文章编号:1671-4598(2015)01-0060-04　　中图分类号:TP273　　文献标识码:A

基于相关分析与最小二乘支持向量机的TE过程多模型建模

胡蓓蓓，李丽娟，熊　路
（南京工业大学 自动化与电气工程学院，南京　211816）

摘要：针对TE化工过程高度非线性、复杂性的特点，文章提出了一种基于相关分析和最小二乘支持向量机对TE过程进行多模型建模方法，以提高模型性能；首先对TE过程采用相关分析法划分为3个子系统，对每个子系统分别采用基于C-均值聚类的最小二乘支持向量机建模和基于K均值聚类的最小二乘支持向量机多模型建模；实验表明，基于K均值聚类的多模型建模能简化计算、提高模型精度、并且能更好的预测模型输出。

关键词：TE过程；相关分析；最小二乘支持向量机；多模型建模

Multi-modeling of Tennessee Eastman Process Based on Correlation Analysis and Least Squares Support Vector Machine

Hu Beibei，Li Lijuan，Xiong Lu
（College of Automation and Electrical engineering，Nanjing University of Technology，Nanjing　211816，China）

Abstract：In order to deal with the nonlinear and complexity of TE，this paper proposes a multi-model modeling method for TE process based on CA and LSSVM which can overcome the disadvantages of single modeling. Firstly，the TE process is divided into three subsystems by CA. Then，LSSVM multi-modeling depended on the C-means and LSSVM multi-modeling depended on the k-means are applied to every subsystem respectively. At the same time，the experimental results also shows this algorithm can simplify the calculation，improve the accuracy of the model and forecast the model output better.

Keywords：TE process；correlation analysis；LS-SVM；multi-model

0　引言

流程工业中的监控、控制都是依靠高性能的模型为前提的。随着工业过程的复杂化程度不断提高，出现了很多非线性、多工况等特点。对整个工业过程采用单模型进行研究，容易出现计算量大、模型精度不高等缺陷，很多专家学者提出了多模型建模方法[1-3]。近年来，由Suykens等人在标准支持向量机基础上提出的最小二乘支持向量机由于其不仅成功克服了参数化函数逼近机制局部极小、不能保证概率意义上收敛等缺点，而且运算速度快于标准支持向量机，因而在建模技术中得到了广泛的应用[4]。

TE过程是基于实际工业过程的仿真案例，这个案例很适合于研究过程控制技术[5]。因而提出后，许多学者对这个案例的过程建模进行了广泛的研究。文献[6]对此提出了一种整体建模方法，此方法计算复杂并且难以保证模型精度。

基于此本文提出了一种相关分析和K-均值聚类与最小二乘支持向量机相结合的多模型建模方法，并采用TE化工过程作为研究对象。为了比较，还采用了基于C-均值聚类的最小二乘支持向量机多模型建模方法进行试验研究。结果表明，本文提出的多模型建模方法不仅简化了计算，而且提高了模型预测精度。

