的浅层学习模式向深度学习和广度学习的拓展. 同时, 从这 3 个方面分别提出支持向量机研究中可以进一步挖掘的优良性质, 并展望未来可能诱导出的理论和方法.

关键词 支持向量机 统计学习 核学习 机器学习 最优化 深度学习

MSC (2010) 主题分类 90C99, 62-07

1 前言

支持向量机 (support vector machine, SVM) [1,2] 是机器学习领域最重要的方法之一, 是借助于统计与优化方法解决机器学习问题的强有力的工具. 它受到了国内外学者和工业界的广泛青睐 (参见文献 [3,4]). 最初的 SVM [1] 是针对两分类问题提出的. 它的基本思想是, 先将在较低维空间中的原始数据映射到近似线性可分的一个高维空间中, 然后在高维空间中进行基于最大间隔原则的线性学习. SVM 利用核函数进行映射, 并使用稀疏损失等技巧, 巧妙地实现了上述思想, 从而获得了巨大成功.

英文引用格式: Shao Y H, Liu L M, Huang L W, et al. Key issues of support vector machines and future prospects (in Chinese). Sci Sin Math, 2020, 50: 1233–1248, doi: 10.1360/SSM-2020-0015

上述思想和做法很快便渗透到机器学习的其他领域, 这不仅包括比较经典的回归问题 [2]、多分类问题 [5]、多示例问题 [6]、多标记问题 [7]、多视角问题 [8]、特征压缩问题 [9]、时间序列问题 [10]、半监督学习问题 [11] 和无监督学习问题 [12], 而且包括深度学习 [13] 和广度学习 [14] 等新的研究热点. SVM 能够迅速发展的主要原因是, 它的原理简单、清晰, 有坚实的统计学习理论基础, 实际计算效果突出, 并在多个应用领域表现出色. 当前关于 SVM 的理论、模型和算法的研究仍在继续, 例如, 从理论上研究如何提高各种 SVM 的泛化能力使其达到期望风险更紧的上界 [2,3], 从模型上研究新的正则化项、损失函数和核函数 [15–17] 的构建及其适用范围. 此外, 由于 SVM 需要求解优化问题, 相应的快速求解算法 [18–20] 也是一个研究方向.

在传统机器学习中, SVM 已被公认为是处理小样本机器学习问题的典范技术. 然而, 到了当今大数据时代, 数据多呈现规模巨大和数据价值密度低等新的特点, 这使得 SVM 的发展遇到了新的重大挑战. 同时, 相对于深度学习和强化学习等新学习范式的蓬勃发展, SVM 的研究陷入了低谷. 本文摒弃将它们对立的观点, 而强调它们的融合与交叉. 我们将从剖析支持向量机的本质出发, 探讨一些新的研究思路. 具体包括支持向量机 3 个方面的问题: 最大间隔思想与稀疏性、核函数与相应的统计学习理论, 以及浅层学习到深度学习与广度学习. 同时, 从这 3 个方面分别提出潜在的挑战性课题, 并展望可能的发展方向.

如前所述, SVM 涉及了机器学习中众多问题. 为了便于讨论, 这里重点以监督学习的两分类问题为例展开论述. 两分类学习问题可描述为: 给定训练集

$$T = \{(x_i, y_i) \mid i = 1, \ldots, m\} \in (\mathbb{R}^n \times \mathcal{Y})^m, \tag{1.1}$$

其中 $x_i \in \mathbb{R}^n$ 为第 i 个数据点的输入向量, $y_i \in \mathcal{Y} = \{+1, -1\}$ 为第 i 个数据点的输出类别标记, $y_i = +1$ 和 $y_i = -1$ 分别对应正类和负类, $i = 1, \ldots, m$, m 是数据点的个数. 试据此寻找 $\mathbb{R}^n$ 空间上的一个实值函数 $g(x)$, 以便用决策函数

$$f(x) = \operatorname{sign}(g(x)), \tag{1.2}$$

推断任一输入 x 对应的输出 y.

图 1 是两分类学习的一个示例, 图 1(a) 绘出了具有 20 个数据点的训练集 $T = \{(x_i, y_i) \mid i = 1, \ldots, 20\} \in (\mathbb{R}^2 \times \mathcal{Y})^{20}$, 其中输入 $x_i = ([x_i]_1, [x_i]_2)^\top$ 是二维向量, 它对应的输出 $y_i = +1$ 和 $y_i = -1$ 分别用星号和圆点表示, $i = 1, \ldots, 20$. 考虑如何构造 $g(x)$ 以推断任一新的输入 x 对应的输出 y.

我们用上述示例说明 SVM 的基本思想. 从图 1(a) 中可以看出, 不能采取用一条直线把两类输入分开的方式, 即不能期望对该数据直接使用线性分类器. 所以, SVM 首先通过一个映射 ϕ 将图 1(a) 中的数据输入 $x_i = ([x_i]_1, [x_i]_2)^\top$ 映射到一个容易线性划分的空间 $\mathbf{x}_i = ([\mathbf{x}_i]_1, [\mathbf{x}_i]_2)^\top$, $i = 1, \ldots, 20$, 如图 1(b) 所示. 然后在映射后的空间中构造一个由两条平行直线构成的能够正确分划两类数据点的带子, 并从中找到带宽最大者, 即两条平行直线之间间隔最大者, 参见图 1(c). 最后用这两条平行直线的中间线作为分类的分划线.

把上述基本思想应用于一般的两分类学习问题 (1.1) 和 (1.2), 便得到最优化模型

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } y_i(\langle w, \phi(x_i)\rangle + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, m, \end{aligned} \tag{1.3}$$

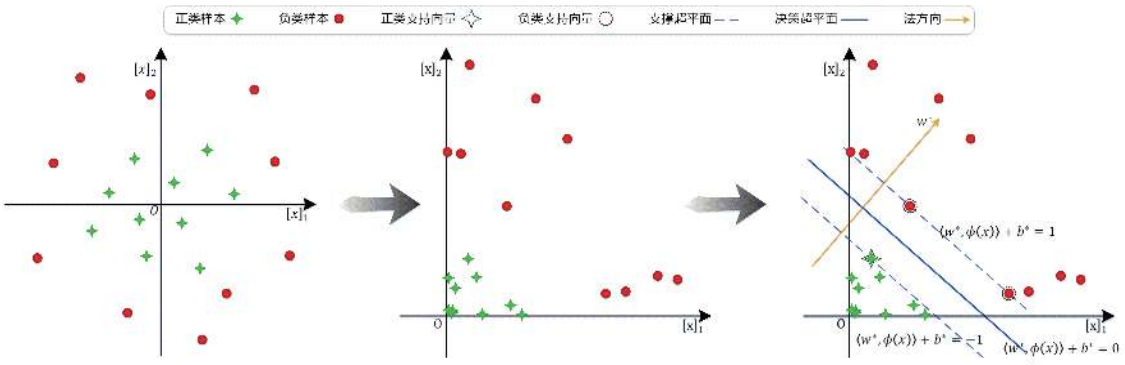


图 1 (网络版彩图) 两分类学习问题 SVM 示意图

其中 $\langle\cdot,\cdot\rangle$ 是内积, $C\geqslant 0$ 是调节参数, $\phi(\cdot)$ 是对输入数据的映射函数. $\langle w,\phi(x)\rangle+b=0$ 对应于示例中的分划直线, 称为决策超平面. $\langle w,\phi(x)\rangle+b=1$ 和 $\langle w,\phi(x)\rangle+b=-1$ 称为支撑超平面. 约束条件是使得给定的输入尽可能地在支撑超平面的两侧. 变量 ξ 的引入是允许训练数据的输入违反必须在两侧的限制. 极小化目标函数的第二项意味着希望违反的程度尽可能小些. 最后, 极小化目标函数的第一项使得两个支撑超平面之间的距离 ($\frac{2}{\|w\|^2}$) 尽可能大, 它体现了最大间隔原则.

