第 36 卷 第 8 期　　自 动 化 学 报　　Vol. 36, No. 8
2010 年 8 月　　ACTA AUTOMATICA SINICA　　August, 2010

多核学习方法

汪洪桥[1,2]　孙富春[2]　蔡艳宁[1]　陈 宁[2]　丁林阁[2]

摘 要　多核学习方法是当前核机器学习领域的一个新的热点. 核方法是解决非线性模式分析问题的一种有效方法, 但在一些复杂情形下, 由单个核函数构成的核机器并不能满足诸如数据异构或不规则、样本规模巨大、样本不平坦分布等实际的应用需求, 因此将多个核函数进行组合, 以获得更好的结果是一种必然选择. 本文根据多核的构成, 从合成核、多尺度核、无限核三个角度, 系统综述了多核方法的构造理论, 分析了多核学习典型方法的特点及不足, 总结了各自的应用领域, 并凝炼了其进一步的研究方向.

关键词　核方法, 多核学习, 合成核, 多尺度核, 支持向量机, 模式识别, 回归

DOI　10.3724/SP.J.1004.2010.01037

On Multiple Kernel Learning Methods

WANG Hong-Qiao[1,2]　SUN Fu-Chun[2]　CAI Yan-Ning[1]　CHEN Ning[2]　DING Lin-Ge[2]

Abstract　Multiple kernel learning is a new research focus in the current kernel machine learning field. The kernel method is an effective approach for non-linear pattern analysis problems. But in some complicated cases, researchers find that the kernel machines with a single kernel function can not meet some practical requirements such as heterogeneous information or unnormalised data, large scale problems, non-flat distribution of samples, etc. Therefore, it is an inevitable choice to consider the combination of kernel functions for better results. According to the composition of multiple kernels, the construction theories of multiple kernel methods are systematically reviewed, the learning methods of multiple kernel with the corresponding characteristics and disadvantages are also analyzed, and the respective applications are summarized from three aspects, which are the composite kernels, the multi-scale kernels, and the infinite kernels. In addition, the paper generalizes the conclusions and some new directions for future work.

Key words　Kernel method, multiple kernel learning, composite kernel, multi-scale kernel, support vector machine (SVM), pattern recognition, regression

人们对核方法[1−3] 的关注, 得益于支持向量机 (Support vector machine, SVM)[4−5] 理论的发展和应用, 核函数的采用使得线性的 SVM 很容易推广到非线性的 SVM. 其核心在于利用相对简单得多的核函数运算, 既避免了特征空间中复杂的内积计算, 又避免了特征空间 (学习机器) 本身的设计[6−7]. 实际上, 有关核函数的研究很早就开始了. 早在 1909 年, Mercer[8] 就研究了正负类型的函数与积分等式理论的联系. 1950 年前后, Aronszajn[9] 发展了再生核希尔伯特空间的理论. 1964 年, Aizerman 等[10] 在研究模式识别的势函数方法时, 利用 Mercer 理论, 把核函数解释为一个特征空间的内积, 并引入到机器学习中. 但是, 当时核方法的潜能并没有被完全挖掘. 直到 1992 年, Boser 等[11] 提出 SVM 方法. SVM 的成功促进了核方法的迅速普及和发展, 逐渐渗透到了机器学习的诸多领域, 如回归估计[12]、模式分类[13]、概率密度估计[14]、子空间分析等[15]. 典型的如 Schölkopf 等[15] 提出了核主成分分析 (Kernel principal component analysis, KPCA), Mika 等[16] 实现了核 Fisher 判别 (Kernel Fisher discriminant, KFD), Baudat 等[17] 提出的核判别分析 (Kernel discriminant analysis, KDA), Lai 等[18] 提出了核规范相关分析 (Kernel canonical correlation analysis, KCCA), Bach 等[19] 提出了核独立分量分析 (Kernel independent component analysis, KICA) 等. 此后, 核方法又得到了大量的改进和推广, 在多个领域得到了广泛应用.

尽管上述的核方法在众多的应用领域有效并且实用, 但这些方法都是基于单个特征空间的单核方法. 由于不同的核函数具有的特性并不相同, 从而使得在不同的应用场合, 核函数的性能表现差别很大, 且核函数的构造或选择至今没有完善的理论依

收稿日期 2009-07-13　录用日期 2010-04-01
Manuscript received July 13, 2009; accepted April 1, 2010
国家重点基础研究专项基金 (G2007cb311003), 国家自然科学基金 (60625304, 60621062) 资助
Supported by National Key Project for Basic Research of China (G2007cb311003) and National Natural Science Foundation of China (60625304, 60621062)
1. 第二炮兵工程学院指挥自动化系 西安 710025　2. 清华大学计算机科学与技术系 智能技术与系统国家重点实验室 清华信息科学与技术国家实验室 北京 100084
1. Department of Command Automation, the Second Artillery Engineering Institute, Xi′an 710025　2. Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084

据. 此外, 当样本特征含有异构信息 (Heterogeneous information)[20−26], 样本规模很大[27−30], 多维数据的不规则 (Unnormalised data)[31−32] 或数据在高维特征空间分布的不平坦 (Non-flat)[33−34], 采用单个简单核进行映射的方式对所有样本进行处理并不合理. 针对这些问题, 近年来, 出现了大量关于核组合 (Kernel combination) 方法的研究, 即多核学习方法[23, 31, 35−40].

多核模型是一类灵活性更强的基于核的学习模型, 近来的理论和应用已经证明利用多核代替单核能增强决策函数的可解释性 (Interpretability), 并能获得比单核模型或单核机器组合模型更优的性能[41−42]. 构造多核模型, 最简单也最常用的一种方法就是考虑多个基本核函数的凸组合, 其形如: $K = \sum_{j=1}^{M}\beta_j K_j$, $\beta_j \geq 0$, $\sum_{j=1}^{M}\beta_j = 1$, 这里 K_j 是基本核函数, M 是基本核的总个数, β_j 是权系数. 因此, 在多核框架下, 样本在特征空间中的表示问题转化成为基本核与权系数的选择问题. 在这个由多个特征空间构建的组合空间中, 由于组合利用了各基本核的特征映射能力, 很好地解决了核函数的选择以及与核目标度量 (Kernel target alignment, KTA)[43−44] 相关的变量与模型的选择难题[31, 45−46]. 同时, 通过将异构数据的不同特征分量分别输入对应的核函数进行映射, 使数据在新的特征空间中得到更好的表达, 能显著提高分类正确率或预测精度. 但是, 这里最重要的问题就是如何得到这个组合的特征空间, 也就是如何学习得到权系数. 针对这一问题, 近来出现了多种有效的多核学习理论及方法. 如早期的基于 Boosting[21, 47] 的多核组合模型学习方法, 基于半定规划 (Semidefinite programming, SDP)[41] 的多核学习方法, 基于二次约束型二次规划 (Quadratically constrained quadratic program, QCQP)[36] 的学习方法, 基于半无限线性规划 (Semi-infinite linear program, SILP)[24, 37] 的学习方法, 基于超核 (Hyperkernels)[31] 的学习方法, 以及近来出现的简单多核学习 (Simple MKL)[27, 29] 方法和基于分组 Lasso 思想的多核学习方法. 在权系数与核函数的组合方面, 研究人员也对多核方法进行了一些改进, 如非平稳的多核学习方法[23], 局部多核学习方法[40], 非稀疏多核学习方法[30] 等. 此外, 基于一类具有多尺度表示特性的核函数, 多核学习方法向多尺度核方法方向又出现了众多的扩展[32−34, 48−52]. 前述的这些多核学习方法都是在有限个基本核函数组合假设下进行的, 容易看到, 有限核的组合受限于选择的有限性, 为了将其扩展到大量核组合的情形, 近来又出现了基于无限核的学习方法[39, 53−54].

从最初在生物信息学领域[55−56] 被提出, 多核学习通过与 SVM 方法相结合, 在众多领域都得到了研究人员的关注, 如模式分类[45, 57−61]、多类目标检测与识别[25, 62−65]、模式回归[22, 33−34, 46]、特征提取[66] 等, 给核机器学习提供了更丰富的设计思路和更广泛的应用前景.

本文系统综述了多核方法的构造和学习, 在分析各种典型方法的特点、总结该领域的研究现状和应用的基础上, 凝炼了其进一步的发展方向. 本文后续内容如下: 首先, 简单介绍核方法原理和基本核函数的一些简单扩展; 然后, 从合成核方法、多尺度核方法、无限核方法三个方面综述了各自的构造理论、学习方法及其优缺点, 以及相应的应用领域; 最后, 总结了多核学习方法目前存在的关键问题和难点, 指出了该领域进一步的研究方向.

1 核方法原理与容许核的构造

1.1 核方法原理

给定一个监督的机器学习问题 $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$, $\cdots$, $(\boldsymbol{x}_l, y_l) \in X \times Y$, 其中输入空间 $X \subseteq \mathbf{R}^n$, 输出空间 $Y \subseteq \mathbf{R}$ (回归问题) 或 $Y = \{-1, +1\}$ (两类分类问题). 可以通过一个非线性映射

$$\begin{aligned} \Phi:\ \ & X \to F \\ & \boldsymbol{x} \mapsto \Phi(\boldsymbol{x}) \end{aligned} \tag{1}$$

把输入数据映射到一个新的特征空间 $F = \{\Phi(\boldsymbol{x}) | \boldsymbol{x} \in X\}$, 其中 $F \subseteq \mathbf{R}^n$. 然后, 利用新的数据表示方法考虑原来的学习问题

$$(\Phi(\boldsymbol{x}_1), y_1), (\Phi(\boldsymbol{x}_2), y_2), \cdots, (\Phi(\boldsymbol{x}_l), y_l) \in F \times Y \tag{2}$$

核函数把非线性映射和特征空间中两个向量的内积这两步结合起来, 使得非线性映射隐式地进行, 而且线性超平面可以由所有训练样本与一个测试样本在特征空间中的内积项的线性组合表示, 从而避免了维数灾难.

