文章编号: 1001-0920(2006)05-0563-04

一种离群数据挖掘新方法的研究与应用

闫 伟[1,2], 张 浩[3], 陆剑峰[1]
(1. 同济大学CMS研究中心, 上海 200092; 2. 山东大学 机械工程学院, 济南 250061; 3. 上海电力学院 信息控制系, 上海 200092)

摘 要: 离群数据挖掘是数据挖掘的重要内容. 利用蚁群算法鲁棒性强的优点, 改进了聚类方法. 在此基础上, 将聚类分析和蚁群算法某些参数相结合, 提出一种基于聚类的离群指数新定义, 成功地实现了离群数据挖掘过程并编程实现. 采用此方法对流程企业的大量历史数据进行分析, 从而起到了对设备运行优化和故障预警的作用.

关键词: 离群挖掘; 蚁群算法; 聚类分析; 流程企业
中图分类号: TP311 **文献标识码**: A

Study and Application of a Novel Method of Outlier Mining

YAN Wei[1,2], *ZHANG Hao*[3], *LU Jian-feng*[1]
(1. CMS Center, Tongji University, Shanghai 200092, China; 2. School of Mechanical Engineering, Shandong University, Ji'nan 250061, China; 3. Department of Information and Control Engineering, Shanghai University of Electric Power, Shanghai 200092, China. Correspondent: YAN Wei, E-mail: yanwei@sdu.edu.cn)

Abstract: Outlier mining is an important part of data mining. Combining ant colony algorithm with *k*-means algorithm, excellent results of clustering analysis are obtained. Then, a novel measure for identifying the physical significance of an outlier is designed by some permanents of ant colony algorithm and *k*-means algorithm, which is called cluster-based outlier index. The FindAnt-*CO*(*t*) algorithm for discovering outliers is proposed. Based on the models, the equipment process is optimized and the faults are monitored.

Key words: Outlier mining; Ant colony algorithm; Clustering analysis; Process industry

1 引 言

数据挖掘就是从大型数据库或数据仓库中提取出隐含的、先前未知的、对决策有潜在价值的知识和规则的前沿方法. 离群数据挖掘(outlier mining)是从大量的数据中挖掘出明显偏离其他数据、不满足数据的一般行为或模式、与存在的其他数据不一致的数据. 对离群数据挖掘的研究往往可使人们发现一些潜在的有用信息.

离群数据的发现已在统计学领域得到广泛研究[1], 它是根据已知的数据分布模型, 使用假设检验来确认离群数据的存在, 这种检验需要知道数据的分布以及分布的参数等. 郑斌祥等人[2]提出了基于第k个最近邻居的离群数据挖掘方法, 但此算法只针对时序数据, 而且运算性能不好; Breuning等人[3]提出一种基于偏离的离群数据发现方法, 但需要先确定相异函数再进行离群数据挖掘, 若其相异函数的选取不合适, 则得不到满意的结果.

聚类分析是一种将物理或抽象的对象, 按照对象间的相似性进行区分和分类的方法. 人们在实施聚类行为时, 会首先确定"核心"(即聚类中心), 然后将周围的对象"吸引"到该"核心"周围, 从而完成聚类过程. 蚁群算法是由意大利学者Dorigo等人[4]提出的一种新型的模拟进化算法, 利用蚁群在搜索食物源过程中所体现出的寻优能力来解决一些离散系统优化中的困难问题. 目前已用该方法求解了旅行商问题、指派问题和调度问题等, 并取得了一系列较

收稿日期: 2005-02-28; **修回日期**: 2005-07-19.
基金项目: 国家863计划项目(2002AA412410).
作者简介: 闫伟(1973—), 男, 济南人, 讲师, 博士生, 从事数据挖掘、故障诊断等研究; 张浩(1963—), 男, 江苏无锡人, 教授, 博士生导师, 从事CMS技术、远程服务等研究.

好的实验结果. 可以发现, 蚂蚁在寻找食物过程中也遵循聚类的原则, 即发现食物源(可类比为聚类中心)后, 蚂蚁就会被"吸引"到食物源周围.

