基于改进 k-最近邻回归算法的软测量建模

叶 涛 [1] 朱学峰 [1] 李向阳 [1] 史步海 [1]

摘 要 机器学习回归方法被广泛应用于复杂工业过程的软测量建模. k-最近邻 (kNN) 算法是一种流行的学习算法, 可用于函数回归问题. 然而, 传统 kNN 算法存在运行效率低、距离计算忽略特征权值的缺点. 本文引入了二次型距离定义和样本集剪辑算法, 改进了传统 kNN 回归算法, 并将改进的算法用于工业过程软测量建模. 仿真实验得到了一些有益的结论.

关键词 k-最近邻算法, 二次型距离, 软测量, 纸浆 Kappa 值
中图分类号 TP181

Soft Sensor Modeling Based on a Modified k-Nearest Neighbor Regression Algorithm

YE Tao[1] ZHU Xue-Feng[1] LI Xiang-Yang[1] SHI Bu-Hai[1]

Abstract Recently, machine learning regression algorithms are widely applied to soft sensor modeling for complex industrial processes. The k-nearest neighbor (kNN) algorithm is a popular learning algorithm for solving regression problems. However, the traditional kNN algorithm has low efficiency and ignores the feature weights in distance computing. Using a quadratic distance definition and a data set editing algorithm, we have modified the traditional kNN regression algorithm. The modified algorithm is applied to soft sensor modeling and some useful conclusions are reached.

Key words k-nearest neighbor algorithm, quadratic distance, soft sensing, pulp Kappa number

1 引言

信息技术的发展正使仪器仪表经历智能化革命. 王大珩院士指出: "在当今以信息技术带动工业化发展的时代, 仪器仪表与测试技术是信息科学技术重要的组成部分." 软测量技术 (Soft sensing technology) 在这一背景下诞生. 软测量技术一经产生就倍受过程控制界的广泛关注. 上世纪 90 年代以来, 有关软测量技术的研究十分活跃, 在理论研究和实际应用方面均取得了一定发展. 但正如著名专家 McAvoy[1] 所指出的: 软测量具有广阔的应用前景, 但缺乏系统的开发思路. 可见, 在软测量这一研究领域中还有很多开拓性的工作要做. 文献 [2, 3] 对软测量建模及校正方法进行了综述, 并指出影响软仪表性能的主要因素. 从某种意义上说, 软测量技术是数据挖掘技术在工业过程控制领域中的应用, 而机器学习是实现数据挖掘技术最主要的理论方法. 因此, 多数解决回归问题的学习方法均可用于实现软测量技术. k-最近邻 (k-nearest neighbor, kNN) 算法是应用广泛的学习算法之

收稿日期 2006-3-8 收修改稿日期 2006-6-15
Received March 8, 2006; in revised form June 15, 2006
国家自然科学基金 (60274033, 60404013) 和广东省自然科学基金 (04300048) 资助
Supported by National Natural Science Foundation of China (60274033, 60404013) and Natural Science Foundation of Guangdong Province (04300048)
1. 华南理工大学自动化科学与工程学院 广州 510640
1. College of Automation Science and Engineering, South China University of Technology, Guangzhou 510640
DOI: 10.1360/aas-007-0996

一. 它是一种非参数学习方法, 可用于解决分类问题和回归问题[4]. 本文使用数据集剪辑方法和二次型距离度量对传统的 kNN 回归算法进行改进. 然后将改进的 kNN 回归算法用于软测量建模, 并将该软测量方法应用于制浆蒸煮过程纸浆 Kappa 值的终点预报仿真.

第 2 节将介绍 kNN 算法的基本思想; 第 3 节阐述改进的 kNN 回归算法和基于该算法的软测量建模方法; 第 4 节给出仿真实验和实验结果; 第 5 节是结论.

2　传统 kNN 回归算法

kNN 算法最初由 Cover 和 Hart[5] 提出, 用于解决文本的分类问题. 此后, 它被广泛用于模式识别和基于内容相似性的信息检索[5,6], 而回归问题相关的应用研究则较少. kNN 算法是一种基于实例的学习方法. 它基于向量空间模型 (Vector space model, VSM), 将每个实例视为 $\mathbf{R}^n$ 空间中的一个点 (向量). kNN 回归算法的基本思想是: 给定一个查询实例 $\boldsymbol{x}_q$, 在训练集中找出它的 k 个最近邻, 对 k 个最近邻的目标值求均值, 将其作为算法的输出估计值. 算法隐含的假设是最近邻域内目标值变化平缓. 传统 kNN 算法中, 实例间的近邻度由标准欧氏距离定义.

