基于支持向量机回归的曲线拟合

西安工业大学计算机科学与工程学院　娄小燕　刘白林

[摘　要] 在科学实验研究中，经常需要实验的观测数据，来寻求两个物理量之间近似的解析函数关系和曲线方程，这就是人们常说的数据拟合或曲线拟合，而且经常要从这些已知数据中总结规律，用以预报未知。本文引入支持向量机作为背景进行曲线拟合。此法能满足在小样本情况研究统计学习规律的理论，通过引入结构风险最小化准则来控制学习机器的容量，从而刻画了过度拟合与泛化能力之间的关系。

[关键词] 支持向量机　曲线拟合

1.引言

在西气东输中，必须对投入使用的钢管性能进行各项试验研究，一旦某个环节出现问题，后果都不敢想象，所以在对这些钢管性能进行测试的时候，需要测试它各个方面的因素，得出一些测试数据，通过所得数据对其进行安全检测，一个有效地方法就是描述试验数据的拟合曲线，对曲线的拟合得出它们的规律，并进行预测研究。

科学实验中常用的数据处理算法有以下几种，当数据的规律接近线性时，用线性回归总结规律，通常认为是标准的，最可靠的方法。如果规律偏离线性，则通常用人工神经网络总结规律，或在线性方程中添加平方或者其他高阶项作非线性回归。除传统的线性回归外，人工神经网络和各种模式识别技术都在广泛使用，并已取得许多成果。但是传统的方法、神经网络经过人们的大量应用呈现了其不如人意之处，其大样本、泛化能力不强、过拟合等缺点。

实践中，人们经常发现用上述各种方法总结的数学模型对已知数据（即所谓的训练集）常能拟合较好，而在预报未知样本时，偏差往往较多，在小样本问题中此问题尤其严重。在数学上将这种现象称为数学模型的“推广能力”不足的问题，即泛化能力差。如何提高算法和数学模型的推广能力，以确保我们预报结果的可靠性，显然是科学试验数据处理中非常重要的课题。这其实就是如何避免“过拟合”和“欠拟合”现象的问题。

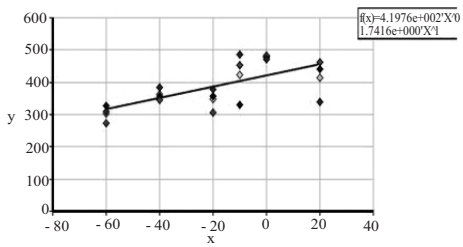

图 1 欠拟合

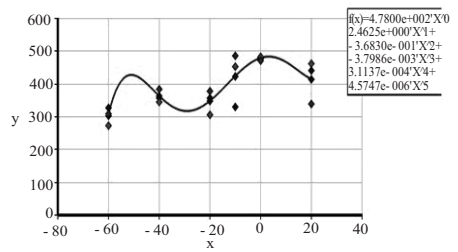

图 2 过拟合

比如以西气东输二线用管材数据为依据，温度对冲击试验数据的影响和实测的冲击功的真实值，采用线性拟合的回归曲线，有些时候如果实验中某些影响实验的因素没有考虑到也会引起欠拟合。如图 1 所示，很容易看到拟合的偏差很大，这样的欠拟合在试验中经常出现，如果采用二次拟合或者更高次的拟合会更好，但是如若采用再高次的拟合很容易出现过拟合现象如图 2 所示，人们就很难从中总结规律了，对于“欠拟合”，传统的拟合方法通常是在线性方程后面加高阶项。此法诚然有效，但由此增加的可调参数未免增加了过拟合的风险。或者采用人工神经网络拟合，因为理论上神经网络可以逼近任何连续函数，如果隐层包含足够多的神经元，它可以逼近任何只有断点的非连续函数，但是正是因为神经网络对训练数据拟合的程度的提高，它的推广能力是不断下降的，神经网络会出现过拟合的情况，因此要合理的选择神经网络的隐藏节点数，在获得高训练精度的同时，也能有好的推广能力，神经网络有能力拟合任何函数，但是由于神经网络中自动产生网络权值，这就导致每次训练结果不同，可能为局部最小值。

