文章编号: 1006-4710(2012)01-0098-04

一种基于 PSO-SVR 的软件可靠性评估方法

何增郎, 张毅坤, 杨凯峰, 张保卫
(西安理工大学 计算机科学与工程学院,陕西 西安 710048)

摘要: 软件可靠性建模时,如果简单地利用支持向量回归机制建模,就有可能由于支持向量回归(SVR)自身参数选择难以及实验数据本身的不确定性,从而导致预测结果不理想、精度低等缺陷。因此,借鉴粒子群优化算法(PSO)多参数寻优的优势,将 PSO 与 SVR 优化算法相结合,利用分层聚类算法对初始实验数据进行归一化处理,剔除异常数据,构建基于 PSO-SVR 的软件可靠性评估方法,从而提高软件模型的预测精度。实验结果表明,基于 PSO-SVR 方法的预测模型其预测精度高,更适应实际软件应用环境。

关键词: 软件可靠性评估模型; 向量回归; 粒子群优化
中图分类号: TP311 **文献标志码**: A

A Kind of Software Reliability Assessment Method Based on PSO-SVR

HE Zenglang, ZHANG Yikun, YANG Kaifeng, ZHANG Baowei
(Faculty of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: In the process of software reliability modeling, if the support vector regression mechanism is simply used, it is likely that the difficulty in parameter selection and the uncertainty in experimental data themselves may lead to the realistic predicted results and low accuracy. For this reason, it is necessary to reference the multi-parameter optimization advantages of particle swarm algorithm to combine PSO and SVR optimization algorithm, to normalize the initial experimental data with hierarchical algorithms, and to reject the abnormal data and construct a software reliability assessment method based on PSO-SVR. Experimental results show that the prediction model based on PSO-SVR has a high precise and a suitable environment for software application.

Key words: software reliability assessment model; vector regression; particle swarm optimization

软件可靠性是衡量软件产品的一个重要指标,对于软件可靠性的研究是评估软件性能、控制软件开发过程、提高软件产品质量的基础,软件可靠性模型在这个过程中起到的作用是至关重要的。通过合理预测可以得出一些积极的软件可靠性指标,很好地进行相关的预测控制活动。目前常用的软件可靠性模型有概率类模型(如 Jelinski-Moranda 模型、Goel-Okumoto 模型)、Bayesian 模型(以 L-V 模型为代表),以及人工神经网络模型等[1-2]。但是,概率类模型常常存在假设条件难以满足的弊端,Bayesian 模型存在先验知识获取困难的问题;人工神经网络模型对建模数据要求较高,而又存在结构难以确定、学习不足、拟合不足或过拟合以及容易陷入局部极小点等问题,模型本身也过于复杂,训练时间较长。

针对上述不足,本文提出一种基于粒子群组合寻优的支持向量回归的软件可靠性评估方法,该方法利用 PSO 组合寻优的优势,对 SVR 参数进行组合寻优,寻找最优参数组合。再结合分层聚类算法对原始数据进行预处理,剔除异常数据,使数据在时间域上分布更均匀[3];利用反馈校正方法对预测值进行校正,以得到更加准确的预测值,从而获得较好的评估结果。

收稿日期: 2011-10-12
基金项目: 陕西省自然科学基金资助项目(2009JM8003-1);陕西省教育厅专项基金资助项目(09JK679)。
作者简介: 何增郎(1983-),男,陕西蓝田人,硕士生,研究方向为软件工程。E-mail: he_zl@qq.com。
张毅坤(1958-),男,陕西西安人,教授,研究方向为软件工程。E-mail:ykzhang163@163.com。

1 PSO-SVR 软件可靠性评估方法

1.1 SVR 参数选择

支持向量回归寻求的是一个估计指示函数，用该函数对测试样本进行类别判别。将问题由寻求指示函数估计推广到寻求实值函数估计后，就可以得到用于函数估计（回归）的支持向量机（SVM）。SVM 有效地解决了小样本、高维数、非线性等问题。然而，作为一种新的学习机器，也存在一些有待完善的地方，其参数（误差 ε、误差惩罚因子 C 和核函数参数 y）选取便是亟待完善的问题之一[4-5]。

误差 ε 控制函数拟合误差的大小，从而控制支持向量的个数和泛化能力，它反映模型对输入变量所含噪声的敏感程度。ε 选择得小，则回归估计精度高，但支持向量数量会增多。

误差惩罚因子 C 是在确定的数据子空间中调节学习机器置信范围和经验风险的比例，以使学习机器具有最好的推广能力。最优的 C 根据不同的数据子空间而不同。在确定的数据子空间中，C 的取值小表示对经验误差的惩罚小，学习机器的复杂度小而经验风险值较大；反之亦然，当 C 超过一定值时，SVM 的复杂度达到了样本空间允许的最大值。此时经验风险和推广能力几乎不再变化。每个数据子空间至少存在一个最优的 C，使 SVM 的推广能力达到最好。