若把任意实例 $\boldsymbol{x}$ 表示为下面的特征向量 $([\boldsymbol{x}]_1,[\boldsymbol{x}]_2,\cdots,[\boldsymbol{x}]_r,\cdots,[\boldsymbol{x}]_n)^{\mathrm{T}}$, 其中 $[\boldsymbol{x}]_r$ 表示实例 $\boldsymbol{x}$ 的第 r 个属性值. 那么两个实例 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的标准欧氏距离 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 定义为

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)\triangleq(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\cdot(\boldsymbol{x}_i-\boldsymbol{x}_j)=\sqrt{\sum_{r=1}^{n}([\boldsymbol{x}_i]_r-[\boldsymbol{x}_j]_r)^2} \quad (1)$$

在最近邻学习中, 目标函数值可以是离散值 (分类问题), 也可以是连续值 (回归问题). 考虑连续值目标函数 $f:\mathbf{R}^n\rightarrow\mathbf{R}$, 传统 kNN 回归算法的步骤如下:

1) 对于每个训练样例 $\langle\boldsymbol{x},f(\boldsymbol{x})\rangle$, 将其加入训练样例列表, 记为 D_{Trn};

2) 给定一个查询实例 $\boldsymbol{x}_q$, 在 D_{Trn} 中找出实例 $\boldsymbol{x}_q$ 的 k-最近邻子集 $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k\}$, 记为 D_{kNN};

3) 计算 k 个最近邻目标值的均值

$$\hat{f}(\boldsymbol{x}_q)\leftarrow\frac{(\sum_{i=1}^{k}f(\boldsymbol{x}_i))}{k} \quad (2)$$

算法的返回值 $\hat{f}(\boldsymbol{x}_q)$ 为对 $f(\boldsymbol{x}_q)$ 的估计.

对 kNN 回归算法的一个明显的改进是对 k 个最近邻的贡献加权, 将较大的权值赋给较近的近邻, 相应的算法称为距离加权 kNN 回归算法, 其计算公式为

$$\hat{f}(\boldsymbol{x}_q)\leftarrow\frac{(\sum_{i=1}^{k}w_i f(\boldsymbol{x}_i))}{(\sum_{i=1}^{k}w_i)} \quad (3)$$

其中, 距离权值 w_i 和距离成反比关系, 例如, $w_i\triangleq 1/d(\boldsymbol{x}_q,\boldsymbol{x}_i)^2$. 当距离 $d(\boldsymbol{x}_q,\boldsymbol{x}_i)$ 非常接近于 0 时, 将其取为某一预先设定的小正数值 ε.

3　基于改进 kNN 回归算法的软测量建模

kNN 算法的突出特点是, 给定一个训练集 D_{Trn}, 它不是在整个训练集上一次性地建立目标模型, 而是针对每个待估实例作出局部 (最近邻子集 D_{kNN}) 的估计. 传统 kNN 算法是一种简单而有效的学习方法, 学习过程只是简单地存储已知训练样本. 对于大容量数据集和高维数向量空间的应用, 传统 kNN 算法也存在一些问题: 一是算法需要大量的计算和存储开销, 导致运行效率较低; 二是定义距离时均等考虑所有属性值, 导致近邻距离被大量不相关属性所支配. 为了解决传统 kNN 算法存在的上述缺陷, 领域学者提出了诸多改进方法. 例如, 压缩近邻法则[7], 用于 kNN 的分支定界算法[8], 参考样本集的最优选择[9] 和数据索引结构[10] 等. 本节结合具体问题对传统 kNN 回归算法进行改进, 并将改进的算法用于软测量建模.

