文章编号：2095-2163(2019)01-0188-04 中图分类号：TP18 文献标志码：A

多核支撑向量回归方法研究

陈 博，郑凯东，王家华
（西安石油大学 计算机学院，西安 710065）

摘 要：近些年来，支撑向量回归方法在减少泛化误差方面表现出了出色的性能。然而，传统的支撑向量机或者支撑向量回归方法是基于单个核函数的，在高维空间中解决非线性问题。但随着应用领域不断扩展，在一些复杂情形下，由单个核函数构成的支撑向量回归方法并不能满足数据异构、输入空间维度过高等实际问题。针对此问题，人们在单核学习的基础上提出了多核学习，即将多个核函数进行线性组合，以此来提高模型精度，并逐渐成为当下机器学习领域研究的热点。文章综述了支撑向量回归算法与多核学习算法理论知识，并分析了各自的特点以及应用领域。总结了多核支撑向量回归方法下一步的研究趋势。

关键词：支撑向量机；支持向量回归；多核学习

Research on Multiple-Kernel Support Vector Regression Method

CHEN Bo , ZHENG Kaidong, WANG Jiahua
(School of Computer Science, Xi'an Shiyou University, Xi'an 710065)

【Abstract】 In recent years, the Support Vector Regression method has shown excellent performance in reducing generalization errors. However, traditional Support Vector Machines or Support Vector Regression methods are based on a single kernel function to solve nonlinear problems in high-dimensional space. However, with the continuous expansion of the application field, in some complicated situations, the Support Vector Regression method consisting of a single kernel function can not meet the practical problems of data heterogeneity and input space dimension is too high. Aiming at this problem, multiple kernel learning has been put forward on the basis of single-core learning, which is to linearly combine multiple kernel functions to improve the accuracy of the model, and gradually become a hot research topic in the field of machine learning. This paper reviews the theory of Support Vector Regression algorithm and multiple kernel learning algorithm, and analyzes their respective characteristics and application fields. At the end of the article, the research trends of the multiple Kernel Support Vector Regression method are summarized.

【Key words】 Support Vector Machine; Support Vector Regression; multiple kernel learning

0 引 言

支撑向量机(SVM)是一种实现了结构风险最小化和 VC 理论的学习方法，于 1995 年首次引入(Cortes and Vapnik,1995)。由于支撑向量机在文本分类中显示出了卓越的性能(Joachims,1998)，很快就成为了机器学习的主流技术。支持向量机有 2 个主要类别：支持向量分类(SVC)和支持向量回归(SVR)[1]。SVM 是一种使用高维特征空间的学习方法，其在支撑向量的子集上扩展预测函数。SVM 可以通过很少的支持向量推广复杂的灰度级结构，从而为图像压缩提供了一种新的机制[2]。1997 年，Vapnik、Steven Golowich 和 Alex Smola 提出了一种 SVM 用于回归的版本。该方法称为支持向量回归(SVR)。是支撑向量机最常见的应用形式。

传统的统计回归过程通常是得到一个函数 $f(x)$ 的过程，这个函数对于所有实验样本预测和实验观察到的数据之间具有最小误差。支撑向量回归(SVR)的一个主要特点是，其不是用来最小化观察到的训练误差，而是尝试最小化泛化误差边界，从而实现泛化性能。

1 支撑向量回归

1.1 线性支撑向量回归

一般的回归估计可以表述为如下问题，给定输入样本 $\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$，$x \in R$，其中，$x$ 代表输入空间，y 代表对应目标数据，回归估计的目标是去寻找一个函数 $y=f(x)$，使得函数对于所有输入数据 x 与实际获得的目标 y 数据具有最小的误差，并且尽可能平滑。其中 $f(x)$ 可以表达为如下

作者简介：陈 博(1993-)，男，硕士研究生，主要研究方向：智能计算、可视化技术；郑凯东(1964-)，男，副教授，主要研究方向：图形学与虚拟现实、程序设计、计算机基础教育；王家华(1945-)，男，教授，主要研究方向：地质统计学算法、油气藏建模、油气田地质图形可视化。