Mercer 条件[67]**.** 设 X 是 $\mathbf{R}^n$ 的一个紧子集, $k: X \times X \to \mathbf{R}$ 是一个连续的对称函数, 如果它在希尔伯特空间上的积分算子满足积分正定条件:

$$\forall f \in L_2(X),\ \int_{X \times X} k(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{z} \tag{3}$$

那么一定存在一个特征空间 F 和一个映射 $\Phi: X \to F$, 使得

$$k(\boldsymbol{x}, \boldsymbol{z}) = \Phi(\boldsymbol{x}) \times \Phi(\boldsymbol{z}) \tag{4}$$

针对不同的应用, 可以设计不同的核函数. 常用的核函数主要有线性核、多项式核、径向基核、Sigmoid 核、Fourier 核等.

1.2 容许核的构造

利用核函数可以大大简化计算, 但如何针对具体的问题设计出最适当的核函数却是一个难点. 实际上, 经常采用的方法是直接定义核函数, 从而隐含地定义了特征空间. Mercer 条件是检验核函数是否定义了一个特征空间的充分条件, 我们称满足 Mercer 条件的核函数为容许核[68].

容许核满足一些闭包性质或条件[69], 这使得我们可以从一些简单的核函数设计出复杂的核函数.

性质 1. 容许核的正系数线性组合是容许核.

性质 2. 容许核的乘积是容许核.

性质 3. 函数乘积的积分是容许核.

设 $s(\boldsymbol{x},\boldsymbol{x}')$ 是一个定义在 $X \times X$ 上的函数, 使得 $k(\boldsymbol{x},\boldsymbol{z}) = \int_X s(\boldsymbol{x},\boldsymbol{x}')s(\boldsymbol{z},\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'$ 存在, 则 $k(\boldsymbol{x},\boldsymbol{z})$ 是一个容许核.

性质 4. 平移不变核是容许核的充要条件.

一个平移不变核 $k(\boldsymbol{x},\boldsymbol{z}) = k(\boldsymbol{x}-\boldsymbol{z})$ 是容许核, 当且仅当其傅里叶变换 $F(\boldsymbol{\omega}) = (2\pi)^{-\frac{n}{2}} \int_X k(\boldsymbol{x}) \times \mathrm{e}^{-i(\boldsymbol{\omega}\cdot\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$ 是非负的.

性质 5. 内积型核是容许核的必要条件.

若一个内积型核 $k(\boldsymbol{x},\boldsymbol{z}) = k(\boldsymbol{x}\cdot\boldsymbol{z})$ 是容许核, 则它必满足 $\forall\xi \geq 0$, $k(\xi) \geq 0$, $\partial_\xi k(\xi) \geq 0$ 且 $\partial_\xi k(\xi) + \xi\partial_\xi^2 k(\xi) \geq 0$.

性质 6. 内积型核是容许核的充要条件.

一个内积型核 $k(\boldsymbol{x},\boldsymbol{z}) = k(\boldsymbol{x}\cdot\boldsymbol{z})$ 是容许核, 当且仅当其幂级数展开式 $k(\boldsymbol{t}) = \sum_{n=0}^{\infty} a_n \boldsymbol{t}^n$ 中所有系数 $a_n \geq 0$. 对于有限维的空间, 条件可以稍微减弱一些.

当前已经存在较多的满足 Mercer 条件的核函数, 常见核函数通常可分为两类: 局部核和全局核[6], 而局部核选择不同的核参数, 又可分为大尺度核与小尺度核. 在一些复杂情形下, 同时考虑核机器分类、回归性能和泛化能力, 将不同核组合使用, 将是更合理的选择.

2 基本多核学习: 合成核方法

将不同特性的核函数进行组合, 获得多类核函数的优点, 可以得到更优的映射性能. 并且, 典型的学习问题经常涉及到多种或者异构的数据, 多核方法可以提供更佳的灵活性. 此外, 它可以作为一种巧妙的方法来解释学习结果, 使得应用问题可以得到更深入的理解. 这就是多核学习的一类基本方法, 即合成核方法.

2.1 合成核的构造

1) 多核线性组合合成方法

多核学习最早从生物信息学领域得到应用和认同. 如 Pavlidis 等[20] 在 2001 年就研究了基于异构数据的基因功能分类问题, 其中就讨论了前期、中期和后期三种集成方式. 早期集成是指数据的集成, 后期集成是指分类器结果的集成, 而中期集成就是核矩阵的组合, 它通过对多个基本核矩阵进行合成得到, 基于这种多核矩阵直接求和方式, 可以实现异构数据源的融合, 用来训练分类器.

此后, 在蛋白质功能预测[56] 与定位[70], 蛋白质分子间的交互预测[22], 蛋白质折叠识别和远端同源性检测[25] 等方向, 由于涉及到多特征空间或有效属性的集合 (Groups of attributes available) 问题, 来自异构源的数据具有不同特性, 如全局特性、局部特性等, 这就需要核矩阵在集成时可以评估这些潜在的异构目标描述子各自的贡献. 因此出现了一些加权的多核合成方法. 这类多核方法都是通过多个核函数的线性组合得到的, 图 1 所示的就是其构成的示意图.

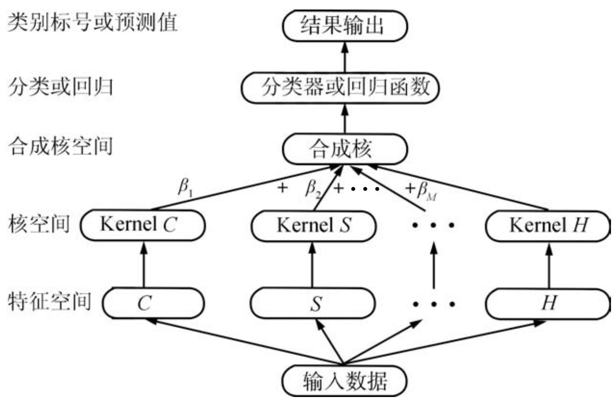

图 1 多核函数线性组合合成示意图

Fig. 1 Sketch map of composition using multiple kernel linear combination

下面采用公式的形式对上述线性组合合成核进行描述. 假定 $k(\boldsymbol{x},\boldsymbol{z})$ 是已知核函数, $\hat{k}(\boldsymbol{x},\boldsymbol{z})$ 是它的归一化形式, 例如核函数 $k(\boldsymbol{x},\boldsymbol{z})$ 可以采用如下方法进行归一化: $\sqrt{k(\boldsymbol{x},\boldsymbol{x})k(\boldsymbol{z},\boldsymbol{z})}$. 采用以上引入的符号, 可以定义以下几种合成核:

a) 直接求和核 (Direct summation kernel)

$$k(\boldsymbol{x},\boldsymbol{z}) = \sum_{j=1}^{M} \hat{k}_j(\boldsymbol{x},\boldsymbol{z}) \tag{5}$$

b) 加权求和核 (Weighted summation kernel)

$$k(\boldsymbol{x},\boldsymbol{z}) = \sum_{j=1}^{M} \beta_j \hat{k}_j(\boldsymbol{x},\boldsymbol{z}),\ \ \beta_j \geq 0,\ \ \sum_{j=1}^{M} \beta_j = 1 \tag{6}$$

c) 加权多项式扩展核 (Weighted polynomial extended kernel)

$$k(\boldsymbol{x},\boldsymbol{z}) = \alpha\hat{k}_1^p(\boldsymbol{x},\boldsymbol{z}) + (1-\alpha)\hat{k}_2^q(\boldsymbol{x},\boldsymbol{z}) \tag{7}$$

其中, $k^p(\boldsymbol{x},\boldsymbol{z})$ 是 $k(\boldsymbol{x},\boldsymbol{z})$ 的多项式扩展.

近来, 这类合成核方法又得到一些改进, 在图像目标的识别领域得到广泛应用. 如在金字塔框架下, 对目标形状进行多核表示[57], 或采用多核方法, 自动获得基于决策的一种相应目标类别的稀疏依赖图, 实现了多类目标联合检测[62], 提高了目标的识别率. 通过同时考虑多核线性组合的稀疏性和分类器的强判别力, 将多核学习问题转化为不同的优化问题[58, 63], 或通过多对象描述子、多特征空间的整合, 并进行快速求解[64]. 此外, 合成核方法在特征提取[66]、处理及应用[71]、分类[59, 72−74]、图像分割[60]、系统辨识[75] 等方面又得到了一些成功应用.

2) 多核扩展合成方法

上述合成核方法都是试图通过一种求和 "平均化" 的思想[42] 来实现不同核矩阵融合. 然而, 这里存在一个丢失原始核矩阵信息的风险. 比如, 如果数据集的局部分布是多变的, 不同的核处理不同的区域会得到更好的结果, 对不同核函数采用平均的方法将丢失刻画这些局部分布的性能. 为了实现核矩阵的组合而不丢失任何原始信息, 可以考虑将多核矩阵进行扩展合成[42], 新的核矩阵由原核矩阵和其他不同的核矩阵共同构成, 在这个更大的核矩阵中原核矩阵仍然存在. 因此, 原始核函数的性质得以保留. 该合成核矩阵的形式为

$$\begin{bmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,s} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ K_{s,1} & K_{s,2} & \cdots & K_{s,s} \end{bmatrix}_{s\times n, s\times n} \tag{8}$$

可以看出, 原始核矩阵位于新矩阵的对角线上. 其他所有元素是定义为 $(K_{p,p'})_{i,j} = K_{p,p'}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 的两个不同核矩阵的混合, 可由如下公式求得 (以两个高斯核为例):

$$K_{p,p'}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \left(\frac{2\sigma_p\sigma_{p'}}{\sigma_p^2+\sigma_{p'}^2}\right)^{\frac{d}{2}} \mathrm{e}^{-2\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma_p^2+\sigma_{p'}^2}} \tag{9}$$

很明显, 当 $p = p'$ 时, $K_{p,p} \equiv K_p$.