优化问题 (1.3) 可以通过其对偶问题求解, 其对偶问题为

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\alpha^{\top}G\ \alpha-e^{\top}\alpha\\&\text{s.t.}\ \ \alpha^{\top}\boldsymbol{y},\quad 0\leqslant\alpha\leqslant C,\end{aligned}\tag{1.4}$$

其中 $\alpha=(\alpha_1,\ldots,\alpha_m)^{\top}$ 是对偶变量; e 是 m 维的全 1 向量; $\boldsymbol{y}=(y_1,\ldots,y_m)^{\top}$; 矩阵 G 定义为 $G_{ij}=y_iy_j[\phi]_{ij}=y_iy_j\langle\phi(x_i),\phi(x_j)\rangle$, $i=1,\ldots,m$, $j=1,\ldots,m$. SVM 的一个精彩之处在于往往并不需要给出 $\phi(x)$ 的具体形式. 注意到 $\phi(x)$ 在 (1.4) 中总是以内积的形式出现, 所以引入核函数 $K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle$, 就可以用核函数替代 $\phi(x)$. 从而得到 SVM 的对偶问题的另一形式

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\alpha^{\top}\bar{G}(K)\ \alpha-e^{\top}\alpha\\&\text{s.t.}\ \ \alpha^{\top}\boldsymbol{y},\quad 0\leqslant\alpha\leqslant C,\end{aligned}\tag{1.5}$$

其中矩阵 $\bar{G}(K)$ 定义为

$$\bar{G}_{ij}(K)=y_iy_j[K]_{ij}=y_iy_jK(x_i,x_j),\quad i=1,\ldots,m,\quad j=1,\ldots,m.$$

在得到对偶问题 (1.5) 的最优解 $\alpha^*=(\alpha_1^*,\ldots,\alpha_m^*)^{\top}$ 后, 决策函数可写为

$$f(x)=\operatorname{sign}\left(\sum_{i=1}^{m}y_i\alpha_i^*K(x_i,x)+b^*\right),\tag{1.6}$$

其中

$$b^*=y_j-\sum_{i=1}^{m}y_i\alpha_i^*K(x_i,x_j),$$

这里 j 是位于开区间 $(0,C)$ 内的 α^* 的某个分量 α_j^* 的下标. 显然, (1.6) 就是所要的 (1.2) 的形式.

以上介绍了两分类学习问题 SVM 的基本思想和实施步骤. 下面将从 3 个方面对 SVM 进行分析和展望.

2 最大间隔原则与稀疏性

2.1 最大间隔原则与低密度区域

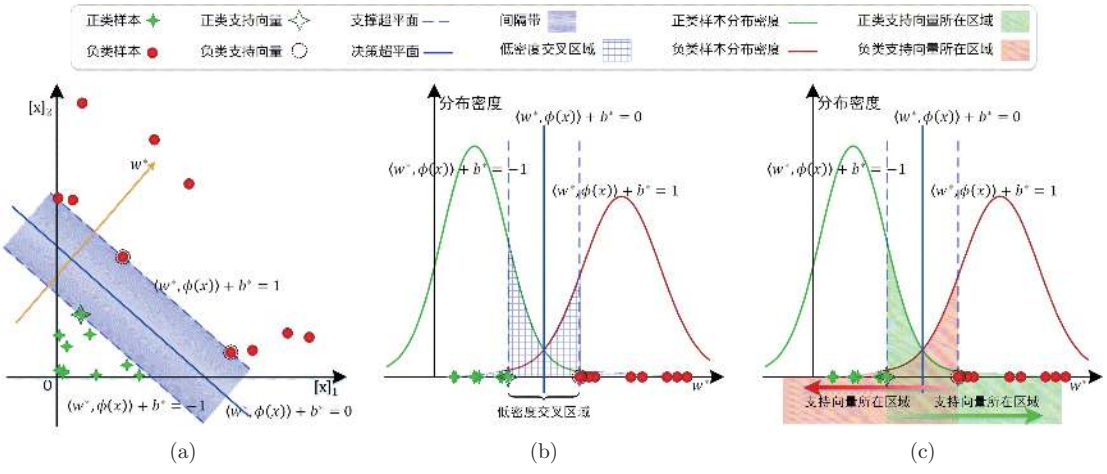
首先, 再明确一下最大间隔原则的几何意义. 仍以图 1 中的数据为例, 图 2(a) 给出了 SVM 构造的一对最大间隔支撑超平面 $\langle w^*, \phi(x)\rangle + b^* = 1$ 和 $\langle w^*, \phi(x)\rangle + b^* = -1$, 它们的法方向为 w^*, 它们之间的间隔为 $\frac{2}{\|w^*\|}$. 两条支撑超平面所夹区域即蓝色区域, 没有数据点. 最大化支撑超平面之间的距离即使得蓝色区域达到最大, 它们中间的决策超平面距离两类数据点的距离也是最大化的. 可以想象, 这是最稳妥的方案.

现在从概率分布角度考察最大间隔的含义. 假设数据服从 Gauss 分布. 如果将数据点向法方向 w^* 投影, 同时标记出支撑超平面对应的位置, 应该大体可以得到图 2(b) 所示的情形. 容易看出对应两个支撑超平面之间的部分是两类数据点分布交叉且都属于低密度的区域. 因此, 最大化间隔法可以理解为寻找造成最低密度交叉区域的投影, 然后选取适当的中间位置作为决策超平面, 从而使得潜在的犯错风险最小. 应该指出, 这一解释不仅适用于图中所示的线性可分情形, 而且对一般情形也成立.

下面指出低密度交叉区域与支持向量所在区域的关系. 这里支持向量定义为对偶问题 (1.4) 最优解 α^* 中非零分量 α_i^* 对应的输入 x_i, 即在决策函数 (1.6) 中起作用的数据点. 可以证明, 在图 2(b) 中, 每类的支持向量都落在它的分布曲线形成的网格区域的下方. 图 2(c) 明确描述了这一现象. 由此可见, 支持向量所在区域即为支撑超平面对应的低密度区域一侧, 它们也是容易被错分的输入.

下面列出最大间隔原则与低密度区域的性质.

性质 1 最大间隔原则得到的间隔带子对应着两类数据交叉且都是低密度的区域, 该区域是容易错分的区域, 在该区域决策即找到犯错尽可能少的区域进行决策 $\Rightarrow$ 最大化间隔对应着寻找数据交叉且低密度区域.



(a) (b) (c)

图 2 (网络版彩图) 最大间隔的投影示意图

由以上性质可知, 处理两分类问题的最大间隔原则即寻找数据分布最低密度的区域. 这一思想可以推广到其他的机器学习中. 这里以图 3 所示的回归问题为例, 说明支持向量回归机是如何联系到“低密度区域” 的. 图 3(a) 中红色的圆数据点构成了训练集 $T=\{(x_1,y_1),\ldots,(x_{10},y_{10})\}$, 其中 $x_i\in\mathbb{R}$ 是输入, $y_i\in\mathbb{R}$ 是输出, $i=1,\ldots,10$. 支持向量回归机构造一对函数 $\underline{f}^*(x)=\langle w^*,\phi(x)\rangle+b^*-\varepsilon$ 和 $\overline{f}^*(x)=\langle w^*,\phi(x)\rangle+b^*+\varepsilon$, $\varepsilon>0$, 两个函数构成 ε- 带, 如图中蓝色区域所示. 希望 ε- 带尽可能包含所有数据点, 同时希望 ε 不能太大. 最终得到回归函数 $f^*(x)=\langle w^*,\phi(x)\rangle+b^*$. 与图 2(b) 对应, 可画出数据点向法方向 w^* 上投影的概率分布图, 如图 3(b) 所示. 显然, ε- 带应该是数据点分布的高密度区域, 而 ε- 带外面应该是数据点分布的低密度区域, 也是支持向量的所在区域.

