

基于粒子群优化算法的支持向量机参数选择及其应用

邵信光, 杨慧中, 陈 刚

(江南大学 控制科学与工程研究中心, 江苏 无锡 214122)

摘要: 参数选择是支持向量机 (SVM) 研究领域的重要问题, 它的本质是一个优化搜索过程, 考虑到进化算法在求解优化问题上的有效性, 提出了以最小化 k -fold 交叉验证误差为目标, 粒子群优化 (PSO) 算法为寻优技巧的 SVM 参数调整方法. 通过仿真例子验证该方法的有效性后, 用其建立了聚丙烯腈生产过程中数均分子量的软测量模型, 结果表明该方法有效.

关键词: 支持向量机; 参数选择; 粒子群优化; 聚丙烯腈; 软测量

中图分类号: TP274 **文献标识码:** A

Parameters selection and application of support vector machines based on particle swarm optimization algorithm

SHAO Xin-guang, YANG Hui-zhong, CHEN Gang

(Research Center of Control Theory and Control Engineering, Southern Yangtze University, Wuxi Jiangsu 214122, China)

Abstract: Parameters selection is an important problem in the research area of support vector machines (SVM), and its nature is an optimization problem. Motivated by the effectiveness of evolution algorithm on optimization problem, a new automatic searching methodology, based on particle swarm optimization (PSO) algorithm, is proposed in this paper. Each particle indicates a group of SVM parameters, and the population is a collection of particles in this method. Furthermore, the k -fold cross-validation error is used as the fitness function of PSO. After having been validated its effectiveness by two artificial data experiments, the proposed method is then applied to establish a soft-sensor model for average molecular weight in polyacrylonitrile productive process. Finally, real data simulation results are also given to show the efficiency.

Key words: support vector machines; parameters selection; particle swarm optimization; polyacrylonitrile; soft-sensor

1 引言(Introduction)

支持向量机^[1] (support vector machines, SVM) 是一种基于统计学习理论的机器学习方法, 能较好地解决小样本、非线性等实际问题, 已成为智能技术领域研究的热点^[2~3]. 实践表明, SVM 的性能与核函数的类型、核函数的参数以及惩罚系数 C 有很大关系. 因此, 研究 SVM 参数选择的方法, 对支持向量机的发展有很重要的实际意义.

通常, 人们采用既直观又简单的试验定参法 (通过大量的试验) 来获得较优的参数, 这种方法比较费时, 而且获得的参数也不一定是最优的. Chapelle^[4] 提出用梯度下降 (gradient descend, GD) 的方法来完成 SVM 参数选择, 虽然在计算时间上得到了明显的改善, 但是 GD 对初始点要求较高, 而且是一种线性搜索法, 因此极易陷入局部最优.

Chen^[5] 和 Zheng^[6] 采用不同的推广能力估计作为遗传算法 (genetic algorithm, GA) 的适应度函数, 提出了两种基于 GA 的 SVM 参数选择方法, 结果表明利用 GA 对 SVM 参数进行优选不仅缩小了计算时间, 而且还降低了对初始值选取的依赖度. 但是 GA 操作往往比较复杂, 对不同的优化问题都需要设计不同的交叉或变异方式.

粒子群优化^[7~8] (particle swarm optimization, PSO) 算法是群体智能一个新的分支, 由 Kennedy 和 Eberhart 于 1995 年首次提出, 源于对鸟群捕食行为的研究. 与达尔文“适者生存, 优胜劣汰”进化思想的 GA 不同, PSO 是通过个体间的协作来寻找最优解, 它的概念更简单, 效率更高, 更容易实现, 目前已广泛应用于函数优化、模式分类、神经网络训练等领域^[9].

