基于一类分类的线性规划支持向量回归算法

孙德山 赵 君 高采葵 郑 平 刘小菲
(辽宁师范大学数学学院 大连 116029)

摘 要 根据一类分类思想,提出一种基于线性规划的支持向量回归算法,该算法揭示了一类分类和回归之间的关系。实验在一个正弦函数、一个混沌时间序列和一个实际的数据上进行。实验结果表明,所给算法的泛化性能优于标准的支持向量回归算法(ε-SVR)、线性规划支持向量回归算法(LP-SVR)和最小二乘支持向量回归算法(LS-SVR),实验结果也说明了所给算法的有效性和可行性。

关键词 一类分类,支持向量机,回归算法,核函数
中图法分类号 TP18 **文献标识码** A

Linear Programming Support Vector Regression Method Based on One-class Classification

SUN De-shan ZHAO Jun GAO Cai-kui ZHENG Ping LIU Xiao-fei
(Department of Math, Liaoning Normal University, Dalian 116029, China)

Abstract A new support vector regression algorithm based on linear programming was proposed according to one-class classification, which can more uncover the relation between one-class classification and regression. The tests were perform on sine function, chaos time series and real world data sets. Experiments show that the new method has comparable or better generalization performance than ε-insensitive Support Vector Regression (ε-SVR), Linear Programming Support Vector Regression (LP-SVR) and Least squares support vector regression (LS-SVR), and also show that the proposed method is feasible and valid.

Keywords One-class classification, Support vector machine, Regression algorithm, Kernel function

1 引言

在模式识别领域中,分类和回归算法是最常用的技术。基于统计学习理论产生的支持向量机算法[1-3]是解决分类和回归问题的优秀算法,目前已经在很多领域得到广泛应用。该算法根据结构风险最小化原则,采用核函数技术巧妙地解决了非线性问题。最常用的支持向量回归算法有标准的支持向量回归算法(ε-SVR)、基于线性规划的支持向量回归算法(LP-SVR)、最小二乘支持向量回归算法(LS-SVR)[4]等。

文献[5]给出了一种基于闭凸包收缩的最大边缘分类算法,并且将该分类算法推广到解决回归问题[6]。一类分类是一种特殊的分类技术,是针对异常值检测提出的[7,8]。实际问题中的异常样本通常不容易获得,而更多情况下是获得正常样本(正类)。一类支持向量机正是在这种背景下提出的,该算法已经在很多领域得到了广泛的应用[9-12]。本文借助一类分类的思想,将一类分类算法推广到解决回归问题,并通过几个具体实例与常用的几个支持向量回归算法的比较,揭示了所给算法的有效性。本文的另一个目的是揭示一类分类与回归之间的关系,从而更直观地理解回归的几何意义,同时也能使分类和回归技术相融合,使得优秀的分类算法也能应用于解决回归问题。

2 一类分类算法

在模式识别中,异常值检测是一类特殊的分类问题。实际问题中的很多情况只能获得更多的正常样本,而异常样本不容易获得,一类分类方法正是在这种情况下产生的。一类分类的一种实现途径是寻找一个包含正类样本的最小超球,若新样本位于超球体内,就认为它是正常样本,否则就是异常样本。为了能将一类分类方法推广到回归情况,这里介绍一类分类的另外一种实现途径。

设给定一个正类样本集$\{x_i, i=1,\cdots,l\}$,$x_i \in R^d$,采用非线性映射ϕ将样本点映射到高维特征空间。一类分类的目的是在高维空间中寻找一个超平面,使之以最大的距离将样本从原点分开,即估计一个函数$f_w(x)=\langle w,\phi(x)\rangle$,当一个样本$x$满足$f_w(x)\geqslant\rho$时,它被确定属于该类。原点到超平面$f_w(x)=\rho$的距离为$\frac{\rho}{\|w\|_2}$,如图1所示,这里$\|w\|_2$表示欧式范数。为了求得$w$和$\rho$的值,根据结构风险最小化原则,并允许一定的误差存在,问题可以归结为下面的优化:

$$\min \frac{1}{2}\|w\|_2^2-\rho+C\sum_{i=1}^{l}\xi_i \tag{1}$$

约束为

$$\langle w,\phi(x_i)\rangle\geqslant\rho-\xi_i,\xi_i\geqslant 0,i=1,\cdots,l \tag{2}$$

