Scientific Journal of Control Engineering
December 2013, Volume 3, Issue 6, PP.384-390

The Process Monitoring Based on Orthonormal Function Approximation for Synchronization of Multivariate Trajectories

Xiang Gao†, Wenfeng Yi, Yumei Sun

Computer & Electrical Automation School, Yantai Nanshan University, Longkou Shandong 265713, P.R. China

†Email: wowisee@163.com

Abstract

On batch process, Multiway Independent Component Analysis can construct multivariate statistical model from three-dimensional non-gaussian distribution data of different batches to extract the independent information components from the model to find out latent faulty information after projection calculation so as to improve the ability of process monitoring. However, in fact, in most cases, some asynchronous batches have no same duration so that the MICA model cannot be built correctly as imagined. The approach of Orthonormal Function Approximation (OFA) changes all the variables of reference trajectories of batches in turn into a series of coefficients with same numbers to form two-dimensional model for monitoring. After having setting several faulty batches, the Penicillin fermentation process simulator (Pensim V2.0) is experimented and calculate indices of SPE and I^2 to show that the MICA monitoring based on OFA could detect the abnormal of process easier than the one of the MICA monitoring on the original unsynchronized model.

Keywords: Batch Process; Synchronization; MICA; OFA; Pensim V2.0

基于正交方程近似的多元轨迹同步化过程监控*

高翔，衣文凤，孙玉梅

烟台南山学院 计算机与电气自动化学院，山东 龙口 65713

摘　要：在间歇过程中，多向独立元分析方法（MICA）能够对不同批次的非正态分布的各变量三维数据列出多元统计模型，从建立的统计模型中提取出互为独立的信息元，经过计算发现其中的隐含的故障信息，从而改善过程监控能力。然而，事实上，在某些反应时间比较长的间歇过程中，由于无法保证各批次的反应速率都一致，这样的异步导致无法准确建立 MICA 模型。正交函数近似（OFA）方法，将各批次轨迹变量经过正交变换为按每一变量都相同个数的一系列系数的二维模型，并将其应用于监控中。用青霉素的发酵模拟软件（Pensimv2.0）进行仿真，通过某些批次设置故障与计算两种指标 SPE 和 I^2 比对，结果证明，与未同步的简单 MICA 建模相比，基于 OFA 的 MICA 过程监控能相对准确地发现过程中的异常。

关键词：间歇过程；同步化；多向独立元分析法；正交函数近似；青霉素发酵模拟软件 2.0 版

引言

现代大型过程工业生产具有数据繁杂，变量多且相互关联等特点，使得操作人员不易对每个变量逐个进行监视，即便照顾到每个变量忽视它们之间的关系也很难发现早期异常[1]。多元统计过程控制方法中，具有代表性的主元分析法（Principal Component Analysis, PCA）将过程数据从高维数据空间投影到低维特征空

*基金资助：受山东省教育厅科研项目资助（J11LG73）。

间，摒弃了冗余信息，抓住相关关系是否发生变化的主要矛盾，从而能发现隐藏在表征过程特征的如温度、压力等变量内部的问题。即使这些变量正常，如果经计算相互关系异常，则有可能是存在故障隐患[2]。多向主元分析（Multiway PCA）将主元分析法扩展到间歇过程，对待激烈变化的时变过程，MPCA 不需要精确数学模型，只需要历史数据，利用测试数据投影于历史数据的模型的方式鉴定有无差别[3]。但 MPCA 默认条件是基于两种假设：1）建模数据服从正态分布；2）间歇过程的各批次反应时间长度相同。

而各数据批次之间的关系并不一定满足正态分布。独立元分析法（Independent Component Analysis, ICA）很好地解决了这个问题：与 PCA 算法不同的是：没有把原始变量简单地组合为相互正交的潜隐的主元变量；而是把它们更细地划分为相互独立的独立元变量，这样就更能接近问题的实质。PCA 方法从二阶统计量均值和方差中寻找信息；而 ICA 则能包含更高阶的统计量，因而独立元数目比主元数目更少，统计量计算结果令判断更直观[4][5]。而且，独立元分析算法在近年来成为各学科研究的热点，不仅用于故障检测[6]，而且还用于信号分离等[7] [8]。MICA 将 ICA 扩展到间歇过程中，同样收到了良好的效果[9]。

