基于 KPCA 与 LS－SVM 的化工过程故障诊断算法研究

解　庆[1]，杨　武[2]，赵小强[2]

（1. 甘肃蓝科石化高新装备股份有限公司，兰州 730070；
2. 兰州理工大学 电气工程与信息工程学院，兰州 730050）

摘要： 针对核主元分析方法（KPCA）在复杂化工在线监控过程中初始故障源难以辨识的问题，该文提出了一种基于核主元分析和最小二乘支持向量机的集成故障诊断方法。该方法首先运用 KPCA 对数据进行预处理，在特征空间构建 T^2 和 *SPE* 来检测故障的发生，然后计算样本的非线性主元得分向量，将其作为最小二乘支持向量机的输入值，通过最小二乘支持向量机的分类进行故障类型的识别。将上述故障诊断方法应用到 Tennessee Eastman（TE）化工过程，多种故障模式下的仿真结果表明，该方法不但能有效地辨识故障，而且提高了故障检测和故障诊断的速度。

关键词： 化工过程；故障诊断；核主元分析；最小二乘支持向量机

中图分类号： O212.4　　**文献标志码：** A　　**文章编号：** 1000－0682（2012）05－0003－05

A fault diagnosis algorithm of chemical industry process based on KPCA and LS－SVM

XIE Qing[1]，YANG Wu[2]，ZHAO Xiaoqiang[2]

（1. *Lanpec Technologies Limited*，*Lanzhou* 730070，*China*；
2. *College of Electrical and Information Engineering*，*Lanzhou University of Technology*，*Lanzhou* 730050，*China*）

Abstract: When kernel principal component analysis（KPCA）method is applied to the on-line monitoring of complex chemical industrial process，primary fault sources are usually difficult to identify. So an integrated fault identification method based on KPCA and least squares support vector machine（LS－SVM）is proposed. First the data are analyzed by using KPCA，T^2 and *SPE* are constructed in the feature space for detecting faults. If T^2 and *SPE* exceed the predefined control limits，a fault may have occurred. Then nonlinear principal component score vectors of the samples are calculated and inputted into least squares support vector machine to identify the faults through least squares support vector machine classification. The proposed method is applied to the Tennessee Eastman（TE）chemical industry process. Simulation results of multiple fault modes demonstrate that the proposed method can not only effectively identify various types of fault sources，but also improve speed of fault detection and diagnosis.

Key words: chemical industry process；fault diagnosis；kernel principal component analysis；least squares support vector machine

0　引言

化工生产对于国民经济和人们的日常生活有着举足轻重的作用，同时化工过程也存在着许多不安全因素和职业危害，因此故障诊断是化工过程中的一个重要问题。对设备的运行状态进行准确、有效地检测和故障诊断，具有十分重要的意义。早期检测和诊断过程故障，主要是通过工厂确保在一个可控制的范围内开车，从而避免异常事件的发生和降低生产过程的损失。美国统计发现，每年石化产业损失大约 20 亿美元，因此故障诊断是迫切需要解决

收稿日期：2012－03－23

基金项目：甘肃省自然科学基金项目（1112RJZA028）；甘肃省教育厅硕士生导师项目（1003ZTC085）

作者简介：杨武（1987），男，江西吉安人，硕士研究生，研究方向为故障诊断。

的问题。现在,大量的科研人员对这一领域的研究产生了相当大的兴趣。

故障诊断技术经过几十年的发展,已经出现了基于各种不同原理的众多方法。其中基于数据驱动的故障诊断技术是比较重要的一种。主元分析方法是研究和应用最多的一种基于数据驱动的化工过程故障诊断方法,并在非线性、动态、多尺度等方面被不断地改进和扩展。Scholkopf[1] 等人提出的核函数主元分析法,能有效地提取故障特征,实现故障的检测。Cho[2] 和 Choi[3] 对非线性系统中的故障识别进行了一定研究。SangWook Choi[4] 进一步提出了基于动态 KPCA 的非线性过程的监控方法。然而在故障检测方面,存在难以直接诊断故障源的问题。薄翠梅[5] 和刘晶晶[6] 将 KPCA 与 PNN 结合对故障进行分类识别,虽然取得了一定的研究成果,然而神经网络的统计规律只有当训练样本接近无限大时才能准确地被表达,在处理故障诊断等实际问题时,只能得到非常有限的故障样本。支持向量机(SVM)由于遵循了结构风险最小化原理,且其可以将非线性问题转化为线性问题并能得到全局最优解,避免了人工神经网络等方法网络结构难以确定、过学习、欠学习、局部最小化等问题,近年来在故障诊断领域得到了广泛的应用[6-8]。

