维普资讯 http://www.cqvip.com

非对称 DNA 序列混合识别模型研究*)

罗泽举[1] 宋丽红[2] 李艳会[3] 朱思铭[3]

(重庆工商大学计算机科学与信息工程学院 重庆 400067)[1]

(重庆工商大学实验实习中心 重庆 400067)[2] (中山大学数学与计算科学学院 广州 510275)[3]

摘 要 建立了一种改进的不对称支持向量机(MISVM)和隐马尔可夫模型结合的混合学习模型,对于实际中具有的不对称样本数据集,采用调整 Hessian 矩阵对角参数的策略,增大数据量少的样本离超平面的距离,再结合隐马尔可夫谱变换,以达到更加精确地分离不对称样本的目的。实验发现,不能简单利用正负两类样本所占百分比或固定参数来改变核函数矩阵的对角参数,而必须加之以可调整的权系数才能控制错分的样本数;经改进后的混合不对称学习算法比标准 SVM 具有更高的分辨率,对启动子序列进行识别,平均识别率达到 91.8%。

关键词 非对称 DNA 序列,核参数,隐马尔可夫谱

A Research on Mixed Recognition Model for Imbalanced DNA Sequence

LUO Ze-Ju[1] SONG Li-Hong[2] LI Yan-Hui[3] ZHU Si-Ming[3]

(School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067)[1]

(Center of Experiment and Practice, Chongqing Technology and Business University, Chongqing 400067)[2]

(School of Mathematics Computational Science, Sun Yat-Sen University, Guangzhou 510275)[3]

Abstract Set up a modified imbalanced SVM(MISVM) mixed learning models associated with HMM, for the imbalanced data set in practice, use the strategy for adjusting the Hessian matrix diagonal parameter, increase the distance between the few samples and the hyperplane, associated with the spectrum transform of hidden markov, this realizes our intention of separating the imbalanced samples more precisely. The experiment indicates that we can't simply use the percentage what the positive samples and the negative samples have or fix the diagonal parameter of kernel function, and must add proper weight coefficient which can be adjusted to control the number of error-divided samples, this modified mixed imbalanced learning algorithm obtain higher recognition rate than that of standard SVM, recognize promoter sequence, the average recognition rate come to 91.8%.

Keywords Imbalanced DNA sequence, Kernel parameter, Hidden markov spectrum

1 不对称支持向量机

样本的不对称性是指在实际中需要分类的两类样本,出现一类样本的数目明显多于另一类样本数目的现象。例如,信息系统领域中,一些偶尔出现的异类信号和大多数正常出现的信号;气象领域中,异常天气只占正常天气的一小部分;在生命科学领域,少数基因的碱基只占整个 DNA 序列的微小部分等等。这些小量的样本数量虽少,但却起着十分重要的作用。当用支持向量机进行学习时,出现伪正和伪负的现象,负类样本混在正类样本中,训练间隔非常小,错分率高。

如果我们调整超平面的位置,使得它尽量靠近小量的正类样本或尽量靠近类多的负类样本,就可以达到尽可能多地分离少数样本的目的。

1.1 不对称支持向量机算法

根据不对称样本的分布情况,我们将小数样本定为正类,而将多数样本定为负类。惩罚常数 C 也因此分为两部分:一部分是针对小的正类样本的 C^+,另一部分是针对负类的 C^-。于是,需要解决的优化问题修改为

$$\min_w \Phi(w)=\frac{1}{2}\|w\|^2+C^+\Big(\sum_{\substack{i=1\\ y_i=+1}}^{k}\xi_i\Big)+C^-\Big(\sum_{\substack{i=1\\ y_i=-1}}^{k}\xi_i\Big).$$

$$\text{s.t. } -[y_i(w\cdot x_i+b)-(1-\xi_i)]\leqslant 0, i=1,2,\cdots,k. -\xi_i \leqslant 0, i=1,2,\cdots,k \quad (1)$$