因为初始的配方差异，气温的变化及容器结垢等使容量尺寸变化种种客观原因，有时候甚至是操作者有意调整反应时间以使产品达到某种标准，各批次的持续时间并不一样。为了解决不同操作时间的建模和监控等问题，已经有许多方法：Shah 等人提出的基于 PCA 的方法，这种方法只有在 PCA 轨迹都具有全局线性这个假设下成立；Lakshminarayanan 等人提出扩展每一族多元轨迹与最长的多元轨迹的持续时间相匹配，短的轨迹部分用该轨迹最后采集到的数据值补齐，这样的做法就是片面认为系统在最后基本不再发生变化；Kourti 提出指示变量法，但很难找到这样一个满足要求的单变量来代表所有的变量；Kassidas 等人将 DTW 理论应用于 MPCA[10]，但其基于动态规划的路径寻优算法实现起来，相当耗用计算时间；Junghui Chen 和 Jialin Liu 提出的 OFA 理论，将多元轨迹按每变量投影到正交函数，得到的同样数量的近似系数，以这些历史轨迹的历史系数建立 MPCA 模型，以测试数据的系数投影到模型中，如同传统 MPCA 一样，以计算统计量分析有无异常[11][12]。其特点是在准确的基础上降低了计算量。

本文将 OFA 方法和 MICA 方法结合，对于各批次异步的青霉素的发酵过程，正交投影后的近似系数组建 MICA 模型并用新批次来测试性监控，并以此结果与未同步的简单处理的 MICA 模型相比，以证明 OFA 方法的同步化优点。

1 多向独立元分析方法

1.1 独立元分析原理

1.1.1 独立元计算

在独立元分析中，原始数据矩阵 X（$d\times n$），d 为变量的个数，n 为系统采样个数。独立元分析的任务是从 X 数据矩阵中估计出独立元矩阵 S（$m\times n$）或者混合矩阵 A（$d\times m$），m 为独立元的个数。三个矩阵的关系为[4] [5]：

$$X = AS + E \tag{1}$$

$E \in R^{d\times n}$ 为未被独立元包括的部分，是残差矩阵。独立元分析方法的首要目标是估计出分离矩阵 W，近而推测出独立元矩阵 $\hat{S}$，以使每个独立元尽量独立，有式（2）的关系：

$$\hat{S} = WX \tag{2}$$

独立元分析的初始化是对建模数据 X 进行“漂白”处理，其协方差矩阵可表示为：$R_x = X^T X(n-1)$，对其进行特征值分解：$R_x = U\Lambda U^T$。漂白后的数据为：

$$Z = QX = QAS = BS \tag{3}$$

其中，$Q = \Lambda^{-1/2}U^T$。由于漂白后的数据矩阵 Z 的协方差矩阵是单位阵，故 B 是是一个正交矩阵。如果 B 给定，可以估计出独立元矩阵：

$$\hat{S} = B^T QX \tag{4}$$

从公式（3）和（4）我们可以得到 W 和 B 的关系：

$$W = B^T Q \tag{5}$$

因此，独立元分析的焦点问题是要估计出 B。为估计有效地偏离非高斯分布的特性，需要计算负熵 J(y)。Hyvarinen 提出一种简单近似的方法：

$$J(y) \approx [E\{G(y)\} - E\{G(v)\}]^2 \tag{6}$$

其中，列向量 y 假设其均值为零，方差为单位方差，v 是具有零均值单位方差的高斯变量，G 是任意非二次方的函数。通过广泛的选择 G 可以获得很好的负熵近似值。Hyvarinen 提供了一种函数比较适合本文的运算：

$$G_1(u) = \frac{1}{a_1} \log \cosh(a_1 u) \tag{7}$$

其中，$1 \le a_1 \le 2$。Hyvarinen 还介绍了一个用于 ICA 的快速简单有效的迭代方法(Fast ICA)，这种方法通过首先随机地设置独立元 B 的列向量，通过迭代，计算正交和归一化，将使 B 逐步满足负熵的条件。计算过 。可以用一种非参数实验密度估计－－核密度估计，去定义一个名义上的操作区域。核密度估计可以表示为：

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{x - x_i}{h}\right\} \tag{8}$$

