第 40 卷第 10 期 机 械 工 程 学 报 Vol.40 No.10
2004 年 10 月 CHINESE JOURNAL OF MECHANICAL ENGINEERING Oct. 2004

# 时间序列中异常值检测的负向选择算法*

董永贵 孙照焱 贾惠波
(清华大学精密测试技术及仪器国家重点实验室 北京 100084)

**摘要**：针对转子振动时间序列中异常数据的检测问题，采用欧氏距离进行匹配计算，在实数域实现了负向选择算法。通过对抗体库中元素增加一个描述其覆盖半径的参数，可更有效地发挥每个抗体元素的检测作用，显著提高了抗体库对异常数据集合的覆盖范围。计算结果表明，这种算法可有效地检测出时间序列中的异常值，且抗体库中元素数量少而可望用于信号的在线监测。

**关键词**：人工免疫系统 负向选择算法 时间序列 异常检测 欧氏距离
**中图分类号**：TP202

## 0 前言

在实际工程应用中，对系统运行状态的刻画与描述往往是建立在对系统一系列参数的观测与分析基础之上的。因此，系统异常状态的监测诊断问题可视为对观测到的时间序列中异常数据的发现与提取。随着状态监测理论及技术的发展，出现了各种针对异常信号检测的系统状态监测方法，如基于规则的方法、基于知识的专家系统、基于模式识别的方法等。这些方法或需要大量的异常特征信号作为先验知识进行训练，或需要建立精确的状态监测理论模型。而先验知识的获取困难、理论模型的不完善及计算实时性往往限制了这些方法在实际监测中应用。

实际上，虽然人们经常将各种状态监测系统与生物体的某个器官或系统做形象类比，但比较起传统的监测系统构成形式，自然界中的生物体是以一种几乎完全不同的方式对体外环境进行感知及应答。人体的神经系统通过五种知觉——视觉、听觉、嗅觉、味觉及触觉即可完全获知体外环境的全部信息。具体到系统状态监测方面，由生物体运行机制启发而来的人工神经网络及遗传算法已经得到广泛的研究与应用。近年来，随着人们对生物体免疫调节机制研究的不断深入，由免疫系统启发而来的人工免疫系统引起了各方面研究者的关注。如美国的 NASA 成立了生物启发技术与系统研究小组[1]，新墨西哥大学则将人工免疫系统成功应用于计算机系统的安全监测[2]，并开始进一步探索人工免疫系统在时间序列异常信号检测方面的潜在应用[3,4]。英国的 University of York 更将生物启发工程与生物医学工程并列，率先在硬件电路的容错研究方面取得了实质性进展[5]。我国也有部分研究者开始了人工免疫系统的研究，目前的主要研究方向集中在计算机系统安全[6,7]及异常信号的检测算法方面[8]。

将结合时间序列中异常信号的检测问题，对人工免疫系统中的负向选择算法进行探讨。首先介绍人工免疫系统的基本知识及目前负向选择算法的实现方案，重点针对负向选择算法中本体库/抗体库的建立算法进行分析讨论。在此基础上，结合转子系统碰摩信号的分析对所提出的在实数域进行匹配计算的负向选择算法的实用性进行讨论。

## 1 人工免疫系统与负向选择算法

人体系统几乎每时每刻都处于各种有害因素的围攻之中，各种病原体如细菌、病毒的入侵都可能导致疾病。免疫系统作为人体防护的主要屏障，需要具备两方面的功能：一是对病原体的检测，二是对病原体或有害物质的有效清除。对病原体的检测机制一般描述为对本体 (人体本身成分) 及异体 (人体外部成分) 的鉴别。由于某些病原体对人体无害，对其作出免疫应答不仅无益且可能伤害到人体，因此也有学者将免疫检测描述为对有害异体的检测。一旦检测到有害异体——抗原，则启动免疫应答进行清除。

在免疫系统中，对有害异体的检测与免疫应答是由淋巴细胞完成的。淋巴细胞有两类：B 细胞和 T 细胞。B 细胞主要功能是产生抗体，在整个生命过程中，持续地从骨髓产生,它执行特异体液免疫功能。T 细胞由胸腺产生，其功能为识别抗原及发起