收稿日期:2014-05-23；　**修回日期**:2014-06-25。

基金项目:国家自然科学基金项目(2007DA690071)；江苏省六大人才高峰项目；工业控制技术国家重点实验室开放课题(ICT1234)。

作者简介:胡蓓蓓(1988-)，女，硕士，主要从事流程工业建模与仿真方向的研究。

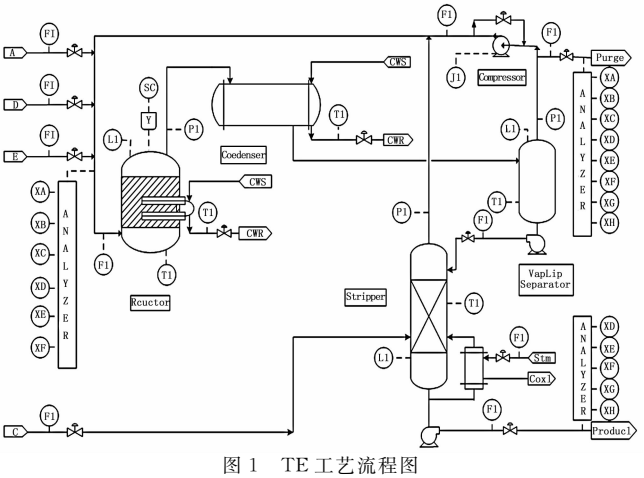

图1　TE工艺流程图

1　TE化工过程

TE过程包括5个主要操作单元：反应器、冷凝器、气液分离器、循环压缩机和汽提塔，其工艺流程如图1所示。该过程包括12个操作变量，41个测量变量。

TE过程共有4个反应，包括4种反应进料A、C、D、E，生成2种产物G、H，和1种副产物F，进料中包含少量的惰性气体B。反应方程式如下：

$$A(g)+C(g)+D(g)\rightarrow G(l),product1$$

$$A(g)+C(g)+E(g)\rightarrow H(l),product2$$

$$A(g)+E(g)\rightarrow F(l),byproduct$$
$$3D(g)\rightarrow 2F(l),byproduct$$

由于该过程的变量太多，采用单模型进行建模计算繁琐，并且难以保证模型精度。针对上述问题本文提出了一种基于 K 均值聚类和最小二乘支持向量机的多模型建模方法。

2　基于相关分析和最小二乘支持向量机多模型建模方法

基于相关分析和最小二乘支持向量机的算法总体结构如图 2 所示。首先对生产过程中采集到的数据进行相关分析，将大系统划分为若干子系统；然后分别对每个子系统进行 LS-SVM 建模。

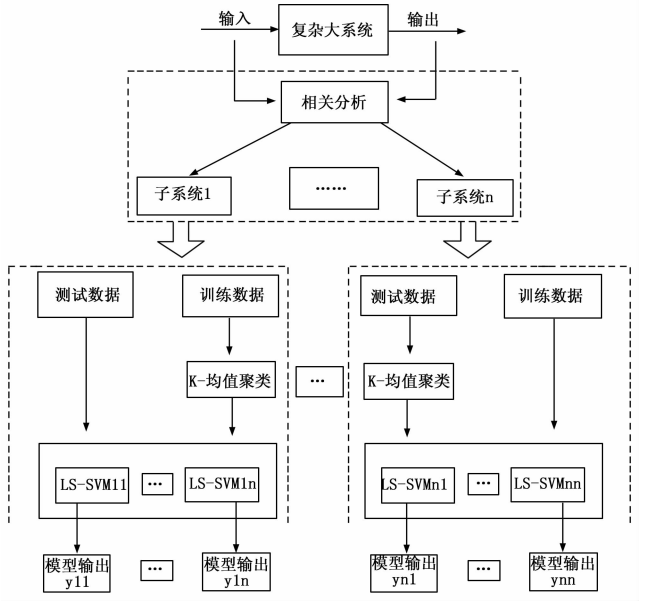

图 2　算法的总体结构

2.1 相关分析

相关分析[7]是一种典型的数据分析方法，它可以消除多维随机变量的多重线性相关性，并且可以将变量系统降维。目前数据分析的方法很多，如主元分析[8]、方差分析[9]等。但是相关分析可以在兼顾两个多维随机变量的线性依赖关系的前提下，消除两个变量系统各自的多重线性相关性和降维。

设 p 维随机量 $\boldsymbol{X}=[X_1,X_2,\cdots X_p]^T$，$q$ 维随机变量 $\boldsymbol{Y}=[Y_1,Y_2\cdots Y_q]^T$，记

$$\boldsymbol{Z}=(X_1,X_2,\cdots X_p,Y_1,Y_2,\cdots Y_q)^T=\begin{bmatrix}X\\Y\end{bmatrix} \tag{2}$$

则 $\boldsymbol{Z}$ 的协方差矩阵为：

$$\boldsymbol{Var}(Z)=\begin{pmatrix}var\ (X) & cov\ (X,Y)\\ cov\ (Y,X) & var\ (Y)\end{pmatrix}=\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}=^{\Delta}V \tag{3}$$

其中 $V_{21}'=V_{12}$，且 $V_{11}\geqslant 0, V_{22}\geqslant 0$

设 $u=(u_1,u_2,\cdots u_p)', v=(v_1,v_2,\cdots v_q)'$ 是两个常向量，令 $z_1=u'x, w_1=v'y$

令 $\mathrm{var}(z_1)=1,\mathrm{var}(w_1)=1$ (4)

又 $\mathrm{cov}(z_1,w_1)=\mathrm{cov}(u'x,v'y)=u'V_{12}v$ (5)