3.1　改进的 kNN 回归算法

鉴于距离度量在 kNN 算法中的重要性, 首先对距离定义进行改进. 式 (1) 定义的距离是标准欧氏距离, 各属性对距离计算的贡献权重一样. 然而, 实际情况是各属性对于实例间相似程度的权重并不一样. 因此, 有必要引入二次型距离对相似度进行度量, 其定义如下

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)\triangleq(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\cdot W\cdot(\boldsymbol{x}_i-\boldsymbol{x}_j) \quad (4)$$

其中 W 为半正定对称矩阵. 通常, 矩阵 W 取一对角阵, 称为权值矩阵, 对各属性 (特征) 进行加权. 特别地, W 取单位阵 E 时, 二次型距离即为标准欧氏距离. kNN 算法的关键操作是搜索 D_{kNN} 子集, D_{kNN} 子集的质量将影响算法的泛化性能, 而 D_{kNN} 子集的获得取决于距离定义, 或者说, 取决于权值矩阵 W.

kNN 回归算法可分为学习过程和新实例估计过程两个阶段, 学习过程表现为 D_{Trn} 的初始建立和维护; 新实例估计过程表现为 D_{kNN} 的获得和目标函数值的估计. 学习过程虽然简单却很重要, 它将极大地影响算法的计算开销. 主要的计算开销 (D_{kNN} 搜索操作) 虽然发生在后一过程, 却由前一过程决定. 以下就学习过程对传统算法进行改进.

给定任意查询实例 $\boldsymbol{x}_q$, 搜索其 D_{kNN} 子集需要计算它和 D_{Trn} 中所有训练样本的距离, 并进行排序. 训练集 D_{Trn} 的容量和其中向量维数大小将决定算法的计算开销. 一般从工业过程数据库提取的原始数据集容量都很大, 而且其中有不少矛盾样本和冗余样本. 因此有必要对原始数据集进行剪辑, 以获得容量尽量小的、反应原始数据分布的一致子集作为训练集. 求取剪辑训练集的算法分剔除矛盾样本和减少冗余样本两大步骤, 具体步骤如下:

1) 在原始数据集 D 中任取一个训练样本 $\boldsymbol{x}$, 根据 kNN 规则估计其函数值 $\hat{f}(\boldsymbol{x})$. 由 $err(\boldsymbol{x})=\hat{f}(\boldsymbol{x})-f(\boldsymbol{x})$ 估计误差, 若 $|err(\boldsymbol{x})|$ 大于阈值 θ, 则剔除该样本; 否则保留该样本.

2) 重复步骤 1), 直至遍历原始数据集 D 中的所有训练样本, 得到中间数据集 D_{Tmp}.

3) 对数据集 D_{Tmp} 中的样本进行聚类, 在样本数较多的类中适当减少冗余样本, 以减少训练集容量. 最终得到剪辑数据集, 记作 D_{Etd}.

算法中的阈值 θ 可由凑试法选取. 在获得剪辑数据集 D_{Etd} 后, 若数据集容量仍较大或实时性要求较高, 可通过建立索引结构进一步减少 D_{kNN} 子集搜索时间. 效率较高的索引结构是树形结构[10,11], 常用于 kNN 算法的索引结构有 R-tree、R*-tree 和 SS-tree 等.

3.2　软测量建模

软测量方法是一种利用较易在线测量的辅助变量和离线分析信息去估计不可测或难测变量 (主导变量) 的方法. 软测量系统的核心是表征辅助变量和主导变量关系的软测量模型. 从软测量的基本原理可知, 软测量建模属于多元函数回

归问题. kNN 回归算法是一种非参数回归算法, 在实施算法之前无需预先知道数据的概率分布模型及其参数. 基于此优点, 它已被应用于分布复杂的系统的回归问题. 本节将改进 kNN 回归算法用于制浆蒸煮过程纸浆 Kappa 值的终点预报建模 (稳态建模).