* 国家重点基础研究 973 资助项目(G19990330)。20031007 收到初稿，20040430 收到修改稿

免疫应答并记忆历史上遭遇过的抗原特征。对异体的识别即由 T 细胞表面的抗体分子与异体细胞表面抗原蛋白分子的特异性结合完成的。这种抗原—抗体识别方式有如下特点。

(1) 抗体是随机产生的，但并不是所有随机产生的抗体均能够进入体液中进行免疫识别检测。抗体的选择是通过一种负向选择机制完成的：抗体产生后，如其能够与本体细胞分子结合，则该抗体在胸腺中即被取消，反之则被释放到体液中去。抗体分子的随机产生机制保证了抗体分子的多样性，而负向选择机制则保证了体液中的抗体仅对可能的异体抗原敏感，而不会因与本体细胞的结合导致假阳性的结果。

(2) 不完全匹配的特征检测机制。抗原/抗体的特异性结合实际上是抗体分子上特异性识别位点与抗原决定基之间的结合。由于抗原表面一般有多个抗原决定基，而一种抗体仅提供一种识别位点，因此对一种抗原的识别实际上是通过多种不同抗体共同完成的。当抗体识别位点与抗原决定基之间的结合数达到一定阈值，即可认定该抗原。这样既可避免仅靠单个抗体进行识别易出现假阳性的缺点，又不会因苛刻的完全匹配原则导致漏检情况的发生。

(3) 对抗原特征识别有学习及记忆功能。由于 T 细胞的记忆功能，当发生同类型抗原再次入侵时，免疫应答的响应时间会大大缩短。

新墨西哥大学的 Stephanie Forrest 基于上述机制提出的人工免疫系统[2~4, 9]需要维护两个数据库，即代表正常数据特征的本体库与代表异常数据特征的抗体库。具体构建及运行过程主要有如下三个步骤。

(1) 学习及抗体库的建立。对抗原的检测基础在于本体与有害异体的识别。抗体库的建立采取的是一种负向选择算法。即首先为系统提供足够多的本体信息特征，构成本体数串集合，即本体串库，再通过负向选择建立抗体库。具体操作过程如图 1 所示。

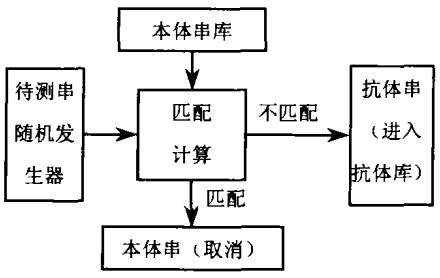


图 1　抗体库的建立

(2) 异常特征识别。人工免疫系统开始运行，将系统运行过程中涉及的代表系统状态参数的待测数串作为输入，与抗体库中的抗体数串进行匹配计算，实现异常特征识别。具体过程如图 2 所示。

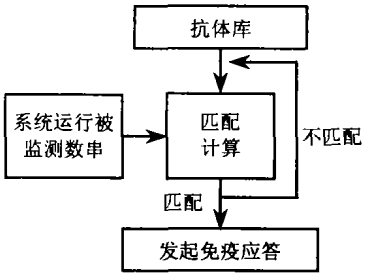


图 2　异常特征识别

(3)免疫应答。主要包括出错/修复进程的发起，抗原特征的采集及抗体库的更新等。

从对时间序列中异常数据检测的角度看，负向选择算法的采用应该是人工免疫算法与其他如人工神经网络等需要提供先验知识进行训练的算法区别最大之处。系统的异常状态信息需要在系统真正出现故障时才可获得，而故障的出现是人们最不希望甚至在重要工程应用中是不允许的，因此这种仅需要提供正常状态特征进行训练的人工免疫系统显得尤为实用。当然，由于对正常状态的描述往往比异常状态要复杂，本体串库及相应抗体库的建立所需要的先验知识及计算量要大些。Stephanie Forrest 所提出的负向选择算法是建立在二进制数串的匹配计算基础之上的。

(1) 将实数域的时间序列进行二进制编码，如取二进制数长度为 $m$ 比特，则时间序列中的每个实数值经编码后均落于对应整数 0(最小值)及 $2^m-1$(最大值)区间的 $m$ 比特长的二进制整数集合内。