由于 KPCA 和 SVM 结合能充分发挥各自的优势,提高故障诊断的实时性。而 LS-SVM 相比传统的 SVM 在保证识别精度的同时能够降低计算的复杂性,加快识别的速度。该文提出了基于核主元分析(KPCA)和最小二乘支持向量机(LS-SVM)的集成故障诊断方法,并在 TE 化工过程的多故障模式下验证了该方法的有效性。

1 核主元分析(KPCA)

传统的主元分析法是基于原始特征的一种非线性变换,当原始数据存在非线性属性时,使用 PCA 分析后留下的显著成分就不能反映这种非线性特性。而核主元分析利用非线性变换将输入空间映射到高维特征空间,转化为求核矩阵的特征向量和特征值,输入数据在特征向量上的投影转化为求核函数的线性组合,大大简化了计算。

设样本集 $X=\{x_1,x_2,\cdots,x_N\}$,其中 $x_k\in R^m$,N 为样本总数,通过非线性映射 Φ 将输入数据从原空间映射到高维特征空间 F,记为 $\Phi(x_k)$,假设满足 $\sum_{k=1}^{N}\Phi(x_k)=0$,则 KPCA 可看作是在高维特征空间 F 中对协方差矩阵:

$$\boldsymbol{C}^F=\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)\,\Phi(x_i)^{\mathrm{T}} \tag{1}$$

对矩阵 $\boldsymbol{C}^F$ 做特征矢量分析,设其特征值为 λ,特征矢量为 $\boldsymbol{V}$,则 $\lambda\boldsymbol{V}=\boldsymbol{C}^F\boldsymbol{V}$。

$\boldsymbol{C}^F$ 的特征矢量 $\boldsymbol{V}$ 可表示为:

$$\boldsymbol{V}=\sum_{i=1}^{N}a_i\Phi(x_i) \tag{2}$$

通过计算映射数据在特征矢量 $\boldsymbol{V}_k$ 上的投影来计算主元,即:

$$t_k=\langle\boldsymbol{V}_k,\Phi(x)\rangle=\sum_{k=1}^{N}a_i^k\langle\Phi(x_i),\Phi(x)\rangle \tag{3}$$

这里 $\langle x,y\rangle$ 表示 x 与 y 的点积,为了避免直接计算非线性映射,在特征空间定义核函数矩阵 $\boldsymbol{K}=[h_{ij}]_{N\times N}$,其中 $h_{ij}=\langle\Phi(x_i),\Phi(x_j)\rangle$,核函数的选择完全决定映射 Φ 和特征空间 F。核函数主元分析的详细求解过程参见文献[9]。

2 最小二乘支持向量机(LS-SVM)

2.1 支持向量机(SVM)

支持向量机是建立在结构风险极小化的基础上,能较好地解决小样本、非线性、高维数和局部极小点等实际问题的一项有效技术[10]。

设有训练样本集 $D=\{(x_i,y_i)\}$,$(i=1,2,\cdots,m)$,$x_i\in R^m$,$y_i\in R$,x_i 为支持向量机的输入数据,y_i 为对应的输出数据。利用非线性映射 $\Phi(x_i)$ 将样本从原空间映射到高维特征空间 F,在特征空间中构造最优分类的超平面。为了保证分类的准确性,引入松弛变量 $\varepsilon_i\geqslant 0$。要实现最优超平面,对所有的训练样本数据能够正确分类,必须满足:

$$y_i\left[\sum_{i=1}^{N}\omega_i\Phi(x_i)+b\right]\geqslant 1-\varepsilon_i \tag{4}$$

