文章编号: 1001-0920 (2015) 04-0698-05 **DOI:** 10.13195/j.kzyjc.2014.0112

基于云 PSO 的 RVM 入侵检测

李国栋[1], 胡建平[1], 夏克文[2]

(1. 天津城建大学 计算机与信息工程学院，天津 300384；2. 河北工业大学 信息工程学院，天津 300401)

摘 要: 入侵检测可为计算机网络信息提供安全保障, 在其方法研究中, 由于相关向量机 (RVM) 具有高稀疏性且预测中使用概率因素, 在网络入侵检测中优于支持向量机. 然而 RVM 的核函数参数是经验估计的, 为此, 提出一种基于云模型的粒子群优化算法的 RVM 方法, 即采用云粒子群算法确定 RVM 的核参数, 构建 RVM 分类模型, 再采用一对一分类方法进行多类检测分类. 经入侵检测实验研究, 所得结果表明所提出的方法优于基于常规相关向量机的检测方法, 且具有更高的入侵检测精度.

关键词: 入侵检测；相关向量机；云粒子群优化

中图分类号: TP273 **文献标志码:** A

Intrusion detection using relevance vector machine based on cloud particle swarm optimization

LI Guo-dong[1], *HU Jian-ping*[1], *XIA Ke-wen*[2]

(1. School of Computer and Information Technology，Tianjin Chengjian University，Tianjin 300384，China; 2. School of Information Engineering，Hebei University of Technology，Tianjin 300401，China. Correspondent：LI Guo-dong，E-mail：lgd@tcu.edu.cn)

Abstract: Intrusion detection can protect computer network information. In the research based on this method, due to the relevance vector machine(RVM) has high sparseness and uses probability factor in predict, which is superior to the support vector machine(SVM) in the network intrusion detection. However, the kernel function parameters of RVM are estimated by experience. Therefore, a kind of RVM method based on the cloud particle swarm optimization(PSO) algorithm is proposed, which adopts the CPSO algorithm to determine the kernel parameter of RVM, then builds RVM model and uses the one-against-one classification method to finish multi-class intrusion detection. The experimental researches on intrusion detection show that the proposed method is superior to the common RVM-based detection method and has high prediction accuracy in intrusion detection.

Keywords: intrusion detection；relevance vector machine；cloud particle swarm optimization

0 引 言

随着互联网技术的全面普及, 信息量的爆炸式增加, 网络的传输和接入量呈指数形式增长, 同时网络的入侵、攻击手段也是层出不穷. 入侵检测 (ID) 就是对系统的攻击企图和攻击行为实时监测, 对正常网络行为和异常入侵行为进行区分, 并将异常入侵行为进行分类. 分类方法, 尤其是智能分类法, 近年来得到了很好的发展, 由于具有较高的分类精度等优点, 智能分类法得到了高度重视和广泛应用. 因此, 研究先进的智能分类方法在入侵监测领域的应用技术, 具有科学的理论基础和现实意义. 目前, 基于统计学习理论的支持向量机 (SVM)[1]以其良好的非线性处理能力和泛化能力, 以及全局收敛性优于一般的学习机等优点, 并能有效解决非线性、小样本问题, 已在入侵检测系统中得到了广泛的应用. 但其也存在一些缺陷, 例如稀疏性差、预测中缺乏概率因素、核函数受 Mercer 条件的限制等[2-5]. 为此, 相关向量机 (RVM)[6-9]方法可以弥补上述 SVM 的缺陷, 其方法和技术得到了国

收稿日期: 2014-01-18；**修回日期:** 2014-04-18.

基金项目: 国家自然科学基金项目(51208168)；国家星火计划项目(2014GA610018)；天津市自然科学基金项目(11JCYBJC00900)；河北省引进留学人员基金项目(JFS-2012-13001)；天津市高等学校科技发展基金计划项目(20110814).

作者简介: 李国栋(1980−), 男, 副教授, 博士, 从事智能信息处理的研究；胡建平(1957−), 男, 教授, 博士生导师, 从事通信技术等研究.

内外诸多学者的高度关注. 然而, 传统的 RVM 的核函数参数是基于经验估计[10]而不是优选的. 为此, 本文在入侵检测中, 提出一种基于云模型的粒子群优化 (CPSO) 算法[11-12]的 RVM 方法, 采用 CPSO 算法优化 RVM 核参数来建模, 进而探索相关向量机的快速算法, 并提高入侵检测精度.