本文提出一种基于 PSO 算法的 SVM 参数选择方法, 利用 PSO 强劲的全局搜索能力实现模型的优化, 通过两个常用的仿真例子验证该方法的有效性后, 用其建立了聚丙烯腈生产过程中数均分子量的软测量模型, 仿真效果令人满意。

2 支持向量回归机(Support vector regression)

支持向量回归算法的基本思想是通过一个非线性映射 ϕ , 将数据 x 映射到高维的特征空间 F , 并在该空间进行线性回归。因此, 对于给定的观测数据集 $G = \{(x_i, y_i)\}_{i=1}^m$ 可以用下式进行回归估计:

$$f(x) = \langle \omega, \phi(x) \rangle + b. \quad (1)$$

其中 ω 和 b 表示回归因子, 可通过最小化以下的正则化风险泛函获得:

$$R(\omega) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m L(f(x_i), y_i). \quad (2)$$

上式中的常数 C 称为惩罚系数, $L(\cdot)$ 表示损失函数, 通常取一次不敏感损失函数

$$L_\varepsilon(f(x), y) = \max\{|f(x) - y| - \varepsilon, 0\}, \quad (3)$$

最小化 $R(\omega)$ 可得 $\omega = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(x_i)$, 其中 α_i , α_i^* 是最小化 $R(\omega)$ 对偶问题的解。将 ω 代入式 (1) 得

$$f(x) = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x) \rangle + b = \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(x_i, x) + b. \quad (4)$$

其中 $k(x_i, x) = \langle \phi(x_i), \phi(x) \rangle$ 称为核函数, 它是对称的正实数函数, 同时满足 Mercer 条件。常用的核函数有径向基核函数 $k(x_i, x) = \exp\{-|x - x_i|^2 / (2\sigma^2)\}$ 和多项式核函数 $k(x_i, x) = [(x, x_i) + c]^q, q \in N, c \geq 0$ 等。

当 b 取在边界上的一点, 便可根据 KKT 最优条件定理进行计算, 但出于稳定性考虑, 一般取边界点上的平均值

$$b = \text{average}_k \{ \delta_k + y_k - \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(x_i, x_k) \}. \quad (5)$$

其中 δ_k 为预测误差。

上述可见, 在支持向量回归估计算法中, 核函数的类型、核函数的参数以及惩罚系数都是很重要的参数 (核函数的选取决定了输入空间到特征空间映射的方式, 惩罚系数 C 用于平衡训练误差和模型复杂度), 为获得更好的推广能力有必要对这些参数进行调整。

3 基于 PSO 算法的 SVM 参数选择 (Parameters selection of SVM based on PSO)

SVM 参数选择的过程实际上是一个优化搜索过程, 搜索空间的每个点都是最佳模型的一个潜在解, 并可由推广能力估计值做出评估, 因此, 求解最优参数的过程实质上就是最小化泛化误差的过程。

PSO 算法是近几年兴起的一种进化算法, 可应用于一切 GA 能应用的场合, 而且在编码和寻优策略上, PSO 要比 GA 更加简单、有效。

3.1 PSO 算法原理 (The principle of PSO)

PSO 求解优化问题时, 问题的解对应于搜索空间中一只鸟的位置, 称这些鸟为“粒子” (particle)。每个粒子都有自己的位置和速度 (决定飞行的方向和距离), 还有一个由被优化函数决定的适应值。各个粒子记忆、追随当前的最优粒子, 在解空间中搜索。

PSO 初始化为一群随机粒子 (随机解), 然后通过迭代找到最优解。在每一次迭代中, 粒子通过跟踪两个“极值”来更新自己。第一个就是粒子本身所找到的最优解, 这个解叫做个体极值 p_{Best} ; 另一个极值是整个种群目前找到的最优解, 这个极值是全局极值 g_{Best} 。

在找到这两个最优值时, 粒子根据如下两个公式来更新自己的速度和位置:

$$v = w \cdot v + c_1 r_1 (p_{\text{Best}} - p) + c_2 r_2 (g_{\text{Best}} - p), \quad (6)$$

$$p = p + \beta \cdot v. \quad (7)$$

其中 p 是粒子当前的位置表示 SVM 参数 $\{C, \sigma\}$ 的当前取值, $v \in [-v_{\text{max}}, v_{\text{max}}]$ 是粒子的速度, 决定下一代 $\{C, \sigma\}$ 的更新方向和大小, β 称为约束因子, 控制速度的权重, 通常取 1, c_1, c_2 是学习因子, 通常取 $c_1 = c_2 = 2$, r_1, r_2 是介于 (0,1) 之间的随机数, w 是非负数, 称为惯性因子。

3.2 基于 PSO 的 SVM 参数调整算法 (Parameters tuning of SVM based on PSO)