2.2 损失函数与稀疏性

SVM 通过极小化结构风险 [3] 来近似极小化期望风险是有理论依据的. 从原始优化问题 (1.3) 出发, 有

$$\xi_i=(1-y_i(\langle w,\phi(x_i)\rangle+b))_+=(1-y_ig(x_i))_+,$$

其中 $(\cdot)_+$ 表示取正函数. $\sum_{i=1}^m\xi_i$ 可推广为一般的经验风险或训练数据的损失 $\sum_{i=1}^m l(g(x_i),(x_i,y_i))$, $\frac{1}{2}\|w\|^2$ 可推广为一般的正则项 $\frac{1}{2}\|g\|_{\mathcal{F}}$, 于是, SVM 模型可表示为如下极小化结构风险 (正则项 + 训练损失) 的一般形式:

$$\min_g\frac{1}{2}\|g\|_{\mathcal{F}}+C\sum_{i=1}^m l(g(x_i),(x_i,y_i)). \tag{2.1}$$

就损失而言, SVM 最主要的特点是具有稀疏性, 即训练数据的损失函数值 $l(g(x_i),(x_i,y_i))$ 会有大量的 0 元素或取值相同的元素的情形. 如果损失值为 0, 则相应数据点无损失; 如果取值相同, 则相应数据点都可集中到同一个数据点. 两分类问题 SVM 常用的几个损失函数 $l(g(x),(x,y))$ 可表示为图 4(a) 中的形式, 包括铰链激活函数、L2 铰链激活函数、斜坡激活函数和硬间隔激活函数, 其中横坐标为 $y(\langle w,\phi(x)\rangle+b)$, 纵坐标是损失值的大小. 从图 4(a) 可以看出, 训练数据点落入某些区域将不被损失函数惩罚或具有相同的惩罚. 稀疏性使得损失函数对某些训练数据点不起作用或具有相同的作用,

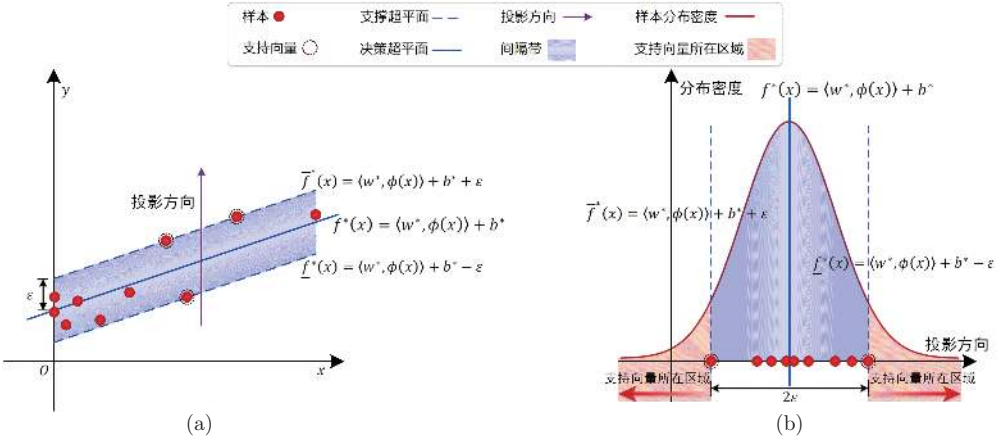


图 3 (网络版彩图) 回归问题 SVM 最大间隔 (低密度区域) 的示意图

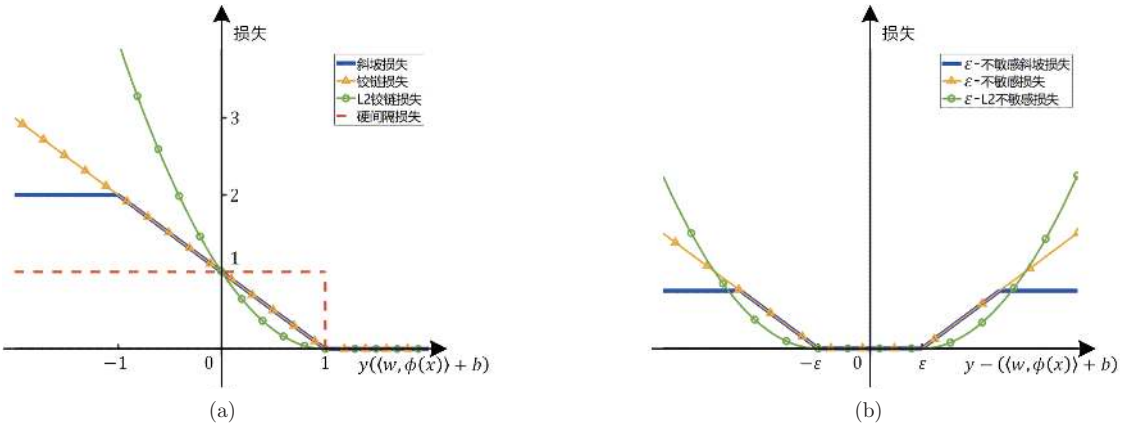


图 4 (网络版彩图) SVM 稀疏损失函数示例

从而, 这些训练数据对 SVM 的决策没有影响或具有一致的影响. 这些没有影响的训练输入也称之为非支持向量, 删去非支持向量对 SVM 来说也是允许的.

下面给出 SVM 损失函数的性质.

性质 2 稀疏损失函数是对大量训练数据点出现 0 损失或具有相同损失值的损失函数 $\Rightarrow$ 损失为 0 或具有相同损失值的数据点对模型无效 (可剔除) 或效用相同.

具有稀疏性的损失函数不仅可以识别出对模型无效的训练数据点, 增强模型的可解释性, 而且剔除无效数据点后, 可提高模型的预测或训练速度. 因此, 稀疏损失函数在机器学习中被广泛研究并应用到各个领域. 以回归问题为例, 图 4(b) 给出了支持向量回归机常用的稀疏损失函数, 包括 ε- 不敏感损失、ε-L2 不敏感损失和 ε- 不敏感斜坡损失, 其中横坐标为 $y-(\langle w,\phi(x)\rangle+b)$, 纵坐标是损失值的大小. 更多稀疏损失函数的研究参见文献 [21–24].

2.3 展望

如何在大规模和数据价值密度低的数据中找到有价值的信息, 是当今大数据时代的主要课题之一. 因此, SVM 最大间隔 (寻找数据边界低密度区域思想) 和稀疏损失函数的进一步研究意义非凡. 展望的研究工作如下:

(i) 经典 SVM 最大间隔是对两类数据中点与点之间而言的, 最大间隔还可考虑数据粒度之间的间隔、数据局部的间隔、数据分布之间的间隔和不同度量意义下的间隔等多种形式 [25]. 利用不同的最大间隔思想识别不同数据边界低密度区域的研究 [26] 很有意义, 特别是对大规模数据的处理.

(ii) 支持向量是数据边界稀疏一侧且不易充分正确划分的数据点输入, 支持向量具有构造决策和重构数据边界等功能, 价值密度较高. 因此, 在大数据中支持向量及其价值值得深入研究. 例如, 利用支持向量进行数据压缩和数据表达, 利用支持向量寻找或构造重要数据和特征等问题. 此外, 构造更加稀疏的支持向量的模型和寻求高效算法也是很有价值的问题.

(iii) 构造并研究新的稀疏损失函数也是当前研究的热点, 这里包括将稀疏损失函数与最大间隔思想融合的研究, 更加稀疏的鲁棒非凸稀疏损失函数的研究, 定义在簇、局部和分布上的稀疏损失函数的研究, 以及大规模或超大规模数据问题的稀疏损失模型与算法的研究.