受此启发, 本文将蚁群算法应用于聚类分析, 将聚类分析与蚁群算法的某些参数相结合, 得出一种基于聚类的离群数据的离群指数表示方法, 成功地实现了离群数据挖掘过程, 编程实现后在某石油天然气公司得到应用, 结果证明了算法的有效性.

2 理论分析

2.1 蚁群算法优化理论

经过大量研究发现, 蚂蚁个体之间是通过一种称之为外激素的物质进行信息传递, 从而能相互协作, 完成复杂的任务. 蚂蚁在运动过程中, 能在它所经过的路径上留下该种物质, 并能感知这种物质的存在及其强度, 以此指导自己的运动方向, 蚂蚁总是倾向于朝着该物质强度高的方向移动. 因此, 由大量蚂蚁组成的蚁群集体行为便表现出一种信息正反馈现象: 某一路径上走过的蚂蚁越多, 则后来者选择该路径的概率就越大. 蚂蚁就是通过这种信息的交流达到搜索食物的目的.

下面以TSP问题为例说明基本蚁群算法的框架. 设有m个城市, $d_{ij}(i, j = 1, 2, \cdots, n)$表示城市$i$和$j$间的距离, $\tau_{ij}(t)$表示在t时刻城市i和j之间的信息量, 以此来模拟实际蚂蚁的外激素. 设共有m只蚂蚁, 用$p_{ij}(t)$表示在t时刻蚂蚁k由城市i转移到城市j的概率, 即

$$p_{ij}(t) = \frac{\tau_{ij}^{a}(t)\eta_{ij}}{\sum_{(i,k)\in S, k\in U}\tau_{ik}^{a}(k)\eta_{ik}}. \tag{1}$$

其中: U为蚂蚁已经搜索过的部分路径, S为蚂蚁k下一步允许走过的城市的集合, a为路径上的信息量对蚂蚁选择路径所起的作用大小, η_{ij}为由城市i转移到城市j的期望程度(例如可以取$\eta_{ij} = 1/d_{ij}$). 当$a = 0$时, 算法就是传统的贪心算法; 而当$b = 0$时, 就成了纯粹的正反馈启发式算法. 经过n个时刻, 蚂蚁可走完所有的城市, 完成一次循环. 每只蚂蚁所走过的路径就是一个解, 此时, 要根据下式对各路径上的信息量作更新:

$$\tau_{ij}^{new} = \rho\tau_{ij}^{old} + \Delta\tau_{ij}, \tag{2}$$

其中: $\rho \in (0, 1)$, 表示$\tau_{ij}(t)$随时间推移衰减的程度; 信息增量表示为$\Delta\tau_{ij} = \sum_{k=1}^{m}\Delta\tau_{ij}^{k}$, 而$\Delta\tau_{ij}^{k}$表示蚂蚁$k$在本次循环中在城市$i$和$j$之间留下的信息量, 它根据计算模型而定.

由上述可知, 蚁群算法的优化过程的本质在于: 1) 选择机制, 信息量越大的路径, 被选择的概率越大; 2) 更新机制, 路径上的信息量会随蚂蚁的经过而增长, 同时也随时间的推移逐渐减小; 3) 协调机制, 蚂蚁之间实际上是通过信息量来互相通讯、协同工作的, 这样的机制使得蚁群算法具有很强的发现较好解的能力.

2.2 基于蚁群的聚类算法

用蚁群算法改进的聚类算法有两种思路: 一是基于蚁堆形成原理来实现数据聚类[5], 它实际上是一种基于网格和密度的聚类方法; 另一种是运用蚂蚁觅食的原理, 它以k-means算法的思想为基础[6].