1　相关向量机及其改进

1.1　相关向量机描述

相关向量机 (RVM) 是 Tipping 等[13]于 2003 年提出的一种快速的边际似然算法, 它是建立在支持向量机 (SVM) 上的稀疏概率学习模型, 其大大提高了算法的运算性能. 相关向量机应用于分类的主要步骤如下:

1) 已知样本 $\{x_i,t_i\}_{i=1}^n, x_i\in R^n$, RVM 的模型输出定义为

$$y(x;w)=\sum_{i=1}^{M} w_i k(x,x_i)+w=\Phi(x)w. \tag{1}$$

其中: $w=[w_0,w_1,\cdots,w_M]^{\mathrm{T}}$ 为模型的权值; $\Phi(x)$ 为 $M\times(M+1)$ 阶的由核函数构成的矩阵, 有

$$\Phi(x)=\begin{bmatrix} 1 & K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_M)\\ 1 & K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_M)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & K(x_M,x_1) & K(x_M,x_2) & \cdots & K(x_M,x_M) \end{bmatrix}. \tag{2}$$

2) 样本概率预测式为

$$p(t_*|t)=\int p(t_*|w)p(w,\alpha|t)\mathrm{d}w\mathrm{d}\alpha. \tag{3}$$

其中

$$p(w,\alpha|t)=$$

$$(2\pi)^{-\frac{N+1}{2}}|\Psi|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}(w-\mu)^{\mathrm{T}}\Psi^{-1}(w-\mu)\Big\}p(\alpha|t),$$

$$\Psi=(\Phi^{\mathrm{T}}\Phi+A)^{-1},$$

$$\mu=\Psi\Phi^{\mathrm{T}}t,$$

$$A=\mathrm{diag}(\alpha_0,\alpha_1,\cdots,\alpha_N).$$

3) 由 delta 近似函数并积分, 得到 RVM 的模型

$$y_*=\Phi(x_*)\mu.$$

4) 采用最大似然计算

$$p(\alpha|t)\propto p(t|\alpha)p(\alpha),$$

$$(\alpha_{MP})=\arg\max_{\alpha} p(t|\alpha).$$

得到 α 的迭代公式

$$\alpha_i^{\mathrm{new}}=\frac{\gamma_i}{\mu_i^2},$$

其中 $\gamma_i=1-\alpha_i\Psi_{i,i}$.

相关向量机的训练过程就是迭代求解 α, 最终求得模型的权值以求得式 (1), 再根据式 (1) 进行相关向量机的分类.

1.2　相关向量机的多分类方法

相关向量机最初是用来解决两类分类问题的, 但实际生活中常常会遇到多分类问题. 目前, 用来解决将相关向量机推广到多分类问题的方法有很多, 主要有一对多、GAD[14]和一对一等多分类方法.

1) 一对多分类方法.

假设共有 K 个类别, 则 K 个二分类模型便可构造出来. 把第 i 类样本作为第 i 个模型的正样本, 其他类别的样本合起来作为负样本进行训练, 其最终决策函数为

$$f(x)=\mathrm{sgn}(\omega^i\phi(x_j)+b^i). \tag{4}$$

判别时, 经过 K 个分类机分类后, 输入样本就能得到 K 个输出值, $f_i(x)=\mathrm{sgn}(g_i(x)), i=1,2,\cdots,K$. 若 $+1$ 只出现一次, 则输入样本类别将取 $+1$ 所对应的类别; 若没有一个输出为 $+1$, 或 $+1$ 不只出现一次, 则比较 $g(x)$ 的输出值, 输入样本的类别将取最大者所对应的类别.

2) 一对一分类方法.

构造 $K(K-1)/2$ 个二分类模型, 每个二分类模型仅训练 K 个类别中的两类样本, 因而形成了 $K(K-1)/2$ 个决策函数. 第 i 类与第 j 类之间的决策函数为

$$f(x)=\mathrm{sgn}(\omega^{ij}\phi(x_j)+b^{ij}). \tag{5}$$

在判别时, 本文采用投票方法, 即对于一个未知样本 x, 利用构造的 $K(K-1)/2$ 个分类模型对其分别分类. 若 x 被判定为第 i 类, 则在相应类的投票上加 1; 否则把这一票投到第 j 类上. 最终得票最多的类即为该样本 x 的类别.