3 核与统计学习理论

尽管线性学习以其简单、易实现而被广泛应用, 但非线性学习问题更为普遍. 因而, 非线性学习一直是机器学习研究的热点. SVM 通过引入核函数, 巧妙地实现了从非线性学习到线性学习的过渡. 本节的重点是从非线性学习引出核学习的问题.

3.1 非线性学习与核函数

前面简要地介绍了使用核函数进行非线性学习的算法, 现在对核函数和非线性学习做进一步的分析. 首先, 以简单的示例说明核函数如何把非线性学习转化为线性学习. 以图 1(a) 中二维输入的两分类数据为例, 显然, 它不能用线性学习得到的直线进行分划, 而应该用非线性学习得到的曲线进行分划, 例如, 选择分划曲线为圆: $[x]_1^2+[x]_2^2=r^2$, 其中 $[\cdot]_1$ 和 $[\cdot]_2$ 分别表示 "$\cdot$" 的第 1 和 2 个分量, 如图 5(a) 所示. 下面希望把寻找分划曲线 (圆) 的问题转化为寻找分划直线的问题. 为此只需引进从原来的 x 空间到 $\mathbf{x}$ 空间的变换

$$\mathbf{x}=\phi(x):\Rightarrow [\mathbf{x}]_1=[x]_1^2,\quad [\mathbf{x}]_2=[x]_2^2,$$

这样就把 x 空间的圆 $[x]_1^2+[x]_2^2=r^2$ 变为 $\mathbf{x}$ 空间的直线 $[\mathbf{x}]_1+[\mathbf{x}]_2-r^2=0$ 了.

我们知道, SVM 从 x 到 $\mathbf{x}=\phi(x)$ 的变换是通过核函数实现的. 现在讨论核函数引起的度量上的差异. 事实上, 度量两个输入 x 与 x' 之间的距离, 在未引入核函数之前, 采用的是它们本身所在的 x 空间中的 Euclid 距离 $\|x-x'\|$; 而在引入核函数 $K(x,x')$ 之后, 采用的是变换后的 $\mathbf{x}$ 空间中的 Euclid 距离 $\|\mathbf{x}-\mathbf{x}'\|$. 注意到

$$\begin{aligned}\|\mathbf{x}-\mathbf{x}'\|^2&=((\mathbf{x}-\mathbf{x}')\cdot(\mathbf{x}-\mathbf{x}'))=(\mathbf{x}\cdot\mathbf{x})+(\mathbf{x}'\cdot\mathbf{x}')-2(\mathbf{x}\cdot\mathbf{x}'),\\&=(\phi(x)\cdot\phi(x))+(\phi(x')\cdot\phi(x'))-2(\phi(x)\cdot\phi(x')),\\&=K(x,x)+K(x',x')-2K(x,x').\end{aligned}\tag{3.1}$$

可见引进核函数 $K(x,x')$ 后, 认定 x 与 x' 的距离的平方 $d^2(x,x')$ 已经变为

$$d^2(x,x')=K(x,x)+K(x',x')-2K(x,x'),\tag{3.2}$$

这里并没有用到 ϕ 具体是什么.

下面再用简单的例子具体说明核函数引起的度量变化. 考虑两分类问题, 给定只有两个输入点的训练集

$$T=\{(x_1,y_1),(x_2,y_2)\}=\{(x_+,1),(x_-,-1)\}\in(\mathbb{R}^2\times y)^2,$$

其中 $x_+,x_-\in\mathbb{R}^2$, $y\in\{1,-1\}$, 如图 6 所示. 考察以下 3 种不同的核函数产生的分类结果.

(i) 取核函数 $K(x,x')=\langle\phi(x)\cdot\phi(x')\rangle=\langle x\cdot x'\rangle$.

求解对偶问题 (1.5), 当 $C\geqslant 2/\|x_+-x_-\|^2$ 时, 可得分划直线 $((x_+-x_-)\times(x-(x_++x_-)/2))=0$, 如图 6(a) 所示. 此时距离度量为通常的 Euclid 距离

$$d^2(x,x')=K(x,x)+K(x',x')-2K(x,x')=\langle x,x\rangle+\langle x',x'\rangle-2\langle x,x'\rangle=\|x-x'\|^2.\tag{3.3}$$

(ii) 取核函数 $K(x,x')=\langle\phi(x)\cdot\phi(x')\rangle=\langle[x]_1\cdot[x']_1\rangle$, 这里 $[\cdot]_1$ 表示向量 "$\cdot$" 的第一个分量.

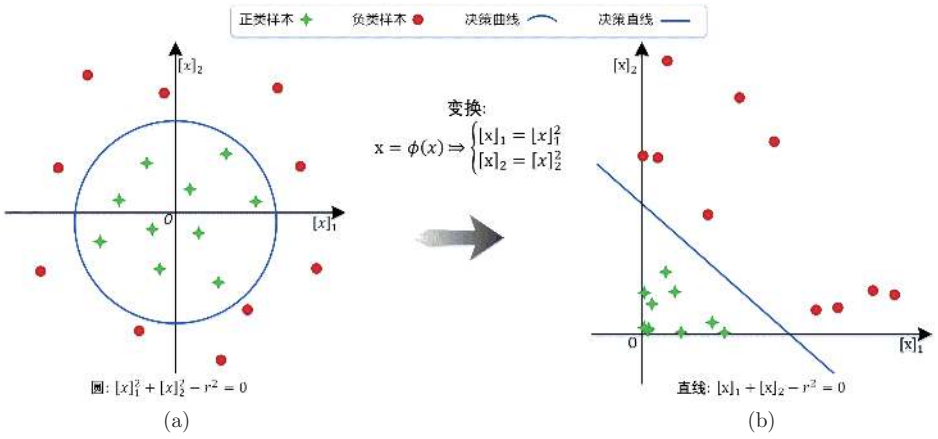


图 5 (网络版彩图) 从非线性到线性映射的示例

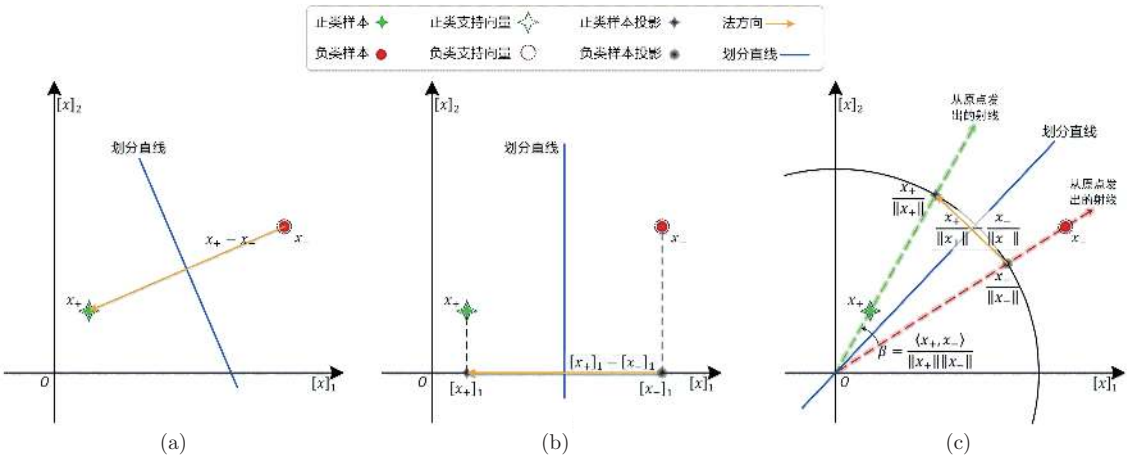


图 6 (网络版彩图) 核与度量之间关系的示例

求解对偶问题 (1.5), 当 $C \geqslant 2/\|[x_+]_1 - [x_-]_1\|^2$ 时, 可得分划直线 $(([x_+]_1 - [x_-]_1) \times (x - ([x_+]_1 + [x_-]_1)/2)) = 0$, 如图 6(b) 所示. 此时距离度量为

$$\begin{aligned} d^2(x, x') &= K(x, x) + K(x', x') - 2K(x, x') \\ &= \langle [x]_1, [x]_1 \rangle + \langle [x']_1, [x']_1 \rangle - 2\langle [x]_1, [x']_1 \rangle \\ &= ([x]_1 - [x']_1)^2. \end{aligned} \tag{3.4}$$