根据结构风险最小化原则,通过优化式(5)的二次规划进行分类

$$\begin{aligned}&\min_{\omega,b,\varepsilon}L=\min_{\omega,b,\varepsilon}\left\{\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{N}s_i\varepsilon_i\right\}\\&\text{s.t. }y_i\left[\sum_{i=1}^{N}\omega\Phi(x_i)+b\right]-1+\varepsilon_i\geqslant 0\end{aligned} \tag{5}$$

式中:c 为对超出控制误差的样本惩罚程度;s_i 为加权系数。

利用 Lagrange 方法进行求解,建立 Lagrange 函数

$$\begin{aligned}L=&\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{N}s_i\varepsilon_i-\beta\varepsilon_i-\\&\sum_{i=1}^{N}\alpha_i\{y_i\left[\omega^{\mathrm{T}}\Phi(x_i)+b\right]-1+\varepsilon_i\}\end{aligned} \tag{6}$$

式中:α_i,β_i 为 Lagrange 乘子。

对上式中的 ω、b 求导，最终求解得到最优分类函数 $f(x)$

$$f(x) = \mathrm{sgn}\left[\sum_{i=1}^{N} \alpha_i y_i \Phi(x, x_i) + b\right] \tag{7}$$

该算法就是通过某种非线性映射，将输入向量映射到一个高维特征空间，在这个特征空间中构造最优分类超平面。

2.2　最小二乘支持向量机(LS－SVM)

最小二乘支持向量机[10]是标准支持向量机的一种扩展。LS－SVM 是基于正规化理论对标准 SVM 的改进，在目标函数中采用二次损失函数代替 SVM 中的不敏感损失函数，将 SVM 中的二次规划问题转化为线性方程组求解，在保证精度的同时大大降低了计算复杂性，加快了求解速度。

设有训练样本集 $D = \{(x_i, y_i)\}$，在原始空间中的函数问题可以描述为求解下面问题:

$$\min_{\omega,b,\varepsilon} L = \min_{\omega,b,\varepsilon}\left\{\frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{N}\varepsilon_i^2\right\}$$

$$\text{s.t. } y_i\left[\sum_{i=1}^{N}\omega_i\Phi(x_i) + b\right] = 1 - \varepsilon_i \tag{8}$$

式中: γ 为惩罚因子，ε_i 为误差变量，b 是偏差量。

引入 Lagrange 系数 α_i，定义如下的 Lagrange 函数:

$$L = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{N}\varepsilon_i^2 - \sum_{i=1}^{N}\alpha_i\{y_i[\omega^{\mathrm{T}}\Phi(x_i) + b] - 1 + \varepsilon_i\} \tag{9}$$

根据 Mercer 条件，存在映射 Φ 和核函数 K，使得 $K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}}\Phi(x_j)$。令 L 对变量 ω，b，ε_i，α_i 的偏导数等于零，并将得到的等式代入式(9)，可以得到矩阵方程:

$$\begin{bmatrix} 0 & 1_V^{\mathrm{T}} \\ 1_V & \Omega + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{10}$$

其中: $y = [y_1, \cdots, y_N]$，$1_V = [1, \cdots, 1]$，$\alpha = [\alpha_1, \cdots, \alpha_n]$，$\Omega$ 中的元素为 $\Omega_{ij} = K(x_i, x_j)$。

求解矩阵方程(10)，得到最后的最小二乘支持向量机的分类决策函数为:

$$y(x) = \mathrm{sgn}\left\{\sum_{i=1}^{N}\alpha_1 K(x, x_i) + b\right\} \tag{11}$$

其中: α，b 由式(10)求解。

LS－SVM 通过解决线性等式集代替了经典 SVM 二次规划。分类面主要是由支持向量决定，而且支持向量的个数一般也是远少于训练向量总数的，因此 LS－SVM 算法一方面使求解难度大大降低，另一方面减少了数据运算量，大幅度提高了算法的运行速度。