1.3　基于 CPSO 的相关向量机

基本粒子群算法的速度和位置更新公式为

$$v_i=v_{i-1}+c_1r_1(\mathrm{pbest}-x_{i-1})+ c_2r_2(\mathrm{gbest}-x_{i-1}), \tag{6}$$

$$x_i=x_{i-1}+v_i. \tag{7}$$

标准的粒子群算法是通过控制权重系数的下降使得种群收敛. 由于惯性权重系数 ω 是固定的, 较大的 ω 值虽然在算法收敛速度上有所缺陷, 但是有利于达到全局最优值; 相反, 较小的 ω 值具有收敛速度快的优点, 从而更有利于局部寻优. 为了克服粒子群算法存在的缺陷, 本文将云模型与 PSO 算法进行有机结合, 形成云粒子群优化算法 (CPSO). 该算法将粒子群分成 3 个子群[15], 依据不同的 ω 值对各个子群生成策略.

设粒子群在第k次迭代中粒子X_i的适应度为f_i^k, 用$f_{\rm avg}^k = \dfrac{1}{N}\sum\limits_{i=1}^{N} f_i^k$表示粒子群的平均适应度, 优于$f_{\rm avg}^k$的粒子的平均适应度用$f'_{\rm avg}$表示, 次于$f_{\rm avg}^k$的粒子的平均适应度用$f''_{\rm avg}$表示, 粒子的最优适应度用$f_{\rm bset}^k$表示. 本文采用云模型来提高选择策略, 依据不同的粒子适应度来确定不同的惯性权重系数, 具体方法如下.

1) 当$f_i^k > f'_{\rm avg}$成立时, 该部分粒子的适应度较好, 接近最优值, 为了加快局部收敛, ω取较低值.

2) 当$f_i^k > f''_{\rm avg}$且$f'_{\rm avg} > f_i^k$成立时, 该部分粒子的适应度不高, 采用云模型对其进行改进. 首先设定粒子的数学期望值

$$\mathrm{Ex} = f_{\rm best}^k, \tag{8}$$

粒子的熵为

$$\mathrm{En} = (f'_{\rm avg} - f_{\rm best}^k)/b_1. \tag{9}$$

可设粒子的超熵与熵的关系值为

$$\mathrm{He} = \mathrm{En}/b_2, \tag{10}$$

其中b_1和b_2为调整系数. 则ω的取值为

$$\omega = 0.9 - 0.5 \times \mathrm{e}^{\frac{-(f_i^k - \mathrm{Ex})^2}{2(\mathrm{En}')^2}}. \tag{11}$$

其中: $\mathrm{Ex} = f_{\rm best}^k$为粒子的数学期望, $\mathrm{En}' = \mathrm{normrnd}(\mathrm{En}, \mathrm{He})$.

3) 当$f_i^k < f_{\rm avg}{}''$成立时, 该部分粒子适应度较差, 为了达到全局范围内重新搜索, ω取较大值.

另外, 对于第2)、第3)种情况, 为了提高其精度和加快收敛速度, 可分别引入交叉变异操作. 交叉操作为

$$x_i^1 = px_{i-1}^1 + (1-p)x_{i-1}^2, \tag{12}$$

$$x_i^2 = px_{i-1}^2 + (1-p)x_{i-1}^1, \tag{13}$$

$$v_i^1 = \frac{v_{i-1}^1 + v_{i-1}^2}{|v_{i-1}^1 + v_{i-1}^2|}|v_{i-1}^1|, \tag{14}$$

$$v_i^1 = \frac{v_{i-1}^1 + v_{i-1}^2}{|v_{i-1}^1 + v_{i-1}^2|}|v_{i-1}^1|, \tag{15}$$

其中p为交叉概率.

云粒子群优化算法(CPSO)的具体步骤如下.

Step 1: 粒子群的初始化. 随机给出每个粒子的初始位置和速度, 并确定ω、迭代次数以及加速系数等参数.

Step 2: 计算适应度值. 根据适应度函数计算每个粒子的适应度值, gbest的初始化是粒子中具有最好函数值的粒子, gbest_i的初始值与前一个粒子的相同.

