

自回归滑动平均和相关向量机的信息系统安全态势估计

李欣蔚¹, 陈 辉², 张亚红³

(1 桂林医学院, 广西 桂林 541004; 2 桂林电子科技大学 计算机与信息安全学院,
广西 桂林 541004; 3 桂林航天工业学院 计算机科学与工程系, 广西 桂林 541004)

摘 要: 为了改善信息系统的安全态势估计效果, 提出了自回归滑动平均和相关向量机的信息系统安全态势估计模型 (ARIMA-RVM). 该模型收集大量信息系统安全的历史数据, 然后采用自回归滑动平均和相关向量机描述信息系统安全态势变化特点, 建立信息系统安全态势估计模型, 最后与其他信息系统安全态势估计模型进行了对比测试, 结果表明, ARIMA-RVM 可以发现信息系统安全态势的变化趋势, 获得比其他模型更好的信息系统安全态势估计结果, 估计结果可以帮助信息系统安全管理者制定防范措施.

关键词: 信息系统; 数据挖掘; 安全态势; 估计模型

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2017)09-0141-04

Information System Security Situation Assessment Based on Data Mining

LI Xin-wei¹, CHEN Hui², ZHANG Ya-hong³

(1 Guilin Medical University, Guilin 541004, China; 2 School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; 3 Department of Information Engineering, Guilin University of Aerospace Technology, Guilin 541004, China)

Abstract: In order to estimate the effect of improving security situation information system, put forward the estimation model of information system based on data mining security situation. The method of collecting a lot of information system security of the historical data, and then using data mining combined kernel function to dig out the characteristics of information system security situation from the historical data, the establishment of information system security situation assessment model, finally model were compared, and the other information system security situation assessment results show that the model can be found to change the trend of information system security situation, get information system security situation better contrast model estimation results, the estimation results can improve the security of the information system and formulate corresponding preventive measures.

Key words: information system; data mining; security situation; estimation model

1 引言

随着信息技术的发展, 各行各业都有自己的信息系统, 而信息安全是信息系统一项关键技术, 由于管理人员安全意识不同, 以及非法攻击手段的多样化, 如何对信息系统的安全态势进行准确估计变得越来越重要^[1-3].

当前信息系统安全态势估计主要采用时间序列

估计方法进行研究, 首先收集不同时间段的信息安全数据, 然后采用数据挖掘技术, 如自回归滑动平均 (ARIMA)、神经网络、相关向量机 (RVM) 等^[4-7] 对信息系统安全态势建立估计模型, 在所有模型中, ARIMA 属于线性模型, 而信息系统安全态势估计属于非线性估计问题, 因此 ARIMA 常对数据进行平稳化处理, 导致信息系统安全态势估计精度较低^[8-9]. 对于 ARIMA, 神经网络的非线性估计能力

更强,但要求训练样本多,而且收敛速度慢,这会导致信息系统安全态势估计结果不可靠^[10-12]. RVM 的估计性能要优于神经网络,但其性能与核函数类型及参数直接相关,采用单一核函数建立信息系统安全态势估计模型,估计精度有待进一步提高^[13].

为了提高信息系统安全态势估计精度,针对当前信息系统安全态势估计模型存在的不足,并结合信息系统安全态势的变化特点,提出了基于自回归滑动平均和相关向量机的信息系统安全态势估计模型(ARIMA-RVM),利用 ARIMA 和 RVM 的优势,分别建立信息系统安全态势估计模型,针对它们对信息系统安全态势估计趋势项和随机项分别进行建模与估计,最后对它们的估计结果进行叠加得到信息系统安全态势的最后估计结果,并通过仿真实验测试其性能.

2 自回归滑动平均和相关向量机

2.1 ARIMA

ARIMA 是一种经典的线性建模方法,可以对信息系统安全态势的趋势项建模,主要步骤如下:

(1) 信息系统安全态势数据具有非平稳性,首先对信息系统安全态势数据进行差分处理,变为平稳数据,设差分处理后的平稳数据为 $\{x_t'\}$.

(2) 根据信息系统安全态势样本计算自相关和偏自相关函数,建立信息系统安全态势估计的初步模型类型, $\{x_t'\}$ 的 ARIMA 模型可以表示为

$$x_t' = \sum_{j=1}^p a_j x_{t-j}' + \sum_{k=0}^q b_k e_{t-k} \quad (1)$$

(3) 采用 AIC 估计式(1)参数 p, q 的值.