其中，x 是在考虑范围内的数据点，x_i 是数据集当中的一个观测值，h 为窗口高度（也称平滑参数），n 为观测值的数量，K 为核函数，本文选择指数函数。选择多大的平滑参数 h 在估计中至关重要。如果 h 太大则过于平滑，将失去细节。如果太小则太不平滑，将不能滤掉尖峰。可以用交叉检验法选取 h。

1.1.4 多向独立元分析

不同批次的各变量的有限操作周期的数据可以按批次×变量×时间的顺序堆放在一起形成三维数据矩阵 $\underline{X}$。然后用 Nomikos & Macgregor 的方法将三维数据展开成二维：垂直于时间轴，将三维数据切成多个批次×变量的数据片，并沿时间轴按时间顺序从左至右依次展开，形成一个超宽的矩阵。MICA 将三维的数据矩阵 $\underline{X}$ 分解为独立元向量 s_r、负荷矩阵 A_r 和残差矩阵 $\underline{E}$ 的乘积并使得独立元矩阵中的独立元互相独立[6]：

$$\underline{X} = \sum_{r=1}^{d} s_r \otimes A_r + \underline{E} = \sum_{r=1}^{d} s_r a_r^T + E = \hat{X} + E \tag{9}$$

其中，$\otimes$ 表示 Kronecker 乘积，d 表示保留的独立元个数。I^2 统计量是独立得分矩阵的平方和，定义如下：

$$I^2(i) = \hat{s}_d(i)^T \hat{s}_d(i) \tag{10}$$

其中，$\hat{s}_d(i)$ 为 $\hat{s}_d$ 的第 i 个列向量。独立元矩阵的第 i 个元素描述了第 i 个矩阵并且将这一批次的所有变化加和起来，此批次在整个运行时间里与其它在历史数据库中的批次存在某种关系。

对于某一批次 i 的 SPE（平方预测误差）统计量，定义如下：

$$SPE(i) = e(i)^T e(i) = (x(i) - \hat{x}(i))^T (x(i) - \hat{x}(i)) \tag{11}$$

其中，$\hat{x}$ 可以用下列公式计算：

$$\hat{x}(i) = Q^{-1} B_d \hat{s}(i) = Q^{-1} B_d W_d x(i) \tag{12}$$

这里，B_d 和 W_d 分别是 B 和 W 取了 d 个独立元的结果。

2 正交函数近似

然而，多向独立元分析算法的假设是各批次的运行时间长度是一致的。当假设不成立时，不能直接建

模。可以将这些多元轨迹经过正交基投影变换为一系列正交系数来近似表征原有轨迹的过程变化，以达到各变量每个变量的系数数目相同。设各批次的多元轨迹的每个变量轨迹可以看成是一个正交函数 $F(t)$，这正交函数可以用一组正交基$\{\varphi_n\}$线性表示[8][9]。

$$F(t) \cong F_n(C,t) = \sum_{n=0}^{N-1} \alpha_n \varphi_n(t) \tag{13}$$

其中，$C = \{\alpha_n\}$是正交系数，$\alpha_n = \int F(t)\varphi_n(t)dt$是函数 $F(t)$在每一个基函数$\{\varphi_n\}$上的投影值。而正交分解的步骤可以表示为：

$$\begin{aligned} &E_0(t_k) = F(t_k) \\ &\alpha_n = [\Phi_n^T \Phi_n]^{-1} \Phi_n^T \mathbf{E}_n \\ &E_{n+1}(t_k) = E_n(t_k) - \alpha_n \varphi_n(t_k) \\ &k = 1,2,\ldots,K_i \\ &n = 0,1,\ldots,N-1; i = 1,2,\ldots,I \end{aligned} \tag{14}$$

这里，$\mathbf{E}_n = [E_n(t_1), E_n(t_2), \ldots, E_n(t_{K_i})]^T$，为第 n 次正交后的余差，$\Phi_n = [\varphi_n(t_1), \varphi_n(t_2), \ldots, \varphi_n(t_{K_i})]^T$，为第 n 次的正交基。

勒让德多项式基函数法可以计算各阶正交基$\{\varphi_n\}$。而判断正交分解的收敛性指标为：

$$G(N) = \alpha_{N-1}^2 / \sum_{n=0}^{N-1} \alpha_n^2 \tag{15}$$

可以人为设定 $G(N)$的百分比，在该变量 j 的多个批次共同正交分解时，某些批次还未满足式（15），而其它批次早已符合式（15），但必须随之继续运行，直到所有批次共同满足收敛性指标为止。

$$N_j = \max_i \{N_{ij}\} \tag{16}$$

求出各变量的正交系数后，可以将它们列成一个矩阵方式：

$$\Theta = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,J} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,J} \\ & & \vdots & \\ C_{I,1} & C_{I,2} & \cdots & C_{I,J} \end{bmatrix} \tag{17}$$