制浆蒸煮过程是一个复杂的物理化学过程, Kappa 值是纸浆的重要质量指标, 稳定 Kappa 值是提高纸浆质量的关键. 目前, 国内外研究表明纸浆 Kappa 值波动主要受 H 因子、有效碱浓度 (EA) 和硫化度 (S) 等过程变量的影响. 使用 kNN 回归算法实现其软测量, 可将每一训练样本表示为 $\langle (H, EA, S)^{\mathrm{T}}, Ka\rangle$, 其中 $\boldsymbol{x} = (H, EA, S)^{\mathrm{T}}$, $f(\boldsymbol{x}) = Ka$. 使用改进 kNN 回归算法进行软测量建模的步骤如下:

1) 首先确定属性权值矩阵 W, 剪辑阈值 θ 和最近邻样本数 k;

2) 使用式 (4) 定义的距离对原始数据集进行剪辑, 获得剪辑训练集 D_{Etd};

3) 对于查询实例 $\boldsymbol{x}_q$, 使用改进 kNN 算法搜索其 D_{kNN} 子集, 按式 (3) 进行泛化估计.

在此, 需要注意式 (3) 中距离权值和式 (4) 中属性 (特征) 权值的区别.

4 应用实例仿真

本节在 Matlab 6.5 平台 (运行于 1.8 GHz CPU 256 MB RAM 的 PC 机) 上就基于改进 kNN 回归算法的软测量模型进行实验, 以检验该软测量建模方法的有效性. 实验使用的数据取自某造纸厂制浆蒸煮过程的监控数据库, 提取 2004 年 6 月至 2005 年 7 月六号蒸煮锅共 653 条蒸煮数据. 排除明显故障锅次后, 剩下 620 锅次作为原始数据集. 对原始数据集进行零均值单位方差标准化处理, 标准化后原始数据集呈一种近似的混合三维正态分布. 下面基于该标准化数据集进行两个实验并给出实验结果.

首先, 用剪辑算法对原始数据集进行剪辑得到剪辑数据集 D_{Etd} (402 个样本), 其中剪辑算法的参数为 $k = 3$, $\theta = 4.5$, $W = \mathrm{diag}(0.45, 0.35, 0.2)$, 运行耗时 3 小时 32 分钟. 取原始数据集中的前 600 个样本作为训练集一 D_{Trn1}, 将剪辑数据集 D_{Etd} 作为训练集二 D_{Trn2}, 随机选取原始数据集中的 60 个样本作为检验集 D_{Test}.

实验一对训练集 D_{Trn1} 和 D_{Trn2} 分别运行 kNN 回归算法 ($k = 3$), 比较剪辑前后的训练集对算法泛化能力和运行效率的影响. 实验统计数据见表 1, 其中运行时间为每个样本点耗时, Ka 为实际值、Kp 为预测值; 针对检验集 D_{Test}, 图 1 给出了两个训练集的预测结果. 由表 1 知, 剪辑后的数据集计算效率比剪辑前快一倍多, 且均方根误差 ($RMSE = \sqrt{\sum_{i=1}^{N_{Test}} (Ka_i - Kp_i)^2}$) 有所改善, 其中 N_{Test} 是检验集 D_{Test} 的样本点数. 然而, $|Ka - Kp| \le 3$ 的准确率略有下降, 究其原因可能是原始数据集所含过程信息量不足或剪辑过度导致.

表 1 实验一统计数据

Table 1 Statistical data of experiment 1

训练集	训练集 D_{Trn1}		训练集 D_{Trn2}	
运行时间 (s)	20.1		9.0	
近邻权值 w_i	$1/d_i$	$1/d_i^2$	$1/d_i$	$1/d_i^2$
$\lvert Ka - Kp\rvert \le 3(\%)$	90	91.7	88.3	88.3
$\lvert Ka - Kp\rvert \le 5(\%)$	95	95	98.3	98.3
$RMSE$	2.157	1.928	1.923	1.750

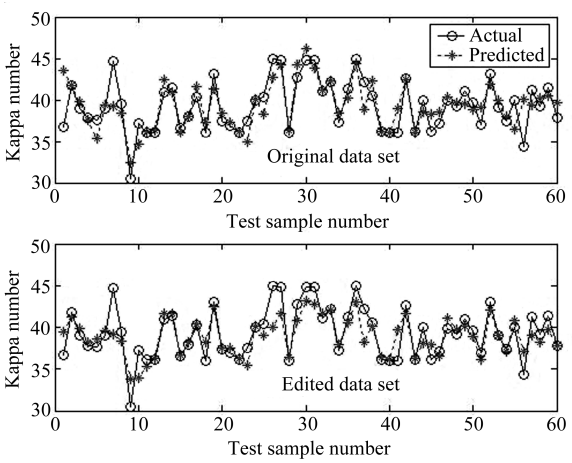

图 1 两个数据集上的预测结果 ($w_i = 1/d_i^2$)