(4) 采用矩阵估计法确定式(1)参数 a_j, b_k 的值,建立信息系统安全态势估计的 ARIMA 模型.

2.2 相关向量机

设信息系统安全态势训练样本为 $\{x_i, t_i\}_{i=1}^n$, ($i=1, 2, 3, \dots, n$), x_i 与 t_i 分别为信息系统安全态势输入和输出,那么相关向量机(RVM)可以描述两者之间的关系,则有

$$t_i = y(x_i, \omega_i) + \omega_i \quad (2)$$

式中, ω_i 表示数据中包含的噪声.

引入核函数 $K(x, x_i)$, 可以得到

$$y(x; \mathbf{w}) = \sum_{i=1}^n w_i K(x, x_i) + w_0 \quad (3)$$

式中, $\mathbf{w}$ 表示权向量.

第 i 个训练样本的概率分布计算公式为

$$P(t_i | x_i) = N(t_i | y(x_i; \mathbf{w}), \sigma^2) \quad (4)$$

采用超参数 β 建立全部信息系统安全态势的训练样本最大似然函数,得到

$$p(\mathbf{t} | \mathbf{w}, \beta) = \left(\frac{\beta}{2}\right)^{N/2} \exp\left\{-\frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2\right\} \quad (5)$$

式中, $\mathbf{t} = [t_0, t_1, \dots, t_N]^T$, $\Phi \in R^{N \times (N+1)}$ 表示矩阵.

a_j^{-1} 表示 w_j 的方差,那么其高斯分布定义为

$$p(\mathbf{w} | \mathbf{a}) = \prod_{j=0}^n N(w_j | 0, a_j^{-1}) \quad (6)$$

式中, $\mathbf{a} = [a_0, a_1, \dots, a_N]^T$ 为超参数.

采用贝叶斯公式计算 $\mathbf{w}$ 后验分布,得到

$$p(\mathbf{w} | \mathbf{t}, \mathbf{a}, \mathbf{a}) = \frac{p(\mathbf{w} | \mathbf{a}) p(\mathbf{t} | \mathbf{w}, \beta)}{p(\mathbf{t} | \mathbf{a}, \beta)} \quad (7)$$

对式(7)进行变换,可以得到

$$p(\mathbf{w} | \mathbf{t}, \mathbf{a}, \mathbf{a}) = N(\mathbf{w} | \mu, \Sigma) \quad (8)$$

式中, Σ 和 μ 的计算公式分别为

$$\Sigma = (\beta \Phi^T \Phi + \mathbf{A})^{-1} \quad (9)$$

$$\mu = \beta \Sigma \Phi^T \mathbf{t} \quad (10)$$

式中, $\mathbf{A}$ 表示对角矩阵.

a_j 和 β 的值为

$$a_j = \frac{1}{\mu_j^2 + \sum_j \mu_j^2} \gamma_j, \quad j = 0, 1, \dots, n \quad (11)$$

$$\beta = (n - \sum_{j=0}^n (1 - a_j \sum_{jj})) / \|\mathbf{t} - \Phi \mathbf{w}\|^2 \quad (12)$$

式中, μ_j 为 μ 的第 j 个元素, $\sum_{jj}$ 是矩阵 Σ 的第 j 个对角元.

对于新输入信息系统安全态势样本 x_* , 可以通过式(13)得到它的估计值(t_*).

$$t_* = \varphi(x_*) \mu \quad (13)$$

RVM 核函数直接影响信息系统安全态势估计结果,当前满足条件的核函数很多,最常用的多项式核函数和 RBF 核函数,它们具体为

$$K_{\text{poly}} = ((x \cdot x_i) + 1)^d \quad (14)$$

$$K_{\text{rbf}} = \frac{\exp(-|x - x_i|^2)}{2\sigma^2} \quad (15)$$

单核函数的缺陷比较明显,因此本文构造组合核函数实现信息系统安全态势估计,具体为

$$K_{\text{mixed}} = \rho_1 K_{\text{poly}} + \rho_2 K_{\text{rbf}} \quad (16)$$

式中, ρ_1, ρ_2 为权重.