3　基于 KPCA 和 LS－SVM 的集成故障诊断算法

该集成故障诊断过程包括两部分: 离线训练建模部分和在线实时监控故障检测诊断部分。

3.1　离线训练建模

离线训练建模主要用来建立正常状态下 KPCA 主元模型和建立历史故障分类模型。

3.1.1　建立正常状态下 KPCA 主元模型的步骤

1) 对正常工况下的数据利用小波去噪，并将去噪后的数据进行标准化。

2) 计算给定的一序列 m 维标准化的正常操作条件的数据 $X_k \in R^m (k = 1, \cdots, N)$ 的内核矩阵 $\boldsymbol{K} \in R^{N\times N}$

$$K_{ij} = [K]_{ij} = \langle\Phi(x_i), \Phi(x_j)\rangle = [k(x_i, x_j)]$$

3) 在特征空间为了使 $\sum_{k=1}^{N}\Phi(x_k) = 0$，做如下变换:

$$\breve{K} = K - 1_N K - K1_N + 1_N K 1_N \tag{12}$$

4) 解决 $N\lambda\alpha = K\alpha$ 的特征值问题，并标准化 α_k，使 $\langle\alpha_k, \alpha_k\rangle = 1$:

5) 从正常的操作数据 X_k 中，提取非线性成分:

$$t_k = \langle V_k, \Phi(x)\rangle = \sum_{I=1}^{N} a_i^k \langle\Phi(x_i), \Phi(x)\rangle = \sum_{I=1}^{N}\alpha_i^k \breve{k}(x_i, x) \tag{13}$$

6) 计算正常操作数据的统计量，并按照一定的置信度确定出 T^2 和 SPE 的控制限。

3.1.2　建立历史故障 LS－SVM 故障诊断模型步骤

1) 对每一种历史故障，采集对应的故障样本;

2) 将历史故障数据输入 LS－SVM 分类器中进行训练;

3) 选择合适的核函数，这里选择高斯核函数;

4) 选择合适的模型参数，包括 2 个参数，即惩罚因子 γ 和高斯核参数 σ;

5) 用最小二乘法求解式(10)，得出 Lagrange 系数 α_i 和分类超平面阈值 b;

6) 建立训练样本的最优决策超平面。

3.2　在线监控和故障诊断

1) 对过程采集的测试数据进行去噪并标准化处理;

2) 对给定的 m 维标准化的测试数据 $X_k \in R^m$，通过 $[K]_{ij} = \langle\Phi(x_i), \Phi(x_j)\rangle = [k(x_i, x_j)]$ 计算核矩阵 $\boldsymbol{K}$;

3) 对测试核向量 $\boldsymbol{K}$ 进行均值中心化;

4) 对测试数据 X_k，按式(13)提取非线性成分;

5) 计算测试数据的监控统计量 T^2 和 SPE;

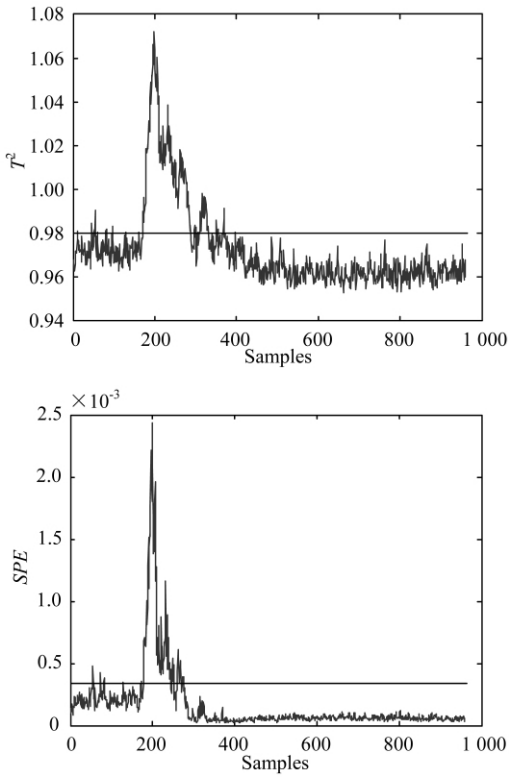
6) 监控 T^2 和 SPE 是否超过控制限,若超过则将故障数据的核主元送到 LS - SVM 分类器进行分类识别;