本文采用后一种方式改进聚类分析. k-means算法的思想是, 把n个向量$x_i(i = 1, 2, \cdots, n)$分成m个类$G_i(i = 1, 2, \cdots, m)$, 并求每类的聚类中心, 使得非相似性(或距离)指标的目标函数达到最小. 当选择第i个类G_i中向量x_k与相应的聚类中心c^{p_i}间的度量为欧基里德距离时, 目标函数可定义为

$$J_p = \sum_{i=1}^{m} J^{p_i} = \sum_{i=1}^{m}\sum_{k, x_k\in G_i}\| x_k - c^{p_i}\|^2. \tag{3}$$

其基本思想是将数据视为具有不同属性的蚂蚁, 聚类中心看作是蚂蚁所要寻找的"食物源", 所以数据聚类便可看作是蚂蚁寻找食物源的过程. 具体过程如下: 每只蚂蚁模拟某个数据点, 以一定条件达到某个聚类中心, 在整个解空间中模拟下一个数据点后, 再达到某个聚类中心, 当搜索到数据点为该聚类样本点总数后, 就认为蚂蚁完成了一个路径的搜索. 为使蚂蚁在同一路径的搜索中不重复搜索同一个样本点, 给每只蚂蚁设置一个禁忌表tabu(N). 同时规定: 如果tabu(j)为1, 则结点j是可以选择的搜索样本点, 当蚂蚁选择了结点j后, 便将tabu(j)置为0, 此时蚂蚁就不能选择结点了.

当所有蚂蚁都完成一次路径搜索后, 称算法进行了一个搜索周期. 第t个搜索周期内, 路径选择概率定义为

$$p_{ij}^{k}(t) = \frac{\tau_{ij}^{a}(t)\eta_{ij}}{\sum_{c}^{m}\tau_{ik}^{a}(t)\eta_{ik}}. \tag{4}$$

其中: c是现在的聚类中心, m是聚类中心的数目, 如果$p_{ij}(t)$比到其他中心的概率大, 蚂蚁k就把数据点i归于j类, 搜寻完所有数据点后, 根据下式求出聚类中心:

$$\bar{c_j} = \frac{1}{k}\sum_{i=1}^{k} X_i, \tag{5}$$

以及$\bar{c_j}$的半径r_j, 其中$X_i \in C_j$.

在基本蚁群算法的信息素增量分配中, 对同一路径的不同路段分配相同大小的信息素增量, 而路段影响蚁群向最佳路径搜索的作用显然不同. 因此, 合理的策略应为: 对于路径较短的路段, 分配较大的

信息素增量; 而对较长路段则分配较小的信息素增量. 考虑到相关系数法和欧氏距离能够相互补偿, 例如, 当 2 矢量在一条直线上而又不完全相等时, 余弦距离为 0, 不能区分这 2 矢量; 而欧氏距离不为 0, 则能很好地区分它们. 于是可以对它们进行线性组合. 这样既可保留上次搜索得到的有效信息, 在较好区域内进行更精细的搜索, 加快算法的收敛; 又可以保证大范围搜索的有效性, 使算法能找到全局最佳路径. 因此路径长度对增量的影响为

$$\Delta\tau_{ij1}^{k}=\begin{cases}(r_j-d_{ij})/r_j, & d_{ij}\leq r_j;\\ 0, & d_{ij}>r_j.\end{cases}$$

数据点之间的相似度对增量的影响为

$$\Delta\tau_{ij2}^{k}=\begin{cases}\dfrac{\sum_{k=1}^{m}(x_{ik}-\overline{x_i})(x_{jk}-\overline{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\overline{x_i})^2}\times\sqrt{\sum_{k=1}^{m}(x_{jk}-\overline{x_j})^2}}, & d_{ij}\leq r_i;\\ 0, & d_{ij}>r_i.\end{cases}$$

其中: $\overline{x_i}=\frac{1}{m}\sum_{k=1}^{m}x_{ik}$, $\overline{x_j}=\frac{1}{m}\sum_{k=1}^{m}x_{jk}$, 当 x_j 是簇团中心时, $\Delta\tau_{ij}^{k}$ 就是簇团内其他参数和中心参数的相似度度量. 最后得到如下公式:

$$\Delta\tau_{ij}^{k}=(\Delta\tau_{ij1}^{k}+\Delta\tau_{ij2}^{k})/2 \tag{6}$$

2.3　离群指数的确定

目前基于聚类的离群数据挖掘是研究的热点之一, 文献[7]给出了较新成果, 其基本思想如下.