Step 3: 确定惯性权重ω. 根据每次迭代的适应度值, 应用云模型确定不同情况下的惯性权重.

Step 4: 更新速度和位置. 根据式(6)对每个粒子的速度进行更新, 根据式(7)对每个粒子的位置进行更新, 且粒子的个体极值和群体极值根据新种群粒子适应度值进行更新.

Step 5: 判断迭代是否满足精度要求或者达到最大迭代次数. 如果满足, 则终止迭代; 若不满足, 则转Step 3继续执行.

由上可知, 为了高精度优化RVM的模型参数, 采用CPSO算法来优化RVM中高斯核函数的核宽度参数, 具体优化步骤如下.

Step 1: 粒子群的初始化. 随机给出每个粒子的初始位置和速度, 并确定ω、迭代次数以及加速系数等参数, 每个粒子向量代表一个RVM模型中对应的核函数的参数σ.

Step 2: 计算适应度值. 根据下式对粒子的适应度值进行计算:

$$F_{\rm fitness} = \sum_{i=1}^{N}\left(\frac{y - y_i}{N}\right)^2. \tag{16}$$

其中: y是样本的测量值, y_i是样本的预测值, N是样本的数量. gbest的初始化是粒子中具有最好函数值的粒子, pbest_i的初始值与前一个粒子的相同.

Step 3: 应用云模型确定惯性权重. 对于第k次迭代的第i个粒子, 如果fit_i^k优于$\mathrm{fit}_{\rm best}^k$, 则更新$\mathrm{pbest}_i$为粒子的当前位置; 如果$\mathrm{fit}_i^k$优于$\mathrm{fit}_{\rm gbest}^k$, 则更新$\mathrm{gbest}_i$为粒子的当前位置. 根据每次迭代的适应度值, 确定不同情况下的惯性权重, 应用云模型增加选择策略和收敛速度.

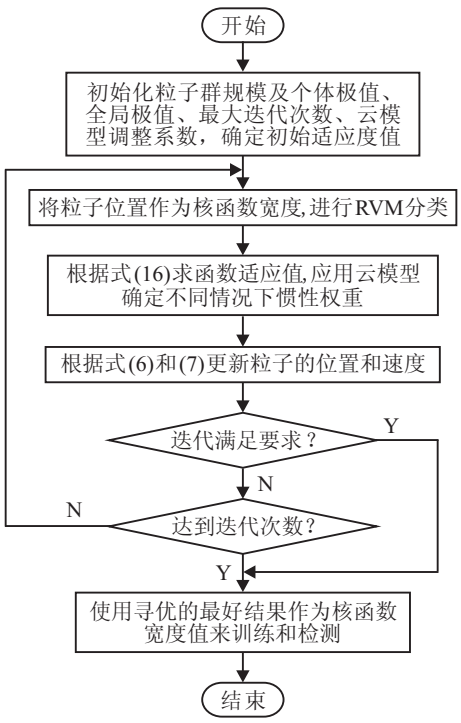

图 1 CPSO-RVM 流程

Step 4: 更新粒子位置和速度. 根据式 (6) 对每个粒子的速度进行更新, 根据式 (7) 对每个粒子的位置进行更新; 计算出每个粒子的适应度; 更新最好位置和全局最优.

Step 5: 判断迭代是否满足终止条件. 如果满足, 则终止迭代, 并将寻优得到的最好的核参数代入相关向量机模型中, 用于相关向量机分类; 反之, 则转 Step 3 继续执行.

图 1 给出了 CPSO-PVM 流程图.

2　实际应用

为了分析入侵检测技术的性能, 本实验采用的技术路线如图 2 所示.

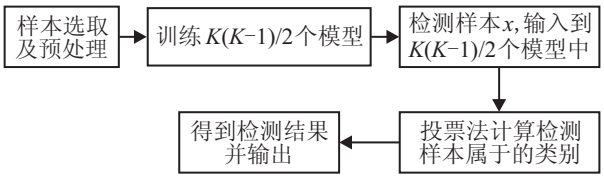

图 2　入侵检测模型

本文实验样本取自 MIT Lincoln 实验室的入侵检测数据集 KDD Cup99 的 data_10_percent 子集. 在该数据集中, 每项 42 个属性中的前 41 个为特征, 第 42 个 label 属性为攻击模式, 其中一条数据表现形式为 "0, udp, private, SF,105,146, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 255, 254, 1.00, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, normal".