核函数参数 y 影响样本数据在高维特征空间中分布的复杂程度。核函数参数的改变实际上是隐含地改变映射函数，从而改变样本空间的维数。样本空间的维数决定了能在此空间构造的线性分类面的最大 VC 维[6]，也就决定了线性分类面能达到的最小经验误差。

SVR 参数寻优需要一种组合寻优算法，能对误差 ε、误差惩罚因子 C 和核函数参数 y 进行组合寻优，而且效率、泛化能力方面以及随着问题规模及复杂程度增加时优化算法的增长趋势等方面具有很好的优势。通过对 PSO 优化算法与其他优化算法的比较分析，笔者发现 PSO 在组合寻优等多方面具有很大的优势，符合 SVR 参数选取的条件。

1.2 SVR 参数的 PSO 寻优

由于 SVR 回归模型的泛化性能和预测精度依赖于误差 ε、误差惩罚因子 C 和核函数参数 y 等 3 个参数，不同参数的组合对最终评估效果有很大的影响。因此，对 (ε, C, y) 参数进行寻优优化显得十分关键。本研究采用改进的 PSO 算法来寻找参数 ε、C、y 的最优组合，即采用速度 - 位置搜索模型来寻找最优参数组合。

设群体中的每个粒子由 3 维参数向量 (ε, C, y) 组成，第 i 个粒子在 3 维解空间的位置为 $x_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$，其速度为 $v_i = (v_{i1}, v_{i2}, v_{i3})^{\mathrm{T}}$，当前时刻的个体极值记为 $p_i = (p_{i1}, p_{i2}, p_{i3})$，全局极值记为 $g_i = (g_{i1}, g_{i2}, g_{i3})$。在每次迭代中，粒子跟踪个体极值、全局极值和自己前一时刻的状态来调整当前时刻的位置和速度，迭代公式如下：

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 rand_1()(p_{id} - x_{id}(t)) c_2 rand_2()(g_{id} - x_{id}(t)) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

式中，$rand()$ 是 $[0,1]$ 之间的随机数；d 是解空间的维数；c_1 和 c_2 是学习因子，通常取为 2；ω 是权重因子。

为了确保其收敛性，避免基本粒子群优化算法在解空间内搜索时，粒子在全局最优解附近“振荡”的现象，借鉴文献[7]的改进方法，随着叠代进行，速度更新公式中的权重因子 ω 由最大权重因子 $\omega_{\max}$ 线性减小到最小权重因子 $\omega_{\min}$，即：

$$\omega = \omega_{\max} - N \cdot \frac{\omega_{\max} - \omega_{\min}}{N_{\max}} \quad (3)$$

式中，N 为当前迭代次数，$N_{\max}$ 为总的迭代次数。

1.3 输入输出校正

由于实际收集到的失效数据存在误差，以及环境等因素干扰，获得的预测值有可能偏离实际值，如果不及时处理，进一步的优化就可能建立在虚假数据的基础上，因此必须对实际收集到的数据进行预处理，并建立反馈机制加以事后校正。本文利用分层聚类算法对初始数据进行归一化处理以剔除异常点[3]。

本文采用的校正方法为：

$$y_p(k+1) = \hat{y}(k+1) + he(k) \quad (4)$$

式中，h 为补偿系数（即误差修正系数），根据实际应用的效果进行调整；$e(k)$ 为 k 时刻实际输出与模型预测之间的误差，即：

$$e(k) = y(k) - \hat{y}(k) \quad (5)$$

式中，$y(k)$ 为实际输出值，$\hat{y}(k)$ 为预测值。

2 PSO-SVR 算法实现

PSO-SVR 算法的实现通常采用以下步骤。

1）收集整理软件失效实验数据；

2）对收集到的失效数据进行预处理（利用分层聚类算法对初始数据进行归一化处理以剔除异常点[3]）；

3）对实验数据进行分组（训练组数据和对照组

数据),并利用粒子群算法对 SVR 的 ε、C、y 参数寻找最佳组合参数;

4) 将获得的 ε、C、y 的最佳组合参数作为模型参数;

5) 确定输出函数:

$$f(x) = \omega \cdot \phi(x) + b = \sum_{i=1}^{m} (a_i - a_i^*) k(x, x_i) + b \quad (6)$$