定义 1　给出数据库 D, 假设对 D 聚类的结果表示为 $C=\{C_1,C_2,\cdots,C_k\}$, 这里 $C_i\cap C_j=\varnothing$, $C_1\cup\cdots\cup C_k=D$, 而且聚类的次序为 $|C_1|\geq|C_2|\geq\cdots\geq|C_k|$, 给出两个数字参数 α 和 β; 设 b 为大小簇团的边界, 则

$$\begin{cases}(|C_1|+|C_2|+\cdots+|C_b|)\geq|D|\cdot\alpha,\\ |C_b|/|C_j|\geq\beta\end{cases} \tag{7}$$

于是, 大簇团定义为

$$LC=\{C_i\,|\,i\leq b\},$$

小簇团定义为

$$SC=\{C_j\,|\,j>b\}.$$

定义 2　假设 $C=\{C_1,C_2,\cdots,C_k\}$ 是聚类的次序: $|C_1|\geq|C_2|\geq\cdots\geq|C_k|$, 而且 α, β, b, LC 和 SC 的定义同上, 则对于任何数据点 t, 得出的离群指数为

$$\mathrm{CBLOF}(t,C_j)=|C_i|\cdot(\mathrm{distance}(t,C_j)), \tag{8}$$

其中: $t\in C_i, C_i\in SC, C_j\in LC, j=1,\cdots,b$.

式(8)的意义是: 某个小簇团内某个数据点的离群指数是由它所在的小簇团的大小和这个数据点与和它最近的某个大簇团边缘点的距离决定. 分析图 1, 如果离群点 b 与大簇团 C_3 的距离等于离群点 a 与大簇团 C_1 的距离, 计算出的 CBLOF(b) 等于 CBLOF(a), 但实际上, 离群点 b 要比离群点 a 更加离群, 即 CBLOF(b) > CBLOF(a). 一个离群的离群指数不仅和此离群点与最近大簇团的距离有关, 而且和这个大簇团内数据点的个数多少有关.

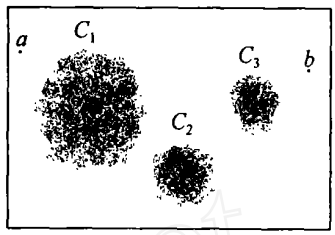

图 1　某个 2 维数据集

从蚁群算法的角度分析聚类过程可以发现, 对于小簇团(离群数据源)而言, 由于蚂蚁数量比较少, 到其中心路径上的信息素 τ_{ij} 也较少, 而对于大簇团则反之. 使其附近的数据点选择此簇团而不选择数据点多的簇团的原因是: 1) 根据式(6)可知, 当路径较短时, $\Delta\tau_{ij}^{k}$ 比较大, 经过蚂蚁几次运动, 使 τ_{ij}^{new} 变大; 2) η_j 比较大(即 $1/d_{ij}$ 大), 综合结果使 $p_{ij}(t)$ 增大, 这样 $p_{ij}(t)$ 也可以作为衡量离群数据的一种指标. 基于此, 本文提出一个新的离群数据的定义.

定义 3　假设 $C=\{C_1,C_2,\cdots,C_k\}$ 是聚类的次序: $|C_1|\geq|C_2|\geq\cdots\geq|C_k|$, 而且 $p_{ij}(t)$, α, β, b, LC 和 SC 的定义同上, 并设 LC 中簇团的个数为 b, 则对于数据点 t, 得出的离群指数为

$$\mathrm{Ant\_CO}(t)=\begin{cases}|C_i|\cdot\sum_{l=1}^{b}p_{di}(t)/p_{dl}(t),\\ \quad t\in C_i, C_i\in SC;\\ 0,\ t\in C_i, C_i\in LC.\end{cases} \tag{9}$$

这种离群指数的定义有明显的优点, 如上所述, 一个离群点的离群指数不仅和此离群点与最近大簇团的距离有关, 而且和此大簇团内数据点的个数多少有关. 蚁群算法中的 $p_{ij}(t)$ 综合考虑了这些情况: τ_{ij} 考虑到簇团内数据点的个数, η_j 考虑到簇团内中心点的距离. 因此 Ant_CO(t) 是一种考虑综合结果的离群指数.