实验结果显示, 当数据集具有变化的局部数据分布时, 这种合成核方法将是更好的选择. 此外, 通常核组合方法在很大程度上依靠训练数据, 并且必须通过学习获取一些权系数, 以标识每个核的重要性. 而在扩展合成核方法中, 这些核函数的重要性可以直接从支持向量机的训练过程中导出. 由此, 分别对应不同核的权系数可以通过一个单独的分类器优化过程整体得到. 并且该优化过程不会像其他加权核方法那样, 由于在优化权系数和训练分类器过程中两次使用训练数据而产生训练数据的过拟合. 但是, 该合成核矩阵的大小为 $(s\times n)\times(s\times n)$, 而原始核矩阵的大小都是 $n\times n$, 由于合成核矩阵是原始核矩阵规模的 s 倍, 因此样本特征必须被复制, 使运算量成倍增加.

3) 其他改进合成核方法

近年来, 针对多核学习中核函数的选取以及权值系数的改进, 又出现了一些新的多核合成方法, 如:

a) 非平稳多核学习

前述的多核线性组合方法都是对核函数的平稳组合, 即对所有输入样本, 不同的核对应的权值是不变的, 无形中对样本进行了一种平均处理. Lewis 等[23] 提出了一种多核的非平稳组合方法, 对每个输入样本配以不同的权值系数. 如常规 SVM 判别函数为

$$f(\boldsymbol{x}) = \sum_t^S y_t\alpha_t k(\boldsymbol{x}_t,\boldsymbol{x}) + b \tag{10}$$

引入不同的加权系数, 典型的合成核 SVM 的判别函数可以改写为

$$f(\boldsymbol{x}) = \sum_t^S y_t\alpha_t \sum_j^M \beta_j k_j(\boldsymbol{x}_t,\boldsymbol{x}) + b \tag{11}$$

而对于非平稳的合成核 SVM, 其判别函数改进为

$$f(\boldsymbol{x}) = \sum_t^S y_t\alpha_t \sum_j^M \beta_j(\boldsymbol{x}) k_j(\boldsymbol{x}_t,\boldsymbol{x}) + b \tag{12}$$

在最大熵判别 (Maximum entropy discrimination, MED) 框架下, 通过使用一种大间隔隐变量生成模型, 使得隐参数估计问题可以通过变化边界和一个内点优化过程来表示, 并且相应的参数估计可以通过快速的序列最小优化 (Sequential minimal optimization, SMO) 算法实现. 通过多种数据集的实验验证, 非平稳的多核学习方法具有更好的通用性.

b) 局部多核学习

此后, 仍旧是针对多核学习在整个输入空间中对某个核都是分配相同权值的问题, Gönen 等[40, 76] 利用一种选通模型 (Gating model) 局部地选择合适核函数, 提出了一种局部多核学习算法. 在 SVM 框架下, 其判别函数形如:

$$f(\boldsymbol{x}) = \sum_t^S y_t\alpha_t \sum_j^M \eta_j(\boldsymbol{x}) k_j(\boldsymbol{x}_t,\boldsymbol{x})\eta_j(\boldsymbol{x}_t) + b \tag{13}$$

其中, $\eta_j(\boldsymbol{x})$ 是选通函数, 其定义形式为

$$\eta_j(\boldsymbol{x})=\frac{\exp\left(\langle \boldsymbol{v}_m,\boldsymbol{x}\rangle+\boldsymbol{v}_{m0}\right)}{\sum\limits_{j}^{M}\exp\left(\langle \boldsymbol{v}_j,\boldsymbol{x}\rangle+\boldsymbol{v}_{j0}\right)} \tag{14}$$

这里的 $\boldsymbol{v}_m$ 和 $\boldsymbol{v}_{m0}$ 是选通模型参数, 可以在多核学习过程中通过梯度下降法获得. 将局部选通模型和基于核的分类器相结合, 优化问题可以用一种联合的方式加以解决. 局部多核学习方法获得了与多核学习近似的精度, 但只需要存储更少的支持向量. 基于此, Christoudias 等[77] 又提出了一种基于 Bayesian 的局部权值求取方法, 以使学习过程能适应大规模的数据集.

c) 非稀疏多核学习

大部分合成核方法都有式 (6) 的形式, 即对多核系数的约束是一种 ℓ_1 范数的形式, 以提高核组合的稀疏性. 稀疏性的提高在一些情况下可以减少冗余, 提高运算效率. 但当某个问题多个特征编码间具有正交性, 稀疏性可能导致有用信息的丢失和泛化性能变弱. Kloft 等[30] 通过对系数引入一种 ℓ_2 范数约束, 即 $\|\beta\|_2=1$, 提出了非稀疏的多核学习方法. 虽然在此约束下, 多核组合形式是非凸的, 但通过使用二范数 $\|\beta\|_2=1$ 边界上的值, 可以得到一个紧致的凸近似, 这就保证了核矩阵的严格正定性. 通过在大规模数据集下与 ℓ_1 范数和常用多核学习 (Multiple kernel learning, MKL) 方法进行对比实验, 仿真实验结果显示 ℓ_2-MKL 在抗噪声和特征集冗余方面具有较强的鲁棒性. 此后, Kloft 等[78] 又将 ℓ_2 范数约束推广到任意 ℓ_p 范数, $p>1$, 进一步增强了核机器的通用性和鲁棒性.

2.2 合成核机器的学习方法

为了求取合成核的参数, 通常是将合成核与支持向量机方法相结合, 然后将目标函数转化成不同的优化问题, 如不同的正则化形式[79] 或对训练样本的一些约束, 通过不同的优化方法进行求解. 基于此, 出现了多种合成核机器的学习方法.

1) Boosting 方法

早期受集成思想和 Boosting 方法的启发, Bennett[21] 提出了一种多自适应回归核 (Multiple additive regression kernels, MARK) 算法. MARK 定义了一种异构核模型, 并考虑一个大规模核矩阵库 (Library), 这个库由不同的核函数和其参数构成. 通过使用一种梯度 Boosting 列生成方法, MARK 构建出异构核矩阵的每一列, 然后将其添加到合成核中. 算法的目标就是在这个核矩阵库的基础上, 找到一种构建推广模型的方法. 这种方法推广性强, 不需要存储大量的数据来应对后续的预测, 提高了预测的效率. 在此基础上, 通过与 SVM 结合, Bi 等[47] 详细阐述了应用于合成核的列生成 Boosting 方法, 并成功推广到分类和回归问题.

2) 二次约束型二次规划

从数学形式上看, 二次约束型二次规划是一类目标函数和约束同为二次函数的优化问题:

$$\begin{aligned}\min\quad & \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}P_0\boldsymbol{x}+q_0^{\mathrm{T}}\boldsymbol{x}\\ \text{s.t.}\quad & \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}P_i\boldsymbol{x}+q_i^{\mathrm{T}}\boldsymbol{x}+r_i\le 0,\ \ i=1,2,\cdots,m\\ & A\boldsymbol{x}=b\end{aligned} \tag{15}$$

这里, $P_0,P_1,\cdots,P_m$ 是 $n\times n$ 矩阵, 优化变量 $\boldsymbol{x}\in\mathbf{R}^n$; 如果 $P_1,\cdots,P_m$ 均为 0 矩阵, 则约束变为线性的, 该问题实际变为一个二次规划问题.

Bach 等[36] 针对多核矩阵和分类器系数锥组合问题的联合优化, 提出了 QCQP 的一种新对偶形式, 把它作为一个二阶锥规划, 可以利用 Moreau-Yosida 正则化来生成 SMO 方法的适用形式. 实验结果显示这种基于 SMO 的算法比常用工具箱中应用的内点法更有效, 广泛应用于支持向量回归问题[46].

3) 半定规划

通过在一个核矩阵中综合考虑训练数据和测试数据, Lanckriet 等[41] 通过半定规划技术实现了核矩阵的学习问题, 也为合成核模型提供了一种功能强大的渐进直推式算法. 该算法被成功应用并推广到蛋白质功能预测[56]. 其考虑的核矩阵具有如下形式:

$$\begin{bmatrix} K_{\mathrm{tr}} & K_{\mathrm{trt}}\\ K_{\mathrm{trt}}^{\mathrm{T}} & K_{\mathrm{t}}\end{bmatrix} \tag{16}$$

其中, $K_{ij}=\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle$, $i,j=1,\cdots,n_{\mathrm{tr}},n_{\mathrm{tr}}+1,\cdots,n_{\mathrm{tr}}+n_{\mathrm{t}}$, 这里 n_{tr} 和 n_{t} 是有标号的训练样本个数和无标号的测试样本个数. 我们的目标是通过优化关于训练数据块 K_{tr} 的损失函数, 学习得到最优的混合数据块矩阵 K_{trt} 和测试数据块矩阵 K_{t}. 即利用有标号的训练样本来预测测试样本的类别, 也就是说, 作者认为在训练的过程中同时考虑训练样本和测试样本, 可以找出最佳的核矩阵. 但这样产生的问题是, 求解核矩阵的搜索空间也相对变大, 为了避免过学习 (Overfitting), Lanckriet 利用限制核矩阵的迹为一常数来控制, 于是有了约束式 $\mathrm{tr}(K)=c$.