收稿日期：2018-10-24

形式：

$$f(x) = \omega x + b \quad x \in R \quad b \in R \tag{1}$$

与传统回归方法不同的是，支撑向量回归是假设可以容忍 $f(x)$ 与 y 之间最多有 ε 的偏差，即仅当 $f(x)$ 与 y 之间差别的绝对值大于 ε 时才计算损失。如图 1 所示，以 $f(x)$ 为中心，构建一个宽度为 2ε 的间隔带，若训练样本落入次间隔带，则认为是被预测正确的。要使公式(1)更加平滑，则意味着要最小化 ω。为此，需要最小化欧几里得范数，即 $\| \omega \|^2$，所以该问题可以转化为凸优化问题，表达如下：

$$\begin{aligned} &\text{minimize } \frac{1}{2} \| \omega \|^2 \\ &\text{s.t.} \quad y_i - \omega x - b \leqslant \varepsilon \\ &\qquad \omega x + b - y_i \leqslant \varepsilon \end{aligned} \tag{2}$$

在一些情况下，误差是允许存在的，不可能对所有训练数据都有极高的精度，为此，可以引入松弛向量 ξ 处理那些不符合要求的样本，公式可以转换为：

$$\begin{aligned} &\text{minimize } \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ &\text{s.t.} \quad y_i - \omega x - b \leqslant \varepsilon + \xi_i \\ &\qquad \omega x + b - y_i \leqslant \varepsilon + \xi_i^* \quad \xi_i \xi_i^* \geqslant 0 \end{aligned} \tag{3}$$

其中，C 为正则化参数，用来平衡模型的复杂度与训练误差，SVR 一般使用 ε 不敏感损失函数。如图 2 所示，其定义为：

$$| \xi | = \begin{cases} 0 & \text{if } | \xi | < \varepsilon \\ | \xi | - \varepsilon & \text{Otherwise} \end{cases} \tag{4}$$

引入拉格朗日乘子，得到该问题的对偶形式：

$$\begin{aligned} L = &\frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \sum_{i=1}^{l} (\alpha_i + \xi_i - \\ &y_i + \omega x_i + b) - \sum_{i=1}^{l} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \omega x_i - \\ &b) - \sum_{i=1}^{l} (\eta_i \xi_i - \eta_i^* \xi_i^*) \end{aligned} \tag{5}$$

求解上述优化问题，令公式(5)中对偶函数 $L = L(\omega, b, \alpha_1, \alpha_1^*, \xi_1, \xi_1^*, \eta^i, \eta_i^*)$ 对 ω、b、ξ_i 和 ξ_i^* 的偏导为 0，则线性支撑向量回归最终可求得如下形式：

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) (x_i \cdot x) + b \tag{6}$$

1.2　非线性支撑向量回归

实际应用中，大多数的训练样本都为高维向量，是线性不可分的。求解公式(6)中涉及到计算输入向量 x 在特征空间的内积。由于特征空间维数可能很高，甚至可能是无穷维。因此直接计算内积是非常困难的。为了避开这个问题，可以设想如下函数：

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = (x_i \cdot x_j) \tag{7}$$

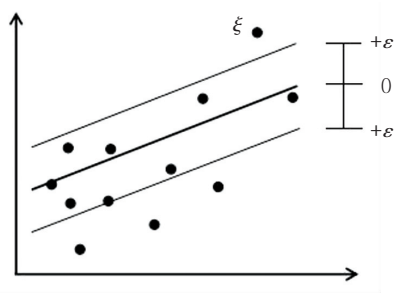

图 1　支撑向量回归示意图

Fig. 1　Support Vector Regression

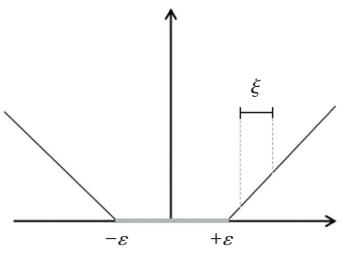

图 2　ε−不敏感损失函数

Fig. 2　ε−insensitive loss function

即 x_i 与 x_j 在特征空间的内积等于它们在原始样本空间中通过函数 $k(\,)$ 计算的结果。有了这样的函数，就不必直接去计算高维甚至无穷特征空间中的内积。这样的函数称为核函数。于是，引入核函数，将原优化问题转换为：

$$\text{minimize } \frac{1}{2} \| \omega \|^2 + \frac{C}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{8}$$

$$\text{s.t.} \begin{cases} (\varphi(x_i) + b) - y_i \leqslant \varepsilon + \xi_i \\ y_i - (\varphi(x_i) + b) \leqslant \varepsilon + \xi_i^* \\ i = 1, \cdots, l \end{cases}$$