2.4　具体算法流程

算法 FindAnt_CO(t):

设样本集为 $D(X_1,X_2,\cdots,X_n)$, 聚类数目为 m, 给定精度为 ϵ, 参数 $\rho=0.7$, $a=1$, $b=1$.

Step 1: 在样本空间中, 随机地大致均匀选择 m 个聚类中心, 并按欧几里德法则确定 m 个聚类, 计算初始聚类中心 $\overline{c_j}$ ($j=1,2,\cdots,m$), 设有 L 个蚂蚁;

Step 2: 对所有路段上的信息素进行初始化, 确定 $\tau_{ij}(t) = 1/M$ (M 为路径总数);

Step 3: L 个蚂蚁分别随机从数据点到达具体聚类中心 $\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_j$, 然后在整个样本空间并行进行蚁群搜索 T 个周期(T 一般选 3～ 5 个);

Step 4: 用式(2) 计算 τ_{ij};

Step 5: 用式(4) 计算 $p_{ij}(t)$, 判断 $\max\limits_{j=1,\cdots,m}\{p_{ij}(t)\}$ 并把 X_i 归于 C_j;

Step 6: 用式(5) 计算该类的聚类中心;

Step 7: 用式(3) 计算 J_p;

Step 8: 如果 $(|J_p| - |J_{p-1}|)/|J_p| < \epsilon$, 则算法终止; 否则, 在 Step 4 和 Step 5 聚类的基础上返回 Step 3;

Step 9: 根据参数 α 和 β 得到 LC 和 SC;

Step 10: 对每个数据点: 如果 $t \in C_i, C_i \in SC$, 则 $Ant\text{-}CO(t) = |C_i| \cdot \sum_{l=1}^{b} p_{dl}(t)/p_{dl}(t)$; 否则 $t \in C_i$ 且 $C_i \in LC$, 则 $Ant\text{-}CO(t) = 0$.

3 应用分析

3.1 原始数据预处理

流程企业的设备参数有温度、压力、流量等参数, 而且由于采用的单位不同, 数量级之间相差很大, 所以首先对数据进行无量纲化处理.

3.2 实际应用

分析生产参数可以发现, 比如图 2 所示的 SHPCGP-CRY: Z2PI260 压力曲线, 靠近中心的数据所占比例大, 而远离中心的数据比例小, 这说明设备长期运行在性能优化状况下. 当设备参数一旦偏离优化状况, 工作人员立刻回调, 使其恢复, 所以对流程企业的大量历史数据进行聚类时, 大簇团内的数据是设备的常规优化数据, 小簇团内的数据往往是设备非常规运行时的数据. 这样便可根据分析数据的离群指数, 找到设备运行异常的参数, 并分析引起异常的原因, 从而起到对设备运行优化和故障预警的作用.

采用上述方法对流程企业的重要设备 —— 冷箱和低温分离器中的关键参数: SHPCGP-CRY: Z2PI260, SHPCGP-CRY: Z2TI260B, SHPCGP-CRY: Z2TI260A, SHPCGP-CRY: Z2TI261A 进行分析, 采集 4 万多个数据点标准化处理, 然后存入数据库中.