3 信息系统安全态势估计模型的构建

受到多种因素的综合作用,信息系统安全态势具有周期性和随机性变化特点,采用单一模型无法

进行准确描述,因此,采用小波变换对其进行分解,得到趋势项和随机项,然后分别进行建模和估计,具体工作步骤如下:

(1)采用小波变换对原始信息安全态势数据(W)进行 j 层分解,得到趋势项(X_{j+1})和随机项(XZ_{j+1})分别为

$$\begin{cases} X_{j+1} = HX_j \\ Z_{j+1} = GZ_j, j = 0, 1, \dots, J \end{cases} \quad (17)$$

式中, H 和 G 分别为高通滤波器和为低通滤波器.

(2)最后对 ARIMA 和 RVM 的信息系统安全态势的估计结果,即有

$$x_j = H^* x_{j+1} + G^* z_{j+1}, j = J - 1, J - 2, \dots, 0 \quad (18)$$

式中, H^* 和 G^* 表示 H 和 G 的对偶算子.

基于 ARIMA-RVM 的建模思想是:首先 ARIMA对信息安全态势的趋势项进行估计,然后采用 SVM对信息安全态势的随机项进行估计,最后进行叠加得到信息安全态势估计结果,工作原理如图 1 所示.

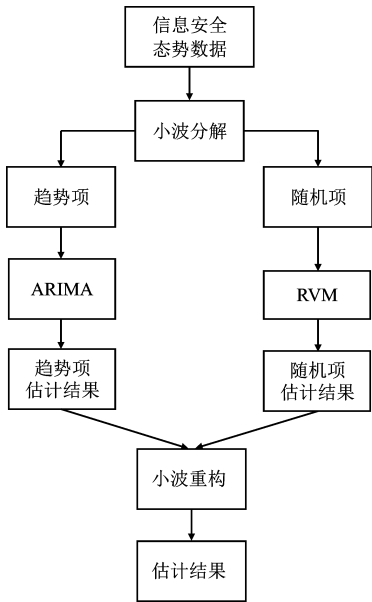


图 1 信息安全态势估计的工作流程

4 性能测试实验结果

4.1 数据来源

为了测试 ARIMA-RVM 的信息系统安全态势估计模型,选取某一个企业的信息安全数据作为研究对象,选择前 200 个样本数据作为训练样本集,其余样本数据测试模型的泛化能力,具体如图 2 所示.并选择 ARIMA、RVM 进行信息安全安全态

势估计对比测试.采用均方根误差(RMSE)对信息安全态势的估计效果进行评价,计算公式为

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n}} \quad (19)$$

式中, x_i 是原始信息安全态势的值, $\hat{x}_i$ 信息安全态势的估计值, n 为样本数量.

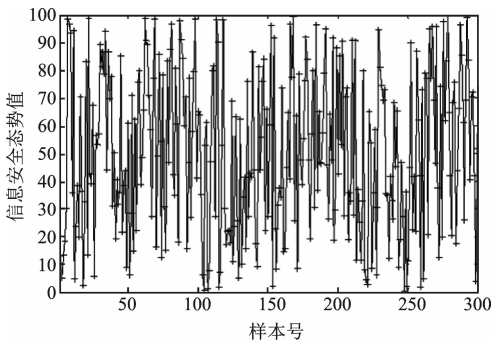


图 2 信息安全态势样本

4.2 核函数的选择

将前 200 个信息安全态势数据输入 RVM 进行训练,分别采用不同的核函数,对其他样本进行估计,统计它们的 RMSE,结果如表 1 所示.对表 1 的信息系统安全态势估计误差进行比较和分析发现,多项式核函数的信息安全态势估计误差最大,这是因为其无法全面描述信息安全态势变化特点,而组合核函数的信息安全态势估计误差要小于单一核函数,结果表明本文选择组合核函数的 RVM 建立信息安全态势估计模型是可行的.

表 1 不同核函数的 RMSE 比较

核函数类型	RMSE
多项式核函数	4.95
RBF 核函数	3.03
组合核函数	2.15

4.3 结果与分析

为了进一步提高信息安全态势的估计精度,ARIMA-RVM 对其进行估计,得到的结果如图 3 所示,同时采用 ARIMA、RBF-SVM 对信息安全态势进行建模和估计,它们的估计结果如图 4 和图 5 所示.对图 3~5 的信息系统安全态势的估计结果进行分析,可以发现,ARIMA-RVM 的信息系统安全态势的估计结果最优,可以更好地模拟出未来信息安全变化趋势,主要是因为采用 ARIMA 对信息安全态势趋势部分进行捕捉,克服

了单一模型的局限性,提高了信息系统安全态势的估计精度,单 ARIMA 和 RBF-RVM 的信息系统安全态势的估计精度较低,有许多点估计误差仍然很大.