由此将上述问题转化为求解下列条件极值问题：

$$\begin{aligned}&Max:\quad u'V_{12}v\\ &s.t.\quad uV_{11}u=1\\ &\quad\ vV_{22}v=1\end{aligned} \tag{6}$$

利用拉格朗日乘子法，令

$$\varphi=u'V_{12}v-\frac{\lambda}{2}(uV_{11}u-1)-\frac{\mu}{2}(vV_{22}v-1) \tag{7}$$

将 φ 对 $u.v$ 求偏导数，令其为 0

$$\frac{\partial\varphi}{\partial u}=V_{12}v-\lambda V_{11}u=0 \tag{8}$$

$$\frac{\partial\varphi}{\partial v}=V_{21}u-\mu V_{22}v=0 \tag{9}$$

由 (3)，(8)，(9) 可知 $\lambda=\mu$

$$V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}u=\lambda^2u \tag{10}$$

$$V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}v=\lambda^2v \tag{11}$$

由此可得到 z_1 和 w_1 使得其相关系数 ρ 达到最大

$$\rho=\frac{\mathrm{cov}(z_1,w_1)}{\sqrt{\mathrm{var}(z_1)\mathrm{var}(w_1)}} \tag{12}$$

2.2　K 均值聚类

K-means 算法[10]是一种典型的聚类方法。它是以确定的聚类数 k 和随机选定的初始聚类中心为前提对数据进行聚类的。但在实际中，事先无法确定聚类数，并且初始聚类中心随机选择也会使结果不稳定。对此文献 [11] 提出了一种改进的 k-means 聚类方法。首先确定聚类数 k 的搜索范围，将由 AP[12]算法产生的聚类数作为上界 $K_{\max}$；而当 $K_{\min}$ 取 1 时表示样本均匀分布没有实际意义所以 $K_{\min}$ 取 2。其次确定初始聚类中心，从 N 个样本中任取一个作为第一聚类中心 z_1，从余下的 N-1 个样本中找出离 z_1 距离最大的作为第二个样本 z_2，计算其他样本 x_i 与 z_1 和 z_2 的距离，并求它们的最小距离 d_i，即

$$d_{ij}=\|x_i-z_j\|, j=1,2 \tag{13}$$

$$d_i=\min(d_{i1},d_{i2}), i=1,2,\cdots,n \tag{14}$$

若 $D_i=\max\{d_i\}>\theta\|z_1-z_2\|$（$\theta$ 为选定的比例系数），则相应的样本 x_i 作为第 3 个聚类中心 z_3，重复同样的步骤，直到再找不到符合条件的新聚类中心。

为了评价聚类效果，文献 [13] 定义了 Silhouette 指标，它反映了类间可分性和类内紧密性。设一个数据集，有 n 个样本，则某个样本 t 的 Silhouette 为：

$$Sil(t)=\frac{b(t)-a(t)}{\max\{a(t),b(t)\}} \tag{15}$$

其中：$a(t)$ 为样本 t 与类内所有其他样本的平均距离，$b(t)$ 为样本 t 与其他每个类中样本平均距离的最小值。Silhouette 指标值在 [−1，1] 之间，在实际数据中其值越大越好，越大表明具有最佳聚类数。

2.3　最小二乘支持向量机

支持向量机[14]是一种机器学习方法，将求解的非线性回归问题转化为高位空间的线性函数问题。它基于结构风险最小化的原则，具有泛化能力强的优点。最小二乘支持向量机是支持向量机的一种改进，用等式约束代替原先的不等式约束，简化了计算，提高了求解速度并且提高了收敛精度。LS-SVM 将优化问题描述如下：

$$\min J=\frac{1}{2}\omega^T\omega+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2 \tag{16}$$

约束条件：$y_k=\omega^T\varphi(x_k)+b+e_k$ (17)

定义拉格朗日函数：

$$L(\omega,b,e,\alpha)=J-\sum_{k=1}^{N}\alpha_k\{\omega^T\varphi(x)+b+e_k-y_k\} \quad (18)$$

求解上述优化问题的 Lagrange 函数，得最优解

$$\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0 & \boldsymbol{E}^T\\ \boldsymbol{E} & \boldsymbol{QQ}^T+\dfrac{\boldsymbol{I}}{\boldsymbol{\gamma}}\end{bmatrix}^{-1}\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \quad (19)$$