式中,$f(x)$ 为拟合函数;ω 为 x 的权重系数向量;$k(x, x_i) = \Phi(x) \cdot \Phi(x_i)$ 为核函数,选择不同形式的核函数就可以生成不同的 SVR 回归模型,采用径向基函数来建立SVR模型;b 是阈值,a_i、a_i^* 通常只有一部分不为 0,被称为支持向量。该输出函数的目标是寻找合适的 ω 和 b,使得用 $f(x_i)$ 估计 y_i 时的估计误差最小,即回归风险最小。

6) 采用上一次的预测偏差修正当前预测值;

7) PSO-SVR 模型的预测结果分析。

3 实验分析与对比

采用度量元(SRE) 指标衡量模型的预测能力,从模型预测值和实际观测值计算 SRE,判断 PSO-SVR 模型的预测精度。

$$SRE = \frac{\sum_{i=1}^{n-1} \frac{| x_r(i+1) - x_p(i+1) |}{x_r(i+1)}}{n-1} \quad (7)$$

式中,$x_r(i+1)$ 表示实际观察到的下一失效间隔时间,$x_p(i+1)$ 表示应用前 i 个故障建立的模型预测的下一失效间隔时间。

SRE 的值越小,说明预测值与实际值的偏差越小,模型的短期预测能力也就越强。

选取在可靠性分析中应用比较广泛的概率类模型(J-M 和 G-O 模型)和经典贝叶斯模型(L-V 模型)与新模型进行比较。对美国海军舰队计算机程序设计中心(U. S. Navy Fleet Computer Programming Center)的海军战术数据系统 NTDS(Naval Tactical Data System)开发过程(前 26 组数据)、测试过程(27 ~ 31 组数据)中的错误统计数据和来自 Musa 公开发表的、从实际工程项目中收集来的 SYSl、SYS2、SYS3 数据进行分析对比。

3.1 NTDS 系列数据实验

将数据集中的前 26 个样本作为训练组,后 5 个数据作为对照组。

利用 PSO-SVR 算法的参数寻优过程和 PSO-SVR 的软件可靠性评估算法的实施步骤,可以得到基于 PSO-SVR 模型的 NTDS 系统的最优参数组合为(0.218,136.38,2.168)。分别计算各个模型的 SRE,对于 NTDS 中的数据,得到短期预测能力 SRE 的值,见表 1。

表 1 NTDS 数据预测结果

Tab. 1 Prediction results of NTDS data

编号	模型			
	PSO-SVR	J-M	G-O	L-V
SRE	1.15(1)	2.488(4)	2.189(3)	1.389(2)

将基于 PSO-SVR 模型预测结果与现有的 J-M 模型、G-O 模型、L-V 模型的预测值,以及对照组数据进行比较,如图 1 所示。

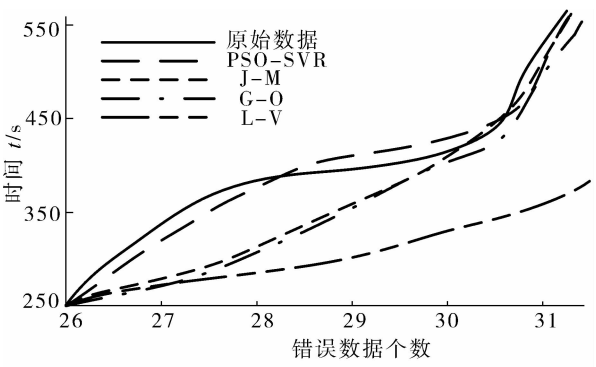

图 1 各个模型预测值与原始数据的比较

Fig. 1 The contrast diagram of each model's forecasts data and original data

从表 1 可看出,基于 PSO-SVR 模型的度量元(SRE)排在最前面,说明 PSO-SVR 新模型的短期预测能力优于其它几个模型。从图 1 可看出,PSO-SVR 模型预测曲线和实际值的接近程度优于其它模型。从表 1 和图 1 可以看出 PSO-SVR 新模型对于这一类软件可靠性预测能力是比较强的。

3.2 SYS 系列数据实验[8]

对模型的通用性进行验证,防止模型对数据的部分有效性和部分依赖性,分别将 SYS1、SYS2、SYS3 每组数据的前 85%、80%、75% 作为训练组,后 15%、20%、25% 作为对照组,对应的组分别为第一组、第二组、第三组。得到的基于 PSO-SVR 算法的新模型的 SYS1、SYS2、SYS3 系列数据的最优参数组合如表 2 所示,根据模型数据计算的各个组对应的各个模型的 SRE 值和预测能力强弱排序见表 3。

从表 3 可以看出,基于 PSO-SVR 算法的新模型的预测能力在大部分数据上都是最优的。与其他模型相比,基于 PSO-SVR 算法的新模型除了在 SYS2 最后 20% 的数据上略低于 L-V 模型外,在其他组数据上预测精度均最优,说明总体上 PSO-SVR 新模型