半定规划是一种凸优化问题 (Convex optimization problem)[80], 它有一个线性的目标函数 (Affine objectives function)、有限个线性矩阵不等式约束 (Linear matrix inequality constraints) 以及有限个线性矩阵等式约束 (Affine matrix equality constraints), 其标准形式如下:

$$\begin{aligned}\min_{\boldsymbol{u}}\quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{u}\\ \text{s.t.}\quad & F^j(\boldsymbol{u})=F_0^j+u_1F_1^j+\cdots+u_qF_q^j\succeq 0,\\ & j=1,\cdots,l\\ & A\boldsymbol{u}=b\end{aligned}\tag{17}$$

其中, 向量 $\boldsymbol{u}$ 是最优化目标, $F_0^j,\cdots,F_q^j$ 是 $n\times n$ 的对称矩阵. $F^j(\boldsymbol{u})$ 是一个半正定矩阵, 上标 j 表示可能有 1 至 l 个约束式; 满足此约束式的 $\boldsymbol{u}$ 所构成的集合是一个凸集合. A 是一个行数与 $\boldsymbol{u}$ 长度相同, 列数与 b 长度相同的矩阵, 表示有限个等式约束式. 因此, 半定规划是在对称且半正定矩阵的凸子集合 (Convex subset) 上求解凸函数的最优化问题.

针对多核支持向量分类问题, 通过定义一种性能指标 (Performance index) $\omega(K)$, 基于原始–对偶变换以及 Schur complement 引理, 原优化问题最终可以转化为一个标准的半定规划形式:

$$\begin{aligned}\min_{\boldsymbol{\beta},t,\lambda,\boldsymbol{v},\boldsymbol{\delta}}\quad & t\\ \text{s.t.}\quad & \mathrm{tr}\left[\sum_{j=1}^{M}\beta_jK_j\right]=c\\ & \sum_{j=1}^{M}\beta_jK_j\succeq 0\\ & \begin{bmatrix} G\left[\sum_{j=1}^{M}\beta_jK_{j,\mathrm{tr}}\right] & \boldsymbol{e}+\boldsymbol{v}-\boldsymbol{\delta}+\lambda\boldsymbol{y}\\ (\boldsymbol{e}+\boldsymbol{v}-\boldsymbol{\delta}+\lambda\boldsymbol{y})^{\mathrm{T}} & t-2\boldsymbol{\delta}C\boldsymbol{e}\end{bmatrix}\succeq 0\\ & \boldsymbol{v}\geq 0\\ & \boldsymbol{\delta}\geq 0\end{aligned}\tag{18}$$

其中, t 是引入的一个替代变量 (Auxiliary variable), $\boldsymbol{v}$, $\boldsymbol{\delta}$, λ 是引入的 Lagrangian 乘子. 至此, 可以通过标准的半定规划求解方法得到 $\boldsymbol{\beta}$ 及相应的 Lagrangian 乘子. 半定规划具有很高的泛化能力, 线性规划 (Linear programming, LP) 以及 QCQP 问题都可以转换推广成半定规划问题. 然后可以很容易地使用内点法 (Interior-point method) 加以解决[80].

4) 半无限线性规划

Sonnenburg 等[37] 在多核矩阵锥组合的基础上, 提出了一种通用而更有效的多核学习算法. 该方法将 Bach 等的 QCQP 对偶形式改写为一种半无限线性规划 (Semi-infinite linear program, SILP) 形式, 新的规划形式可以在标准的 SVM 应用问题中, 利用成熟的线性规划方法进行求解. 并且, 通过将此形式进行推广, 算法能有效解决更多类型的问题, 如回归问题、一类分类 (奇异检测) 问题等. 实验结果显示该算法可以有效增强模型的自动选择能力, 并能提高学习结果的解释性, 同时能有效应用于数十万个样本和数百个核的大规模组合优化问题. 这种半无限线性规划相比其他方法明显提高了学习速度, 适宜于解决大规模问题. 特别是当 SVMs 与一些已出现的字符串核 (String kernel) 相结合, String kernel 也是一种有效的核方法, 它根据两个字符串的所有公共子串计算它们的相似度, 利用这些核对特征的稀疏映射, 使得我们可以训练一种字符串核 SVM, 并应用于计算生物学中的千万级样本的数据片段[24]. 在此基础上, Zien 等[38] 提出了一种应用于联合特征映射的多核学习方法, 为多类分类问题的多核学习提供了一种更方便和原理化的途径. 通过对一种凸 QCQP 以及两种 SILPs 在数据集上进行比较, 实验结果显示 SILPs 比 QCQP 在速度上更有优势.

5) 超核 (Hyperkernels)

对基于核方法的支持向量机而言, 如何选择一个合适的核函数实现自动的机器学习是一个很大的挑战. Ong 等[31] 通过定义一种核空间上的再生核 Hilbert 空间, 即超再生核 Hilbert 空间, 并引入超核的概念及构造方法, 在更广义的层面上实现了这一目标.

定义 1 (超再生核 Hilbert 空间, Hyper reproducing kernel Hilbert space). 设 X 为非空集合, $\underline{X}:X\times X$ 是复合指标集, $\underline{H}$ 为函数 $f:\underline{X}\to\mathbf{R}$ 的 Hilbert 空间, 该函数可表示为该空间中两个向量的内积, 且其范数 $\|f\|=\sqrt{\langle f,f\rangle}$, 则 $\underline{H}$ 被称为超再生 Hilbert 空间, 如果存在一个超核 $\underline{k}:\underline{X}\times\underline{X}\to\mathbf{R}$ 具有如下性质:

a) 再生性: 对所有 $f\in\underline{H}$, 有 $\langle\underline{k}(\underline{\boldsymbol{x}},\cdot),f\rangle=f(\underline{\boldsymbol{x}})$, 特殊地, $\langle\underline{k}(\cdot,\underline{\boldsymbol{x}}),\underline{k}(\cdot,\underline{\boldsymbol{x}'})\rangle=\underline{k}(\underline{\boldsymbol{x}},\underline{\boldsymbol{x}'})$;

b) $\underline{k}$ 张成整个空间 $\underline{H}$, 即 $\underline{H}=\mathrm{span}\{\underline{k}(\underline{\boldsymbol{x}},\cdot)|\underline{\boldsymbol{x}}\in\underline{X}\}$;

c) 对任一固定的 $\underline{\boldsymbol{x}}\in\underline{X}$, 超核 $\underline{k}$ 是关于其第二个输入的核函数, 即对任一固定的 $\underline{\boldsymbol{x}}\in\underline{X}$, 函数 $k(\boldsymbol{x},\boldsymbol{x}')=\underline{k}(\underline{\boldsymbol{x}},(\boldsymbol{x},\boldsymbol{x}'))$, $\boldsymbol{x},\boldsymbol{x}'\in X$ 是一个核函数.

在超再生核 Hilbert 空间上, 可以用类似于正则化品质函数的方法, 得到一个从训练数据对核进行学习的推理框架. 对超核的学习, 可以通过定义一个被称为品质函数 (Quality functional) 的量 (类似于风险函数) 来实现, 这个量可以衡量核函数 "非良 (Badness)" 的程度.

定义 2 (正则化品质函数, Regularized quality functionality). 设 X,Y 分别是训练测试样本组合和样本标签, 对 X 上的一个半正定核矩阵 K, 其正则化品质函数定义为如下形式:

$$\mathcal{Q}_{\rm reg}(k,X,Y)=\mathcal{Q}_{\rm emp}(k,X,Y)+\frac{\lambda_{\mathcal{Q}}}{2}\|k\|_{\underline{H}}^2 \quad (19)$$

这里, $\lambda_{\mathcal{Q}}\geq 0$ 是一个正则化常数, $\|k\|_{\underline{H}}^2$ 表示空间 $\underline{H}$ 中的范数, $\mathcal{Q}_{\rm emp}(k,X,Y)$ 是一种经验品质函数, 它表示核函数 k 与某一特定数据集 X,Y 的匹配程度, 该函数的值常用来调整 k 以使得 $\mathcal{Q}_{\rm emp}$ 最优 (如: 最优核目标度量).

引理 1 (再生核 Hilbert 空间的表示定理, Representer theorem for hyper-RKHS). 设 X 为非空集合, $\mathcal{Q}_{\rm emp}$ 是任意经验品质函数, X,Y 分别是训练测试样本组合和样本标签, 则每一个最小化正则化品质函数 $\mathcal{Q}_{\rm reg}(k,X,Y)$ 的 $k\in\underline{H}$ 具有以下的一种表示形式:

$$k(\boldsymbol{x},\boldsymbol{x}')=\sum_{i,j}^{M}\beta_{ij}\underline{k}((\boldsymbol{x}_i,\boldsymbol{x}_j),(\boldsymbol{x},\boldsymbol{x}')),\quad \forall\boldsymbol{x},\boldsymbol{x}'\in X \quad (20)$$

对每一个 $1\leq i,j\leq M$, 这里 $\beta_{ij}\in\mathbf{R}$.

根据超再生核 Hilbert 空间的表示理论可知, 由超核构造的决策函数不仅由某一个单核构成, 而且还由多核之间的一个线性组合构成, 因此具有更优的性能. 在分类、回归以及奇异检测等方面的实验证实了该方法的有效性[31, 81], 拓展了多核模型选择与合成的研究途径.

6) 简单 MKL

从 Bach 等的多核学习框架[36] 出发, Sonnenburg 等提出了一种通用而更有效的多核学习算法[37], 该方法通过迭代使用现有的支持向量机代码, 从一个新的角度解决了大规模问题的多核学习. 然而, 这种迭代算法在收敛到一个合理解之前, 需要过多的迭代运算. Rakotomamonjy 等[27] 用一种自适应的 ℓ_2 范数正则化方法来考虑多核学习问题, 每个核矩阵的权系数被包含在标准 SVM 的经验风险最小化问题中, 并采用 ℓ_2 约束以提高解的稀疏性. 然后采用了一种基于分块 ℓ_1 范数正则化的算法来解决这一问题, 为多核问题提供了一个新的视角, 并且证明了该方法与 Bach 等的方法是等效的. 从上述描述可以看出, 除了学习合成核外, 该方法解决的是一个标准的 SVM 优化问题, 这里核的定义形式为多个核的线性组合. Rakotomamonjy[29] 称之为简单多核学习 (Simple MKL). 在加权的 2 范数正则化形式下, 同时对多核权系数进行一个额外的 1 范数约束, 为多核学习提供了一种基于混合范数正则化的新思路. 简单多核学习可以从两类分类问题向其他方向扩展, 如回归、聚类、一类分类 (奇异检测) 以及多类分类问题, 具有很强的通用性, 并且与其他多核学习算法相比, 该算法收敛速度更快且效率更高.