(iii) 取核函数

$$K(x, x') = \langle \phi(x) \cdot \phi(x') \rangle = \frac{\langle x \cdot x' \rangle}{\|x\| \|x'\|}.$$

求解对偶问题 (1.5), 当 $C \geqslant 2/(\|\frac{x_+}{\|x_+\|} - \frac{x_-}{\|x_-\|}\|^2)$ 时, 可得分划直线

$$\left(\frac{x_+}{\|x_+\|} - \frac{x_-}{\|x_-\|} \right) \times (x - 0) = 0,$$

如图 6(c) 所示. 此时距离度量为

$$\begin{aligned} d^2(x,x') &= K(x,x)+K(x',x')-2K(x,x') \\ &= \left(\frac{\langle x\cdot x\rangle}{\|x\|\|x\|}\right)+\left(\frac{\langle x'\cdot x'\rangle}{\|x'\|\|x'\|}\right)-2\left(\frac{\langle x\cdot x'\rangle}{\|x\|\|x'\|}\right) \\ &= \left|\frac{x'}{\|x'\|}-\frac{x}{\|x\|}\right|. \end{aligned} \tag{3.5}$$

上面讨论的是对分类问题引进的核函数, 核函数也容易推广到其他机器学习问题中. 如图 7 所示的简单的一维回归问题, 显然不能用直线而应该用曲线作为它的回归函数, 例如, 选择二次函数 $y=wx^2+b$. 为了把二次函数转化为线性函数, 考虑从 x 空间到 $\mathbf{x}$ 空间的变换

$$\mathbf{x}=\phi(x) :\Rightarrow \mathbf{x}=x^2, \tag{3.6}$$

这样, 二次函数 $y=wx^2+b$ 变为线性函数 $y=w\mathbf{x}+b$. 若取核函数 $K(x,x')=\langle\phi(x)\cdot\phi(x')\rangle=\langle x^2\cdot x'^2\rangle$, 则得到该问题的支持向量回归机核模型.

3.2 结构风险最小化与最优核学习

由 (2.1) 可知, SVM 极小化的是结构风险. 下面给出 SVM 极小化结构风险的泛化误差界的估计. 根据统计学习理论 [2,27], 假设在核空间中 SVM 经验风险为 0 时的最大间隔为 $\gamma^*=\frac{2}{\|w^*\|^2}$, 则对任意的 $\delta\in(0,1]$, SVM 的泛化误差 $\mathrm{Error}_{\mathrm{Test}}$ 至少以 $1-\delta$ 的概率满足

$$\mathrm{Error}_{\mathrm{Test}}\leqslant\frac{4\sqrt{\mathrm{tr}(K)}}{m\gamma^*}+3\sqrt{\frac{\ln(2/\delta)}{2m}}, \tag{3.7}$$

其中 m 是训练数据点的个数, $\sqrt{\mathrm{tr}(K)}$ 是核矩阵的迹.

由此可知, 泛化误差不仅与模型的最大间隔 γ^* 有关, 而且与核矩阵的迹 $\sqrt{\mathrm{tr}(K)}$ 有关. 极小化泛化误差需要同时极大化间隔和极小化核矩阵的迹, 于是, 人们考虑规范化的有界映射 SVM 模型 [28–30].

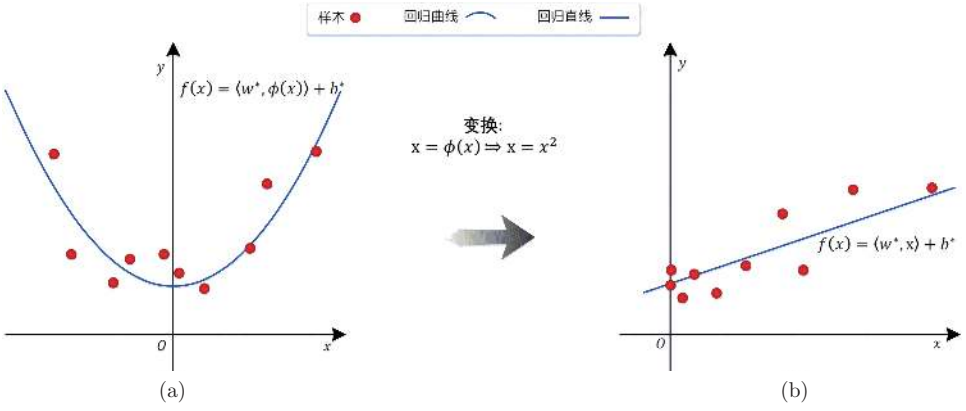


图 7 (网络版彩图) 回归问题的核映射示例

特别地, 为了寻找 SVM 最优的核矩阵, 文献 [31] 提出了如下模型:

$$\begin{aligned} &\min_{K\in\mathcal{K}} \min_{\alpha} \ \frac{1}{2}\alpha^{\top}G(K)\alpha - e^{\top}\alpha \\ &\text{s.t. } \alpha^{\top}\boldsymbol{y}, \quad 0 \leqslant \alpha \leqslant C, \\ &\qquad \mathrm{tr}(K) = r, \end{aligned} \tag{3.8}$$

其中 $\mathcal{K}$ 是所有核函数构成的集合, r 是一个正的参数. 问题 (3.8) 开启了从理论上学习最优核的研究. 然而此问题不易求解, 因此, 该文献假设最优核矩阵可由一组 (k 个) 恰当的核矩阵作为基底线性组合而成, 即设

$$K = \sum_{j=1}^{k} \mu_j K_j, \quad \mu_j \geqslant 0, \quad j \in \{1, \ldots, k\}. \tag{3.9}$$

将上式代入问题 (3.8), 便将其转化为一个半定规划问题了.

3.3 展望

非线性学习是机器学习最重要的学习问题之一, 以 SVM 为代表的核方法为非线性学习提供了一种非常好的实现路径. 然而, 如何寻找最优的核映射, 以及核方法如何有效应用于大规模问题等依然有很长的路要走. 展望的工作如下.

(i) 核函数的构造及理论研究. 以上核函数对应的是 Hilbert 空间中的核映射, 而现实中不同数据结构 (字符串、图、空间数据等) 下可能需要构造相应的更广泛空间映射意义下的核, 如不定核 [32, 33] 和图核 [34, 35] 等, 而对非 Hilbert 空间中的核映射研究还很不充分. 针对不同数据构造具有良好解释性的、适用性更广泛的、更易于计算的核, 并研究其统计学习意义下的性质和理论等有广阔的前景.

(ii) 大型核矩阵的存储与计算. 由于核方法中大多需要核矩阵, 而对于大规模问题而言, 核矩阵往往很大, 这对存储和计算都带来了困难. 如何求解相应的大型核矩阵是核方法发展的一个瓶颈. 构造适用于大规模问题的核模型与算法 (如随机核模型、可并行计算的核模型 [36, 37] 等) 都值得深入研究.

(iii) 寻找或选择最优核映射. 尽管文献 [31] 为学习最优核矩阵提供了良好的开端, 但实践中, 如何选择或构造出合适的核矩阵基底并求得最优核等问题尚未解决. 目前, 利用深度学习和启发式学习 [38, 39] 等方式的研究为最优核学习提供了一个新的方向.

