基于支持向量回归估计算法的小样本集回归分析

邱 彤
(清华大学化学工程系,北京 100084)

摘要:在简要介绍支持向量回归估计(SVR)算法的基础上,以某软测量建模为例,验证了 SVR 算法对于小样本集的回归分析问题可以得到具有良好泛化能力的回归估计函数,进而针对不同的核函数,探讨了设计参数和核函数参数的选择问题。

关键词:支持向量回归估计(SVR);小样本集;泛化能力;核函数
中图分类号:TP273 **文献标识码**:A **文章编号**:0253 - 4320(2004)S2 - 0160 - 03

Small-sample regression analysis based on support vector regression

QIU Tong
(Department of Chemical Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Firstly, an overview of the basic ideas of support vector regression(SRV) is given. Then, there is an example of soft-sensor modeling. Its regression analysis shows that SVR has good generalization ability for small-sample regression problems. Furthermore, the selection of parameters and kernel-based methods are discussed.

Key words: support vector regression (SVR); small-sample; generalization ability; kernel-based method

回归分析方法在科学研究和工程实际中被广泛应用,软测量、分子设计的 QRAR 建模等实际应用中,针对小样本集、非线性的回归分析是研究的重点。1995 年,基于统计学习理论(statistical learning theory, SLT)中的结构风险最小化(structural risk minimization, SRM)准则发展起来的支持向量机(support vector machines, SVMs)算法,综合考虑了模型复杂性与预测性能间的平衡,可以实现全局最优和良好的泛化能力。

传统的基于经验风险最小化(empirical risk minimization, ERM)准则的机器学习算法以训练误差最小为目标,用于多元非线性回归时,依赖较大的数据样本集,过拟合现象的出现会降低回归估计的泛化能力。支持向量回归估计(SVR)算法通过不敏感损失函数和核函数的引入,可以很好地应用于非线性回归分析,并且对小样本集问题具有良好的预测性能。

1 支持向量回归估计算法

给定 r 个样本数据$\{x_i, y_i\}_{i=1}^{r}$,式中 $x_i \in R^n$ 为 n 维样本输入,$y_i \in R$ 为样本输出。所谓非线性回归估计建模,就是要找出一个函数 f,使之通过样本训练后,对于样本以外的 x,通过 f 找出对应的 y。对非线性回归估计问题,利用非线性映射 $\varphi(\cdot)$将训练数据集非线性地映射到一个高维特征空间(Hilbert 空间),使得在输入空间中的非线性回归估计转化为高维特征空间中的线性估计问题。设函数 f 的形式如下:

$$f(x) = w^{\mathrm{T}}\varphi(x_i) + b, \quad \omega \in R^{nh}, b \in R \tag{1}$$

式中,非线性函数 $\varphi(\cdot): R^n \rightarrow R^{nh}$,将输入空间映射到一个高维特征空间,这里特征空间的维数不是固定的,b 为偏置量。根据结构风险最小化准则,f 应使得$\frac{1}{2}\| w \|^2 + C \cdot R_{\mathrm{emp}}[f]$最小。其中 C 是平衡因子,$R_{\mathrm{emp}}[f]$表示经验风险,可用 ε 不敏感损失函数定义 $R_{\mathrm{emp}}[f]$:

$$R_{\mathrm{emp}}[f] = \begin{cases} 0, & | y - f(x) | \leqslant \varepsilon \\ | y - f(x) | - \varepsilon, & | y - f(x) | > \varepsilon \end{cases} \tag{2}$$

为了确保最优化问题有解,引入松弛变量 ζ_i、ζ_i^*,回归预测模型可表示为:

$$\min \quad \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{r}(\zeta_i + \zeta_i^*)$$

$$\text{s.t.} \quad \begin{cases} y_i - w^T\varphi(x_i) - b \leqslant \varepsilon + \zeta_i \\ w^T\varphi(x_i) + b - y_i \leqslant \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geqslant 0 \end{cases} \quad i = 1, \cdots, r \tag{3}$$

采用对偶理论,建立 Lagrange 函数,得到对偶最优化问题:

$$\max \quad J = -\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} - \boldsymbol{d}^{\mathrm{T}}\boldsymbol{u}$$

作者简介:邱彤(1971 -),女,博士,讲师,从事过程系统工程领域教学与科研工作,010 - 62784572, qiutong@tsinghua.edu.cn。

$$\text{s.t.}\quad \begin{cases} \boldsymbol{s}^{\mathrm{T}}\boldsymbol{u} = 0 \\ \boldsymbol{u}_i \in [0, C] \qquad i = 1, \cdots, 2r \end{cases} \tag{4}$$