为了解决上述优化问题,得到 Lagrange 常数,有 $y_i=-1$: $0\leqslant\lambda_i\leqslant C^-$; $y_i=+1$: $0\leqslant\lambda_i\leqslant C^+$。我们改进核函数矩阵的对角调整技术,设训练样本为$(x_1,y_1),\cdots,(x_k,y_k)$,其中正类样本数为 m^+,负类样本数为 n^-,$k=m^++n^-$,则正类样本占的比重是 m^+/k,负类样本占的比重是 n^-/k。调整 Hessian 矩阵的对角线元素为

$$\begin{aligned} &H((i,i)\leftarrow H(i,i),\delta^+\frac{m^+}{k},\text{若 } y_i=1\\ &H(i,i)\leftarrow H(i,i)+\delta^-\frac{n^-}{k},\text{若 } y_i=-1 \end{aligned} \quad (2)$$

其中 $\delta^+\geqslant 0,\delta^-\geqslant 0$。

在这里,关键是条件 $\delta^+\geqslant 0,\delta^-\geqslant 0$,比起以前的方法进行了很大的放松[1,2]。因为在训练中,我们发现,由于 n^-/k 比 m^+/k 大很多,结果是超平面明显往样本数多的负类偏移。

*)重庆市教育委员会科学技术研究项目资助(KJ0707022)。**罗泽举** 博士,主要研究方向为机器学习与模式识别、生物信息学;**宋丽红** 实验师,主要研究方向为机器学习、计算机应用;**李艳会** 博士,主要方向为应用数学、常微分方程、计算机网络;**朱思铭** 教授,博士生导师,主要研究方向为应用数学、常微分方程、计算机应用。

维普资讯 http://www.cqvip.com

如果仅以 n^-/k、m^+/k 为参数或者简单固定参数 $(\delta^+ m^+)/k$，$(\delta^- n^-)/k$，超平面甚至可能会越过某些负类样本而使伪正类样本显著增加，这样反而会使错分的样本更多。因此，我们结合偏移的情况增加权参数 δ^+，δ^-，让机器根据要求自动调整权参数，以适当调整超平面靠向负(正)类一侧，将真正正类的少数样本分离出来，而达到提高分类效果的目的。

模型中 C^+、C^- 是惩罚参数，控制对两类样本错分的惩罚，由于负类样本居多，故主要是对负类偏向正类的惩罚。也就是为了减少伪正样本的数量，而不是为了减少伪负类的样本的数量。因为正类样本本身就很少，所以可以取 C^- 大些，而取 C^+ 小些。

2 隐马尔可夫离散谱变换

对于隐马氏模型[3,4] $\lambda=(S,\Sigma,A,B,\pi)$，设由模型 λ 产生观察序列 O 的概率为 $P(O|\lambda)$，用其自然对数 $L=\log P(O|\lambda)=\ln P(O)|\lambda)$(log likelihood Value，$L$ 值) 计算序列的离散谱范围。设观察序列是 $O=O_1^* O_2^* \cdots O_n^*$，相应的状态序列为 $Q=q_1 q_2 \cdots q_n$，定义变量

$$\alpha_t(i)=P(O_1^* O_2^* \cdots O_t^*, q_t=S_i|\lambda) \tag{3}$$

$$\beta_t(i)=P(O_{t+1}^* O_{t+2}^* \cdots O_n^*|q_t=S_i,\lambda) \tag{4}$$

$$\begin{aligned}\xi_t(i,j)&=P(q_t=S_i,q_{t+1}=S_j|O^*,\lambda)=\alpha_t(i)a_{ij}b_j(O_{t+1}^*)\\&\quad\beta_{t+1}(j)/P(O^*|\lambda)\\&=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1}^*)\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{l=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1}^*)\beta_{t+1}(j)}\end{aligned} \tag{5}$$

$$\gamma_t(i)=P(q_t=S_i|O^*,\lambda)=\sum_{i=1}^{N}\xi_t(i,j) \tag{6}$$

然后采用如下公式[5,6]重估模型的参数：

$$\pi_j^*=\gamma_1(i),a_{ij}^*=\frac{\sum_{t=1}^{k-1}\xi_t(i,j)}{\sum_{i=1}^{k-1}\gamma_t(i)},b_j^*(k)=\frac{s.t.\ O_t^*=v_K}{\sum_{t=1}^{k-1}\gamma_t(j)}\Big/\frac{\sum_{t=1}^{k}\gamma_t(j)}{\sum_{t=1}^{k-1}\gamma_t(j)} \tag{7}$$