其中，$C_{i,j} = [\alpha_{i,0}, \alpha_{i,1} \ldots \alpha_{N_j - 1}]$，式（17）中不同的变量可以有不同个数的正交系数，而同一变量（即每一列）之间，分解的系数按式（16），必须次数相同。这样，正交函数变换将各不等长的变量轨迹提取出等长的正交系数，可以把式（17）中的 Θ 视为 PCA 中的原始数据矩阵 X，运用 PCA 算法解决过程监控问题。

3 控制指标 SPE 和 I^2 统计量

在 MICA 中，从过程模型中推出两种统计量：对于过程变量的系统部分的 D 统计量和对于过程变量残差部分的 Q 统计量。对于某一批次 i 的 Q 统计量也叫做 SPE 统计量，定义如下[6]：

$$SPE(i) = e(i)^T e(i) = (x(i) - \hat{x}(i))^T (x(i) - \hat{x}(i)) \tag{18}$$

其中，$\hat{x}$ 可以用下列公式计算：

$$\hat{x}(i) = Q^{-1} B_d \hat{s}(i) = Q^{-1} B_d W_d x(i) \tag{19}$$

对于某个批次 i 的 D 统计量也叫做 I^2 统计量，它是独立得分矩阵的平方和，定义如下：

$$I^2(i) = \hat{s}_d(i)^T \hat{s}_d(i) \tag{20}$$

其中，$\hat{s}_d(i)$ 为 $\hat{s}_d$ 的第 i 个列向量。独立元矩阵的第 i 个元素描述了第 i 个矩阵并且将这一批次的所有变化加和起来，此批次在整个运行时间里与其它在历史数据库中的批次存在某种关系。

这里，I^2 值是用来监测在 MICA 模型子空间中由于不正常的变化而引起的故障，而，SPE 值是用来监测在 MICA 模型子空间中没有包含进去的新事件故障。

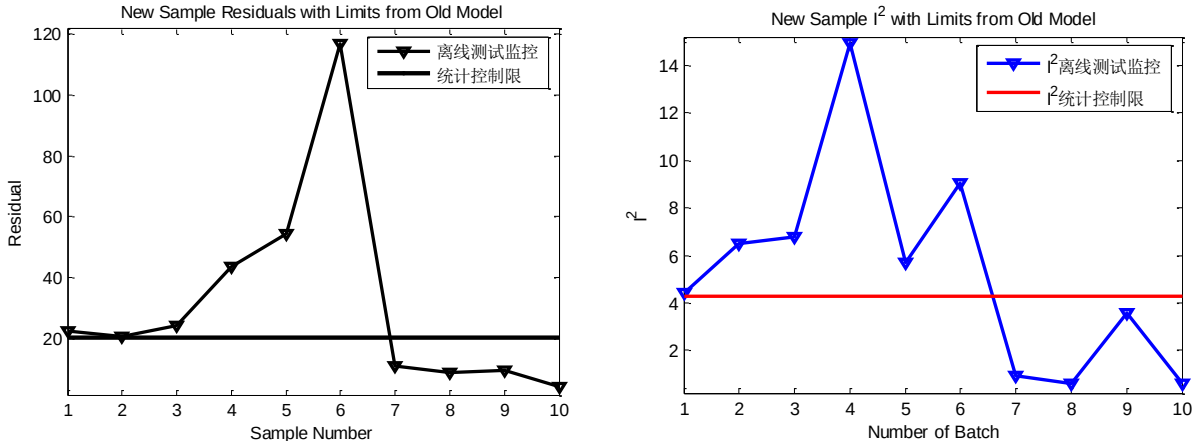
两种控制指标的置信控制限由于不受任何分布限制，所以不能直接从某一分布计算而来。可以用核密度估计来计算，这里从略。