(2) 对时间序列进行加窗处理，设窗长为 $n$，Stephanie Forrest 采用的是无重叠加窗方式，即沿时间轴滑动步长亦为 $n$，则得到长度为 $m\times n$ 的二进制数串。

(3) 匹配计算采用简单的异或方式，如图 3 所示。如两数串中有超过 $r$(阈值)个连续位相同，则匹配成功。

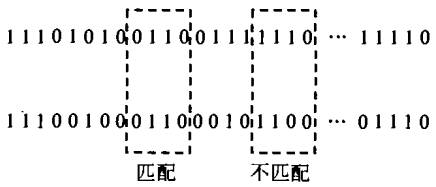


图 3　二进制匹配计算示意

这种基于二进制数串的负向选择算法在计算机网络安全方面的应用是成功的，然而当用于时间序列的异常数据检测时，抗体库的维护恐怕是最大的问题：为保证得到比较理想的检测效果，势必需

要产生大量的抗体，而抗体库中元素数量的增长必然会导致计算量的增长而影响检测的实时性。$m$ 及 $n$ 越大，这个问题也就越突出。

## 2 基于欧氏距离的实数域负向选择算法

综上所述，对时间序列中异常数据的检测可归结为对时间序列中哪段数据正常、哪段数据异常进行判别的问题。首先对时间序列进行归一化，得到 $y_i \in [0.0, 1.0]$，则如窗长为 $n$，加窗处理后的时间序列数据串可表示为 $n$ 维矢量集合 $U \subseteq [0.0, 1.0]^n$，集合中的元素为 $n$ 维矢量 $\boldsymbol{x}^j = (x_1^j, x_2^j, \cdots, x_n^j) \in [0.0, 1.0]^n$。定义代表正常数据串的本体元素集合为 $S \subseteq U$。代表异常数据串的异体元素集合为 $N \subseteq U$，则

$$S \cup N = U \quad S \cap N = \Phi$$

采用负向选择算法所建立的抗体库 $D$ 实际上是 $N$ 的子集，即 $D \subseteq N$。记本体集合 $S$ 中元素为 $\boldsymbol{x}^S$，抗体库 $D$ 中元素为 $\boldsymbol{x}^d$，定义 $d(x,y)$ 为 $[0.0, 1.0]^n$ 上的距离函数，这里选取欧氏距离

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2}$$

参照 Stephanie Forrest 的算法，利用负向选择算法建立抗体库的问题可描述如下。

对于任意随机产生的矢量 $\boldsymbol{x}^r \in [0.0, 1.0]^n$，如果 $\min(d(\boldsymbol{x}^r, \boldsymbol{x}^S)) > \lambda_S \in \mathbf{R}_+$，则 $\boldsymbol{x}^r \in D$，进入抗体库，否则 $\boldsymbol{x}^r \in S$，取消该矢 量，重新产生候选矢量直到抗体库中元素数量达到要求。

相应地，时间序列异常数据的检测问题可描述如下。

对于 $\forall x^j$，如果 $\min(d(\boldsymbol{x}^j, \boldsymbol{x}^d)) < \lambda_d \in \mathbf{R}_+$，则 $\boldsymbol{x}^j \in N$，即数据异常，否则 $\boldsymbol{x}^j \in S$，即数据正常。

记以 $\boldsymbol{x}^S$ 为中心，$\lambda_S$ 为半径的超球面内所有点的集合为 $X_{\lambda_S}^S$，则按照上述算法所定义的本体集合为全部 $X_{\lambda_S}^S$ 的并集，即 $S = \bigcup_{\boldsymbol{x}^S \in S} X_{\lambda_S}^S$。类似地，记以 $\boldsymbol{x}^d$ 为中心，$\lambda_d$ 为半径的超球面内所有点的集合为 $X_{\lambda_d}^D$，则抗体库集合可表示为 $D = \bigcup_{\boldsymbol{x}^d \in D} X_{\lambda_d}^D$。由于检测过程中所有 $X_{\lambda_d}^D$ 的半径均为 $\lambda_d$，因此抗体库中元素对异常数据的覆盖度不尽合理。$\boldsymbol{x}^d$ 理想的覆盖半径无疑应为 $\lambda > \lambda_d$ 如图 4 所示。为解决这一问题，对抗体库中的每个元素增加一个描述覆盖半径的参数 $r_i$，记抗体库中元素 $ab_i = (\boldsymbol{x}_i^d, r_i)$，对应于中心点 $\boldsymbol{x}_i^d$ 的半径为 $r_i$ 的超球面内所有点的集合为 $X_{r_i}^d$，抗体库中元素数为 $M$，则抗体库集合可表示为 $D = \bigcup_{i=1}^{M} X_{r_i}^d$。相应地，利用负向选择算法建立抗体库的问题可描述如下。