到稿日期:2013-05-02 返修日期:2013-10-19 本文受国家自然科学基金项目(61105059)资助。

孙德山(1970-),男,副教授,硕士生导师,主要研究方向为统计学习理论;赵 君,高采葵,郑 平,刘小菲 硕士生,主要研究方向为支持向量机及其应用。

其中，$\frac{1}{2}\|w\|_2^2$为规划项，参数C在误差项和规划项中做出折中。

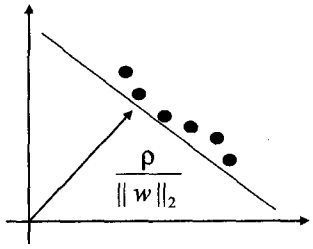

图1 一类分类超平面

若优化问题(1)中的规划项采用l_∞-范数，由文献[13]中的定理2.2可以得到其等价的线性优化问题：

$$\min -\rho+C\sum_{i=1}^{l}\xi_i \tag{3}$$

约束为

$$\langle w,\phi(x_i)\rangle\geqslant\rho-\xi_i,\xi_i\geqslant 0,i=1,\cdots,l \tag{4}$$

$$\|w\|_1=1 \tag{5}$$

将$\langle w,\phi(x_i)\rangle$采用核展开式$\sum_{j=1}^{l}\alpha_j k(x_j,x_i)$表示，于是可得到下面的线性规划形式：

$$\min -\rho+C\sum_{i=1}^{l}\xi_i \tag{6}$$

约束为

$$\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\geqslant\rho-\xi_j,j=1,\cdots,l \tag{7}$$

$$\sum_{i=1}^{l}\alpha_i=1 \tag{8}$$

$$\xi_i\geqslant 0,i=1,\cdots,l \tag{9}$$

这里只限制α_i的求和为1，而每个α_i的正负并不限制，是为了将之推广到后面的回归函数中。解此线性规划可以求得α和ρ的值，于是得到一个决策函数：

$$f(x)=\sum_{i=1}^{l}\alpha_i k(x_i,x) \tag{10}$$

根据优化问题的几何意义，对于大部分训练样本将满足$f(x)\geqslant\rho$，参数C的意义就是控制满足条件$f(x)\geqslant\rho$的样本数量，较大的参数C值将使所有的样本满足条件。得到的超平面为：

$$\sum_{i=1}^{l}\alpha_i k(x_i,x)=\rho \tag{11}$$

3 基于一类分类的回归算法

设回归问题的训练样本集为：

$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\},x_i\in R^n,y_i\in R$

为了将一类分类算法推广到回归情况，将训练样本重新记为：$z_i=(x_i,y_i),i=1,2,\cdots,l$。

求两个平行的超平面$\langle w,\phi(z)\rangle=\rho_1$和$\langle w,\phi(z)\rangle=\rho_2$，使第一个超平面以最大距离$\frac{\rho_1}{\|w\|}$将样本从原点分离开来，令第二个超平面以最小距离$\frac{\rho_2}{\|w\|}$使样本位于靠近原点的一侧，其几何意义如图2所示。这样回归函数就可以看成是两个超平面的垂直平分超平面，记为：