其中：$\boldsymbol{E}=[1,1,\ldots,1]^T$；$\boldsymbol{I}$ 为单位阵；$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\ldots,\alpha_n]^T$；$\boldsymbol{Q}=[\varphi(x_1)^T,\varphi(x_2)^T,\ldots,\varphi(x_N)^T]$；$\boldsymbol{y}=[y_1,y_2,\ldots,y_N]^T$

假定核函数 $K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$，则可以通过式，最终得到最小二乘支持向量机的表达形式

$$y(x)=\sum_{i=1}^{N}\alpha_iK(x,x_i)+b \quad (20)$$

针对本文所提的多模型建模算法，首先对去噪后的工业数据进行相关分析，并依据相关系数的大小将系统分解为若干小系统；对每个小系统的建模训练数据进行 K 均值聚类，得到若干个子类，并对每一个子类采用 LS-SVM 进行建模，为了验证模型的精度，采用测试样本进行测试，当测试样本来时先根据它与每个聚类中心的欧氏距离判断所属类，并根据对应的子模型得到模型输出，直到算法结束。

2.4 算法步骤

Step 1：采集工业过程数据进行去噪处理，得到输入、输出变量集；采用相关分析（CA）算法进行分析，得到各个输入变量 X、输出变量 Y 间的相关系数 ρ。

Step 2：根据相关系数 ρ 的大小，对大系统进行划分。若相关系数大于或等于 0.85 则判定两变量相关，得到若干相互独立的子系统。

Step 3：对每个子系统分别采用基于 K 均值聚类的 LSSVM 多模型建模。首先将每个子系统数据划分为训练数据和测试数据，对训练数据采用 K-means 聚类方法聚为 k 类，对各个类采用 LSSVM 建立各自的子模型。

Step 4：根据训练样本建立好多模型后，确定每个测试数据与各个聚类中心数据的欧式距离，判断测试数据所属子类。

Step 5：将测试数据拿到所属类所对应的子模型进行预测输出。若训练样本同时隶属于多个子模型，则分别拿到多个子模型进行预测输出，最后采用加权方式进行输出。

3 实验仿真与分析

针对第 1 节所述的 TE 化工过程，采用基于相关分析和最小二乘支持向量机的多模型建模方法进行实验建模，实验中核函数选择径向基核函数：

$$K(x,x_i)=\exp\{-\|x-x_i\|_2^2/\sigma^2\} \quad (21)$$

对工业采集到的数据经过去噪处理，再出去搅拌速率等常量后，选择反应器进料量、反应器冷却水温度、流量等 16 个变量为输入变量；选择反应器温度、气液分离器温度、G/H 产率比为输出变量。对输入、输出变量采用相关分析法分析，若相关系数大于或等于 0.85 则判定两变量相关，最后得到 3 个子系统，反应器温度作为子系统 1 的输出、气液分离器温度作为子系统 2 的输出、G 与 H 的产率比作为子系统 3 的输出。对每个子系统数据进行处理得到 300 个样本点，选取前 200 个样本作为训练数据，后 100 个作为测试数据。采用 K 均值聚类的方法将每个子系统训练数据又划分为 4 类，对每个类采用最小二乘支持向量机建立模型，根据欧氏距离确定每个训练样本的所属类，并根据对应的模型求出其输出值。为了比较，本文还采用了基于 C 均值聚类的实验方法进行比较验证。