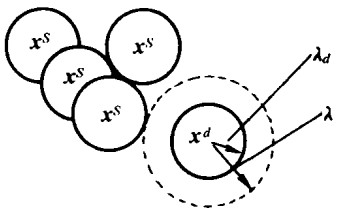


图 4 抗体库中元素的覆盖度示意

对于任意随机产生的矢量 $\boldsymbol{x}^r \in [0.0, 1.0]^n$，如果 $\min(d(\boldsymbol{x}^r, \boldsymbol{x}^S)) < \lambda_S \in \mathbf{R}_+$，则 $\boldsymbol{x}^r \in S$，取消该矢量，重新产生候选矢量，否则计算 $\boldsymbol{x}^r$ 是否落入已有 $X_{r_i}^d$ 内，如否，则 $\boldsymbol{x}^r$ 为新抗体进入抗体库，对应的抗体覆盖半径为 $r^r = \min(d(\boldsymbol{x}^r, \boldsymbol{x}^S)) - \lambda_S$。

时间序列异常数据的检测问题可描述如下。

对于 $\forall \boldsymbol{x}^j$，如果 $\exists ab_i$，满足 $d(\boldsymbol{x}^j, \boldsymbol{x}_i^d) < r_i$，则 $\boldsymbol{x}^j \in N$，即数据异常，否则 $\boldsymbol{x}^j \in S$，即数据正常。

## 3 仿真计算及结果分析

按照上述算法，首先对简单周期性信号进行了计算分析(图 5 所示为 100 点的正弦时间序列)

$$y(i) = \sin\left(\frac{\pi}{10} i\right)$$

在 65～85 点间信号异常

$$y(i) = \frac{1 + \mathrm{rand}(i)}{2} \sin\left(\frac{\pi}{10} i\right)$$

式中 $i$——点数

$\mathrm{rand}(i)$——[0,1]间的伪随机数

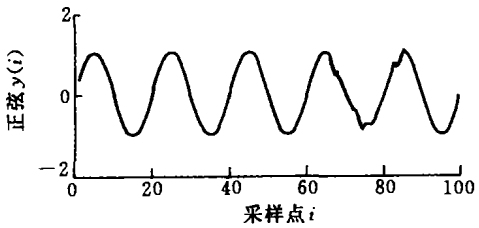


图 5 正弦时间序列：65～85 点间数据异常

利用前 50 点作为本体数据建立本体库，再经负向选择算法生成 20 个元素的抗体库，对全部时间序列进行检测。选取窗长 $n$=5，本体匹配阈值 $\lambda_S = 0.01$，计算如下 4 种情况。