3 不对称 SVM 和 HMM 混合分类迭代模型

结合支持向量机处理不对称样本的上述算法，我们提出以下不对称 SVM 和 HMM 混合分类模型，如图 1 所示。模型算法如下：

Step1：获取数据样本，对数据样本进行清除异常数据和补缺处理及标准化；

Step2：根据预先确定的初值 C^+、C^-，权参数 δ^+，δ^-，经过变换的核函数进行 SVM 分类训练；

Step3：根据训练结果看是否达到风险要求，控制决策面和错分率。如果满足要求，则转入 Step6，进行正式分类预测；否则转入 Step4。让计算机自动调整 Hessian 矩阵对角参数继续训练若干步后，如果满足要求，转入 Step6，否则转入 Step5。进行隐马尔可夫离散谱变换后，转入 Step6；

Step4：调整 Hessian 矩阵对角参数；

Step5：训练隐马尔可夫离散谱范围；

Step6：进行分类测试；

Step7：停止迭代。

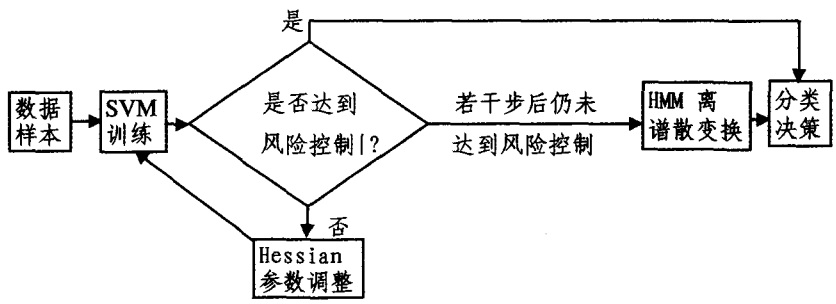

图 1 不对称 SVM 混合识别模型

4 实验结果

数据仍然从从瑞士实验癌研究组织生物信息组的真核生物非沉余启动子数据库(The Eukaryotic Promoter Database，EPD)(当前版本为 Release 88，2006-10-05 前)下载启动子序列数据(http://www.epd.isb-sib.ch/)。下载棘皮类动物 44 个、软体动物 3 个、线虫类 26 个、原核生物质粒体 8 个、脊椎动物动物选取人类 300 个，以这五类为样本进行实验。有的 DNA 序列很少(如软体动物只有 3 个)，有的则很多(如人类 1871)，样本明显是不对称的，以两类和多类分别进行实验。

4.1 实验参数调整结果

我们取棘皮类动物和软体动物进行不对称样本实验，棘皮类动物 10 个、软体动物 3 个作为训练。取各条序列的 230 bps 做统一的长度，相当于每个样本为 230 维。通过调整 Hessian 矩阵对角参数来观察间隔面的变化，用主成分分析法，只取前两个主成分，即将它投影到平面二维空间来观察其超平面的位置变化，正负样本非线性可分，我们用二阶多项式核函数进行分类，参数调整规则是：

$$H(i,i)\leftarrow(x_i\cdot x_i+1)^2+\delta^+\frac{m^+}{k}(y_i=1)$$

$$H(i,i)\leftarrow(x_i\cdot x_i+1)^2+\delta^-\frac{n^-}{k}(y_i=-1) \tag{8}$$

其中 $k=13$，$m^+=3$，$n^-=10$。可以看到，当 $\delta+\frac{m^+}{k}=\frac{12}{13}>\delta^-\frac{n^-}{k}=\frac{1}{13}$ 时，权重太偏向正类一边，使得一个样本远离负类样本，但此时达到了错分两个正类样本的事实(图 2)；当减少正类的权重而达到 $\delta^+\frac{m^+}{k}=\frac{8}{13}>\delta^-\frac{n^-}{k}=\frac{5}{13}$，超平面接近负类，此时 3 个正类样本得到了正确区分。而且，全部负类样本也得到了正确区分(图 3)，边界面接近正类样本，这正是我们需要的情况。但如果此时继续减少正类样本的权重)或者说增加负类样本的权重)，达到 $\delta^+\frac{m^+}{k}=\frac{5}{13}<\delta^-\frac{n^-}{k}=\frac{8}{13}$ 时，最优分类超平面已经越过了 4 个负类样本，使这 4 个负类样本得到错分(图 4)。