此外, 核方法在其他机器学习问题中也大有可为.

4 浅层学习、深度学习和广度学习

4.1 神经网络与 SVM 网络

神经网络 [40] 是实现非线性机器学习的有力工具之一. 它的基本思想是针对数据, 构造并学习出一个具体的网络, 从而利用该网络进行学习和预测等. 我们用图 8 来说明经典的神经网络 [41] 结构. 令 x 为输入层节点, y 为输出层节点, 连接输入层与输出层之间的节点为隐层节点. 神经网络对输入层数据节点 x 进行线性映射后, 由激活函数激活得到新的隐层节点, 然后类似地构造一系列隐层节点, 直到最后的输出层节点为 y. 这样便得到了如图 8 所示的含有输入层、L 个隐层和输出层的神经网络模型. 当学习得到连接各层的权重系数和激活函数后, 即可进行预测.

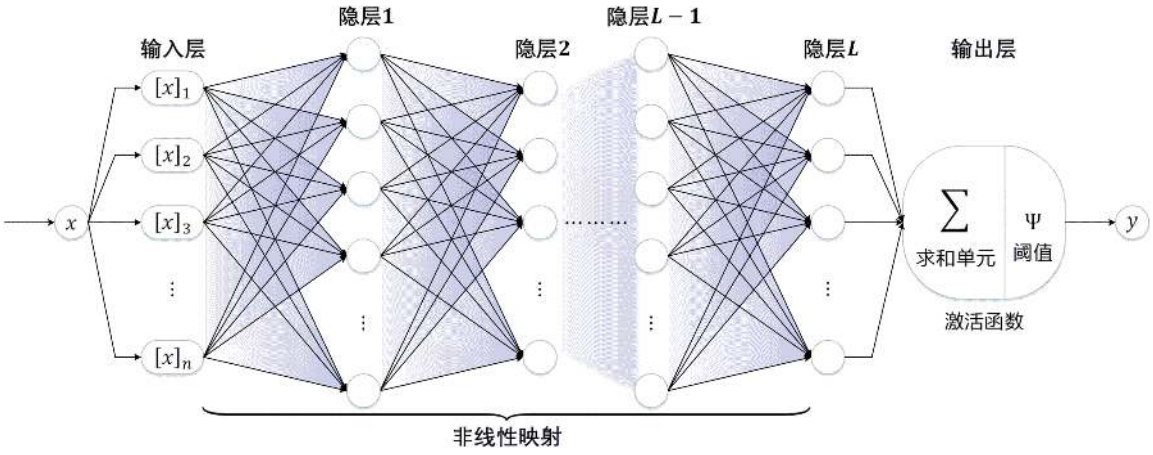


图 8 (网络版彩图) 神经网络多层映射示意图

事实上, SVM 可以看成是一类特殊形式的神经网络模型. 如前所述, SVM 首先对原始数据做映射 $\phi(x)$, 然后在映射后的空间进行线性学习, 得到输出 y. 与图 8 相对应, 在图 9 中, 若令 x 为输入层节点, y 为输出层节点, $\phi(x)$ 为隐层节点, 则可视 SVM 为一个含有输入层、一个隐层和输出层的神经网络模型, 这是对原始空间而言的. 我们知道 SVM 通常不需要给出 ϕ 的具体形式, 而可以用核函数 K 替代, 因此, 隐层节点 $\phi(x)$ 可以用 $K(\cdot,x)$ 替代. 它也是一个只有一个隐层的神经网络模型, 不过是对对偶空间而言的. 特别地, 若 $\phi(x)=K(\cdot,x)=x$, 则对应没有隐层 (单层) 神经网络模型.

传统神经网络的每一层映射都是线性映射, 但允许多个隐层. 它通过构造多层不同的线性映射进行学习, 即利用多个简单映射的组合得到一个相对复杂的非线性学习器. 而 SVM 只有一个隐层映射, 但映射是非线性的, 它对该映射进行线性组合得到学习器. 二者比较而言, 神经网络是多隐层的, 参数很多, 但它是容易处理的线性的; 而 SVM 是单隐层的, 结构简单且解释性好, 但它的映射是难以处理的非线性的. 事实上, 要想取得好的学习效果, 对隐层非线性核映射的选取提出的要求的确很高.

4.2 深度 SVM 与广度 SVM

借助于 SVM 的理论与方法, 构造多隐层网络模型是一个自然的想法. 这方面的研究已经有了一些初步成果 (参见文献 [13,42–44]). 其中一个角度的研究是直接构造多个核映射的隐层, 并实现层与层之间叠加. 图 10(a) 给出了深度非线性映射示例, 这里层与层间是复合映射. 显然随着层数的增加, 非线性映射的复合运算的复杂度和计算量会急剧增加. 对此, 人们尝试利用容易计算的特殊的核映射 [13], 使复合函数尽量简单, 从而实际计算变为可行. 此外, 考虑原始对偶问题的共轭性质, 构造基于原始对偶耦合的途径的深度模型, 也取得了较好的效果 (参见文献 [43]). 然而总的来说, 目前深度核模型的研究只是处于起步阶段. 另一个角度的研究是与 SVM 中稀疏损失函数相对应, 具有稀疏性质的激活函数在深度学习研究中已被广泛应用, 相应具有稀疏性质的深度网络结构模型与算法 [42] 也已受到关注. 图 10(b) 给出了几个深度学习中常用的稀疏激活函数. 对比图 4(a) 的损失函数和图 10(b) 的激活函数可以看出, 它们很相似.

一个不争的事实是, 多层非线性映射依然是一个非线性映射, 如图 10(a) 所示. 因此, SVM 原则上应该可以只找一个非线性映射就能达到目的. 是否有必要增加网络的深度来构造多层非线性模型, 是