图 3 是反应器温度预测对比；图 4 是气液分离器温度预测对比；图 5 是 G 与 H 产率比预测对比。其中 AV 代表实际值，PV 代表预测值。

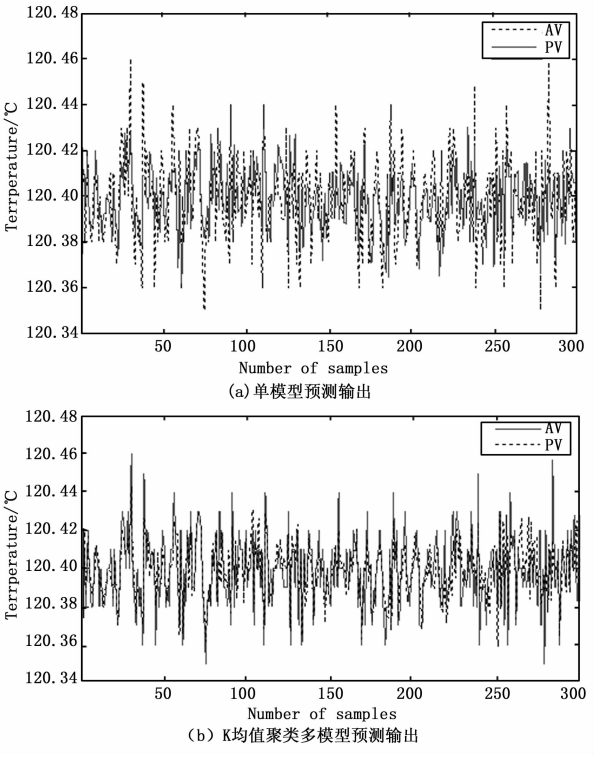

图 3 反应器温度预测对比

采用泛化均方根误差（RMSE）和最大相对误差（MAXE）来评价模型的预测性能。

$$RMSE=\sqrt{\frac{1}{n}\sum_{1}^{n}[y_i-f(x_i)]^2} \quad (22)$$

$$MAXE=MAX\left(\frac{1}{y_i}|(y_i-f(x_i))|\right) \quad (23)$$

其中：y_i 和 $f(x_i)$ 分别为测试数据的实际值和模型预测值，n 为测试数据数。根据实验结果，3 个子系统的对应的模型预测输出泛化均方根误差和最大相对误差如表格 1 所示。

表 1 *RMSE*、*MAXE* 对比

	C-means		K-means	
	RMSE	*MAXE*	*RMSE*	*MAXE*
Subsystem 1	0.017 64	0.045 27	0.015 93	0.041 45
Subsystem 2	0.169 83	0.409 74	0.137 87	0.396 57
Subsystem 3	0.015 76	0.035 92	0.012 24	0.023 07

实验结果表明，采用相关分析将 TE 大复杂化工系统拆分为 3 个子系统，并对每个子系统分别建立模型，简化了计算并且提高了精度。

本文采用的基于 K 均值和 LS-SVM 支持向量机建模方法，分别对每个独立的类建立子模型，相对而言模型精度和预测精

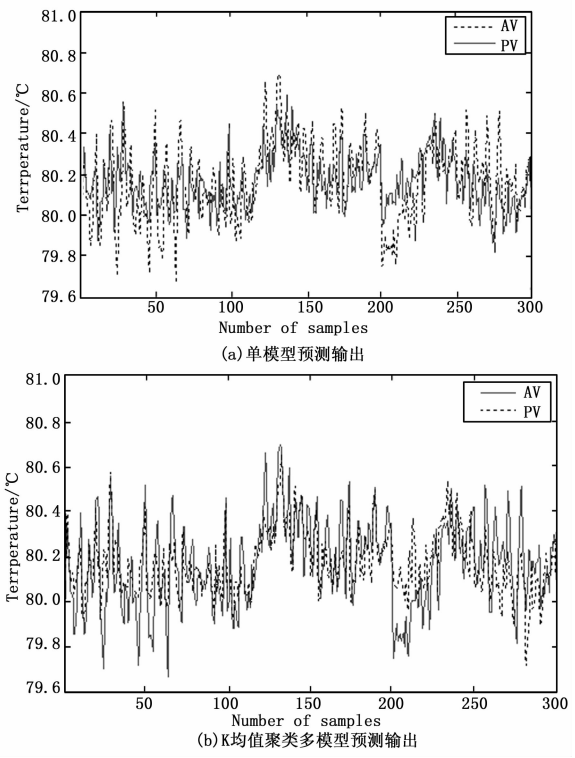

图 4　气液分离器温度预测对比

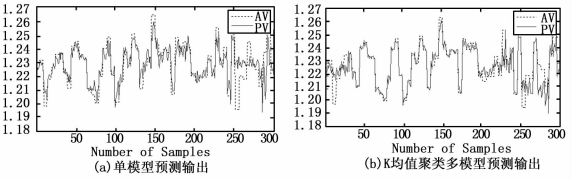

图 5　G 与 H 产率比预测对比

度较单模型与基于 C 均值聚类的多模型建模方法有所提高。相关分析和最小二乘支持向量机多模型建模方法，在 TE 化工过程建模和预测中取得了很好的测试性能。

4　结论

TE 化工过程是一个高度非线性，复杂的化工过程，过程变量繁多，并且工业数据会随着工况的改变成堆聚集的特性。如果采用 MIMO 模型对整体进行建模时需要很大的计算量，并且难以保证模型精度。本文首先采用相关分析法将 TE 过程分为 3 个子系统，对每个子系统分别采用了基于 LSSVM 的单模型建模，基于 C 均值的多模型建模和基于 k 均值的多模型建模进行建模仿真，并且采用泛化均方根误差和最大相对误差对两种模型进行比较，结果表明基于 k 均值的多模型具有更好的泛化能力，提高了整体的预测精度。