本实验选用 5 种模式: normal、back、neptune、satan、portsweep. 其中: normal 为正常网络行为, 其他 4 种为异常入侵行为. 定义 5 种攻击模式的 label 分别为 1、2、3、4、5. 对以上 5 个属性各取 600 个数据, 训练样本取自每个属性的前 100 个数据, 测试样本取其余的 500 个数据, 即本实验有 500 个训练样本、2 500 个测试样本. 本文采用一对一多分类方法, 分别利用 RVM、PSO-RVM、CPSO-RVM 进行入侵检测, 入侵检测结果如图 3 所示.

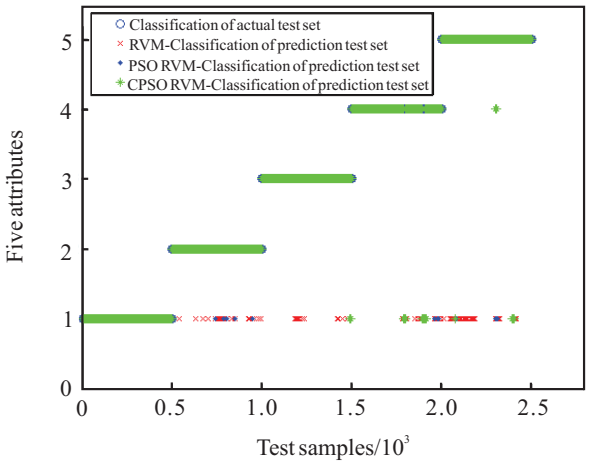

图 3　RVM、PSO _RVM、CPSO_RVM 入侵检测结果

图 4 为 PSO-RVM 和 CPSO-RVM 的迭代曲线比较. 表 1 为实验检测结果比较, 其中最大迭代次数设定为 100, 且粒子群规模设定为 20, 云模型调整系数分别为 2.9 和 10.

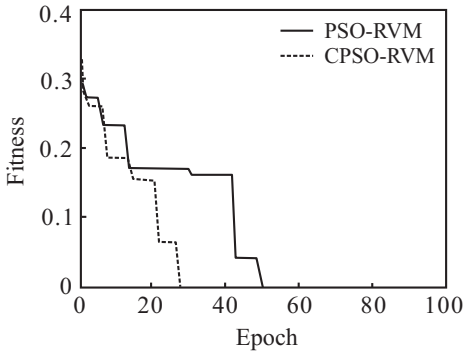

图 4　PSO-RVM 和 CPSO_RVM 的迭代曲线

表 1　实验检测结果比较

分类算法	检测精度/%	误报率/%	分类时间/s
RVM	90.84	9.16	31
PSO-RVM	97.56	2.44	520
CPSO-RVM	98.44	1.56	679

入侵检测系统的性能可以由检测精度、误报率及分类时间等技术指标来衡量, 定义如下:

$$\text{检测精度 (或检测准确率)} = \frac{\text{分类准确的样本数}}{\text{总样本数}},$$

$$\text{误报率 (或检测差错率)} = \frac{\text{存在误报的样本数}}{\text{总样本数}}.$$

从图 4 可以看出, 相对于 PSO, CPSO 不但收敛速度快, 而且克服了 PSO 易陷入局部最优的不足. 在完全相同的实验条件下, 通过表 1 可以看出, 相对于 RVM, PSO-RVM 和 CPSO-RVM 算法在入侵检测上较大地提高了检测精度, 尤其是 CPSO-RVM 优化算法具有很高的检测精度.

3　结　论

本文提出了一种基于云 PSO 的相关向量机网络入侵检测方法, 其中利用云粒子群优化算法可以快速有效地确定出相关向量机的核宽度参数. 采用入侵检测数据集 KDD Cup99 的子集, 经入侵检测实验, 结果表明, 本文的 CPSO-RVM 方法具有很高的检测精度, 优于 RVM 和 PSO-RVM 方法, 在网络入侵检测领域有着广阔的应用前景.