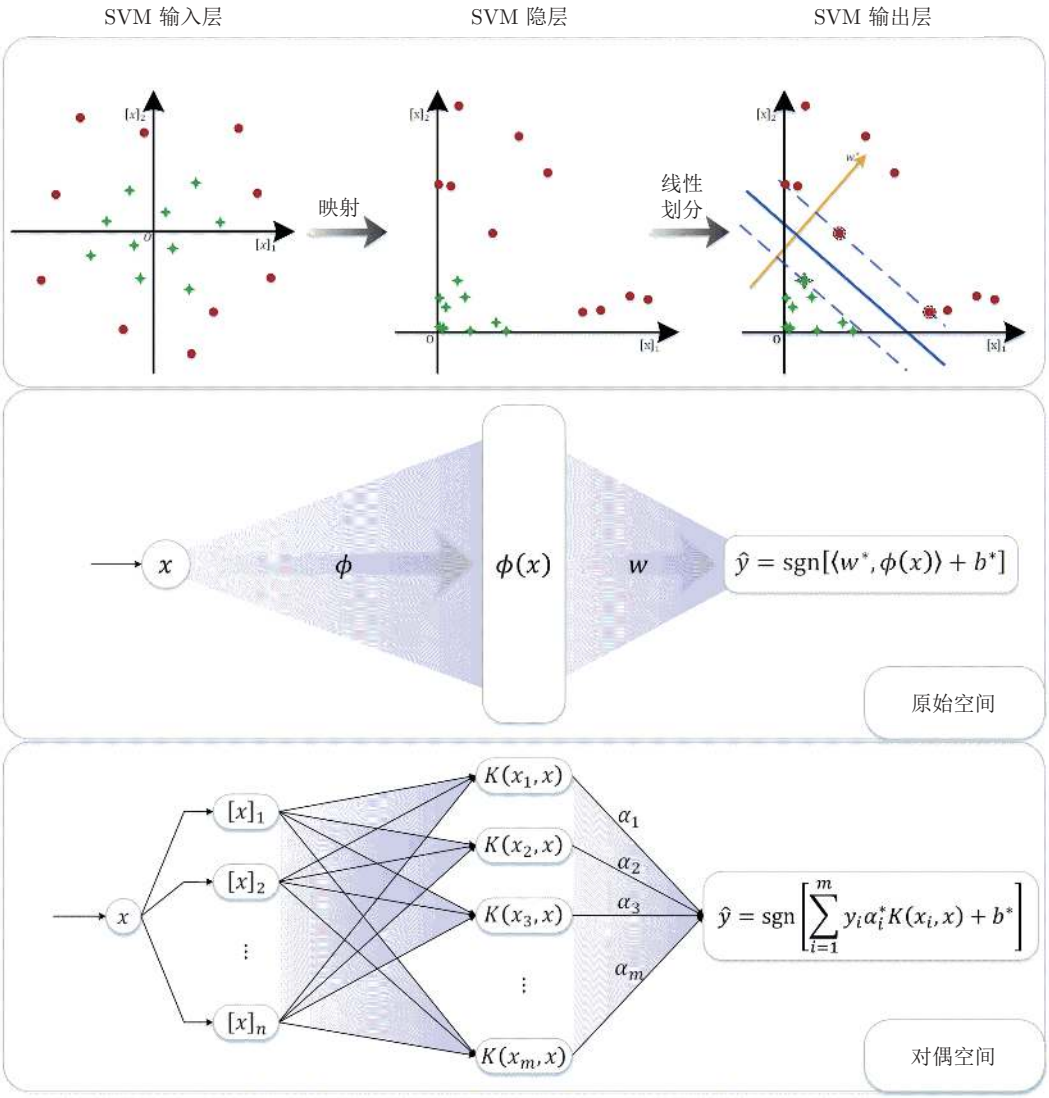


图 9 (网络版彩图) SVM 网络示意图

SVM 研究中的一个争论焦点. 对 SVM 来说, 的确找到一个恰当的非线性映射即可解决问题, 然而实际中要找到恰当的非线性映射往往是非常困难的. 利用深度网络结构进行多层简单非线性映射的组合, 无疑是一种有益尝试.

除了将 SVM 的网络结构向深度学习拓展, 人们还在研究向广度学习拓展, 即将数据映射到不同的核空间再进行组合. 广度 SVM 研究的对象往往是数据来源或学习任务是多样化的情形, 如多视角学习 [45]、多任务学习 [46] 和迁移学习 [47] 等. 此时, 用同一映射往往不能充分学习各种数据或任务的特异性. 对应的广度 SVM [14, 48–50] 进行多对一的核映射或多对多的核映射学习, 每个核映射除了要充分利用自己对应的数据, 还希望用到尽可能多的有关的其他数据. 这里要注意的是各数据源或各任务的一致性和特异性.

多核学习是广度 SVM 的典型代表, 它的基本思想是将数据映射到多个不同核空间并组合学习得

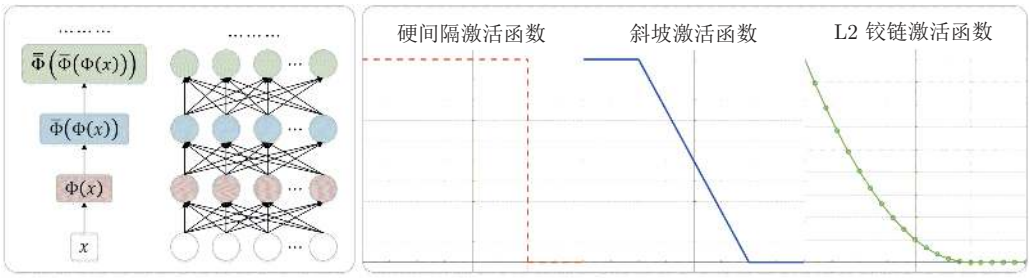


图 10 (网络版彩图) (a) 多层非线性映射示例; (b) 稀疏激活函数示例

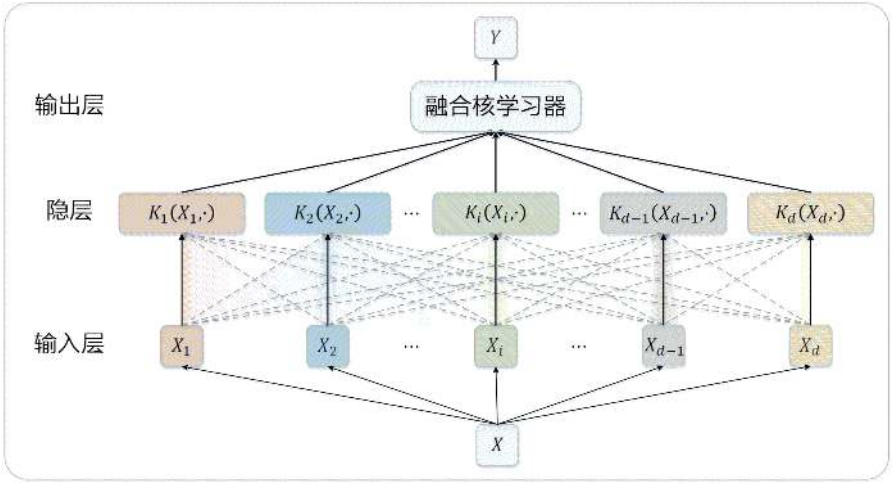


图 11 (网络版彩图) 多核学习示意图

到最终的学习器. 由若干核矩阵为基底线性组合而得到如 (3.8) 中最优核就是多核学习的思想, 但实现多个不同的核映射以及它们的组合方式可根据多视角或多任务学习等任务特点不同而不同. 图 11 给出了多核学习 SVM 的基本网络结构. 假设输入数据 X 可分为 d 组数据, 各组数据之间可以独立, 也可以交叉. 对这 d 个数据源生成 d 个核映射, 最后融合这些核矩阵并得到最终的学习器. 显然, 得到 d 组数据以及组合这些核映射可以有多种方式 (参见文献 [14, 50]).

4.3 展望

深度学习和广度学习是当前大规模复杂数据问题的研究热点. SVM 作为一种特殊结构的神经网络模型, 由于其核映射的非线性和稀疏性等特点, 在深度网络和广度网络的研究中也受到了关注. 但是这方面的研究还很不充分, 许多基础理论问题尚未解决, 有效的模型和算法也不多. 因此, 存在着较大发展潜力. 展望的研究工作如下.

(i) 研究深度学习 SVM 的理论模型. 深度学习 SVM 的理论基础尚缺, 关于是否需要多层 SVM 的争论还在继续. 近年来, 深度学习不同网络结构的研究取得了较大进展, 而不同网络结构的 SVM 模型研究才刚刚起步, 容易计算且可以解决大规模问题的深度 SVM 网络结构缺乏, 这极大限制了 SVM 在深度学习中的发展.

(ii) 研究深度 SVM 和广度 SVM 的算法. 多层非线性映射模型、稀疏损失的深度模型和多核学习模型的复杂度往往都很高, 这是制约深度 SVM 和广度 SVM 研究的现实问题. 如果能在存储和计算方面有所突破, 将为以上模型提供现实可行的路径.

(iii) 研究深度学习 SVM 与广度学习 SVM 的结合. 深度学习 SVM 与广度学习 SVM 模型都构造非线性映射从而构成最终的学习器, 它们都是研究复杂数据问题的有力手段. 如何针对特定的问题, 将二者进行融合等是值得期待的新领域.

(iv) 研究可以实现端到端的深度和广度核模型与算法. 端到端将特征学习和学习器学习融为一体, 是深度学习的主要优点之一. 核映射具有广泛适用性, 构造核映射可以生成不同的特征. 可以期望利用深度学习和广度学习的网络结构, 构造出适当的核和学习器融合, 从而有效地实现端到端的学习过程.