解的编码和适应度函数是利用进化算法解决优化问题的两个重要方面。由于 PSO 算法采用实值编码, 因此只需在解的区间内任意初始化即可; 适应度函数取 SVM 模型的推广能力估计值。

推广能力估计是机器学习中的另一个重要问题, 也是参数选择的基础, 常用的方法有留一法 (leave-one-out), k -fold 交叉验证法、支持向量率法、Jaakkola-Haussler 法、Oppen-Winther 法、V-C 法等等。其中留一法是 k -fold 交叉验证法的特例, 要求 k 等于样本的个数。考虑到 k -fold 交叉验证法的估计无偏性, 本文选用 k -fold 交叉验证误差作

为SVM参数选择的目标值,以径向基核函数为例,利用PSO算法对核参数 σ 和惩罚系数 C 进行优选.

算法具体步骤描述如下:

步骤1 读取样本数据,随机产生一组 $\{C, \sigma\}$ 作为粒子的初始位置;

步骤2 把整个样本平均分割成 k 个互不包含的子集 $S_1, S_2, \dots, S_k$;

步骤3 根据当前的 $\{C, \sigma\}$ 训练SVM,计算 k -fold交叉验证误差;

步骤3.1 初始化 $i = 1$;

步骤3.2 S_i 子集留作检验集,其余的子集合并起来作为训练集,训练SVM;

步骤3.3 计算第 i 子集的泛化误差 $e_i = \text{mean}(S_i - \hat{S}_i)^2$,令 $i = i + 1$,重复步骤3.2直到 $i = k + 1$;

步骤3.4 计算 k 次泛化误差的平均值得到 k -fold交叉验证误差;

步骤4 以 k -fold交叉验证误差作为适应值,并记忆个体与群体所对应的最佳适应值的位置为 p_{Best} 和 g_{Best} ,根据式(6)式(7)搜寻更好的 $\{C, \sigma\}$;

步骤5 重复步骤2直到满足最大迭代次数;

步骤6 结束.

4 仿真示例(Simulation example)

为验证上述算法的有效性,本节选用SVM回归中常用的两个函数来进行仿真,所有试验都是在主频3.0GHz,内存256MB的PC机,MTALAB 7.0软件平台上完成.

4.1 一维函数示例(One dimensional function examples)

取一维 $\sin c$ 函数如下:

$$f(x) = \sin c(x) + \nu, x \in [-3, 3]. \quad (8)$$

其中: ν 是均值为0、方差为0.1的高斯噪声.在输入变量域内取100个数据构成SVM的训练样本,以最小化5-fold交叉验证误差为目标,利用PSO算法对径向基核函数参数 σ 和惩罚系数 C 进行优选.其中SVM采用 $\varepsilon=0.1$ 的一次不敏感损失函数;PSO中的 v_{max} 取100; w 采用线性下降法,从1.2下降到0.3; σ 和 C 分别在 $[0, 5]$ 和 $[1, 1000]$ 内初始化;种群大小取20;最大迭代1000次.图1显示了根据本文方法得到的参数对SVM进行训练的结果,可见拟合效果很好.为便于比较,分别利用正交法(2因素10水平)和遗传算法(GA,交叉概率0.8,变异概率0.2%,种群数和最大迭代次数与PSO法一样)进行多次试验,并随机产生50个数据作为测试集,对所得SVM模型进行测试,平均结果记于表1.从表

中可见,利用本文方法所得的SVM模型支持向量个数最少,测试均方误差(MSE)最小.

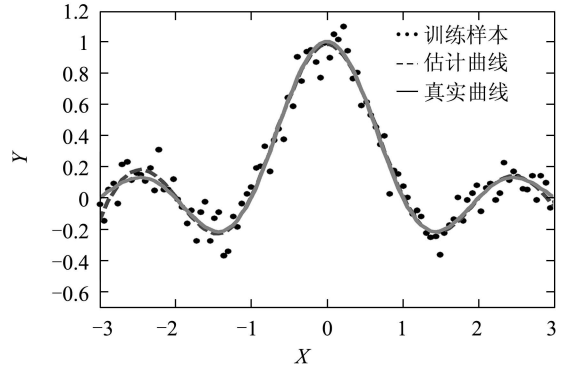


图1 $\sin c$ 函数仿真结果

Fig. 1 Simulation result of $\sin c$ function

表1 不同方法进行SVM参数选择试验的平均结果比较

Table 1 Average results of parameters selection of SVM by diferent methods

优选方法	C	σ	#SV	训练误差	测试误差
正交法+5-fold交叉验证	500	2	32	0.0075	0.0011
GA+5-fold交叉验证	792.14	3.73	28	0.0073	7.9057×10^{-4}
PSO+5-fold交叉验证	750.45	3.99	26	0.0072	5.5782×10^{-4}