7) 分组 Lasso

Lasso 回归是目前处理多重共线性的主要方法之一, 相对于其他方法, 更容易产生稀疏解, 在参数估计的同时实现变量选择, 因而可以用来解决检验中的多重共线性问题, 以提高检验的效率. Lasso 可以推广为分组 Lasso (Group Lasso), 从而使得模型的解可以保持组稀疏性和层次性. Bach[26] 关注于分块 1 范数正则化的最小二乘回归, 即分组 Lasso 问题, 研究了其渐进模型一致性, 推导出了分组 Lasso 一致性在一些实际假设下的充要条件, 如模型误定. 当线性预测器和欧氏范数 (2 范数) 用函数和再生核 Hilbert 范数代替, 这就是常说的多核学习问题. 通过使用函数分析工具和特定的协方差算子, 将上述一致性结果推广到无限维情形, 同时提出了一种自适应方法来获得一致性模型的估计, 即使在非适应方法必要性条件不满足的情况下也能适用, 为多核学习问题提供了一条新的途径.

2.3 其他合成核参数学习方法

从最简单的多个核直接求和到上述的各种改进合成核构造方法, 多核学习经历了从经验性选择到运用多种优化方法求解的过程. 但针对一些具体问题, 对核参数的选取, 多核权系数的设定, 目前还没有形成一个合理统一的模式, 常用的方法只能是凭借经验、实验对比、大范围的搜索或通过交叉验证等进行寻优. 在这种情况下, 也出现了其他的一些方法, 实现了多核学习问题, 典型的有:

1) 基于智能优化方法的多核学习

这类方法主要通过一些比较成熟的智能优化方法, 建立目标函数, 寻找该函数极值的过程就是合成核参数寻优的过程. 如采用多项式核与径向基核的合成核[82] 作为支持向量机的核函数 $k=\rho k_{\rm poly}+(1-\rho)k_{\rm rbf}$, 将其用 SVM 进行预测过程中的参数向量 (d,σ,γ,ρ) 作为粒子, 其中 d 为多项式核参数, σ 为径向基核尺度参数, γ 为 SVM 调整参数, ρ 为合成核的权重参数, 利用粒子群算法对该合成核的参数进行优化, 最终找到最优的预测结果.

2) 基于核目标度量的多核学习

核度量[43−46] 是两个核函数之间或核函数与目标函数间的一个相似性度量, 在多核矩阵信息融合方面得到了应用, 其概念最早由 Cristianini 等提出. 考虑一个两类分类数据集 $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^l$, 其中 $y_i\in\{+1,-1\}$, 则在数据集 S 下, 两个核矩阵之间的核度量定义为

$$\hat{A}(S,K_1,K_2)=\frac{\langle K_1,K_2\rangle_F}{\sqrt{\langle K_1,K_1\rangle_F\langle K_2,K_2\rangle_F}} \quad (21)$$

这里, $\langle K_p,K_q\rangle_F=\sum_{i,j=1}^l K_p(\boldsymbol{x}_i,\boldsymbol{x}_j)K_q(\boldsymbol{x}_i,\boldsymbol{x}_j)$. 通

过上式, 对应于 S 的核矩阵 K 的性能可以通过 $\hat{A}$ 值来量度, 如: $\hat{A}(S,K,G)$, 这里的 G 是基于特定任务的理想核 $G=\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$, 其中 $\boldsymbol{y}=[y_1,\cdots,y_l]^{\mathrm{T}}$. 基于对目标核的度量原理, 通过使用不同的核函数, 或者调节不同的参数值, 可以产生一组核矩阵. 然后, 对该度量值的最大化执行半定规划或其他学习方法, 以得到一个对不同核矩阵加权组合的最优核.

3 多个尺度的多核学习: 多尺度核方法

合成核方法虽然有了一些成功应用, 但都是根据简单核函数的线性组合, 生成满足 Mercer 条件的新核函数; 核函数参数的选择与组合没有依据可循, 对样本的不平坦分布仍无法圆满解决, 限制了决策函数的表示能力. 在此情况下, 出现了多核学习的一种特殊化情形, 即将多个尺度的核进行融合. 这种方法更具灵活性, 并且能比合成核方法提供更完备的尺度选择. 此外, 随着小波理论、多尺度分析理论的不断成熟与完善, 多尺度核方法通过引入尺度空间, 使其具有了很好的理论背景. 这类方法目前也得到了很好的利用, 典型的如 Kingsbury 等[32] 将多个尺度大小的核进行分步分类; Zheng 等[33]、Yang 等[34] 提出了多尺度支持向量回归, 分别用于非平坦函数的估计和时间序列预测. 此外, 通过进一步将多尺度核与支持向量机结合, 多尺度核方法在基于回归的热点检测[48] 和图像压缩[49] 等方面均得到了应用. 近来, 结合多尺度分析方法, 基于 Hilbert 空间中的再生核进行函数重构[50] 得到了重视并进行了相关的应用研究; 此外, 多尺度核方法又逐步推广到了高斯过程的建模与处理[51−52, 77], 这对基于核方法的机器学习又是一次大的扩展.

3.1 具有多尺度表示能力的核函数

多尺度核方法的基础就是要找到一组具有多尺度表示能力的核函数. 在被广泛使用的核函数中, 高斯径向基核

$$k(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right) \tag{22}$$

是最受欢迎的, 因为它们具有通用普遍的近似能力, 同时它也是一种典型的可多尺度化核. 以此核为例, 将其多尺度化 (假设其具有平移不变性):

$$k\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma_1^2}\right),\cdots,k\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma_m^2}\right)$$

其中, $\sigma_1<\cdots<\sigma_m$. 可以看出, 当 σ 较小时, SVC 可以对那些剧烈变化的样本进行分类; 而当 σ 较大时, 可以用来对那些平缓变化的样本进行分类, 能得到更优的泛化能力. 具体实现时, σ 的取值可以借鉴小波变换中尺度变化的规律, σ 可由下式定义:

$$\sigma_i=2^i\sigma,\ \ i=0,1,2,\cdots$$

另一种典型多尺度核为小波核函数 (Wavelet kernel function)[83].

定理 1. 令 $h(x)$ 是一个小波母函数, a 和 c 分别表示伸缩和转移因子, $a,c\in\mathbf{R}$, 如果 $\boldsymbol{x},\boldsymbol{z}\in\mathbf{R}^n$, 则内积型小波核函数可表示为

$$k(\boldsymbol{x},\boldsymbol{z})=\prod_{i=1}^{n}h\left(\frac{x_i-c_i}{a}\right)h\left(\frac{z_i-c_i'}{a}\right) \tag{23}$$

转移不变小波核函数为

$$k(\boldsymbol{x},\boldsymbol{z})=\prod_{i=1}^{n}h\left(\frac{x_i-z_i}{a}\right) \tag{24}$$

定理 2. 考虑具有一般性的小波函数

$$h(x)=\cos(1.75x)\left(-\frac{x^2}{2}\right) \tag{25}$$

如果 $\boldsymbol{x},\boldsymbol{z}\in\mathbf{R}^n$, 则小波核函数为

$$k(\boldsymbol{x},\boldsymbol{z})=\prod_{i=1}^{n}h\left(\frac{x_i-z_i}{a}\right)=\prod_{i=1}^{n}\left\{\cos[1.75(x_i-z_i)]\exp\left(-\frac{\|x_i-z_i\|}{2a^2}\right)\right\} \tag{26}$$

通过伸缩因子 a 的变化, 即可得到不同尺度的小波核函数.

3.2 多尺度核的学习方法

1) 多尺度核序列学习方法

对多尺度核的学习, 很直观的思路就是进行多尺度核的序列学习. 多尺度核序列合成方法[32] 简单理解就是先用大尺度核拟合对应决策函数平滑区域的样本, 然后用小尺度核拟合决策函数变化相对剧烈区域的样本, 后面的步骤利用前面步骤的结果, 进行逐级优化, 最终得到更优的分类结果.

考虑一个两尺度核 k_1 和 k_2 合成的分类问题. 我们要得到合成的决策函数:

$$f(\boldsymbol{x})=f_1(\boldsymbol{x})+f_2(\boldsymbol{x}) \tag{27}$$

这里

$$\begin{aligned} f_1(\boldsymbol{x})&=\sum_{i=1}^{N}\alpha_i k_1(\boldsymbol{x}_i,\boldsymbol{x})+b_1\\ f_2(\boldsymbol{x})&=\sum_{i=1}^{N}\beta_i k_2(\boldsymbol{x}_i,\boldsymbol{x})+b_2 \end{aligned} \tag{28}$$

设想 k_1 是一个大尺度的核函数 (如 σ 较大的径向基函数), 相关的核项系数 α_i 选择那些决策函数

$f(\boldsymbol{x})$ 光滑区域对应的支持向量; 而 k_2 是小尺度核函数, 核项系数 β_i 选择那些决策函数 $f(\boldsymbol{x})$ 剧烈变化区域对应的支持向量. 具体方法是: 首先通过大尺度的单核 k_1 构造函数 $f_1(\boldsymbol{x})$, 这样, 该函数可以很好地拟合光滑区域, 但在其他地方存在显著误差, 可以使用相对较小的松弛因子来求取 α_i; 然后, 在 $f_1(\boldsymbol{x})$ 基础上使用小尺度的核 k_2 构造 $f_2(\boldsymbol{x})$, 使得联合函数 $f_1(\boldsymbol{x})+f_2(\boldsymbol{x})$ 比 $f_1(\boldsymbol{x})$ 具有更好的拟合性能. 这种方法实际上是多次使用二次规划以实现参数的获取, 运算复杂度较高, 同时支持向量的数量大量增加.