其中，$\boldsymbol{u} = \begin{bmatrix} a_1 \\ \vdots \\ a_r \\ a_1^* \\ \vdots \\ a_r^* \end{bmatrix}$，$\boldsymbol{Q} = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}$，$\boldsymbol{h} = \begin{bmatrix} \varphi(x_1) \\ \vdots \\ \varphi(x_r) \\ -\varphi(x_1) \\ \vdots \\ -\varphi(x_r) \end{bmatrix}$，

$$\boldsymbol{d} = \begin{bmatrix} \varepsilon - y_1 \\ \vdots \\ \varepsilon - y_r \\ -\varepsilon - y_1 \\ \vdots \\ -\varepsilon - y_r \end{bmatrix}, \boldsymbol{s} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \tag{5}$$

式(5)是带有等式约束的标准二次规划问题，求解可得 w 和 $f(x)$：

$$\begin{cases} w = \sum_{i=1}^{r} (a_i - a_i^*)\varphi(x_i) \\ f(x) = \sum_{i=1}^{r} (a_i - a_i^*)[\varphi(x_i) \cdot \varphi(x)] + b \end{cases} \tag{6}$$

$(\alpha_i - \alpha_i^*) \neq 0$ 对应的样本构成支持向量。变量 w 反映了函数的复杂度，它是非线性映射 $\varphi(\cdot)$ 的线性组合。引入核函数(kernel function) $k(x, x') = \langle \varphi(x) \cdot \varphi(x') \rangle$，可以简化计算。偏置量 b 可通过任意一个满足 KTT(Karush - Kuhn - Tucker)条件的样本计算得到：

$$f(x) = \sum_{i=1}^{r} (a_i - a_i^*)k(x_i, x) + b$$

$$\begin{cases} b = y_j - \sum_{i=1}^{r} (a_i - a_i^*)k(x_i, x_j) - \varepsilon \qquad a_i \in (0, C) \\ \text{或 } b = y_j - \sum_{i=1}^{r} (a_i - a_i^*)k(x_i, x_j) + \varepsilon \quad a_i^* \in (0, C) \end{cases} \tag{7}$$

核函数 $k(x, x')$ 是满足 Mercer 条件的任意对称函数，常用的核函数有：

(1)多项式核

$$k(x, x') = (x \cdot x' + 1)^p \qquad p = 1, 2, \cdots, n$$

(2)径向核(RBF)

$$k(x, x') = \exp(-\| x - x' \|^2 / 2\sigma^2)$$

(3)感知器核

$$k(x, x') = \tanh(\beta x \cdot x' + b)$$

2　支持向量回归的估计算法应用

在回归估计算法的实际应用中，实际数据来源于工业实际或实验室试验数据，已知样本数量有限的小样本集问题经常遇到。利用 SVR 算法进行回归估计时，所需设计参数少。SVR 算法中的参数 ε 表明了估计函数在样本数据点上误差的期望(误差的要求)，而 SRM 准则要求利用参数 C 折衷考虑是严格地要求训练误差小于 ε，还是充分考虑回归函数的泛化能力。核函数中包括的设计参数的数量依所选函数形式的不同而异。如选择多项式核函数，只涉及一个设计参数——多项式阶数 p，p 取正整数。此外，SVR 算法得到的回归函数的预测精度并非与所有样本有关，而只与支持向量有关。因此，从算法机理出发，SVR 适用于小样本集问题的回归处理。

本算例是一个化工过程软测量问题，某炼油厂希望根据进料量、进料温度、塔底温度、再沸器气相返塔温度、塔顶温度、塔顶压力、回流量和回流温度等 8 个过程参数预测催化裂化(FCC)稳定塔汽油饱和蒸汽压指标 y，$y = f(x_1, x_2, \cdots, x_8)$。

算例中，软测量模型有 8 个输入参数，而实际样本数据只有 28 组。当利用神经网络等算法进行回归估计时，为了建立合理的预测网络拓扑结构，一般依靠加入虚拟样本的方式扩大训练集。因此，本例利用线性差值法产生了 10 组虚拟样本。利用 SVR 法对上述样本进行回归估计计算，选用多项式核函数，所得结果如表 1 所示。

表 1　不同样本数的 SVR 回归估计(多项式核)

训练样本数(虚拟样本数)	15(0)		20(0)		28(8)	
检验样本数(虚拟样本数)	23(10)		18(10)		10(2)	
SVR 参数						
ε	0.10	0.05	0.10	0.05	0.10	0.05
C	20	20	20	20	15	15
p	1	1	1	1	1	1
训练方差	0.15	0.08	0.13	0.06	0.12	0.06
训练最大误差	0.34	0.15	0.26	0.13	0.36	0.18
检验方差	0.28	0.14	0.15	0.08	0.11	0.05
检验最大误差	0.63	0.31	0.37	0.18	0.26	0.13
支持向量个数	14	15	19	20	25	27