在通常的实验中真实值和我们测量出的实验数据往往是有误差的，这就导致计算出的结果有误差，得到的拟合方程已经不是一个精确的方程。在追求高的拟合精度的同时，正是因为有这些因素的存在，才能为一些过拟合的存在提供了条件。针对上述问题，20 世纪 90 年代中期，Vapnik 创建了基于统计学习的一种机器学习算法——支持向量机。

2.支持向量机原理及算法

在科学实验中，针对现有存在的问题，我们不应该像传统的方法那样用拟合求唯一的“精确解”，而应当在承认误差和欠拟合造成的影响的情况下，找寻一种具有过拟合最小，推广能力最强的解的数学模型，因此，如何从小样本集出发，得到预报（推广）能力较好的模型，遂成为模式识别研究领域内的一个难点，即所谓“小样本难题”。数学家 Vladimir N. Vapnik 等通过三十余年的严格的数学理论研究，提出来的统计学习理论（statistical learning theory，简称 SLT）和支持向量机（support vector machine，简称 SVM）算法已得到国际数据挖掘学术界的重视，并在语音识别、文字识别、药物设计、组合化学、时间序列预测等研究领域得到成功应用，这正是统计学系理论，它是在研究小样本的情况下的机器学习规律的理论，在这体系下的统计推理规则不仅考虑了对渐进性能的要求，而且追求在现有有线条件下得到最优结果。统计学系理论被认为是目前针对小样本统计估计和预测学习的最佳理论。它从理论上较为系统地研究了经验风险最小化原则成了的条件，有限样本条件下经验风险与期望风险的关系以及如何利用这些理论找到新的学习原则和方法等问题。

2.1 支持向量机

支持向量机方法(SVM)就是从解决线性可分情况的最优分类面出发的，其思想就是选取使“间隔”达到最大的那个法方向，相应得到的两条极端直线就是最优分类线，所谓最优分类线是能将两类点正确分开的分划线(训练错误率为零)。使分类间隔最大实际上就是对推广能力的控制，这是 SVM 的核心思想之一。最小化训练误差和最大化泛化能力(推广能力)就是体现了支持向量机最小化结构风险的思想。进一步，对于选定的法方向 w，会有两条极端的直线，选取 b 使得要找的直线为两条极端直线“中间”的那条直线。

2.2 核函数

对于 N 维空间中的线性函数，计算的复杂度不是由空间维数决定的，而是由样本数来决定的，支持向量机方法巧妙的避开了高维空间的计算，并不是相似地进行变换计算，而是做训练样本之间的内积运算，这种内积运算由事先定义的核函数来实现，将线性空间中的非线性问题变为非线性空间中的线性问题，从而根本上解决非线性问题。用核函数代替线性方程中的线性项可以使原来的线性算法“非线性化”，即能作非线性回归。与此同时，引进核函数达到了“升维”的目的，而增加的可调参数却很少，于是过拟合仍能控制。并且核函数引入后，使得回归问题的求解绕过特征空间，直接在输入空间上求取，从而避免了计算非线性映射。核函数 K(x,x')是对称正实数函数，同时满足 Mercer 条件：

$$\iint K(x,x')g(x)g(x')dxdx' \geq 0 \quad g \in L_2$$

常用的核函数

(1) 多项式核函数

$K(x,x')=((x\cdot x')+c)^d \quad c \geq 0, d \in N$

(2) 高斯核函数(RBF)