近来, Cortes 等[28] 关注于序列核函数学习问题, 提出了一种学习合理核函数的通用形式, 并且证明这个大的合理核函数族可以通过结合支持向量机目标函数和核脊回归的一种简单二次规划加以解决. 这种方法提高了先前出现的一些学习方法效率, 如 SDP、QCQP. 此外, 对于核脊回归这种特定情形, 给出了基于最优核矩阵封闭解的一个替代解, 该方法成功应用到了分类和回归问题.

2) 基于智能优化的多尺度核学习方法

与合成核类似, 针对多尺度核学习问题, 另一类重要的方法仍是基于多核的整体智能优化. 如: 利用 EM 算法训练多尺度支持向量回归 (Multi-scale support vector regression, MS-SVR), 这种方法起源于非平坦函数的估计问题[33]. 它采用多种尺度的核函数, 使得较小尺度的核可以拟合快速变化, 而较大尺度的核可以拟合平缓变化. 基于此, 文献 [84] 提出了两种迭代的 EM 训练算法, 它们分别以渐进的方式实现了 1 范数和 0 范数; 通过限制基函数的数目从而限制了模型的复杂性, 同时解的稀疏性大大加快了训练过程, 也节约了大量的测试时间. 并且 MS-SVR 的优化目标函数是全局的, 可以直接求得所有 ml 个核项系数 α_{ij}. 文献 [85] 提出了一种多尺度径向基核参数选择的进化策略, 用于支持向量机的训练. 该算法针对多尺度高斯核函数的组合形式:

$$k(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\alpha_i k(\boldsymbol{x},\boldsymbol{y},\gamma_i) \tag{29}$$

$$k(\boldsymbol{x},\boldsymbol{y},\gamma_i)=\exp(-\gamma_i\|\boldsymbol{x}-\boldsymbol{y}\|^2) \tag{30}$$

这里有 $2n$ 个参数需要确定, 分别是 n 个权重参数和 n 个径向基核的宽度 α_i,γ_i, $i=1,\cdots,n$. 这些参数值的确定会影响到组合核的性能, 也最终会影响分类精度. 具体方法是将样本集中的总训练样本分成数量相等的几份, 如 5 份, 任意选取其中的 4 份作为训练样本, 可以训练得到 5 个分类器, 分别以剩余的 1 份作为测试样本, 可以得到 5 个分类结果. 将这 5 个分类结果 (正确率) 的平均值作为目标函数 $f(\boldsymbol{v})$, 其中, $\boldsymbol{v}$ 是由 $2n$ 个组合核参数构成的向量 $(\alpha_1,\gamma_1,\alpha_2,\gamma_2,\cdots,\alpha_n,\gamma_n)$, 我们的目标就是找到一个合适的 $\boldsymbol{v}$, 使得 $f(\boldsymbol{v})$ 最大. 其核心就是通过对解向量进行组合、变异、选择, 在设定的进化代数内不断求取并更新目标函数, 最终得到 $f(\boldsymbol{v})$ 的最大值.

4 从有限向无限核的扩展 : 无限核方法

前述的合成核与多尺度核方法都是在有限个核函数线性组合前提下加以讨论的. 但对一些大规模问题, 基于有限个核的多核处理方法不一定有效, 多核融合的决策函数的表示能力也不能达到处处最优; 此外, 在一个多尺度核函数族中, 有限个核函数的选择并不唯一, 并且其不能完备地表征这个核函数族. 因此, 将有限核向无限核的扩展也是一个重要的方向.

早在 2006 年, Argyriou 等[53] 关注于监督学习中的核学习问题, 提出了一种多核学习的新方法, 该方法可以从由多个基本核函数的合法集合所构成的一个凸壳中找到某个核, 使其能最小化凸正则化函数. 与其他方法相比, 这个方法有一个独有的特征, 即上述基本核的个数可以是无限多个, 仅仅需要这些核是连续参数化的. 例如: 基本核可以是各向同性、方差规定为一定间隔的高斯核, 甚至可以是通过对多个连续参数进行参数化的高斯核. 尽管这个优化问题是非凸的, 但它属于更广范畴的凸函数差分 (Difference of convex functions) 规划问题, 因此, 可以应用最新的凸函数差分优化理论加以解决.

在此基础上, Gehler 等[39] 更侧重于用半无限规划解决来自通用核类型的核函数学习问题, 并设计了一种新的算法, 称为无限核学习 (Infinite kernel learning, IKL). IKL 算法既适用于无限也适用于有限核情形, 并且在大量核情形下, 它比 Simple MKL 算法更快也更稳定. 通过将 IKL 和 MKL 应用于 SVM 对大量数据集进行实验发现: 在一些数据集上, 由于核函数族中核的大幅增加, IKL 可以比 SVM/MKL 大大提高分类正确率, 在这些情况下, IKL 能保持它的实用性, 而交叉验证和 MKL 都是不实用的.

此后, 针对有限核的凸组合受限于选择的有限性问题, Ozogur-Akyüz 等[54] 又提出了一种基于无限和半无限优化的无限核组合新方法, 核的组合可以通过同伦参数 (Homotopy parameter) 或更特殊的参数来实现. 考察所有来自无限核集合中核的极小良凸组合, 可以发现: 由于组合, 间隔最大化是在一个紧指标集 (Compact index set) 下的无穷多个约束和一个额外的整数约束条件下实现的. 通过在概率测度 (Probability measures) 空间的参数化, 该问题变成了一个半无限问题, 可以通过半定规划方

法加以解决.

5 总结与展望

核方法在模式回归、分类以及概率密度估计等方面都有大量的理论分析和成功应用. 针对单核函数的弊端, 多核方法已经成为当前核机器学习的热点. 特别是在解决一些复杂问题时, 多核方法具有更优的性能. 基于此, 也出现了很多相关的多核机器学习方法, 这些方法为多核矩阵的融合提供了多种更优性能的解决途径, 同时一些方法还能有效解决大规模问题. 可以预见, 多核方法必将在核 Fisher 判别分析、支持向量机、相关向量机、奇异检测、核聚类分析等方向得到不断推广, 并在诸如时间序列 (动态系统、财经等) 预测、信号和图像的滤波、压缩和超解析 (Super-resolution)、故障预报、文本分类、图像处理、目标检测、视频跟踪、生物信息学 (双螺旋、基因序列数据的分类、蛋白质功能预测等) 等领域得到广泛应用.

多核学习方法已经得到了广泛认同和成功应用, 必将成为今后的理论研究热点和有效的模式分析工具, 以下总结其进一步的几个研究方向:

1) 除了常规的回归、分类等应用领域, 核方法的热点还包括核密度估计、子空间分析、用于回归或分类的特征选择[86] 等. 同样的, 合成核与多尺度核方法在这些领域也大有用武之地. 如用于两类、多类分类问题最优贝叶斯决策面的获取, 改进的核主成分分析等. 此外, 在当前机器学习的一些热门领域, 如多示例学习[87]、半监督学习、非监督学习等方向, 多核方法也得到了一些研究和应用. 典型的如 Kembhavi 等[88] 提出了一种多核方法的增量学习算法, 并将其用于目标识别. 黎铭等[89] 应用多核理论, 提出了基于多核集成的在线半监督学习方法 OMike (Online multi-kernel ensemble), 能够有效利用大量未标记数据提升在线学习性能. Zhao 等[90−91] 研究了基于最大间隔的多核聚类方法. 牟少敏等[92] 针对实际应用中经常出现的异类数据源和多核函数带来的计算量增加等问题, 提出了一种利用协同聚类对多核支持向量机的训练数据进行简化的方法, 大大减少了支持向量的数目, 从而减少了计算量, 并且还不会影响分类精度.

2) 核函数的构造一直是核机器学习的一个难点. 针对具体应用问题, 如何选择或构造出最合适的核函数, 目前并没有完善的理论基础. 多核方法通过多个或多种核函数的组合, 从另一个角度解决了特定核函数的构造问题, 通过多个参数的学习与调节, 使组合的核函数尽可能满足实际需求. 因此, 多核方法是针对复杂问题, 构造新的核函数与新的高效核机器[93] 的一条实用途径. 此外, 超核理论是 Ong 等[31] 近来提出的一种更广义的核理论, 在核函数的构造方面也是一种重大创新, 必将成为这一领域一个新的热点, 其理论体系还有待进一步完善与扩展.

3) 多核学习方法逐步向大规模数据集的应用领域扩展. 对于一些规模较小的数据集 (虽然通常核机器学习问题考虑的就是小样本情形), 多核方法采用这些组合操作和学习方法并不会遇到太大障碍. 但对于一些大规模问题, 由于涉及到多核矩阵的快速求解, 高维多核扩展矩阵的各种分解等处理, 常规的多核学习方法的学习效率会非常低下. 利用支持向量机以及各种改进方法, 虽然当前已经出现了一些基于大规模问题的学习算法, 如 Boosting 方法、QCQP、SDP 或一些分级方法[94−95] 等, 但这些方法在学习速度方面仍然有待提高.