4 青霉素发酵过程监控仿真

仿真的数据来源于青霉素发酵模拟软件 PenSim v2.0[13]，共采集了 60 个正常运行的批次，各批次的仿真时间介于 350~450h，采样时间是 1h，并从 PenSim2.0 软件所模拟的 18 个过程变量的数据中选取出与产品浓度相关的 10 个变量的数据用于建模。选用的 10 个变量如下：通风速率，鼓风机功率，葡萄糖反馈温度，溶解氧浓度饱和度，培养容积，二氧化碳浓度，PH 值，温度，产生热量，冷水速率。

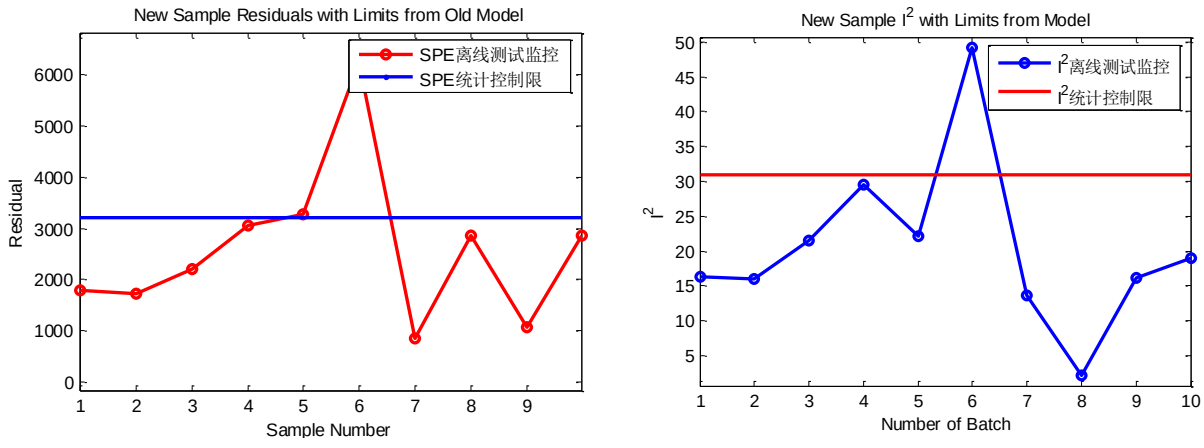
测试批次中，设置了 6 个故障批次：1~6 批次是故障数据，7~10 批次是正常数据。故障批次 1、2 为变量 2 故障，阶跃幅值 10%，斜坡幅值 1%，故障时间从 45h 到 200h，属于早、中期故障；3、4 批次为变量 3 故障，阶跃幅值 65%、斜坡幅值 0.1%，故障时间从 300h 到 400h，属晚期故障；5、6 批次为变量 3 故障，阶跃幅值 70%、斜坡幅值 0.1%，故障时间从 45h 到 150h，属早期故障。

仿真采用了经同步化方法和非同步化的 MICA 监控效果对比：用 OFA 同步后的正交系数的 MICA 方法；简单地将轨迹长度规定到 400 小时的 MICA 方法，其中短的轨迹用采样最后值延续到第 400 小时，长的轨迹就截去。OFA 同步后系数的 MICA 需要 4 个独立元就可以解释过程变化，而不用同步的处理方式的 MICA 则需要 44 个独立元。



(a)SPE 检验　　(b)I^2 检验

图 1 经 OFA 同步的 MICA 的两种统计量



(a)SPE 检验　　(b)I^2 检验

图 2 未同步化的 MICA 的两种统计量

过程监控中某批次如果满足条件 SPE 和 I^2 的计算结果都不超限，则视为正常；否则，该批次就存在异

常。从图 1 看到，OFA 同步化后，正交系数的 MICA 的 SPE 和 I^2 都检验出了 1~6 批次的故障；从图 2 看到，未同步化的简单处理方法的 MICA 的 SPE1~4 批次漏检，I^2 则 1~5 批次漏检。说明未同步化轨迹建立的 MICA 模型并不能通过投影方式发现故障模式；同时也说明 OFA 的正交投影得到的近似系数基本上能够简单表达过程的变化模式。

5 结论

本文中使用到的正交近似方法，将某一批次的其中一个变量的轨迹，通过一个正交函数变换压缩为一系列近似系数，使得各批次同一变量的相互同步，同时，此近似系数表征了该变量在这一批次的变量的运动特征；接续的 MICA 方法将其中相互关联的变量关系转化为数目较少的几个相互独立的独立元，进一步对过程数据降维，以搜寻到过程的主要变化是否超出误差指标。两种方法的结合有利于分析复杂的间歇过程的监控结果。准确快速的监控方法对后续过程的质量指标的判断提供了依据。设想进一步的工作是：基于本文的离线研究工作基础上，对于过程数据暂时缺失的在线间歇过程监控，需要研究合适的方法，以及时发现问题。