$$\langle w,\phi(z)\rangle=\frac{\rho_1+\rho_2}{2} \tag{12}$$

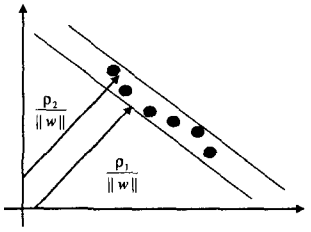

图2 两个平行的超平面

类似优化问题(6)，可得优化问题为：

$$\min -\rho_1+C\sum_{i=1}^{l}\xi_i+\rho_2+C\sum_{i=1}^{l}\xi_i' \tag{13}$$

约束为

$$\sum_{i=1}^{l}\alpha_i k(z_i,z_j)\geqslant\rho_1-\xi_j,j=1,\cdots,l \tag{14}$$

$$\sum_{i=1}^{l}\alpha_i k(z_i,z_j)\leqslant\rho_2+\xi_j',j=1,\cdots,l \tag{15}$$

$$\sum_{i=1}^{l}\alpha_i=1 \tag{16}$$

$$\xi_i,\xi_i'\geqslant 0,i=1,\cdots,l \tag{17}$$

于是回归方程为

$$\sum_{i=1}^{l}\alpha_i k(z_i,z)=\frac{\rho_1+\rho_2}{2} \tag{18}$$

其中常用的核函数的选取有：

多项式核函数：

$$K(z,z_i)=(1+z\cdot z_i)^d \tag{19}$$

高斯核函数：

$$K(z,z_i)=\exp(-\|z-z_i\|^2/\sigma^2) \tag{20}$$

如果直接采用核函数将不易求出相应的回归方程，因为因变量隐藏在核函数中。为了能将因变量分离出来，这里只对自变量x进行非线性映射得到$\phi(x)$，而因变量y不进行映射。高维空间中的点可以表示为$(\phi(x),y)$，此时高维空间中的两个平行的超平面表示为$y-\langle w,\phi(x)\rangle=\rho_1$和$y-\langle w,\phi(x)\rangle=\rho_2$，$\langle w,\phi(x_i)\rangle$仍采用核展开式$\sum_{j=1}^{l}\alpha_j k(x_j,x_i)$表示，于是优化问题变成：

$$\min -\rho_1+C\sum_{i=1}^{l}\xi_i+\rho_2+C\sum_{i=1}^{l}\xi_i' \tag{21}$$

约束为

$$y_j-\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\geqslant\rho_1-\xi_j,j=1,\cdots,l \tag{22}$$

$$y_j-\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\leqslant\rho_2+\xi_j',j=1,\cdots,l \tag{23}$$

$$\sum_{i=1}^{l}\alpha_i=1 \tag{24}$$

$$\xi_i,\xi_i'\geqslant 0,i=1,\cdots,l \tag{25}$$

于是回归方程(18)变为

$$y=\sum_{i=1}^{l}\alpha_i k(x_i,x)+\frac{\rho_1+\rho_2}{2} \tag{26}$$

本文给出的支持向量回归算法是根据一类分类思想得到的，因此记为OSC-SVR。

4 实验分析

实验1 拟合sin函数

取自变量$x\in[0,4\pi]$，按照间隔0.08取158个训练样本，按间隔0.33取39个不同的测试样本，因变量$y=\sin(x)$，并加入高斯白噪声$N(0,0.1)$，如图3所示。