参考文献：

[1] 李修亮，苏宏业，褚　健. 基于在线聚类和关联向量机的多模型软测量建模 [J]. 化工自动化及仪表. 2008，35 (3)：34－37.

[2] 李雅芹，杨慧中. 基于仿射传播聚类和高斯过程的多模型建模方法 [J]. 计算机与应用化学，2010，27 (1)：51－54.

[3] Frey B J，D. Dueck. Clustering by passing messages between data points [J]. Science，2007，315 (5814)：972－976.

[4] 夏梁志，李　华，饶克克，等. 基于 QGA－LSSVM 的醋酸乙烯聚合率软测量建模 [J]. 计算机测量与控制，2012，20 (4)：907－909.

[5] Downs J J，Vogel E F . Aplant wide industrial process control problem [J] . Computers Chem Engng，1993，17 (3) ：245－25.

[6] Ben C. Juricek，Dale E. Seborg，Wallace E. Larimore. Identification of the Tennessee Eastman challenge process with subspace methods [J]. Control Engineering Practice，2001 (9)：1337－1351.

[7] 袁志发，周静芋. 多元统计分析 [M]. 北京. 科学出版社，2004.

[8] 宋　坤. 基于 SVM 多模型建模的软测量研究 [D]. 南京：南京工业大学，2010.

[9] 陈文亮，张　湜，李　晖. 基于 LS-SVM 沼气进化变压吸附过程甲烷浓度建模 [J]. 天然气化工，2013，38 (1)：36－38.

[10] Jain A K，Flynn P J. Image segmentation using clustering [A]. In：Ahuja N，Bowyer K. eds. Advances in Image understanding：A. Festchrift for Azriel Rosenfeld [C]. Piscataway：IEEE press，1996：65－83.

[11] 周世兵，徐振源，唐旭清. 新的 K-均值算法最佳聚类数确定方法 [J]. 计算机工程与应用，2010，46 (16)：27－31.

[12] 李丽娟，潘　磊，张　湜. 基于 AP 聚类算法的跳汰机床层松散度软测量建模 [J]. 化工学报，2012，63 (9)：2675－2680.

[13] Frey B J，Dueck D. Clustering by passing messages between data points [J]. Science，2007，315 (5814)：972－976.

[14] Yang H，Luo F，Xu Y G，et al. New LS-SVM nonlinear predictive controller method based on chaos optimization [J]. Computer Engineering and Applications，2010，46 (5)：229－232.

(上接第 59 页)

想记忆搜索功能，可以有效的提高了航空雷达网络入侵监测诊断效率，且具有较好的适应性与多样性，本文方法诊断准确率达到 93.3%。相比于传统的检测算法，在检测率、误报率、漏报率等方面均有明显改善，并通过仿真实例验证了该方法的有效性与通用性。

参考文献：

[1] 马殿哲，常天庆，陈军伟. 基于 BAM 网络的坦克航空雷达在线故障诊断方法研究 [J]. 计算机测量与控制，2011，19 (12)：3001－3003.

[2] 龙鹏飞，宋　振. 基于 BAM 网络和遗传免疫的入侵检测算法 [J]. 计算机工程与设计，2007，28 (12)：2793－2795.

[3] 文　莹，肖明清，盛　晟，等. 基于概念格的航空雷达故障诊断研究 [J]. 计算机测量与控制，2013，21 (10)：2612－2614.

[4] 戴英侠，连一峰，王　航. 模型安全与入侵检测 [M]. 北京：清华大学出版社，2002.

[5] 刘　赛，许　斌，梁意文. 入侵检测模型中的一种免疫遗传算法 [J]. 计算机工程，2004，30 (8)：63－64.

[6] 徐琰珂，梁晓庚，贾晓洪. 雷达/红外双模导引头信息融合算法研究 [J]. 计算机测量与控制，2013，21 (1)：129－132.