对偶问题转化为：

$$\begin{aligned} \max &\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*) \\ &(\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) \\ \text{s.t.} &\begin{cases} \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ \frac{C}{l} \geqslant \alpha_i, \ \alpha_i^* \geqslant 0, \ i = 1, \cdots, l \end{cases} \end{aligned} \tag{9}$$

支撑向量回归模型为

$$y = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{10}$$

2 多核支撑向量回归学习

2.1 介绍

由于近些年来支撑向量机和支撑向量回归方法的发展。极大地提高了核函数领域的研究。核函数的使用,使得支撑向量算法可以由线性不可分问题转换为线性可分问题,使其快速应用到诸多领域。2008 年,黎铭等人提出了基于多核集成的在线半监督学习方法。之后,牟少敏等人提出了一种基于协同聚类的多核学习方法。2010 年,韩彦军等人提出了一种基于多核学习的双稀疏关系学习算法,随着理论研究的逐渐深入,核方法的使用有效地改进了许多学习算法的性能,极大地扩展了这些算法在文本分类、储层建模、疾病诊断、图像划分、计算机视觉、目标检测以及医疗成像分析等领域的应用[3]。传统的支撑向量回归方法,往往使用单个核函数。核函数的选择直接决定了支撑向量回归的最终性能。由于不同的核函数具有不同的特性,所以,在使用单个核函数时,往往模型的性能会各不相同。如常见的核函数有高斯核函数、线性核函数、多项式核等[4]。

此外,当样本特征含有异构信息、样本规模很大,多维数据的不规则、或者数据在高维特征空间分布不平坦、模型采用单个核函数进行数据映射并不合理。所以,近些年来,人们尝试使用不同的核函数组合来对应不同的相似性概念或者多个来源的信息,即多核学习方法。

2.2 核方法

由公式(9)可看出,给定一个回归模型,其中,x 代表输入向量,y 代表输出向量,若不考虑偏移项 b,在线性不可分的情况下,学得的模型总可以表示成核函数 $K(x_i,x_j)$ 的线性组合。则输入向量 x 可以通过一个非线性映射将数据映射到高维特征空间。然后,在高维特征空间中,将原有的线性不可分问题转换为线性可分问题进行求解。非线性映射表达如下:

$$x \rightarrow \phi(x) \tag{11}$$

将输入数据映射到一个新的特征空间后,原有的模型学习转变为:

$$(\phi(x_1),\ y_1),\ (\phi(x_2),\ y_2),\cdots,(\phi(x_l),\ y_l) \tag{12}$$

由于超平面可以表示为所有训练样本在高维特征空间中与测试样本内积的线性组合,而核函数可以把非线性映射和内积计算 2 个过程结合起来。从而降低了在高维空间中计算内积的复杂度,避免了维数灾难。常用的核函数如下:

线性核:

$$k(x_i,\ x_j) = x_i^{\mathrm{T}} x_j \tag{13}$$

多项式核:

$$k(x_i,\ x_j) = (x_i^{\mathrm{T}} x_j)^d \tag{14}$$

高斯核:

$$k(x_i,\ x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \tag{15}$$

2.3 多核学习

有些情况下,样本规模很大或者数据含有异构信息,即数据来自于不同的数据源,例如在储层建模领域,Demyanov 已经证明了半监督支撑向量回归(支撑向量机的扩展)方法有能力,在流体环境中将岩石物性转换为模型分布,但在高维空间里计算线性回归,对数据映射使用单一的核,会使复制多尺度、不稳定结构变得复杂,不能精确反映出数据的相关性。尽管这种复杂性在一定程度上是可以使用未跟踪数据的半监督学习来解决的,但是却不能在多重尺度的精确反映出相关性。所以,多核学习逐渐成为当下研究的热点。