本文在研究相关向量机模型过程中, 除了运用进化计算优化其核函数参数等参量外, 还可以从核函数的选择等方面作进一步的研究. 由于经典 RVM 是单

一核函数的, 对RVM多核算法的研究将是一个重要方向.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995: 11-13.

[2] 聂盼盼, 臧洌, 刘雷雷. 基于对支持向量机的多类分类算法在入侵检测中的应用[J]. 计算机应用, 2013, 33(2): 426-429.
(Nie P P, Zang L, Liu L L. Application of multi-class classification algorithm based on twin support vector machine in intrusion detection[J]. J of Computer Applications, 2013, 33(2): 426-429.)

[3] Sankar Mahadevan, Sirish L Shah. Fault detection and diagnosis in process data using one-class support vector machines[J]. J of Process Control, 2009, 19(10): 1627-1639.

[4] Duan Qing, Zhao Jianguo, Ma Yan. RVM and SVM for classification in transient stability assessment[C]. 2010 Asia-Pacific Power and Energy Engineering Conf (APPEEC). Chengdu, 2010: 1-4.

[5] Bilgin G, Erturk S, Yildirim T. Segmentation of hyper spectral images via subtractive clustering and cluster validation using one-class support vector machines[J]. IEEE Trans on Geoscience and Remote Sensing, 2011, 49(8): 2936-2944.

[6] 杨国鹏, 周欣, 余旭初. 稀疏贝叶斯模型与相关向量机学习研究[J]. 计算机科学, 2010, 37(7): 225-228.
(Yang G P, Zhou X, Yu X C. Research on sparse Bayesian model and the relevance vector machine[J]. Computer Science, 2010, 37(7): 225-228.)

[7] Subimal Ghosh, Mujumdar P P. Statistical downscaling of GCM simulations to stream flow using relevance vector machine[J]. Advances in Water Resources, 2008, 31(1): 132-146.

[8] Clodoaldo A M Lima, AndréL V Coelho, Sandro Chagas. Automatic EEG signal classification for epilepsy diagnosis with relevance vector machines[J]. Expert Systems with Applications, 2009, 36(6): 10054-10059.

[9] John Flake, Todd K Moon, Mac McKee, et al. Application of the relevance vector machine to canal flow prediction in the Sevier River Basin [J]. Agricultural Water Management, 2010, 97(2): 208-214.

[10] 夏俊杰, 何迪. 基于相关向量机的网络入侵检测算法[J]. 信息安全与通信保密, 2010, 8: 47-51.
(Xia J J, He D. Intrusion detection method based on relevance vector machine[J]. Information Security and Communications Privacy, 2010, 8: 47-51.)

[11] 王立坤, 杨新锋. 一种基于RVM回归的分类方法[J]. 电子科技, 2011, 24(5): 14-17.
(Wang L K, Yang X F. A classification method based on RVM regression[J]. Electronic Science and Technology, 2011, 24(5): 14-17.)

[12] Suresh S, Sujit P B, Rao A K. Particle swarm optimization approach for multi-objective composite box-beam design[J]. Composite Structures, 2007, 81(4): 598-605.

[13] Tipping M E, Faul A C. Fast marginal likelihood maximization for sparse Bayesian models[C]. Proc of the 9th Int Workshop on Artificial Intelligence and Statistics. Key West, 2003: 3-6.

[14] Platt J C, Cristianini N, Shawe Taylor J. Large margin DAG for multiclass classification[C]. Advances in Neural in Formation Processing Systems. Cambridge: MIT Press, 2000: 547-553.

[15] 夏克文, 高峰, 武睿, 等. 云粒子群优化算法在无线传感器网络中的应用[J]. 控制理论与应用, 2011, 28(9): 1175-1178.
(Xia K W, Gao F, Wu R, et al. Optimal wireless sensor network using cloud adaptive particle-swarm-optimization algorithm[J]. Control Theory & Applications, 2011, 28(9): 1175-1178.)

[16] 杨树仁, 沈宏远. 基于相关向量机的机器学习算法研究与应用[J]. 计算技术与自动化, 2010, 29(1): 43-47.
(Yang S R, Shen H Y. Research and application of machine learning algorithm based on relevance vector machine[J]. Computing Technology and Automation, 2010, 29(1): 43-47.)

(责任编辑: 李君玲)