4) 多尺度核方法具有更优的研究基础和前景. 随着尺度空间理论和小波理论的不断完善, 多尺度/多分辨分析方法得到了认同和推广. 多尺度核由于具有较为统一的模型形式和丰富的尺度选择, 使得它在处理复杂问题时有很大的优势, 特别是如高斯核这种具有强映射能力的核函数 (能将样本空间映射到无穷维的特征空间). 因此, 多尺度核方法的一种有效解决方案是引入尺度空间, 构成具有多分辨分析能力的增广特征空间, 获得决策函数的多分辨表示. 期望在这些特征空间所组成的一个增广空间中, 构造新的全局优化问题, 找到一个融合了它们各自优势的更好的解. 它的难点在于如何构建一套完备的多分辨分析的核函数空间理论, 并应用到支持向量机的快速实现. 这里可以设想采用标准化尺度核函数的构造, 期待多尺度特征增广空间解的快速求取. 此外, 多尺度核是有限个向无限个核函数扩展的基础. 对于如何解决多个尺度核分步执行带来的支持向量个数急剧增大的问题, 可以考虑构建新的基于多尺度合成核的全局优化问题, 采用迭代的方法, 获取基于非传统意义下的稀疏的支持向量[96−100]. 除此之外, 还可以将不同多尺度核进行合成, 研究多尺度合成核方法, 如高斯核与小波核, 进一步增强决策函数的多尺度分辨与表达性能.

5) 当前的一些多核学习方法, 还存在大量未解决的问题, 阻碍了多核方法的进一步推广. 如根据简单核函数的凸组合, 生成满足 Mercer 条件的新核函数, 它在核函数的选择与组合上没有依据可循; 如何处理好多个核之间的关系, 采用多核学习方法确定系数时的效率问题, 这些都有待进一步探索; 此外, 通常合成核方法对样本的不平坦分布仍无法很好解决, 限制了回归函数的拟合性能和分类决策函数的表示能力; 多尺度核方法在尺度选取上没有完备有效的理论依据, 并不是真正意义上的多尺度; 采用优化方法进行参数选择和序列多尺度方法, 都会使学

习时间成倍增加. 这些问题的存在, 使得合成核与多尺度核方法的研究还任重而道远.

References

1 Schoelkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 1998, **10**(5): 1299−1319

2 Scholkopf B, Mika S, Burges C J C, Knirsch P, Muller K R, Ratsch G. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Network*, 1999, **10**(5): 1000−1017

3 Muller K R, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel based learning algorithms. *IEEE Transactions on Neural Networks*, 2001, **12**(2): 181−201

4 Vapnik V N. *The Nature of Statistical Learning Theory*. Berlin: Springer, 1995

5 Vapnik V N. *Statistical Learning Theory*. New York: Wiley, 1998

6 Scholkopf B, Smola A J. *Learning with Kernels*. Massachusetts: The MIT Press, 2002

7 Shawe-Taylor J, Cristianini N. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004

8 Mercer J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London, Series A*, 1909, **209**: 415−446

9 Aronszajn N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950, **68**(3): 337−404

10 Aizerman A, Braverman E M, Rozoner L I. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 1964, **25**(5): 821−837

11 Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, USA: ACM, 1992. 144−152

12 Smola A J, Scholkopf B. A tutorial on support vector regression. *Statistics and Computing*, 2004, **14**(3): 199−222

13 Burges C J C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998, **2**(2): 121−167

14 Kerm P V. Adaptive kernel density estimation. *Stata Journal*, 2003, **3**(2): 148−156

15 Schölkopf B, Mika S, Smola A, Ratsch G, Muller K R. Kernel PCA pattern reconstruction via approximation preimages. In: Proceedings of the International Conference on Artificial Neural Networks. Skovde, Sweden: IEEE, 1998. 147−152

16 Mika S, Ratsch G, Weston J, Scholkopf B, Mullers K R. Fisher discriminant analysis with kernels. In: Proceedings of the Conference on Neural Networks for Signal Processing. Washington D. C., USA: IEEE, 1999. 41−48

17 Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 2000, **12**(10): 2385−2404

18 Lai P L, Fyfe C. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2000, **10**(5): 365−377

19 Bach F R, Jordan M I. Kernel independent component analysis. *The Journal of Machine Learning Research*, 2002, **3**(2): 1−48

20 Pavlidis P, Weston J, Cai J, Grundy W N. Gene functional classification from heterogeneous data. In: Proceedings of the 5th Annual International Conference on Computational Biology. Montreal, Canada: ACM, 2001. 242−248

21 Bennett K P, Momma M, Embrechts M J. MARK: a boosting algorithm for heterogeneous kernel models. In: Proceedings of 8th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada: ACM, 2002. 24−31

22 Ben-Hur A, Noble W S. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 2005, **21**(1): 38−46

23 Lewis D P, Jebara T, Noble W S. Nonstationary kernel combination. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM, 2006. 553−560

24 Sonnenburg S, Ratsch G, Schafer C, Scholkopf B. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 2006, **7**(7): 1531−1565

25 Damoulas T, Girolami M A. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 2008, **24**(10): 1264−1270

26 Bach F R. Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 2008, **9**(6): 1179−1225

27 Rakotomamonjy A, Bach F R, Canu S, Grandvalet Y. More efficiency in multiple kernel learning. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon: ACM, 2007. 775−782

28 Cortes C, Mohri M, Rostamizadeh A. Learning sequence kernels. In: Proceedings of the International Conference on Machine Learning for Signal Processing. Washington D. C., USA: IEEE, 2008. 2−8

29 Rakotomamonjy A, Bach F R, Canu S, Grandvalet Y. Simple MKL. *The Journal of Machine Learning Research*, 2008, **9**(11): 2491−2521

30 Kloft M, Brefeld U, Laskov P, Sonnenburg S. Non-sparse multiple kernel learning. In: Proceedings of the Workshop on Kernel Learning: Automatic Selection of Optimal Kernels. Whistler, Canada: The MIT Press, 2008. 1−4

31 Ong C S, Smola A J, Williamson R C. Learning the kernel with hyperkernels. *The Journal of Machine Learning Research*, 2005, **6**(7): 1043−1071

32 Kingsbury N, Tay D B H, Palaniswami M. Multi-scale kernel methods for classification. In: Proceedings of the IEEE Workshop on Machine Learning for Signal Processing. Washington D. C., USA: IEEE, 2005. 43−48

33 Zheng D N, Wang J X, Zhao Y N. Non-flat function estimation with a multi-scale support vector regession. *Neurocomputing*, 2006, **70**(1-3): 420−429

34 Yang Z, Guo J, Xu W, Nie X F, Wang J, Lei J J. Multi-scale support vector machine for regression estimation. In: Proceedings of the 3rd International Symposium on Neural Networks. Chengdu, China: Springer, 2006. 1030−1037

35 Lanckriet G R G, Cristianini N, Bartlett P, Ghaoui L E, Jordan M I. Learning the kernel matrix with semi-definite programming. In: Proceedings of the 19th International Conference on Machine Learning. San Francisca, USA: Morgan Kaufmann Publisher, 2002. 323-330

36 Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004. 41−48

37 Sonnenburg S, Ratsch G, Schafer C. A general and efficient multiple kernel learning algorithm. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2005. 1273−1280

38 Zien A, Ong C S. Multiclass multiple kernel learning. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 1191−1198

39 Gehler P V, Nowozin S. Infinite Kernel Learning, Technical Report 178, Max Planck Institute for Biological Cybernetics, Germany, 2008

40 Gönen M, Alpaydin E. Localized multiple kernel learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 352−359

41 Lanckriet G R G, Cristianini N, Bartlett P, Ghaoui L E, Jordan M I. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 2004, **5**(1): 27−72

42 Lee W J, Verzakov S, Duin R P. Kernel combination versus classifier combination. In: Proceedings of the 7th International Workshop on Multiple Classifier Systems. Prague, Czech Republic: Springer, 2007. 22−31

43 Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. In: Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2002. 367−373

44 Cristianini N, Elisseeff A, Shawe-Taylor J. On Optimizing Kernel Alignment, Technical Report, BIOwulf Technologies, Royal Holloway Universtiy of London, UK, 2003

45 Lin Y Y, Liu T L, Fuh C S. Local ensemble kernel learning for object category recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE, 2007. 1−8

46 Qiu S B, Lane T. Multiple kernel support vector regression for siRNA efficacy prediction. In: Proceedings of the 4th International Conference on Bioinformatics Research and Applications. Atlanta, USA: Springer, 2008. 367−378

47 Bi J B, Zhang T, Bennett K P. Column-generation boosting methods for mixture of kernels. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: ACM, 2004. 521−526

48 Pozdnoukhov A, Kanevski M. Multi-scale support vector algorithms for hot spot detection and modeling. *Stochastic Environmental Research and Risk Assessment*, 2007, **22**(5): 647−660

49 Li B, Zheng D N, Sun L F, Yang S Q. Exploiting multiscale support vector regression for image compression. *Neurocomputing*, 2007, **70**(16-18): 3068−3074

50 Opfer R. Multiscale kernels. *Advances in Computational Mathematics*, 2006, **25**(4): 357−380

51 Zhou Y T, Zhang T Y, Li X H. Multi-scale Gaussian processes model. *Journal of Electronics (China)*, 2006, **23**(4): 618−622

52 Walder C, Kim K I, Scholkopf B. Sparse multiscale Gaussian process regression. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 1112−1119

53 Argyriou A, Hauser R, Micchelli C A, Pontil M. A DC-programming algorithm for kernel selection. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM, 2006. 41−48

54 Ozogur-Akyüz S, Weber G W. Learning with infinitely many kernels via semi-infinite programming. In: Proceedings of the EURO Mini Conference on Continuous Optimization and Knowledge Based Technologies. Neringa, Lithunia: Vilnius Gediminas Technical University Publishing House, 2008. 342−348

55 Pavlidis P, Weston J, Cai J, Grundy W N. Gene functional classification from heterogeneous data. In: Proceedings of the 5th Annual International Conference on Computation Biology. Montreal, Canada: ACM, 2001. 249−255