可以看出，随着训练样本的增加，训练方差和检验方差均有所下降。当取 15 组实际样本作为训练集，13 组实际样本和 10 组虚拟样本作为检验样本

集时,所得到的回归预测结果是:检验方差为 0.14,检验最大误差为 0.31。训练样本数增加到 20 组时,检验方差下降到 0.08,检验最大误差降到 0.18。因此即使不加入虚拟样本,预测结果已可满足实际需要。

3 支持向量回归估计算法的应用讨论

3.1 设计参数的确定

应用 SVR 算法时,设计参数 C 和 ε 以及核函数参数的确定是至关重要的。参数 C 表示对训练误差超出 ε 的范围的样本的惩罚。如表 2 所示,在其他条件不变的情况下,一般 C 越大,训练误差越小,检验误差也会相应减小;但 C 取值过大,即过分强调减小训练误差,使 $\| w \|^2$ 项的权重相对下降,会造成回归函数泛化能力下降。参数 ε 表征回归估计的精度要求。对于小样本集问题,由于训练样本有限,ε 的选择对支持向量数的影响不明显。ε 取值,训练时回归估计的精度会有所提高。与 C 取值过大相类似,训练误差下降并不意味着检验误差也一定下降,盲目追求训练误差下降就相当于回到了以 ERM 准则建立估计函数,只强调经验误差小,而忽视了函数的泛化能力。

目前尚没有一种有效的方法能合理、快捷地确定 SVR 所涉及的各个参数,一般依靠交叉验证方法选取适当的参数。因此当前在软测量等领域应用较广的最小二乘 SVR 法,不仅是因为它能够将二次规划问题转化为求解线性方程组,简化计算,提高计算速度,而且它还使设计参数由 2 个变成 1 个,方便了设计参数的确定。

表 2 SVR 参数变化对回归结果的影响(多项式核)

SVR 参数			训练	训练最大	检验	检验最大
ε	C	p	方差	误差	方差	误差
0.001	40	3	0.005	0.03	0.48	0.89
0.01	10	1	0.08	0.21	0.10	0.26
0.01	10	2	0.08	0.38	0.31	0.61
0.01	15	2	0.04	0.20	0.30	0.58
0.01	60	2	0.01	0.03	0.32	0.74
0.05	10	2	0.07	0.26	0.29	0.56
0.05	10	1	0.18	0.46	0.20	0.53

3.2 核函数及其参数的选择

从表 1、表 2 的计算结果可以看出,利用多项式核函数时,参数 p 的理想取值为 1,当 $p > 1$ 时,训练精度有所提高,但泛化能力明显下降。核函数参数的取值与 C 和 ε 的取值相关。在获得相同的训练和检验精度时,对应的参数选择可能是不同的。尤其在训练样本很少时,不同的参数得到的支持向量及其系数也可能是相同的,此时得到的回归函数形式就是相同的。此外,核函数参数取值不同,计算得到的优化问题中矩阵 $\boldsymbol{Q}$ 各个分量的值差异很大,这将直接影响回归计算结果。

为了得到同样的回归估计结果,并非只有某一种形式的核函数是适用的。表 3 所列的是本算例利用径向核函数(RBF)进行回归估计得到的结果,对不同的训练样本数,都可以找到适宜的设计参数和核函数参数,使其训练和检验误差均与利用多项式核函数时的结果相当。因此,某种核函数在处理实际问题时是否适用,必须对建模中的各个参数进行充分的交叉取值验证,进而选择稳定性好、预测精度高的核函数。

表 3 不同样本数的 SVR 回归估计(RBF 核)

训练样本数(虚拟样本数)	15(0)		20(0)		28(8)	
检验样本数(虚拟样本数)	23(10)		18(10)		10(2)	
SVR 参数						
ε	0.05	0.05	0.01	0.005	0.005	0.005
C	500	400	500	500	500	600
σ	3.5	3	3.5	3.5	3.5	4
训练方差	0.85	0.05	0.03	0.03	0.03	0.02
训练最大误差	0.06	0.07	0.09	0.1	0.11	0.08
检验方差	0.26	0.32	0.05	0.04	0.06	0.05
检验最大误差	0.52	0.71	0.14	0.12	0.12	0.11
支持向量个数	15	15	20	20	28	28

参考文献

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995.

[2] Cortes C, Vapink V.[J]. Machine Learning, 1995, 20(1): 1 - 25.

[3] Colin Campbell.[J]. Neurocomputing, 2002, 48(1 - 4): 63 - 84.

[4] Cawley G C, Talbot N L C.[J]. Neurocomputing, 2002, 48(1 - 4): 1025 - 1031.

[5] 杜树新,吴铁军.[J].浙江大学学报(工学版),2004,28(3):302 - 306.

[6] 朱国强,刘士荣,俞金寿.[J].华东理工大学学报,2002,28(增刊):6 - 10.■