REFERENCES

[1] Kresta James, John F. MacGregor and Thomas E. Marlin. “Multivariate statistical monitoring of Process Operating Performance.” Canadian Journal of Chemical Engineering, February 1991, 69(1): 35-47

[2] J. Edward Jackson and Gavind S. Mudhalkar. “Control Procedures for Residuals Associated with Principal Component Analysis.” Technometrics, August 1979, 21(3): 341-349

[3] Paul Nomikos and John F. MacGregor. “Monitoring of Batch processes Using Multi-Way Principal Component Analysis.” AIChE Journal, August 1994, 40(8): 1361-1375

[4] Anne-Sophie Carpentier, Alessandra Riva, Pierre Tisseur, et al. “The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA.” Computational Biology and Chemistry, 2004, 28: 3-10

[5] Jong-Min Lee, Chang Kyoo Yoo, In-Beum Lee. “Statistical process monitoring with independent component analysis.” Journal of Process Control, August 2004, 14(5): 467-485

[6] Yingwei Zhang, Yang Zhang. “Fault detection of non-Gaussian processes based on modified independent component analysis.” Chemical Engineering Science, August 2010, 65(16): 4630-4639

[7] Tang Ying, Li Jian-Ping, Wu huai. “A simple and Accurate ICA Algorithm for Separating Mixtures of up to Four Independent Components.” Acta Automatica Sinica, 2011, 37(7): 794-799

[8] D. Jouan-Rimbaud Bouveresse, A. Moya-Gonzalez, F. Ammari, D.N. Rutledge. “Two novel methods for the determination of the number of components in independent components analysis models.” Chmometrics and Intelligent Laboratory Systems, March 2012, 112: 24-32

[9] Chang kyoo Yoo, Jong-Min Lee, Peter A. Vanrolleghem, et al. “ On-line monitoring of batch processes using multiway independent component analysis.” Chemometrics and Intelligent Laboratory System, 2004, 71: 151-163

[10] Athanassios Kassidas, John F. MacGregor, Paul A. Taylor. “Synchronization of batch trajectories using dynamic time warping.” AIChE Journal, April 1998, 44(4): 864-875

[11] Junghui Chen and Jialin Liu. “Post Analysis on Different Operating Time Processes Using Orthonormal Function Approximation and Multiway Principal Component Analysis.” Journal of Process Control, 2000, 10: 411-418

[12] Junghui Chen and Jialin Liu. “Multivariate Calibration Models Based on Fuctional Space and Partial Least Square for Batch Processes.” 4th IFAC workshop on On-line fault detection and supervision in the chemical process industries, Jejudo Korea, June 7-9, 2001, published: Seoul Korea, Romagnoli Jose, 2001, 161-166

[13] Cenk Ündey, Gülnur Birol and Ali Cinar, Pensim v2.0: Stand-alone Dynamic Simulator for Fed-Batch Penicillin Production, download site: http://www.chee.iit.edu/~cinar/

【作者简介】



[1] 高翔（1967-），男，汉族，博士学位，副教授，研究方向为复杂工业过程的建模与控制、故障诊断，学习经历：1989 年毕业于吉林化工学院生产过程自动化专业，学士学位；1998 年毕业于东北大学控制理论与控制工程专业，硕士学位；2001 年毕业于中国科学院沈阳自动化研究所机械电子工程专业，博士学位。Email: wowisee@163.com

[2] 衣文凤（1982-），女，汉族，硕士学位，讲师，研究方向为模式识别与人工智能，学习经历：2006 年毕业于山东大学自动化专业，学士学位；2009 年毕业于山东大学，控制理论与控制工程专业，硕士学位。

Email: sucui660870.student@sina.com

[3] 孙玉梅（1974-），女，汉族，硕士学位，讲师，研究方向为数据采集与处理，学习经历：1997 年毕业于成都理工学院应用电子技术专业，学士学位；2009 年毕业于青岛大学材料学专业，硕士学位。Email: sun 5185@163.com