56 Lanckriet G R G, Deng M, Cristianini N, Jordan M I, Noble W S. Kernel-based data fusion and its application to protein function prediction in yeast. In: Proceedings of the Pacific Symposium on Biocomputting. Berkeley, USA: Springer, 2004. 300−311

57 Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval. Amsterdam, The Netherlands: ACM, 2007. 401−408

58 He J F, Chang S F, Xie L X. Fast kernel learning for spatial pyramid matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE, 2008. 1−7

59 Zheng S, Liu J, Tian J W. An efficient star acquisition method based on SVM with mixtures of kernels. *Pattern Recognition Letters*, 2005, **26**(2): 147−165

60 Gustavo C V, Luis G C, Jordi M M, Joan V F, Javier C M. Composite kernels for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing Letters*, 2006, **3**(1): 93−97

61 Nakajima S, Binder A, Muller C, Wojcikiewicz W, Kloft M, Brefeld U. Multiple Kernel Learning for Object Classification, Technical Report on Information-Based Induction Sciences, University of Berlin, Germany, 2009

62 Lampert C H, Blaschko M B. A multiple kernel learning approach to joint multi-class object detection. In: Proceedings of the 30th DAGM Symposium on Pattern Recognition. Munich, Germany: Springer, 2008. 31−40

63 Kumar A, Sminchisescu C. Support kernel machines for object recognition. In: Proceedings of the IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−8

64 Damoulas T, Girolami M A. Pattern recognition with a Bayesian kernel combination machine. *Pattern Recognition Letters*, 2009, **30**(1): 46−54

65 Vedaldi A, Gulshan V, Varma M, Zisserman A. Multiple kernels for object detection. In: Proceedings of the International Conference on Computer Vision. Tokyo, Japan: IEEE, 2009. 1−8

66 Mak B, Kwok J T, Ho S. A study of various composite kernels for kernel eigenvoice speaker adaptation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canada: IEEE, 2004. 325−328

67 Cristianino N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000

68 Smola A, Scholkopf B. A tutorial on support vector regression. *Statistics and Computing*, 2004, **14**(3): 199−222

69 Zheng Da-Nian. Reaseach on Kernel Methods in Machine Learning [Ph. D. dissertation], Tsinghua University, China, 2006
(郑大念. 机器学习中的核方法研究 [博士学位论文], 清华大学, 中国, 2006)

70 Damoulas T, Ying Y, Girolami M A, Campbell C. Inferring sparse kernel combinations and relevance vectors: an application to subcellular localization of proteins. In: Proceedings of the 7th International Conference on Machine Learning and Applications. San Diego, USA: IEEE, 2008. 577−582

71 Fu S Y, Yang G S, Hou Z G, Liang Z, Tan M. Multiple kernel learning from sets of partially matching image features. In: Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: IEEE, 2008. 1−4

72 Fung G, Dundar M, Bi J, Rao B. A fast iterative algorithm for fisher discriminant using heterogeneous kernels. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004. 40−47

73 Damoulas T, Girolami M A. Combining feature spaces for classification. *Pattern Recognition*, 2009, **42**(11): 2671−2683

74 Damoulas T, Girolami M A. Pattern recognition with a Bayesian kernel combination machine. *Pattern Recognition Letters*, 2009, **30**(1): 46−54

75 Gustavo C V, Manel M R, Rojo-Alvarez J, Munoz-Mari J L. Nonlinear system identification with composite relevance vector machines. *IEEE Signal Processing Letters*, 2007, **14**(4): 279−282

76 Gönen M, Alpaydm E. Multiple kernel machines using localized kernels. In: Proceedings of the 4th International Conference on Pattern Recognition in Bioinformatics. Sheffield, UK: University of Sheffield, 2009. 1−10

77 Christoudias M, Urtasun R, Darrell T. Bayesian Localized Multiple Kernel Learning, Technical Report No. UCB/EECS-2009-96, University of California, USA, 2009

78 Kloft M, Brefeld U, Sonnenburg S, Laskov P, Muller K R, Zien A. Efficient and accurate l_p-norm multiple kernel learning. In: Proceedings of the Conference on Neural Information Processing Systems. Berkeley, USA: The MIT Press, 2009. 997−1005

79 Wang Z, Chen S, Sun T. MultiK-MHKS: a novel multiple kernel learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(2): 348−353

80 Nesterov Y E, Nemirovsky A S. *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*. Philadelphia: SIAM, 1994

81 Borgwardt K, Ong C S, Schoenauer S, Vishwanathan S, Smola A, Kriegel H P. Protein function prediction via graph kernels. *Bioinformatics*, 2005, **21**(1): 47−56

82 Jiang T J, Wang S Z, Wei R X. Support vector machine with composite kernels for time series prediction. In: Proceedings of the 4th International Symposium on Neural Networks: Advances in Neural Networks. Nanjing, China: Springer, 2007. 350−356

83 Zhang L, Zhou W D, Jiao L C. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2004, **34**(1): 34−39

84 Zheng D N, Wang J X, Zhao Y N. Training sparse MS-SVR with an expectation-maximization algorithm. *Neurocomputing*, 2006, **69**(13-15): 1659−1664

85 Phienthrakul T, Kijsirikul B. Evolutionary strategies for multi-scale radial basis function kernels in support vector machines. In: Proceedings of the Conference on Genetic and Evolutionary Computation. Washington D. C., USA: ACM, 2005. 905−911

86 Manik V, Rakesh B. More generality in efficient multiple kernel learning. In: Proceedings of the 26th International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 1065−1072

87 Tang L, Chen J, Ye J. On multiple kernel learning with multiple labels. In Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, USA: Morgan Kaufmann Publishers, 2009. 1255−1260

88 Kembhavi A, Siddiquie B, Miezianko R, McCloskey S, Davis L S. Incremental multiple kernel learning for object recognition. In: Proceedings of the International Conference on Computer Vision. Washington D. C., USA: IEEE, 2009. 1−8

89 Li Ming, Zhou Zhi-Hua. Online semi-supervised learning with multi-kernel ensemble. *Journal of Computer Research and Development*, 2008, **45**(12): 2060−2068
(黎铭, 周志华. 基于多核集成的在线半监督学习方法. 计算机研究与发展, 2008, **45**(12): 2060−2068)

90 Zhao B, Kwok J T, Zhang C S. Multiple kernel clustering. In: Proceedings of the 9th SIAM International Conference on Data Mining. 2009. Philadelphia, USA: SIAM, 638−649

91 Zhao B, Kwok J T, Zhang C S. Maximum margin clustering with multivariate loss function. In: Proceedings of the 9th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE, 2009. 637−646

92 Mu Shao-Min, Tian Sheng-Feng, Yin Chuan-Huan. Multiple kernel learning based on cooperative clustering. *Journal of Beijing Jiaotong University*, 2008, **32**(2): 10−13
(牟少敏, 田盛丰, 尹传环. 基于协同聚类的多核学习. 北京交通大学学报, 2008, **32**(2): 10−13)

93 Do H, Kalousis A, Woznica A, Hilario M. Margin and radius based multiple kernel learning. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Bled, Slovenia: Springer, 2009. 330−343

94 Xu Z L, Jin R, King I, Lyu M R. An extended level method for efficient multiple kernel learning. In: Proceedings of the Conference on Neural Information Processing Systems. Hong Kong, China: The MIT Press, 2009. 1825−1832

95 Dinuzzo F. Kernel machines with two layers and multiple kernel learning [Online], available: http://arxiv.org/abs/1001.2709, June 16, 2009

96 Tipping M E. Sparse Bayesian learning and relevance vector machine. *Journal of Machine Learning Research*, 2001, **1**(3): 211−244

97 Figueiredo M A T. Adaptive sparseness using Jeffreys prior. In: Proceedings of the Conference on Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2001. 697−704

98 Figueiredo M A T, Jain A K. Bayesian learning of sparse classifiers. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2001. 35−41

99 Figueiredo M A T. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(9): 1150−1159

100 Subrahmanya N, Shin Y C. Sparse multiple kernel learning for signal processing applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(5): 788−798

汪洪桥 清华大学计算机科学与技术系博士研究生, 第二炮兵工程学院指挥自动化系讲师. 主要研究方向为机器学习与模式识别. 本文通信作者.
E-mail: whq05@mails.tsinghua.edu.cn
(**WANG Hong-Qiao** Ph. D. candidate in the Department of Computer Science and Technology, Tsinghua University; lecturer in the Department of Command Automation, the Second Artillery Engineering Institute. His research interest covers machine learning and pattern recognition. Corresponding author of this paper.)

孙富春 清华大学计算机科学与技术系教授. 主要研究方向为人工智能、智能控制及机器人.
E-mail: fcsun@mail.tsinghua.edu.cn
(**SUN Fu-Chun** Professor in the Department of Computer Science and Technology, Tsinghua University. His research interest covers artificial intelligence, intelligent control, and robotics.)

蔡艳宁 第二炮兵工程学院指挥自动化系讲师. 主要研究方向为机器学习.
E-mail: caiyanning666@gmail.com
(**CAI Yan-Ning** Lecturer in the Department of Command Automation, the Second Artillery Engineering Institute. Her main research interest is machine learning.)

陈 宁 清华大学计算机科学与技术系博士研究生. 主要研究方向为机器学习与模式识别. E-mail:
chenn07@mails.tsinghua.edu.cn
(**CHEN Ning** Ph. D. candidate in the Department of Computer Science and Technology, Tsinghua University. Her research interest covers machine learning and pattern recognition.)

丁林阁 清华大学计算机科学与技术系硕士研究生. 主要研究方向为机器学习与数字图像处理.
E-mail: dlg06@mails.tsinghua.edu.cn
(**DING Lin-Ge** Master student in the Department of Computer Science and Technology, Tsinghua University. His research interest covers machine learning and digital image processing.)