维普资讯 http://www.cqvip.com

图 2　错分两个正类样本

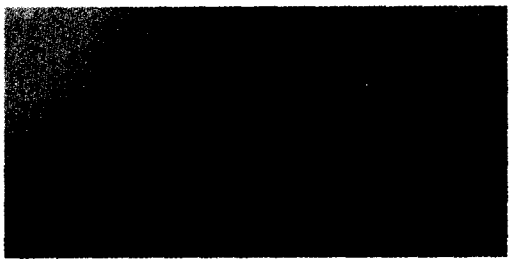

图 3　样本得到正确分类

图 4　错分 4 个负类样本

以伪正类(False Positive, FP)、伪负类(False Negative, FN)、真正正类(True Positive)、真正负类(True Negative)、平均准确率(mean accuracy rate for total, MAR)来比较棘皮类动物、软体动物、线虫类、原核生物质粒体、脊椎动物动物(人类)这五类启动子 DNA 序列的识别情况，SVM 用一对一(one-against-one)投票策略进行两类决策，得到表 1 所示的识别结果。

虽然在上节画图时我们用 PCA 方法将空间投影到了二维，但为了和标准 SVM 进行比较，在识别阶段不进行降维，而直接用 MISVM 和标准 SVM 计算。由于数据点是非线性可分的，核函数选取二阶多项式，计算时以伪正、伪负、真正正、真正负四个指标计算准确率而得出平均值。表 1 是原核质粒体和棘皮类动物的两类比较及线虫与人类启动子 DNA 序列的比较结果，Plasmid 类 8 个，MISVM 可以分离出 7 个正确样本来，而标准 SVM 只能分离出 6 个；线虫类 MISVM 能正确分离出 23 个，而标准 SVM 只能分离出 22 个。从表 2 可以看出，MISVM 比起标准 SVM 来，越是样本少的类其学习算法的识别效率就越显著。例如，对于原核生物质粒体，MISVM 的平均识别率为 87.5%，比标准 SVM 的 75%要高出 12 个百分点；而对于样本数目多的样本(例如人类)，两种方法识别的优势并不明显，MISVM 为 91.3%，标准 SVM 为 90.5%，MISVM 比标准 SVM 只高出 0.8 个百分点，原因是负类中为数少的错分样本比起整个负类样本来说，所占比重是很小的。因此 MISVM 算法是针对不对称样本集的，主要目的是将为数少量的样本能从混合的多数样本中分离出来，而不是将多数样本从少量样本中分离出来，这在实际中有非常重要的作用。例如，从大量 DNA 数据中识别小量 DNA 基因、从大量正常信号中判断少数异常信号、检测异常疾病等。

4.2　实验指标

表 1　MISVM 和标准 SVM 比较

类别 (categories)	样本数 (counts)	训练类别	FP、FN、TP、TN、MAR 各项指标比较									
			MISVM					标准 SVM				
			FP	FN	TP	TN	MAR	FP	FN	TP	TN	MAR
Plasmid	8	P	0	1	7	0	0.87	0	2	6	0	0.75
Echinoderm	44	N	5	0	0	39	0.88	6	0	0	38	0.86
Nematode	26	P	0	3	23	0	0.88	0	4	22	0	0.84
Homo sapiens	300	N	22	0	0	278	0.92	27	0	0	273	0.91

表 2　MISVM、MISVM+HMM 和标准 SVM 比较

类别	样本数	平均准确率(MART)		
		MISVM	MISVM+HMM	标准 SVM
Echinoderm	44	90%	93.1%	88%
Mollusc	3	100%	100%	100%
Nematode	26	87%	88.5%	86%
Prokaryotic plasmid	8	87.5%	100%	75.0%
Homo sapiens	300	91.3%	91.6%	90.5%

注：正类准确率为：$A_T = TP/m^+$；负类准确率为 $A_N = TN/n^-$；总体准确率为 $A_{all} = (TP+TN)/k$；$m^+ = TP+FN$，$n^- = TN+FP$，$k = m^+ + n^-$。

从表 2 的多类识别结果进一步得知，其中 MISVM 栏是用两类识别结果的平均得到的，MISVM+HMM 栏是用 MISVM 结合隐马尔可夫离散谱变换得到的，当 Hessian 参数调整无法达到所需的要求时，我们再结合 HMM 训练，将不同序列映射到不同的谱范围，采用 HMM 识别多类特有的 L 值谱映射优势进行进一步识别，得到了更为理想的识别结果，平均识别率达到了 91.8%。