5 结论

本文总结了支持向量机研究的若干关键问题, 包括 3 个方面: 支持向量机的大间隔原则及其带来的稀疏性、核映射的高维划分技巧及其统计学习理论、支持向量机的浅层学习模式及其到深度学习的拓展. 同时, 从这 3 个方面分别提出潜在的科学问题, 展望支持向量机可能的发展方向. 随着人工智能和大数据的快速发展, 深度学习等新方法不断涌现, 我们希望本文不仅为支持向量机本身的研究提供一些新的思路, 更为支持向量机与其他方法交叉融合提供研究线索.

参考文献

1 Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273–297
2 Vapnik V. Statistical Learning Theory. New York: John Wiley & Sons, 1998
3 Schölkopf B, Smola A J, Bach F. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: MIT Press, 2002
4 Deng N Y, Tian Y J, Zhang C H. Support Vector Machines: Theory, Algorithms, and Extensions. Boca Raton: CRC Press, 2012
5 Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw, 2002, 13: 415–425
6 Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems. Vancouver: Curran Associates, 2003, 577–584
7 Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Advances in Neural Information Processing Systems. Vancouver: Curran Associates, 2002, 681–687
8 Farquhar J, Hardoon D, Meng H, et al. Two view learning: SVM-2K, theory and practice. In: Advances in Neural Information Processing Systems. Vancouver: Curran Associates, 2006, 355–362
9 Tao Q, Chu D, Wang J. Recursive support vector machines for dimensionality reduction. IEEE Trans Neural Netw, 2008, 19: 189–193
10 Müller K R, Smola A J, Rätsch G, et al. Predicting time series with support vector machines. In: International Conference on Artificial Neural Networks. Heidelberg: Springer, 1997, 999–1004
11 Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the Sixteenth International Conference on Machine Learning. Montreal-Quebec: Association for Computing Machinery, 1999, 200–209
12 Ben-Hur A, Horn D, Siegelmann H T, et al. Support vector clustering. J Mach Learn Res, 2001, 2: 125–137
13 Cho Y, Saul L K. Kernel methods for deep learning. In: Advances in Neural Information Processing Systems. Vancouver: Curran Associates, 2009, 342–350

14 Duan L, Tsang I W, Xu D. Domain transfer multiple kernel learning. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 465–479
15 Suykens J A K, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett, 1999, 9: 293–300
16 Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. Cambridge: MIT Press, 1999, 61–74
17 Maji S, Berg A C, Malik J. Classification using intersection kernel support vector machines is efficient. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008, 1–8
18 Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol, 2011, 2: 1–27
19 Shalev-Shwartz S, Singer Y, Srebro N, et al. Pegasos: Primal estimated sub-gradient solver for SVM. Math Program, 2011, 127: 3–30
20 Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification. J Mach Learn Res, 2008, 9: 1871–1874
21 Bartlett P L, Jordan M I, McAuliffe J D. Convexity, classification, and risk bounds. J Amer Statist Assoc, 2006, 101: 138–156
22 Feng Y, Yang Y, Huang X, et al. Robust support vector machines for classification with nonconvex and smooth losses. Neural Comput, 2016, 28: 1217–1247
23 Huang X, Shi L, Suykens J A K. Support vector machine classifier with pinball loss. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 984–997
24 MacKay D J C. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press, 2003
25 Zhang T, Zhou Z H. Optimal margin distribution machine. IEEE Trans Knowl Data Eng, 2020, 32: 1143–1156
26 Elsayed G, Krishnan D, Mobahi H, et al. Large margin deep networks for classification. In: Advances in Neural Information Processing Systems. Montréal: Curran Associates, 2018, 842–852
27 Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press, 2004
28 Shivaswamy P K, Jebara T. Maximum relative margin and data-dependent regularization. J Mach Learn Res, 2010, 11: 747–788
29 Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Comput, 2000, 12: 2013–2036
30 Wu X, Zuo W, Lin L, et al. F-SVM: Combination of feature transformation and SVM learning via convex relaxation. IEEE Trans Neural Netw Learn Syst, 2018, 29: 5185–5199
31 Lanckriet G R G, Cristianini N, Bartlett P, et al. Learning the kernel matrix with semidefinite programming. J Mach Learn Res, 2004, 5: 27–72
32 Ong C S, Mary X, Canu S, et al. Learning with non-positive kernels. In: Proceedings of the Twenty-First International Conference on Machine Learning. Banff: Association for Computing Machinery, 2004, 81:1–81:8
33 Loosli G, Canu S, Ong C S. Learning SVM in Kreĭn spaces. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1204–1216
34 Vishwanathan S V N, Schraudolph N N, Kondor R, et al. Graph kernels. J Mach Learn Res, 2010, 11: 1201–1242
35 Shervashidze N, Schweitzer P, Leeuwen E J, et al. Weisfeiler-Lehman graph kernels. J Mach Learn Res, 2011, 12: 2539–2561
36 Hsieh C J, Si S, Dhillon I. A divide-and-conquer solver for kernel support vector machines. In: International Conference on Machine Learning. Beijing: Microtome Publishing, 2014, 566–574
37 Si S, Hsieh C J, Dhillon I S. Memory efficient kernel approximation. J Mach Learn Res, 2017, 18: 682–713
38 Li C L, Chang W C, Mroueh Y, et al. Implicit kernel learning. In: The 22nd International Conference on Artificial Intelligence and Statistics. Naha: Microtome Publishing, 2019, 2007–2016
39 Song H, Thiagarajan J J, Sattigeri P, et al. Optimizing kernel machines using deep learning. IEEE Trans Neural Netw Learn Syst, 2018, 29: 5528–5540
40 Haykin S S. Neural Networks and Learning Machines. New York: Prentice Hall, 2009
41 McClelland J L, Rumelhart D E. PDP Research Group. Parallel Distributed Processing. Cambridge: MIT Press, 1987
42 Tang Y. Deep learning using linear support vector machines. arXiv:1306.0239, 2013
43 Suykens J A K. Deep restricted kernel machines using conjugate feature duality. Neural Comput, 2017, 29: 2123–2163
44 Bohn B, Rieger C, Griebel M. A representer theorem for deep kernel learning. J Mach Learn Res, 2019, 20: 1–32
45 Xu C, Tao D, Xu C. A survey on multi-view learning. arXiv:1304.5634, 2013
46 Evgeniou T, Pontil M. Regularized multi-task learning. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2004,

109–117
47 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
48 Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the Twenty-First International Conference on Machine Learning. Banff: Association for Computing Machinery, 2004, 6:1–6:8
49 Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res, 2011, 12: 2211–2268
50 Wang W, Wang H, Zhang C, et al. Cross-domain metric and multiple kernel learning based on information theory. Neural Comput, 2018, 30: 820–855

Key issues of support vector machines and future prospects

Yuanhai Shao, Liming Liu, Lingwei Huang & Naiyang Deng

Abstract As one of the main methods of machine learning, the support vector machine (SVM) not only has a solid theoretical foundation of statistical learning, but also shows excellent generalization performance in many fields, so it has received extensive attention. However, in recent years, compared with the vigorous development of deep learning, the research on SVM has fallen into a trough. This paper starts from the essence of SVM, discusses the intersection and fusion of the research methods of machine learning methods, such as deep learning and the SVM, and then puts forward some new ideas. Specifically, it includes three aspects: the principle of large margin with the low density property of SVM, the high-dimensional division technique of kernel mapping and its statistical learning theory, the shallow learning of SVM and its extension to deep learning and broad learning. At the same time, the excellent properties that can be further explored from these three aspects, and the theories and methods that may be induced in the future are expected.

Keywords **support vector machine, statistical learning, kernel learning, machine learning, optimization, deep learning**

MSC(2010) **90C99, 62-07**

doi: 10.1360/SSM-2020-0015