对本体数据加窗处理滑动步长 step_$s$=$n$(无重叠)，对检测数据加窗处理滑动步长 step_$m$=$n$(无

重叠)。

对本体数据加窗处理滑动步长 step_$s$=$n$(无重叠)，对检测数据加窗处理滑动步长 step_$m$=1(重叠 4 点)。

对本体数据加窗处理滑动步长 step_$s$=1(重叠 4 点)，对检测数据加窗处理滑动步长 step_$m$=$n$(无重叠)。

对本体数据加窗处理滑动步长 step_$s$=1(重叠 4 点)，对检测数据加窗处理滑动步长 step_$m$=1(重叠 4 点)。

计算结果如图 6 所示，其中纵坐标值为 1 处为异常点，为零处为正常点。

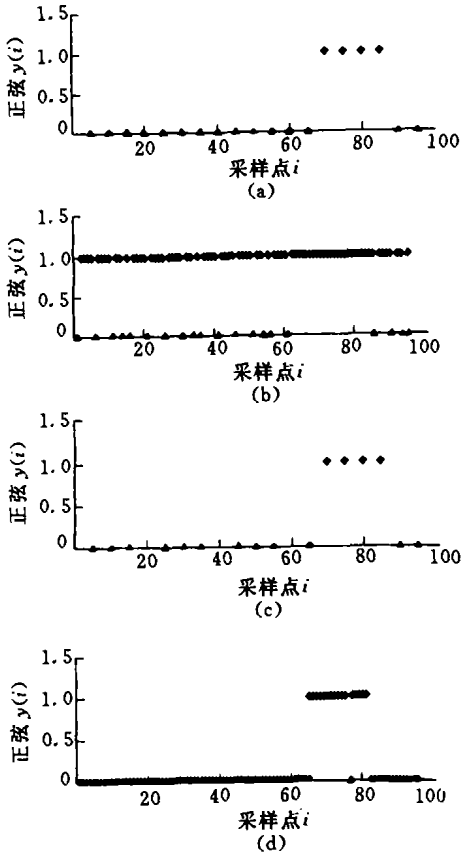


图 6　不同加窗处理方式计算结果比较

由图中可见，图 6a、6c、6d 的结果与预想相同，而图 6b 的结果则将许多正常数据点归为异常，即出现了“假阳性”的结果。造成这种现象的原因在于，采用无重叠的方式对本体数据进行加窗处理后，由于本体数据的周期性，造成了本体库中元素的构成对加窗起点非常敏感，因而所形成的本体库对本体数据特征覆盖度不够，而图 6a 的结果之所以未出现“假阳性”则是由于对本体数据与对检测数据进行加窗处理时，两者起点恰恰相同。因此综合考虑，对含有趋势项的时间序列数据进行处理时，本体库的建立应采取逐点移动加重叠窗的处理方式。

如图 7a、b 分别为利用转子试验台采集到的正常状态与有碰摩发生时的 $x$ 向振动数据(各 1 000 点，已经进行量纲一处理)。同样采用图 6d 情况的方法，取窗长 $n$=5，$\lambda_S$ = 0.01，利用图 7a 的前 500 点作为本体数据建立本体库，并进一步产生 20 个元素的抗体库，利用此抗体库对图 7a、b 进行异常检测，得到的结果分别如图 7c、d 所示。图 7c 中全部数值均为零，表明本体库很好地覆盖了正常数据集合，而图 7d 的结果则很好地说明了本算法对异常数据检测的有效性。

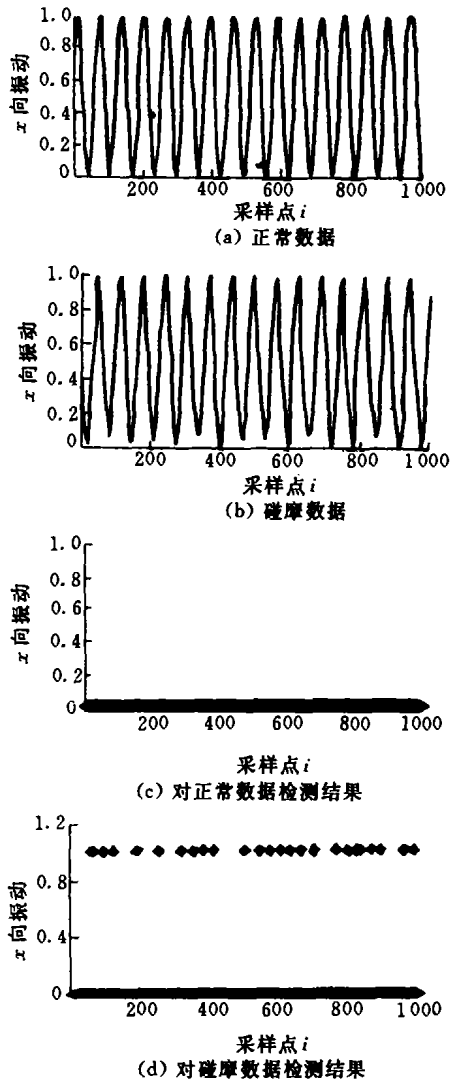


图 7　转子碰摩振动数据计算结果

## 4　结论

以人工免疫系统的兴起为代表的生物启发理论与技术的研究目前已经成为国际上的一个重要研究领域。从 2002 年在英国召开的首届人工免疫系统国际会议[10]来看，人工免疫系统在信息技术领域的应用研究尚处于初级发展阶段，主要的研究方向集中在计算机安全防护及电子系统的异常状态检测等