讨论　通过调整 Hessian 矩阵对角参数使最优分类超平面的位置发生偏移，可以将小量样本从多数样本中分离出来，超平面既可以偏向正样本的类，也可以偏向负样本的类。实际中可以根据需要进行偏向调整，通过迭代逐步调整到所需要的参数。实验还发现，超平面始终偏向对角权重系数大的类，而且权重系数可以是任意大于等于零的实数，于是 Hessian 矩阵对角参数范围可以进一步拓宽到$[0, +\infty)$。这一结论与以往结论不同，按照文[5,6]，核函数矩阵对角加的是固定常数值。实际上，加一固定常数值的弊端是不能确定是否可以真正分离出小量样本，甚至还可能错分更多的样本，使识别率比标准 SVM 更低，因为算法并没有证明加入这个定常数后就可以分离出更多的样本，所以我们认为要根据实际情况进行调整而不是将这些参数值固定。对于多类情况，我们采用隐马尔可夫离散谱映射进行再一次识别，得到比 MIS-

维普资讯 http://www.cqvip.com

VM更为显著的识别效果。

参 考 文 献

1 Taylor J S, Cristianini N. Further results on the margin distribution. In: Proceedings of the 12th Conference on Computational Learning Theory, 1999
2 Karakoulas G J, Taylor J S. Optimizing classifiers for imbalanced training sets. NIPS, 1999, 11: 253～259
3 Schlapbach A, Bunke H. Using HMM based Recognizers for writer identification and verification. Frontiers in Handwriting Recognition, 2004. 167～172
4 Gough J, Chothia C. SUPERFAMILY: HMMS representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research, 2002, 30(1): 268～272
5 Baum L E, Sell G R. Growth functions for transformations on manifolds. Pac J Math, 1968, (27)2: 211～227
6 Baker J K. The dragon system-An overview. IEEE Acoust Speech Signal Processing, 1975, A ASSP-23(1): 24～29

(上接第206页)

$+\theta_2+\theta_3<2.5$ 时，系统同样也不是混沌的。

综合(1)、(2)，可知分数阶 Lü 系统能产生混沌吸引子的最低阶数为 2.5 阶。

3 分数阶 Lü 混沌系统的控制

在本节中，我们研究将分数阶 Lü 混沌系统镇定到它的不稳定平衡点。取 $\theta_1=\theta_2=\theta_3=0.9$，显然，分数阶 Lü 系统处于混沌状态，并且 $E(0,0,0)$ 是它的一个不稳定平衡点。下面我们利用线性反馈控制法将分数阶 Lü 混沌系统的运动轨道镇定到不稳定平衡点。

受控分数阶 Lü 系统为

$$\begin{cases}\dfrac{\mathrm{d}^\theta x}{\mathrm{d}t^\theta}=a(y-x)-u_1\\ \dfrac{\mathrm{d}^\theta y}{\mathrm{d}t^\theta}=-xz+cy-u_2,\\ \dfrac{\mathrm{d}^\theta z}{\mathrm{d}t^\theta}=xy-bz-u_3\end{cases}\tag{9}$$

这里 $\theta=29$，$u_i(i=1,2,3)$，是反馈外部输入控制，可使系统(6)的混沌运动轨道镇定到不稳定平衡点 E 上。设计线性反馈控制器为

$$\begin{cases}u_1=k_1(x-\bar{x})\\ u_2=k_2(y-\bar{y}),\\ u_3=k_3(z-\bar{z})\end{cases}\tag{10}$$

式(10)中$(\bar{x},\bar{y},\bar{z})$代表系统(6)的不稳定平衡点 E，k_1、k_2 和 k_3 是反馈增益。

受控系统(9)在平衡点处的 Jacobi 矩阵为

$$J=\begin{pmatrix}-a-k_1 & a & 0\\ 0 & c-k_2 & 0\\ 0 & 0 & -b-k_3\end{pmatrix},$$