2004 年,Lanckriet 提出使用多个核函数的最优线性凸组合来替代单核,即多核学习。多核学习(MKL)现在已经作为一个特征选择技术使用到不同的应用中。例如,Tuia et.al 使用 MKL 从卫星图中选择相关特征。Takashima et.al 使用 MKL 从声源定位中寻找声音转移函数的相关光谱维度。随后,Dileep et.al 更详细的解释了 MKL 特征选择理论,并把其应用到图像分类当中,且多核 SVR 已经被应用到放射性映射和模拟风场当中。多核学习可表示为如下形式:

$$k(x_i,\ x_j) = \sum_{m=1}^{M} d_m k(x_i,\ x_j;\ \theta_m)$$

$$d_m \geqslant 0 \qquad \sum_{m=1}^{M} d_m = 1 \tag{16}$$

其中,θ_m 代表单核的各个内部参数;d_m 代表每个核权重;M 为核函数的个数。多重核方程可以灵活地代替曾使用的输入特征、核的超参数和数量。每个核函数可能使用不同的特征来描述来自不同数据源的特征,所以与单核学习不同的是,原有的模型优化问题转化为核权重 d 的选择问题,因此,需要在单核中同时学习系数 θ 和权重 d。图 3 所示就是多核支撑向量回归学习过程图。

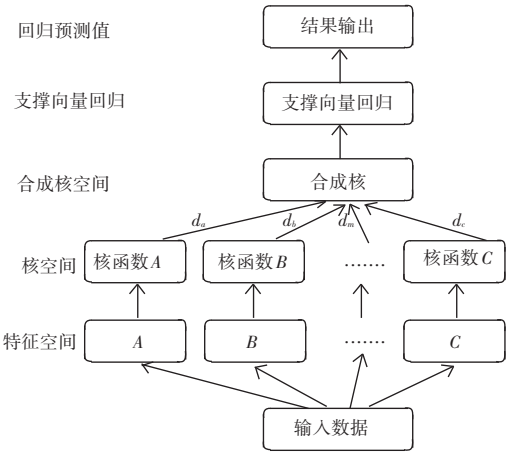

图 3　多核支撑向量回归过程

Fig. 3　Multi-core Support Vector Regression process

3　总结与展望

如前文所述，文章针对支撑向量回归，多核学习的概念、算法以及应用做出了研究综述。多核支撑向量回归方法，在模式识别等诸多领域都有大量的理论分析和成功应用。由于单核支撑向量回归所存在的模型精度不高的弊端，多核支撑向量回归方法已经成为当下机器学习的研究热点。在解决一些复杂问题时，如数据异构、输入数据维度较高，多核支撑向量回归模型会具有更高的精度。可以预见，多核学习必将在支撑向量机、支撑向量回归等方面得到不断推广。

基于支撑向量回归的多核学习现如今已经得到了各个领域的普遍认可，必将成为一种有效的回归分析工具。但当前的学习方法，仍然存在一些未解决的问题。阻碍了多核支撑向量回归学习的进一步发展。

(1)当产生新的核函数时，其满足 Mercer 条件，该如何与之前所使用的核函数进行融合。

(2)如何处理和多个核函数的之间的关系，也就是当使用多核支撑向量回归模型时，确定模型系数的主要步骤是反复学习支撑向量回归模型，这极大地降低了学习效率。

(3)多核支撑向量回归模型拥有诸多参数，如超参数 C、敏感损失函数 ε、核带宽参数 σ，良好的参数可以很大程度上减少模型的训练时间，所以，寻找模型参数的最优组合也是一大研究热点。

这些问题的存在，使得多核支撑向量回归学习研究的道路依然任重而道远。

参考文献

[1] 张亦俊. 针对互联网公共服务的搜索引擎关键技术研究[D]. 南京：东南大学，2016.

[2] 陈健. 基于多变量相空间重构的投资组合策略研究[D]. 广州：华南理工大学，2015.

[3] 陈健. 基于多变量相空间重构的投资组合策略研究[D]. 长沙：国防科学技术大学，2013.

[4] 王梅，李董，孙莺萁，等. 求解大规模问题的多核学习正则化路径算法[J]. 模式识别与人工智能，2018，31(2)：190-196.

[5] 汪洪桥，孙富春，蔡艳宁，等. 多核学习方法[J]. 自动化学报，2010，36(8)：1037-1050.

[6] GÖNEN M，ALPAYDIN E. Multiple kernel learning algorithms [J]. Journal of Machine Learning Research，2011，12：2211-2268.

[7] BASAKD，PAL S ，PATRANABIS D C. Support Vector Regression [J]. Neural Information Processing - Letters and Reviews，2007，11(10)：203-224.

[8] SMOLA A J，SCHOLKOPF B. A tutorial on support vector regression[J]. Statistics and Computing，2004，14：199-222.