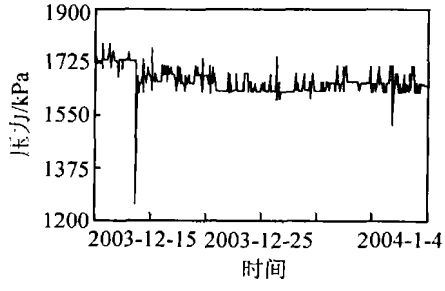

图 2 参数 SHPCGP-CRY: Z2PI260 压力随时间变化曲线

采用算法 FindAnt-CO(t) 建立了设备参数聚类模型, 表 1 是所有数据点聚类的结果, 可见绝大部分数据集中在第 1 到第 4 簇团内, 设参数 $\alpha = 0.95$ 且 $\beta = 5$, 簇团 5 到 10 都属于离群数据簇团, 并根据离群指数 Ant-CO(t) 的大小排列离群数据点, 还原为原始数据点后发现小簇团内的数据往往是设备非常规运行时的数据.

表 1 所有数据点聚类结果

簇团	数据点数	簇团	数据点数
1	15 210	6	367
2	14 840	7	201
3	6 766	8	206
4	3 871	9	310
5	236	10	327

将最离群数据点还原为原始数据点, 与簇团 1 和簇团 2 中心点比较的结果见表 2. 最离群参数点实际上是进入冷箱气体温度、压力很高, 而冷箱冷却效果很差的情况, 这样便可根据分析数据的离群指数, 找到设备运行异常的参数, 从而起到对设备运行优化和故障预警的作用.

表 2 比较结果

	SHPCGP-CRY:			
	Z2PI260 /kPa	Z2TI260A /℃	Z2TI260B /℃	Z2TI261A /℃
簇团 1 中心点	1 552.5	18.7	-17.2	-35.2
簇团 2 中心点	1 673.8	23.4	-21.7	-43.3
最离群参数点	1 937.2	38.5	-3.1	-6.8

4 结 语

离群数据挖掘可使人们发现潜在的有用信息, 本文在 k-means 算法的基础上, 利用蚁群算法对聚类分析进行了改进. 在此基础上, 将聚类分析和蚁群算法某些参数相结合, 提出一种基于聚类的离群指数新定义, 成功地实现了离群数据挖掘过程并编程实现. 通过对某石油天然气公司大量的历史数据分析表明, 该方法为设备运行在优化状态下提供了依据.

参考文献(References)

[1] Markos Markou, Sameer Singh. Novelty Detection: A Review — Part 1: Statistical Approaches[J]. *Signal Processing*, 2003, 83(14): 2481-2497.

(下转第 571 页)

的情况下, 通过式(4) 所构造的参考系统观测向量也相互独立, 且假设两者都服从多元正态分布, 那么利用 Roy-Bose 多元统计法, 令置信度为(1 - r_j) 时的 ci 为

$$(l_j, u_j) = \| \overline{A_j^{(m)}} - \overline{B_j^{(m)}} \| / \sqrt{N} \pm \sqrt{\frac{4(m-1)^2 \| (S_j^{(m)2} - \overline{S_j^{(m)2}}) \|_{T^2_{r_j,\tau,2m-\tau-1}}}{m\sqrt{N}}}, \tag{11}$$

其中 $T^2_{r_j,\tau,2m-\tau-1}$ 是自由度为 τ 和 $2m - \tau - 1$ 的 Hotelling T^2 分布的上 r_j 百分位点. 将式(11) 代入(8) 中, 即可得到仿真系统的同时置信区间[l, u].

若[l, u] ⊂ [L, U], 则认为该次仿真运行有效, 且置信度为 100(1 - r)%; 若[l, u] ⊄ [L, U], 那么增大 m, 或在可接受范围内增加相应的 r_j, 重新计算[l, u]; 在 m 足够大, 或在可接受范围内不允许再增加 r_j 时, 认为该仿真系统运行无效.

4　结　　论

本文结合实际工程特点, 提出了基于隐Markov 模型的复杂仿真系统运行有效性评价方法. 通过采用隐Markov 模型定量描述仿真剧情以及复杂仿真系统运行状态的方法, 有效支持了复杂仿真系统运行有效性定量分析. 在此基础上, 针对评价过程中的关键问题进行了深入研究, 为实现定量分析复杂仿真系统运行有效性提供了新的技术途径.