其特征方程为

$$(\lambda+a+k_1)(\lambda-c+k_2)(\lambda+b+k_3)=0$$

显然，特征方程的特征根为

$$\lambda_1=-a-k_1,\lambda_2=c-k_2,\lambda_3=-b-k_3$$

当特征方程的实特征根均为负值时，受控系统(9)渐近稳定地趋于平衡点 E，可求得 $k_1>-36$、$k_2>20$ 和 $k_3>-3$。

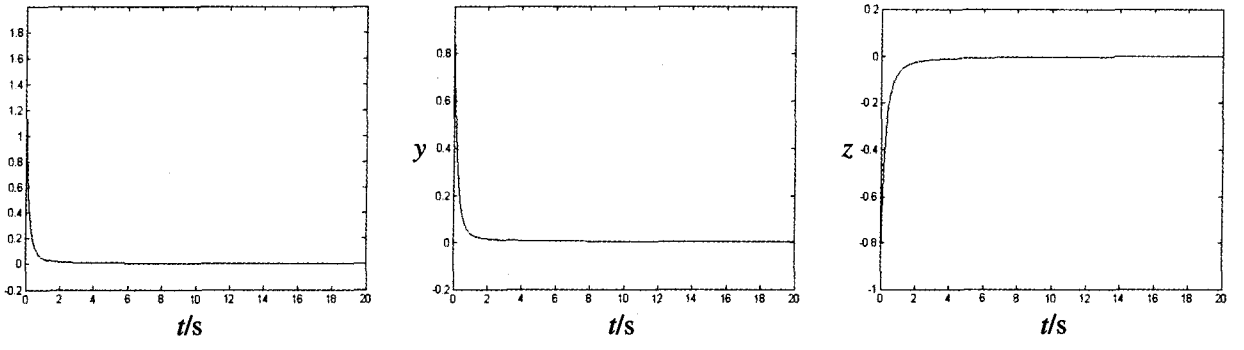

图 6 利用线性反馈法镇定分数阶 Lü 系统到平衡点时 $x(t)$、$y(t)$ 和 $z(t)$ 的变化曲线

图 6 给出了当选取受控系统(9)的初始点为：$x(0)=2$、$y(0)=1$ 和 $z(0)=-1$，取 $k_1=0$、$k_2=25$ 和 $k_3=0$ 时，受控系统(9)镇定到平衡点 E(0,0,0) 上的结果。由图 6 可见：当 t 分别接近 5.8s、7.8s 和 13.8s 时，$x(t)$、$y(t)$ 和 $z(t)$ 分别稳定到了零点，即受控系统(9)被镇定到平衡点 E(0,0,0) 上。

结论 本文研究了分数阶 Lü 系统的混沌动力学特性，数值模拟证明分数阶 Lü 系统确实存在混沌，并且得出分数阶 Lü 系统能产生混沌吸引子的最低阶数为 2.5 阶。作者利用线性反馈控制器成功地将分数阶 Lü 混沌系统镇定到平衡点。

参 考 文 献

1 Podlubny I. Fractional differential equations (New York: Academic Press), 1999
2 Hifer R. Applications of fractional calculus in physics (New Jersey: World Scientific), 2001
3 Grigorenko I, Grigorenko E. Phys. Rev. Lett. 91 34101, 2003
4 Gao X, Yu J B. Chaos Solitons Fract. , 24 1097, 2005
5 Hartly T T, Lorenzo C F, Qammer H K. IEEE Trans. CAS I, 42 485, 1995
6 Li C G, Chen G R. Physica A, 341 55, 2004
7 王发强，刘崇新. 物理学报，55 3922, 2006
8 Li C G, Chen G R. Chaos Solitons Fract. , 22 549, 2004
9 Li C G, Liao X F, Yu J B. Phys. Rev. , E 68 67203, 2003
10 Samko S G, Klibas A A, Marichev O I. Fractional integrals and derivatives: theory and applications (Amsterdam: Gordan and Breach), 1993
11 Caputo M. II. Geophys. J. R. Astron. Soc. , 13 529, 1967
12 Lorenz E N. J. Atmos. Sci. 20 130, 1993
13 Chen G R, Ueta T. Int. J. Bifur. Chaos, 9 1465, 1993
14 Lü J, Chen G R. Int. J. Bifur. Chaos, 12 659, 2002
15 Diethelm K. Elec. Trans. Numer. Anal. , 5 1, 1997
16 Diethelm K, Ford N J. J. Math. Anal. Appl. , 265 229, 2002
17 Diethelm K, Ford N J. Nonlinear Dyn. , 29 3, 2002
18 Wolf A, Swinney J B, Swinney H L. Physica D, 16 285, 1985