4.2 二维函数示例(Two dimensional function example)

取二维函数如下:

$$f(x_1, x_2) = \sin x_1 \cdot \cos x_2, -\pi \leq x_1, x_2 \leq \pi. \quad (9)$$

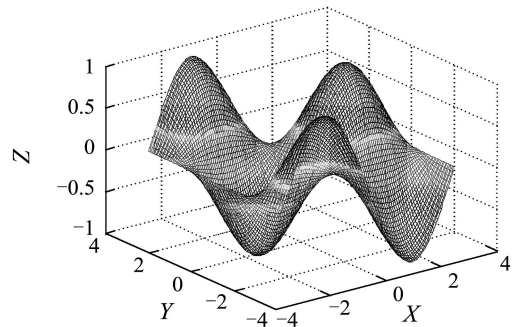


图2 二维函数测试结果

Fig. 2 Testing result of the two-dimensional function

在输入变量域内分别产生100组数据和10000组数据构成SVM的训练样本与测试样本,同样以最小化5-fold交叉验证误差为目标,利用PSO算法对径向基核函数参数 σ 和惩罚系数 C 进

行优选(参数设置同上例一样), 得到 $\sigma^*=0.716$, $C^*=372.30$. 用这两个参数重新训练SVM模型, 并用10000组测试样本对模型进行测试, 得到该二维函数的测试结果如图2所示, 测试误差曲面如图3所示. 从中可见, 经过PSO优选后所得的SVM模型有较好的泛化性能.

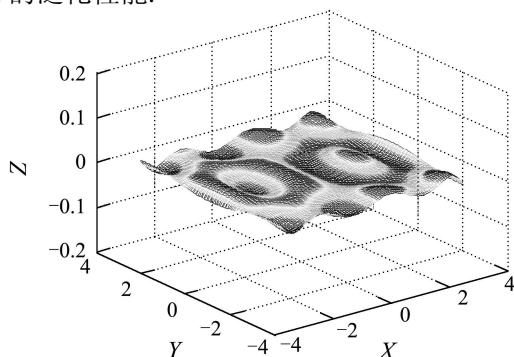


图3 二维函数测试误差曲面

Fig. 3 Testing error surface of the two-dimensional function

5 基于PSO算法的SVM建模应用(Application of SVM based on PSO)

数均分子量是聚丙烯腈生产过程中非常重要的一个质量指标, 由于设备条件的限制, 在运行中仅以一天一次的人工分析值作为质量指标的观测, 时滞太大, 无法对产品质量进行实时监控, 因此有必要开发一套数均分子量的软测量模型, 提供质量指标的

实时显示.

利用SVM建立聚丙烯腈数均分子量软测量模型, 首先要确定影响质量指标的主要因素, 其次要选择具有可观测性的主要生产变量和质量指标构成样本数据集, 对样本数据进行预处理后, 再利用本文的方法对SVM参数进行优选, 最后根据获得的参数训练SVM.

本文在文献[10]基础上将物料停留时间 τ 、引发剂初始浓度 $[I]_0$ 、初始亚硫酸浓度以及聚合温度 T 作为SVM模型的输入变量, 聚合釜出口物的数均分子量AMW作为模型的输出. 采集 $m=121$ 组质量指标人工分析值和与之对应的现场数据作为样本数据集. 为消除各个因子由于量纲和单位不同的影响, 对样本的输入、输出数据分别进行归一化处理, 并进行主成分分析. 将处理后的数据分成两部分, 100组用于训练, 21组用于测试. 选取 $\varepsilon=0.01$ 的一次不敏感损失函数和径向基核函数, 以最小化5-fold交叉验证误差为目标, 用本文的方法进行SVM参数优选, 得到最优的参数分别为 $C=410.70$ 、 $\sigma=0.049$, 再用该参数对SVM重新训练, 得到数均分子量估计结果如图2所示, 从图中可见估计值与人工分析值非常接近, 平均相对误差为1.38%, 较文献[10]的神经网络法(2.3%)和文献[11]的常规支持向量机法(1.83%)有较大的提高.