7) 输出故障信号。

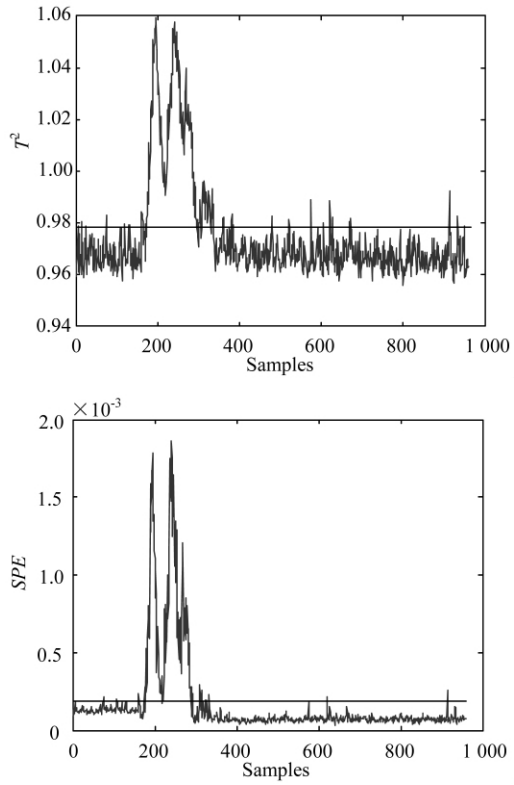
4 仿真结果与分析

将基于 KPCA - LS - SVM 的集成故障诊断方法应用到 TE 化工仿真模型,TE 过程包括 12 个可作为操纵变量的阀门和 41 个可测量变量(包括 22 个连续变量和 19 个分成测量值),过程都包含高斯噪声。TE 过程包括 20 种预先设定好的故障模式,分别代表阶跃、随机变化、慢漂移和粘滞等故障类型,其工艺流程图和详细的过程参见文献[12]。文中采用文献[13]控制方案,进行闭环控制,得到仿真数据(960 × 52 的矩阵)。

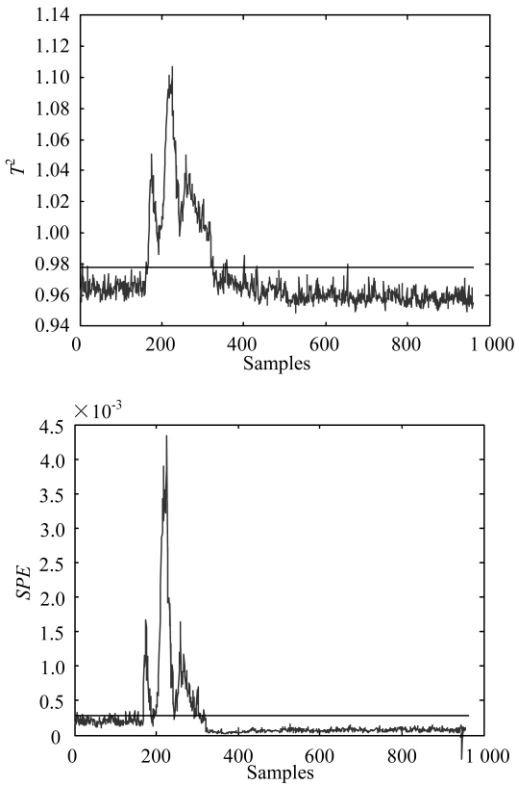
选取故障 1、故障 5、故障 7 作为研究故障模式,分别在 170 采样点附近引入故障,实时监控曲线如图 1 所示。从图中可以看出,统计量 T^2 和 SPE 都可以迅速检测到故障的发生。由于在实际的化工过程中,从故障的排除到工况平稳运行需要一定的时间,所以统计量 T^2 和 SPE 分别在不同程度反映了工况的过渡过程,回落到置信限内的样本点的位置也有所不同。



a 故障 1



b 故障 5



c 故障 7

图 1 故障 1、5、7 模式下的 T^2 和 SPE 监控曲线

如果 T^2 和 SPE 超过控制限,说明有故障发生。选取 52 个变量为条件属性,选取 15 个类别(15 种

不同的故障) 表示决策属性。考虑将故障 1、故障 5 和故障 7 作为故障样本集来分析多类故障识别的情况。试验中标准 SVW 的正则化因子 $C = 10$，高斯核参数 $\sigma_2 = 0.1$，LS - SVM 的惩罚因子 $\gamma = 1$，高斯核参数 $\sigma^2 = 1$。选择决策树型策略作为诊断策略，在不同训练样本和测试样本的条件下，具体诊断结果如表 1 所示。