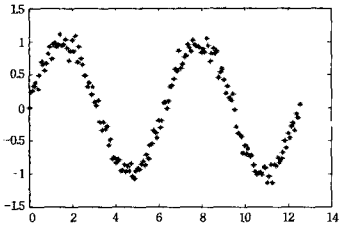

图 3　加入白噪声的正弦函数样本

采用 Matlab 语言编程实现，选取常用的几个支持向量回归算法进行比较分析，它们分别是标准支持向量回归算法(ε-SVR)、线性规划支持向量回归(LP-SVR)、最小二乘支持向量回归(LS-SVR)，其中核函数都取高斯函数。实验中需要选取惩罚参数 C 和高斯核函数中的参数 σ^2。网格搜索是参数寻优的一种最基本的算法。其思想是让待搜索的参数在一定的范围内按照指定的步长划分网格并遍历网格内的所有点进行取值，并将取出的参数带入支持向量回归算法中，使拟合误差(MSE)达到最小的即为最优参数。由于网格搜索算法实际上是一种全局搜索，因此该算法相对比较耗时。另外，我们在实验中发现以上几种算法对参数的选取都有一个敏感区域和不敏感区域。参数在敏感区域内变化时，算法的拟合误差变化明显，而参数在不敏感区域变化时，算法的拟合误差变化很小或者没有变化。因此，为了克服网格搜索算法的复杂性，我们将经验搜索和网格搜索相结合。经验搜索是在较大范围内试选几个参数，然后淘汰掉拟合误差较大的参数，保留拟合误差较小的参数。然后适当选取保留参数的邻近范围作为网格搜索区间，并选取适当的步长应用网格搜索算法进一步选取最优参数。不同的支持向量回归算法选取的搜索区间及步长如表 1 所列。ε-SVR 和 LP-SVR 模型中的参数 ε 同取 0.001，而 LS-SVR 和 OSC-SVR 模型中不含参数 ε。

表 1　不同回归算法的搜索区间及步长

算法	C		σ^2	
	区间	步长	区间	步长
ε-SVR	[50,200]	5	[0.001,0.1]	0.001
LP-SVR	[50,200]	5	[1,10]	0.5
LS-SVR	[50,200]	5	[1,10]	0.5
OSC-SVR	[0.01,0.1]	0.01	[1,10]	0.5

根据搜索算法所选取的每个回归算法的最优参数见表 2。本文所给算法的拟合结果见图 4，其中“—”表示实际值，“*”表示预测值。

表 2　不同回归算法的参数选取

算法	C	ε	σ^2
ε-SVR	100	0.001	0.01
LP-SVR	100	0.001	7.5
LS-SVR	100	—	7.5
OSC-SVR	0.03	—	7.5

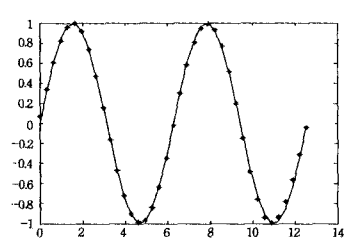

图 4　本文算法的拟合结果

测试指标采用均方误差 MSE，计算方法见表 5。4 个回归算法的比较结果见表 3，本文所给算法的泛化误差小于其它几个算法。

表 3　sin 函数的测试结果

分类方法	ε-SVR	LP-SVR	LS-SVR	OSC-SVR
MSE	0.0921	0.034	0.026	0.0258

实验 2　Henon 混沌时间序列预测

该序列由下式产生：

$$Y_t=1-1.4Y_{t-1}^2+0.3Y_{t-2},t\in Z \tag{27}$$

初始值取[0,0]，然后根据迭代表达式(27)取一维数据 288 个样本，前 280 个作为训练样本，后 8 个作为测试样本。为了验证算法的鲁棒性，在样本中加入高斯白噪声 $N(0,0.2)$。对该序列建立回归模型，需要根据时间延迟进行相空间重构，嵌入维数取为 2，时间延迟取为 1。另外，为了减少训练样本的数量以及提高训练样本的质量，每次预测都选择距离待预测点较近的一些样本作为训练集，选择标准为欧氏距离小于等于 $\delta(>0)$，本实验取 $\delta=0.3$。另外，根据搜索算法选取惩罚参数 C 和核函数参数 σ^2 时，我们发现针对这个数据几种不同回归算法所选取的参数几乎相同，并且在较小的范围内调整参数时几种算法的拟合误差不变。于是我们在不同的回归算法中选取相同的参数。选取的参数为：$\sigma^2=0.05$，$\varepsilon=0.0001$，$C=1000$。预测结果见表 4，本文所给算法的预测误差小于其它几种算法。