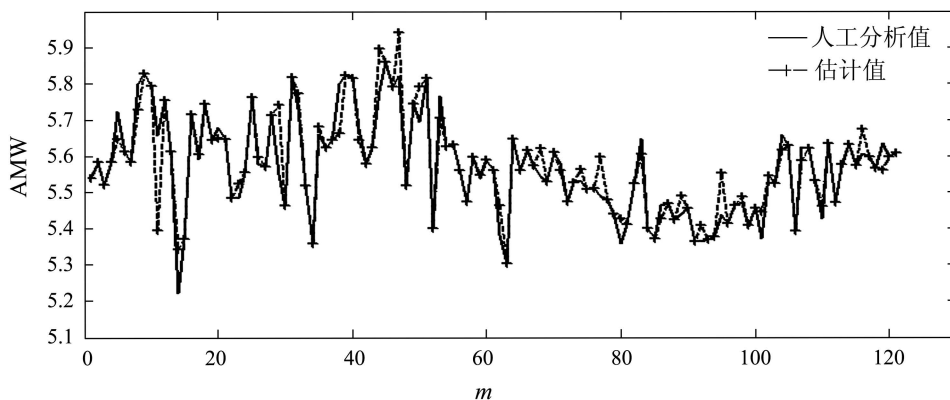


图4 数均分子量仿真结果

Fig. 4 Simulation result of average molecular weight

6 结论(Conclusion)

支持向量机的参数选择是一个既有实际价值又有理论意义的研究课题, 是一以来限制SVM推广应用的关键所在, 本文首次提出一种基于PSO算法的SVM参数选择方法, 通过实验验证该方法的有效性之后, 用其建立了聚丙烯腈

生产过程中数均分子量的软测量模型, 仿真结果表明PSO强劲的全局搜索能力, 对提高模型的准确率有较大的贡献.

另外, PSO的生命力在于工程应用, 开拓粒子群优化算法新的应用领域也是一项具有非常重要意义的工作.

(下转第748页)

应地确定变异率,使算法在初期有较大的变异率保持个体的多样性,避免早熟;在后期逐步降低变异率,保留优良信息,避免最优解遭到破坏,增加了搜索到全局最优解的概率.通过ADE算法和DE算法对典型数学优化问题寻优分析,ADE算法的离线性能和在线性能都有较大的改进,搜索到全局最优解的概率获得较大提高,对算法参数的敏感性低.应用ADE算法获得4-CBA含量软测量模型参数全局最优估计值,其对应的软测量模型对校验样本的预测误差平方和均值仅为0.00364,模型准确预测CTA产品中4-CBA含量.

参考文献(References):

- [1] HOLLAND J H. *Adaptation in Natural and Artificial Systems*[M]. Michigan: The University of Michigan Press, 1975.
- [2] WANG F S, CHIOU J P. Optimal control and optimal time location problems of differential-algebraic systems by differential evolution[J]. *Industrial & Engineering Chemistry Research*, 1997, 36(1): 5348 – 5357.
- [3] MICHALEWICZ Z, JANIKOW C Z, KRAWCZYK J B. A modified genetic algorithm for optimal control problems[J]. *Computer Math Applications*, 1992, 23(12): 83 – 94.
- [4] WRIGHT A H. *Genetic Algorithms for Real Parameter Optimization. Foundations of Genetic Algorithms*[M]. Rawlins G J E, EDS CA: Morgan Kaufmann, 1991: 205 – 218.
- [5] YAN X F, CHEN D Z, HU S X. Chaos-genetic algorithms for optimizing the operating conditions based on RBF-PLS Model[J]. *Computers and Chemical Engineering*, 2003, 27(12): 1393 – 1404.
- [6] YAN X F, DU W L, QIAN F. Development of a kinetic model for industrial oxidation of p-Xylene by RBF-PLS and CGA[J]. *AIChE Journal*, 2004, 50(6): 1169 – 1176.
- [7] RAGHAVENDRACHAR P, RAMACHANDRAN S. Liquid-phase catalytic oxidation of p-xylene[J]. *Ind Eng Chem Res*, 1992, 31(2): 453 – 462.
- [8] 刘建新, 白鹏. 对苯二甲酸工艺技术及生产[J]. 化工科技, 2000, 8(3): 64 – 67.
(LIU Jianxin, BAI Peng. Technology of terephthalic acid and its production [J]. *Science and Technology Chemical Industry*, 2000, 8(3): 64 – 67.)