方面。在时间序列的异常数据识别与系统的故障诊断方面，目前的大多数算法需要异常特征的先验知识，而异常样本数据的收集一直是困扰故障诊断系统应用于工程实际的难题。相比之下，生物体免疫系统对异常的免疫检测机制要高明得多：对异常特征的识别更多地依赖于对正常状态的经验积累。因此从实用的角度看，负向选择算法的采用应该是人工免疫算法与其他需要提供先验知识进行训练的算法区别最大之处。

针对时间序列中异常数据检测的问题，对人工免疫系统的负向选择算法进行了初步探讨。由于实际检测时，待测数据串需要与抗体库中每个元素进行匹配计算，因此抗体库中元素的数量越多，实时性越差。而抗体库中元素数量太少，又可能导致对异常数据集合的覆盖范围过小而造成“假阴性”的结果。所提出的基于欧氏距离的实数域负向选择算法，通过对抗体库中元素增加一个描述其覆盖半径的参数 $r$，可更有效地发挥每个抗体元素的检测作用，从而可望较好地解决上述矛盾。当然，由于这种方式所得到的抗体库实际上是多个互不相交的超球面体子集合所组成的并集，对那些落在超球面体之间的异常数据无检测能力。如何解决这一问题或将这一“漏洞”的影响最小化，将是以后进一步研究改进的方向。

## 参 考 文 献


1 http://cism.jpl.nasa.gov/programs/RCT/BioCompUD.html

2 Hofmeyr S, Forrest S. Architecture for an artificial immune system. Evolutionary Computation Journal, 2000, 8(4)：443～473

3 Dasgupta D, Forrest S. Artificial immune systems in industrial applications. In：the International Conference on Intelligent Processing and Manufacturing Material (IPMM). Honolulu, HI, 1999, http：//issrl.cs. memphis.edu/AIS/

4 Dasgupta D, Forrest S. Novelty detection in time series data using ideas from immunology. In：Proceedings of The International Conference on Intelligent Systems, 1999, http://issrl.cs.memphis.edu/AIS/

5 http://www.elec.york.ac.uk/bio/welcome.html

6 杨晓宇，周佩玲，傅忠谦. 人工免疫与网络安全. 计算机仿真，2000，18(6)：83～85

7 侯朝桢，张雅静. 基于 multi-agent 的仿生物免疫：计算机抗病毒研究新思路. 北京理工大学学报，2002，22(3)：270～273

8 刘树林，王日新，黄文虎，等. 基于反面选择算法的压缩机振动在线监测方法研究. 压缩机技术，2001：(5)：9～10

9 Chao D L, Forrest S. Information immune systems. International Conference on Artificial Immune Systems (ICARIS), 2002：132～140

10 http://www.aber.ac.uk/icaris-2002/icaris-2002.htm


# NEGATIVE SELECTED ALGORITHM FOR ANOMALY DETECTION IN TIME SERIES DATA


*Dong Yonggui Sun Zhaoyan Jia Huibo*

*(State Key Laboratory of Precision Measurment Technology and Instruments, Tsinghua University, Beijing 100084)*



**Abstract:** For the anomaly detection in the vibration time series of the rotor system, a real-valued negative selection algorithm based on Euclidean distance has been implemented. By means of adding the corresponding coverage radius to each antibody elements, the detection efficiency of each antibody element is increased. The coverage scope of the antibody set is significantly enlarged for the anomaly data set. The calculation results indicate that the algorithm can efficiently detect the anomaly in time series data. Moreover, the number of detectors in antibody set is less enough for potential application in online signal monitoring.



**Key words:** Artificial immune system
Negative selection algorithm
Time series Anomaly detection
Euclidean distance



作者简介：董永贵，男，1965年出生，博士，副研究员。主要研究方向为非接触式位移振动传感器、石英晶体化学传感器、振动信号分析与处理、信息存储与传输。

E-mail: dongyg@pim.tsinghua.edu.cn