预测能力是非常好的。另外，通过 SRE 值的比较可知，其他模型在不同数据上的排名不稳定，反映出其它模型的假设条件与实际数据仍有一定偏差。当实际收集到的数据分布与其假设比较符合时，该模型的预测精度就好，反之则差。

表 2 SYS 系列数据参数最优组合

Tab. 2 Parameters optimal combination of SYS series data

组	数据	ε	C	γ
第一组	SYS1	0.226	185.382	2.236
	SYS2	0.336	208.774	2.018
	SYS3	0.276	228.337	1.785
第二组	SYS1	0.234	208.339	2.376
	SYS2	0.267	179.565	1.758
	SYS3	0.358	217.624	2.058
第三组	SYS1	0.197	199.354	1.765
	SYS2	0.301	257.063	2.368
	SYS3	0.286	278.228	1.990

表 3 SYS 系列数据预测结果

Tab. 3 Prediction results of SYS series data

组	数据	PSO-SVR 模型	J-M 模型	G-O 模型	L-V 模型
第一组	SYS1	10.138(1)	13.122(3)	14.586(4)	12.835(2)
	SYS2	1.126(1)	1.365(3)	1.653(4)	1.241(2)
	SYS3	3.336(1)	5.0766(4)	4.876(2)	4.993(3)
第二组	SYS1	8.966(1)	10.276(3)	12.348(4)	10.267(2)
	SYS2	1.163(2)	1.224(3)	1.337(4)	1.109(1)
	SYS3	3.996(1)	4.206(4)	4.0987(2)	4.194(3)
第三组	SYS1	19.253(1)	29.992(2)	35.681(3)	39.887(4)
	SYS2	1.213(1)	1.349(3)	1.632(4)	1.333(2)
	SYS3	3.996(1)	4.502(3)	4.408(2)	4.585(4)

从以上实验结果可以看出，基于 PSO-SVR 方法的新模型具有一定的通用性。即在不需要有假设的前提下，通过 SVR 学习机制和 PSO 寻优优化参数，动态调整模型参数，使得预测结果更加精确。这样就避免了假设条件与实际不符合、模型的静态性以及模型的片面性等问题。

4 结 论

本文将支持向量回归理论引入到软件可靠性建模中，并结合粒子群寻优优化算法的特性，提出了一种基于粒子群算法（PSO）寻优的支持向量回归（SVR）的软件可靠性评估新模型。实验结果和实验数据分析结果表明，该方法在性能上具有比较好的效果，具有一定的通用性。同时，该方法原理简单，易于实现，避免了假设条件过多与实际存在的偏差，具有广泛的应用前景。

参考文献：

[1] 王小丽，徐中伟，杜军威. 改进的 J-M 模型及其在软件安全性评估中的应用[J]. 小型微型计算机系统，2008，29(2)：269-273.
Wang Xiaoli, Xu Zhongwei, Du Junwei. Improved J-M model and application to software safety assessment[J]. Journal of Chinese Computer Systems, 2008, 29(2): 269-273.

[2] 马兴元，郭建英，赵海秋. 软件可靠性增长 G-O 模型的改进[J]. 传感器与微系统，2011，30(5)：69-71.
Ma Xingyuan, Guo Jianying, Zhao Haiqiu. Improvement of software reliability growth G-O model[J]. Transducer and Microsystem Technologies, 2011, 30(5): 69-71.

[3] 许宁. 基于正交分层聚类算法的软件可靠性预测分析[J]，计算机应用，2007，27(3)：635-637.
Xu Ning. Research on reliability prediction model based on orthogonal layer—clustering algorithm[J]. Journal of Computer Applications, 2007, 27(3): 635-637.

[4] Chapelle O, Vapnik V. Choosing multiple paramenters for support vector machines[J]. AT & T Research Labs, 2002, 26(3): 1-3.

[5] Keerthi S S, Lin C J. Asympotic behaviors of support vector machines with Gaussian kernel[J]. Neural Computation, 2003, 15: 1667-1689.

[6] Shi Y, Eberhart R C. A modified swarm optimizer: IEEE International of Evolutionary Computation[C]. Anchorage, Alaska, 1998.

[7] Kennedy J, Eberhart R. Particle swarm optimization: proceedings of IEEE Conference on Neural Networks[C]. Perth, Australia, 1995.

[8] Phillips B T, Kidd A R, King R, et al. Reciprocal asymmetry of SYS-1/Beta-Catenin and POP-1/TCF controls asymmetric divisions in caenorhabditis elegans: proceedings of the National Academy of Sciences of the United States of America[C]. USA, 2007.

（责任编辑 王卫勋）