$K(x,x')=\exp(-\frac{\| x-x' \|^2}{2\sigma^2})$

(3) B 样条核函数

$K(x,x')=B_{2N+1}(\| x-x' \|)$

(4) sigmoid 核函数

$K(x,x')=\tanh(k(x\cdot x')+v) \quad k>0, v>0$

2.3 最小二乘支持向量机

最小二乘支持向量机(简称 LS - SVM)是 SuyKensJ. A. K 在 1999 年提出的一种新型支持向量机算法。是支持向量机引入了最小二乘的思想,LS - SVM 在目标函数中采用平方和误差损失函数替代传统 SVM 中的不敏感损失函数,用等式型约束替代传统 SVM 中的不等式约束,这样使得 SVM 求解二次规划问题转化为求解一组线性关系式, LS- SVM 的优化函数只需解线性等式方程组,计算量较小,最重要的是避免了 SVM 中惩罚因子(常数) C 值的选择问题。从而极大地简化了问题,提高了学习速率。

3.实验研究

支持向量机应用于管材试验的曲线拟合和预测中的最大优点在于可以方便而全面的考虑对管材性能分析中有重要影响因素(如断口剪切面积,冲击功等),而不需要对输入的数据做相关的预测,体现在实际中就是可以方便地将这些影响因素作为输入变量与某输出变量之间的映射关系。本文采用最小二乘支持向量机算法实现曲线拟合。

3.1 特征归一化

在使用 SVM 分类方法之前,对提取的特征进行归一化非常重要。主要的优点是在建立分类超平面时,避免动态范围大的特征淹没了动态范围小的特征,使它们具有同等的作用。另外一个优点是,在特征向量的内积计算时避免大数计算的困难,大的特征值可能引起计算的溢出。因此,常常需要对特征进行归一化处理。经过归一化处理后,特征的范围限制在[- 1,1]之间。在进行特征归一化时,需要对训练集与测试集样本的特征采用同样的方式进行归一化。本文利用分类器对产生的数据进行归一化。

3.2 SVM 核函数和参数的选择

本实验利用 LS - SVMlab 进行评价,采用性能比较好的径向基核函数,因为它有以下特点:

⑴ 表示形式简单,即使对于多变量输入页不增加太多的复杂性;

⑵ 径向对称;

⑶ 光滑性好,任意阶导数均存在;

⑷ 由于核函数表示简单且解性好,因此便于进行理论分析。

LS- SVM 主要有两个参数:一个是正则化参数 γ;另一个是径向基核函数参数 σ, 这两个参数在很大程度上决定了该方法的模式识别能力.γ 参数是权衡拟合曲线的光滑度和拟合误差,γ 越大拟合越好,γ 减小模型复杂度降低;σ 增大可以使得拟合曲线更光滑。经过参数优化处理,将选取参数 σ=0.24847 ,γ=2.6984 。

3.3 SVM 方法的训练及评价结果

LS- SVM 的函数逼近性能

给定训练数据,开始对 LS- SVM 进行训练,经过 0.2030s 的训练时间,得到了如图 3 所示的逼近性能及其误差曲线,整个测量过程中的函数逼近误差均能控制到较高的水平,即均方误差(Mean Square Error, MSE)为 8.4680e- 007,表明 LS- SVM 具有较好的函数逼近能力(在回归算法中通常采用拟合的均方误差(mean square error, MSE)来作为性能指标)。在该图中,描述了任意两组冲击功数据和落锤数据之间的曲线拟合,根据这样的曲线,可以从中发现一些想要的规律。

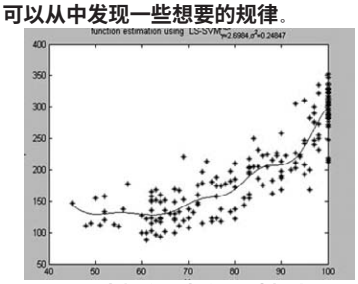

图 3 支持向量机的曲线拟合

LS- SVM 的泛化性能

根据 LS- SVM 的拟合曲线,比较真实值和预测值的泛化性能仿真分析结果可以看出,虽然比上述逼近性能的误差大,但仍表现出优良的性能。同样的,除了在举升行程的起始和终了阶段受系统固有特性的影响引发的误差较大外,总体误差指标(MSE 为 6.8638e- 006)已能很好的符合误差要求的范围。