表 1　TE 过程多类故障诊断结果

训练样本	测试样本	KPCA - LS - SVM		KPCA - SVM	
		分类时间 (s)	识别精度 (%)	分类时间 (s)	识别精度 (%)
360	120	0.62	93.34	4.6	93.24
720	240	5.66	95.33	48.4	94.67
1 200	360	8.04	96.28	189.6	96.33

从表 1 可以看出，当训练样本集增大时，提出的算法在分类时间上要明显优于 KPCA - SVM 算法，诊断速度明显提高。从诊断精度可以看出，在样本不是很大的情况下，两种方法的分类效果差别不大，当训练样本为 360 和 720 时，该算法的诊断精度略有提高。

5　结论

该文针对复杂化工过程实时故障诊断难度加剧问题，提出了一种基于 KPCA - LS - SVM 的故障诊断方法，该方法首先利用 KPCA 提取特征样本，再通过 LS - SVM 对故障类型进行分类识别，有效地解决了 KPCA 在复杂化工过程中故障源难以识别的问题。在 TE 化工模型的多故障模式下的仿真结果表明，上述方法不仅能有效地辨识出初始故障源，而且在保证识别精度的同时大大缩短了故障检测与辨识时间。同时，该方法在核函数的选择及优化策略方面还需进一步研究。

参考文献:

[1] SCHOLKOPF R, SMOLA A J, MULLER K. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998(10): 1299 - 1319.

[2] Cho J H, Lee J M, Choi S W, et al. Fault identification for process monitoring using kernel principal component analysis [J]. Chemical Engineering Science, 2005, 60(1): 279 - 288.

[3] Choi S W, Lee C, Lee J M, et al, Fault detection and identification of nonlinear processes based on kernel PCA [J]. Chemometrics and Intelligent Laboratory Systems, 2005, 75(1): 55 - 67.

[4] Choi S W, Lee I B. Nonlinear dynamic process monitoring based on dynamic kernel PCA [J]. Chemical Engineering Science, 2004, 59(24): 5897 - 5908.

[5] 薄翠梅，王执铨，张广明. 基于 KPCA - PNN 的复杂工业过程集成故障辨识方法 [J]. 信息与控制，2009，38(1): 98 - 109.

[6] 刘晶晶，尹洪胜，张晋虎. 基于 KPCA/PNN 的煤矿主通风机的故障诊断 [J]. 煤矿机械，2011，32(11)，250 - 252.

[7] GE M, DU R, ZHANG G C, et al. Fault diagnosis using support vector machine with an application in sheetmetal stamping operations [J]. Mechanical Systems & Signal Processing, 2004(18): 143 - 159.

[8] ZHANG Y W. Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM [J]. Journal of Chemical Engineering Science, 2009(64): 801 - 811.

[9] SUYKENS J A K, VANDERWALLE J. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293 - 300.

[10] CHO JH, LEE JM, CHOI S W. Fault identification for process monitoring using kernel principal component analysis [J]. Chemical Engineering Science, 2005, 60(1): 279 - 288.

[11] Lee Jongmin, ChangKyoo Yoo, In - beum Lee. Statistical monitoring of dynamic process based on dynamic independent component analysis [J]. Chemical Engineering Science, 2004, 59(11): 2995 - 3006.

[12] Suykens J A K, Gestel T van, Brabanter J, et al. Least Squares Support Vector Machines [M]. Singapore: Word Scientific Pub. Co. Inc., 2002.

[13] CH IANG L H, RUSSELL E L, BRAATZ R D. Fault detection and diagnosis in industrial systems [M]. London: Spring - Verlag London Limited, 2001.

[14] CHEN J H, LIAO C M. Dynamic process fault monitoring based on neural network and PCA [J]. Journal of Process Control, 2002, 12(2): 277 - 289.