表 4　Henon 混沌时间序列的测试结果

分类方法	ε-SVR	LP-SVR	LS-SVR	OSC-SVR
MSE	0.7071	0.3975	0.3619	0.3551

实验 3　实际数据

采用 UCI 数据库中的 Diabetes 数据集，该数据包含 442 个数据，输入维数为 10，输出维数为 1。我们随机取其中的 419 个数据作为训练样本，23 个数据为测试样本。惩罚参数 C 和核函数参数 σ^2 的选取类似实验 2，几个算法的参数同取为：$\sigma^2=0.05$，$\varepsilon=0.001$，$C=100$。本实验采用的比较指标见表 5，预测结果见表 6。由于 ε-SVR 算法的运算时间较长，本实验并未给出其结果。从表 6 的结果可见，本文所给算法的预测误差与最小二乘支持向量回归算法达到同样的效果，而拟合误差是最优的。

表 5　测试指标及计算

度量指标	计算公式
SSE	$SSE=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$
SST	$SST=\sum_{i=1}^{m}(y_i-\bar{y})^2$
SSR	$SSR=\sum_{i=1}^{m}(\hat{y}_i-\bar{y})^2$
NMSE	$NMSE=SSE/SST$
R^2	$R^2=SSR/SST$
MSE	$MSE=\sqrt{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2/m}$

表 6　数据集 Diabetes 的回归结果

分类方法	SSE	NMSE	R^2	MSE
LP-SVR	130.1266	5.3874	4.4681	0.4166
LS-SVR	0.0041	1.7155e−004	0.9814	0.1993
OSC-SVR	4.7194e−023	1.9539e−024	1	0.1993

(下转第 243 页)

的拟合精度要高于文献[8]中所用挖掘方法的精度，且使用AR模型时会出现个别误差较大的情况。而使用灰色-周期外延模型时没有出现误差过大的情况，并且误差分布得比较平稳。再次验证了本方法具有较好的有效性和较高的预测精度。

结束语 本文提出了一种基于灰色-周期外延模型的动态关联规则元规则挖掘方法。通过一个实例说明本方法的一般过程：首先建立序列GM(1,1)模型，然后对残差序列建立周期外延模型，作为灰色GM(1,1)模型的残差补偿。通过比较得出此模型比直接用灰色模型预测要可靠、合理，总体效果要优于灰色GM(1,1)模型，能更准确地把握规则的变化趋势，从而使动态关联规则挖掘在合理的元规则指导下得到更精确的结果。今后的主要研究方向是将灰色-周期外延模型应用到动态关联规则趋势度的挖掘中[15]；当原始数列波动性过大且无明显规律时如何保证预测的精度；对象为大数据库或者海量数据时，如何提高挖掘算法的效率和准确性[16]。

参 考 文 献

[1] Agrawal R, Imeielinksi T, Swami A. Mining Assocaiation Rules Between Sets of Items in Large Databases [C]// Proceedings of the 1993 ACM SIGMOD Conference Washington DC. Washington, 1993: 207-216

[2] Au W-H, Chan K C C. Mining changes in association rules: a fuzzy approach[J]. Fuzzy Sets Syst, 2005, 149(1): 87-104

[3] 荣冈，刘进锋，顾海杰. 数据库中动态关联规则的挖掘[J]. 控制理论与应用，2007，24(1)：129-133

[4] 沈斌，姚敏. 一种新的动态关联规则及其挖掘算法[J]. 控制与决策，2009，24(9)：1310-1315

[5] Shen Bin, Yao Min, Wu Zhao-hui, et al. Mining dynamic association rules with comments [J]. Knowledge and Information Systems, 2010, 23(1): 73-98