Fig. 1 Prediction over two datasets ($w_i = 1/d_i^2$)

实验二使用训练集 D_{Trn2} 对不同 k 值 ($k = 1 \sim 5$) 分别运行 kNN 回归算法, 研究 k 的取值对算法泛化能力的影响. 实验统计数据见表 2, 下标 g 表示只统计训练集 D_{Trn2} 外的 12 个样本点; 图 2 给出了不同 k 值的算法的泛化能力. 由表 2 看出, k 值对算法运行效率影响不大; 对整个测试集 D_{Test} 进行统计的性能指标中, $k = 1$ 和 $k = 2$ 的性能偏高, 原因是测试集中的大部分样本点 (48 个) 属于训练集 D_{Trn2}, 即未被剪辑掉, 只有 12 个样本点落在训练集 D_{Trn2} 之外 (可由图 2 (a) 看出). 图 3 (见下页) 为 k 值对算法性能指标的影响. 由图 3 可知, 针对训练集 D_{Trn2} 外 12 个测试样本的性能指标, 当 $k = 3$ 时算法性能最好; 针对测试集 D_{Test} 的 60 个测试样本的性能指标, $k = 1$ 和 $k = 2$ 的性能偏高. 实际上, 针对训练集 D_{Trn2} 外测试样本进行统计的性能指标 (带下标 g) 更能反映算法的真实泛化能力.

表 2 实验二统计数据

Table 2 Statistical data of experiment 2

k 值	1	2	3	4	5
运行时间 (s)	9.03	8.92	9.0	8.95	8.99
$\lvert Ka - Kp\rvert \le 3(\%)$	91.7	90	88.3	85	86.7
$RMSE$	1.485	1.617	1.750	1.947	1.982
$\lvert Ka - Kp\rvert_g \le 3(\%)$	58.3	75	83.3	66.7	75
$RMSE_g$	3.320	2.468	2.290	2.591	2.584

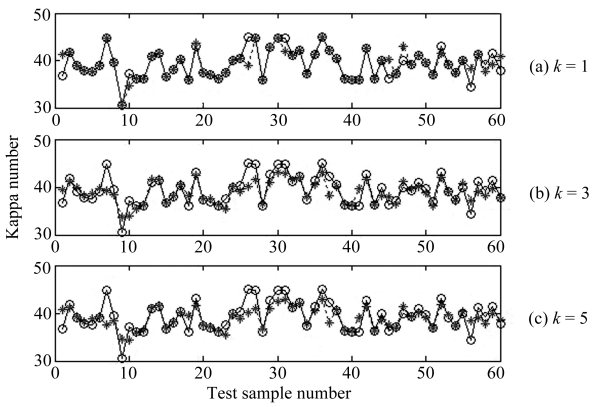

图 2 数据集 D_{Trn2} 对于不同 k 值 ($k = 1, 3, 5$) 的预测结果

Fig. 2 Prediction over dataset D_{Trn2} with $k = 1, 3, 5$

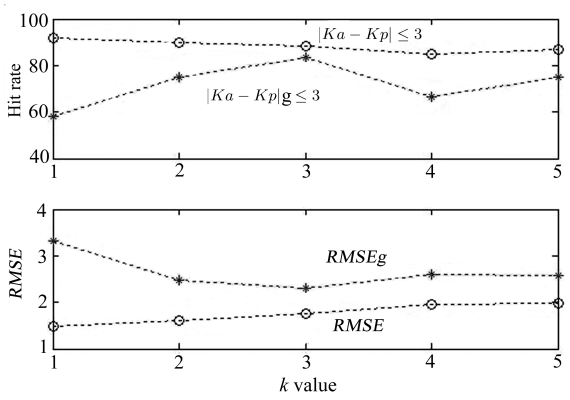

图 3　算法取不同 k 值 ($k=1\sim5$) 的性能分析

Fig. 3　Performance analysis of the algorithm with different k ($k=1\sim5$) values

5　结论

本文首先介绍了 kNN 算法的基本原理, 在此基础上引入了二次型距离的定义和数据集剪辑算法, 对传统 kNN 回归算法进行改进, 并将改进的 kNN 回归算法用于制浆蒸煮过程软测量建模. 通过仿真实验得到以下结论:

1) 剪辑后的数据集运行效率比剪辑前的快一倍多, 且 $RMSE$ 有所改善, 但 $|Ka-Kp|\leq 3$ 的准确率略有下降;

2) 基于改进 kNN 回归算法的软测量建模方法的计算效率和预测精度均达到了实际应用要求, 实例中当 $k=3$ 时算法的泛化能力最好.