作者简介:

颜学峰 (1972—), 男, 副研究员, 博士, 研究方向为化工系统工程、智能信息处理, E-mail: yan_xuefeng@hotmail.com;
余娟 (1973—), 女, 助理研究员, 研究方向为系统工程;
钱锋 (1961—), 男, 教授, 博士生导师, 研究方向为化工系统工程、智能控制.

(上接第743页)

参考文献(References):

- [1] VAPNIK V N. *The Nature of Statistical Learning Theory*[M]. New York: Springer-Verlag, 1995.
- [2] SMOLA A J, SCHOLKOPF B. *A tutorial on support vector regression*[R]. NeuroCOLT2 Technical Report NC2-TR-1998-030. London: Royal Holloway College, University of London, 1998.
- [3] SANCHEZ A D. Advanced support vector machines and kernel methods[J]. *Neurocomputing*, 2003, 55(1): 5 – 20.
- [4] CHAPPELLE O, VAPNIK V, BOUSQUET O, et al. Choosing multiple parameters for support vector machines[J]. *Machine Learning*, 2002, 46(1): 131 – 160.
- [5] CHEN Pengwei, WANG Jungying, LEE Hahnming. Model selection of SVMs using GA approach[C] // *Proc of 2004 IEEE Int Joint Conf on Neural Networks*. Piscataway, NJ: IEEE Press, 2004: 2035 – 2040.
- [6] ZHENG Chunhong, JIAO Licheng. Automatic parameters selection for SVM based on GA[C] // *Proc of the 5th World Congress on Intelligent Control and Automation*. Piscataway, NJ: IEEE Press, 2004: 1869 – 1872.
- [7] KENNEDY J, EBERHART R. Particle swarm optimization[C] // *Proc of IEEE Int Conf on Neural Networks, Perth, Australia*. Piscataway, NJ: IEEE Press, 1995: 1942 – 1948.
- [8] EBERHART R, KENNEY J. A new optimizer using particle swarm theory[C] // *Proc of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan*. Piscataway, NJ: IEEE Press, 1995: 39 – 43.
- [9] VANDEN B F, ENGELBRECHT A P. Training product unit networks using cooperative particle swarm optimizers[C] // *Proc of the Third Genetic and Evolutionary Computation Conference (GECCO)*, San Francisco, USA. Piscataway, NJ: IEEE Press, 2001: 126 – 132.
- [10] 杨慧中, 张素贞, 陶振麟. 聚丙烯腈质量指标软测量混合模型的应用[J]. 化工自动化及仪表, 2002, 29(6): 11 – 13.
(YANG Huizhong, ZHANG Suzheng, TAO Zhenlin. The application of soft measurement hybrid model for the quality performance of polyacrylonitrile[J]. *Control and Instruments in Chemical Industry*, 2002, 29(6): 11 – 13.)
- [11] 邵信光, 杨慧中, 石晨曦. ε -不敏感支持向量回归在化工数据建模中的应用[J]. 东南大学学报, 2004, 34(增刊): 215 – 218.
(SHAO Xinguang, YANG Huizhong, SHI Chenxi. Chemical process data modeling based on ε -insensitive support vector regression[J]. *Journal of Southeast University*, 2004, 34(Suppl): 215 – 218.)

作者简介:

邵信光 (1981—), 男, 主要从事机器学习、支持向量机、智能建模的研究, E-mail: shaoxinguang@yahoo.com.cn;
杨慧中 (1955—), 女, 教授, 博士生导师, 主要从事复杂过程建模和优化控制的研究, E-mail: yhz.jn@163.com;
陈刚 (1972—), 男, 工程师, 主要从事过程控制研究.