4.结语

本文首先提出了传统的曲线拟合的最小乘、神经网络等各种方法的局限性,但是 SVM 方法用于曲线拟合比较有优势, 其训练的样本量少,先验干预比较少,学习性很强,简单易用,模型经过样本训练,确定支持向量,数据只要归一化后,直接输入模型即可。最后,结合最小乘支持向量回归算法,把该算法应用于管材数据的数据拟合和预测中,通过仿真试验验证,取得了比较满意的结果。

参考文献

[1] 侯振雨,蔡文生,邵学广.主成分分析 - 支持向量回归建模方法及应用研究[J].分析化学,2006.

[2] 邓乃扬,田英杰.数据挖掘中的新方法—支持向量机[M].北京:科学出版社,2004.

[3] 求是科技.MATLAB7.0 从入门到精通[M].北京:人民邮电出版社,2006.

[4] Vapnik V N.The nature of statistical learning theory [A].Jordan M, Lauritzen S L,Lawless J F,et al. Statistics for Engineering an Information Science [C].New York:Springer,1999.

(上接第 439 页) 波达方向向量,这是 MUSIC 算法所不能比拟的。

参考文献

[1] 张贤达,保铮.通信信号处理[M].北京:国防工业出版社,2002, 320- 326

[2] 张贤达.现代信号处理(第二版)[M].北京:清华大学出版社, 2002,126- 146

[3] 邵玉斌.MATLAB/ SIMULINK 通信系统建模与仿真实例分析[M].北京:清华大学出版社,2008,221- 330

[4] 王世一.数字信号处理(修订版)[M].北京:北京理工大学出版社,2006275- 314

[5] 王立宁.MATLAB 与通信仿真[M].北京:人民邮电出版社,2000, 50- 98

[6] 朱德文,张崇庆,孙建伟.信号与系统(第二版)[M].北京:北京理工大学出版社,2000,175- 266

[7] Serpedin E, Giannakis G B. Blind channel identifications and equalizations with modulation- induced cyclostationarity [C].IEEE Trans. Signal Processing,1998, 46: 1930- 1944

[8] Roy R, Paulraj A, Kailath T. ESPRIT—A subspace rotation approach to estimation of parameter of cissoids in noise [C].IRRR Trans.Acoust., Speech, Signal Processing, 1989, 37: 984- 995

[9] 肖先赐.现代谱估计——原理与应用[M].哈尔滨:哈尔滨工业大学出版社,1991,96- 143

[10] 王永良,陈辉,彭应宁,万群.空间谱估计算法与原理[M].北京:清华大学出版社,2004,83- 106

[11] 司伟建.MUSIC 算法多值模糊问题研究[J].系统工程与电子技术,2004,Vol.26:960- 963

[12] 张志涌.精通 MATLAB6.5 [M].北京:北京航空航天大学出版社,2003

(上接第 440 页) 分析,结果表明:粮食和棉花的产量 Hurst 指数 H 值都介于 1/2 到 1 之间,相关函数 C(t)大于零,存在明显的赫斯特规律,即粮食和棉花的产量发展呈现显著持续规律性,从而得出未来我国农产品粮食和棉花产量将继续持续增加的趋势。根据时间序列分维数与 Hurst 指数 H 值的关系计算出 44 年来粮食和棉花产量的分维 D 的值很接近 1,说明了粮食和棉花的产量随时间发展具有分形特征,呈现很好的规律性。

参考文献

[1] 孙霞等.分形原理及应用[M].中国科学技术大学出版社,2003.

[2] Peters E E.Chaos and Order in the Capital Markets [M].New York: John WielySons Inc,1991.

[3] Mandelbrot,BB,etal Fractional Brow nian motion,fractional nosie and application [J].SIAM Rev,1968.

[4] 黄登仕,李后强.分形几何学 R/S 分析与分式布朗运动[J].自然杂志,1990,13(8):477- 482.

[5] 国家统计局.《新中国五十五年统计资料汇编》[M],中国统计出版社,2005.