参考文献(References)

[1] Chew J, Sullivan C. Verification, Validation, and Accreditation in the Life Cycle of Models and Simulations[A]. *Proc of Winter Simulation Conf* 2000 [C]. Orlando, 2002: 813-818.

[2] 李伯虎, 王行仁, 黄柯棣, 等. 综合仿真系统研究[J]. ***系统仿真学报***, 2002, 12(5): 429-434.
(Li B H, Wang X R, Huang K D, et al. The Research of Synthetics Simulation System [J]. *J of System Simulation*, 2002, 12(5): 429-434.)

[3] Young S J. Competitive Training in Hidden Markov Models [A]. *Proc ICASSP* [C]. Cambridge: Cambridge University, 1990: 681-684.

[4] 谢锦辉. ***隐Markov 模型(HMM) 在语音处理中的运用*** [M]. **武汉: 华中理工大学出版社**, 1995.
(Xie J H. *The Application of HMM in Speech Processing* [M]. Wuhan: Huazhong University of Science and Technology Press, 1995.)

[5] Satish L, Gururaj B I. Use of Hidden Markov Models for Partial Discharge Pattern Classification [J]. *IEEE Trans on Electrical Insulation*, 1993, 28(2): 172-182.

[6] Gwendolyn H Walton, Robert M Patton, Douglas J Parsons. Usage Testing of Military Simulation Systems [A]. *Proc of the Winter Simulation Conf* 2001 [C]. Arlington, 2001: 771-779.

[7] Don Caughlin. An Integrated Approach to Verification, Validation, and Accreditation of Models and Simulations[A]. *Proc of Winter Simulation Conf* 2000 [C]. Orlando, 2000: 872-881.

[8] Seong Kyu Yoon, John F MacGregor. Fault Diagnosis with Multivariate Statistical Models, Part Ⅰ: Using Stead State Fault Signatures[J]. *J of Process Control* Ⅱ, 2001, 11(4): 387-400.

[9] Balci O. How to Access the Acceptability and Credibility of Simulation Results[A]. *Proc of Winter Simulation Conf* [C]. Orlando, 1989: 62-71.

[10] Ephrainm Y, Rabiner L R. On the Relation between Modeling Approaches for Information Source [A]. *Proc ICASSP* [C]. New York, 1988: 24-27.

[11] 李健, 王作英. HMM 转移概率的新的重估算法[J]. ***电子学报***, 2001, 29(12A): 1832-1835.
(Li J, Wang Z Y. A New Estimation Algorithm of HMM's Transition Probability [J]. *Acta Electronica Sinica*, 2001, 29(12A): 1832-1835.)

(上接第566 页)

[2] 郑斌祥, 席裕庚, 杜秀华. 基于离群指数的时序数据离群挖掘[J]. ***自动化学报***, 2004, 30(1): 70-77.
(Zheng B X, Xi Y G, Du X H. Outlier Mining for Time Series Data Based on Outlier Index [J]. *Acta Automatica Sinica*, 2004, 30(1): 70-77.)

[3] Breuning M, Kriegel H, Ng R. Optics of: Identifying Local Outliers [A]. *Proc of the 3rd European Conf on Principles and Practice of Knowledge Discovery in Databases*[C]. Prague, 1999: 262-270.

[4] Dorigo M, Maniezzo V, Colorni A. Ant System: Optimizationby a Colony of Cooperating Agents [J]. *IEEE Trans on System, Man and Cybernetics B*, 1996, 26(1): 29-41.

[5] Tsai C F, Tsai C W, Wu H C. Acodf: A Novel Data Clustering Approach for Data Mining in Large Databases [J]. *J of Systems and Software*, 2004, 73(3): 133-145.

[6] Kuo R J, Kuo Y P, Chen K Y. Developing a Diagnostic System through Integration of Fuzzy Case-based Reasoning and Fuzzy Ant Colony System [J]. *Expert Systems with Applications*, 2005, 28(6): 783-797.

[7] He Z Y, Xu X F, Deng S C. Discovering Cluster-based Local Outliers [J]. *Pattern Recognition Letters*, 2003, 24(12): 1641-1650.