[6] 刘俊，谢彦峰，张忠林. 基于灰色Markov模型动态关联规则元规则挖掘[J]. 计算机应用，2008，28(9)：2353-2356

[7] 张忠林，许凡. 基于小波变换的动态关联规则元规则GM(1,1)挖掘[J]. 计算机科学，2013，40(5)：209-212，246

[8] 张忠林，刘俊，谢彦峰. AR-Markov模型在动态关联规则挖掘中的应用[J]. 计算机工程与应用，2010(14)：135-137，147

[9] 任峰，李伟，丁超. 基于灰色-周期外延组合模型的电力负荷预测[J]. 电网技术，2007，24：52-54

[10] 杨俊祥，程盛芳. 灰色-周期外延组合模型在煤炭需求预测中的应用[J]. 统计与决策，2010，13：162-163

[11] 刘思峰，党耀国，方志耕，等. 灰色系统理论及其应用[M]. 北京：科学出版社，2004，142-146

[12] Zhang Yi, Wei Yong, Zhou Ping. Improved Approach of Gray Derivative in GM (1,1) Model [J]. The Journal of Grey System, 2006, 116(10): 160-162

[13] Sun Yan-na. Optimization of Grey Derivative in GM(1,1) Based on the Discrete Exponential Sequence, 2009 [C]// Proceeding of the 2nd International Symposium on Information Processing(ISTP2009). Huangshan, P. R. China, 2009: 313-315

[14] 刘思峰，邓聚龙. GM(1,1)模型的适用范围[J]. 系统工程理论及实践，2000，20(5)：121-124

[15] 张忠林，曾庆飞，许凡. 动态关联规则的趋势度挖掘方法[J]. 计算机应用，2012，32 (1)：196-198

[16] Angel K-M, Fatima R-E. A search space reduction methodology for data mining in large databases[J]. Engineering Applications of Artificial Intelligence, 2009, 22(1): 57-65

(上接第 232 页)

从上面的几组实验中可以看到，所给算法对非线性回归问题具有较好的拟合和预测能力。

结束语 分类和回归技术在实际中具有广泛的应用领域，比如医学中的疾病诊断就是一种模式分类问题，而根据历史的经济数据建立回归模型来预测未来的经济趋势是统计工作者经常应用的方法。本文根据一类分类思想和核方法提出一种新的非线性支持向量回归算法，该算法归结为求解一个线性规划，因此其运算速度要明显优于基于二次规划的标准支持向量回归算法。另外，该算法还展示了一类分类和回归之间的关系，从而拓宽了建立支持向量回归的途径，算法有助于直观地理解支持向量回归的几何意义。在两个人工数据和一个实际数据上的实验结果显示了所给算法具有良好的泛化能力。

参 考 文 献

[1] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer-Verlag, 1995

[2] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Knowledge Discovery and Data Mining, 1998, 2 (2): 34-38

[3] Smola A J, Scholkopf B. A tutorial on support vector regression [R]. NeuroCOLT TR NC-TR 98-030. Royal Holloway College University of London, UK, 1998

[4] Suykens J A K, Vandewalle J. Least squares support vector machines classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300

[5] 陶卿，孙德敏，范劲松，等. 基于闭凸包收缩的最大边缘线性分类器[J]. 软件学报，2002，13(3)：404-409

[6] 陶卿，曹进得，孙德敏. 基于支持向量机分类的回归算法[J]. 软件学报，2002，13(5)：1024-1028

[7] Scholkopf B, Williamson R C, Smola A J, et al. Support vector method for novelty detection[J]. Advances in Neural Information Processing Systems, 2000, 12(3): 582-588

[8] Scholkopf B, Platt J C, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471

[9] 徐磊，赵光宙. 基于模糊一类支持向量机的核聚类算法[J]. 控制与决策，2008，23(9)：1030-1034

[10] 田江，顾宏. 孤立点一类支持向量机算法研究[J]. 电子与信息学报，2010，32(6)：1284-1288

[11] 张庆，徐光华，华成，等. 一类支持向量机的设备状态自适应报警方法[J]. 西安交通大学学报：自然科学版，2009，11：61-65

[12] 陈伟，余旭初，张鹏强，等. 基于一类支持向量机的高光谱影像地物识别[J]. 计算机应用，2011，31(8)：2092-2096

[13] Mangasarian O L. Arbitrary-norm separating plane[J]. Operations Research Letters, 1999, 1(24): 15-23