后续工作将集中在剪辑算法中阈值 θ 和二次型距离定义中权值矩阵 W 对回归算法性能的影响; 建立合适的相似索引结构, 以进一步提高算法的运行速度.

References

1 McAvoy T J. Contemplative stance for chemical process control. *Automatica*, 1992, **28**(2): 441∼442

2 Yu Jing-Jiang, Zhou Chun-Hui. Soft-sensing techniques in process control. *Control Theory and Applications*, 1996, **13**(2): 137∼144
(于静江, 周春晖. 过程控制中的软测量技术. 控制理论与应用, 1996, **13**(2): 137∼144)

3 Li Hai-Qing, Huang Zhi-Yao. *Soft Sensing Technology: Principles and Applications*. Beijing: Chemical Industry Press, 2000
(李海清, 黄志尧. 测量技术原理及应用. 北京: 化学工业出版社, 2000)

4 Mitchell T M. *Machine Learning*. Beijing: McGraw-Hill Education and China Machine Press, 2003

5 Cover T M, Hart P E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, **13**(3): 21∼27

6 Luo Ming, Bai Xue-Sheng, Xu Guang-You. Hierarchical similarity indexing for quadratic distance based on SVD technology. *Journal of Tsinghua University (Science & Technology)*, 2002, **42**(1): 36∼39
(罗明, 白雪生, 徐光祐. 基于 SVD 的二次型距离相似索引层次算法. 清华大学学报 (自然科学版), 2002, **42**(1): 36∼39)

7 Hart P E. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 1968, **14**(3): 515∼516

8 Fukunaga K, Narendra P M. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, 1975, **24**(7): 750∼753

9 Zhang Hong-Bin, Sun Guang-Yu. Optimal selection of reference subset for nearest neighbor classification. *Acta Electronica Sinica*, 2000, **28**(11): 16∼21
(张鸿宾, 孙广煜. 近邻法参考样本集的最优选择. 电子学报, 2000, **28**(11): 16∼21)

10 Guttman A. R-trees: a dynamic index structure for spatial searching. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Boston, USA, 1984. **13**(2): 47∼57

11 Zelenko D. Machine Learning for Information Extraction [Ph. D. dissertation], University of Illinois at Urbana-Champaign, 2003

叶 涛　华南理工大学自动化科学与工程学院控制理论与控制工程专业博士研究生. 主要研究方向为智能检测与智能控制, 机器学习和数据挖掘. 本文通讯作者. E-mail: towerye@21cn.com
(**YE Tao**　Ph. D. candidate in control theory and control engineering at College of Automation Science and Engineering, South China University of Technology (SCUT). His research interest covers intelligent sensing and intelligent control, machine learning, and data mining. Corresponding author of this paper.)

朱学峰　华南理工大学自动化科学与工程学院教授. 主要研究方向为智能检测与智能控制, 软测量建模和模式识别.
E-mail: xfzhu@scut.edu.cn
(**ZHU Xue-Feng**　Professor at College of Automation Science and Engineering, SCUT. His research interest covers intelligent sensing and intelligent control, soft sensor modeling, and pattern recognition.)

李向阳　华南理工大学自动化科学与工程学院副教授. 主要研究方向为机器学习和模式识别. E-mail: xyangli@scut.edu.cn
(**LI Xiang-Yang**　Associate professor at College of Automation Science and Engineering, SCUT. His research interest covers machine learning and pattern recognition.)

史步海　华南理工大学自动化科学与工程学院副教授. 主要研究方向为过程控制和模式识别. E-mail: bhshi@scut.edu.cn
(**SHI Bu-Hai**　Associate professor at College of Automation Science and Engineering, SCUT. His research interest covers process control and pattern recognition.)