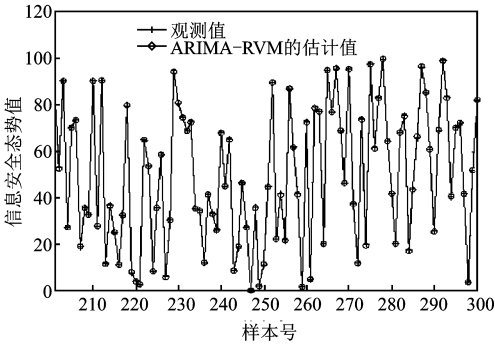


图3 ARIMA-RVM 的信息系统安全态势估计结果

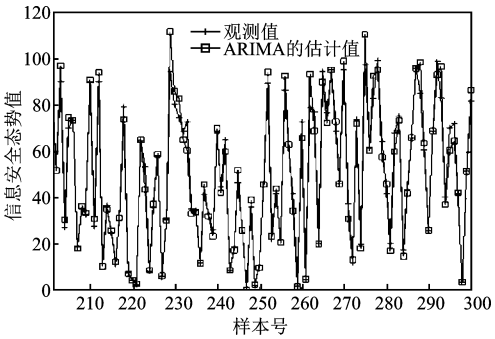


图4 ARIMA 的信息系统安全态势估计结果

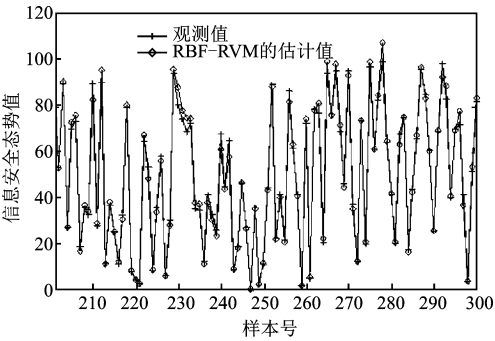


图5 RBF-RVM 信息系统安全态势估计结果

5 结束语

针对传统信息系统估计方法的安全态势估计精度低,无法满足信息系统实际应用的难题,提出了基于数据挖掘的信息系统安全态势估计模型,结果表明,本文模型可以发现信息系统安全态势的变化趋势,获得比其他模型更好的信息系统安全态势估计结果,具有较好的实际应用价值.

参考文献:

[1] 张勇,谭小彬,崔孝琳,等. 基于 Markov 博弈模型的信息安全态势感知方法[J]. 软件学报, 2011, 22(3): 495-508.

[2] 贾焰,王晓伟,韩伟红,等. YHSSAS:面向大规模信息的安全态势感知系统[J]. 计算机科学, 2011, 38(2): 4-8.

[3] 徐茹枝,常太华,吕广娟. 基于时间序列的信息安全态势估计方法的研究[J]. 数学的实践与认识, 2010, 40(12):124-133.

[4] 孟锦,马驰,何加浪,等. 基于 HHGA-RBF 神经网络的信息安全态势估计模型[J]. 计算机科学, 2011, 38(7): 70-73.

[5] 王庚,张景辉,吴娜. 信息安全态势估计方法的应用研究[J]. 计算机仿真, 2012, 29(2):98-101.

[6] 唐作其,陈选文,戴海涛,等. 多属性群决策理论信息安全风险评估方法研究[J]. 计算机工程与应用, 2011, 47(15):104-106.

[7] 王帆,霍明奎,王晓婷. 基于模糊灰度的信息系统安全风险评价与对策[J]. 情报科学, 2014, 32(1): 110-114.

[8] 王桢珍,姜欣,武小悦,等. 信息安全风险概率计算的贝叶斯网络模型 [J]. 电子学报, 2010, 38(2):18-22.

[9] 付钰钰,吴晓平,叶清,等. 模糊推理与多重结构神经网络在信息系统安全风险评价中的应用 [J]. 海军工程大学学报, 2011, 23(1): 10-16.

[10] 李方伟,郑波,朱江,等. 一种基于 RBF 神经网络的信息安全风险估计模型[J]. 重庆邮电大学学报(自然科学版), 2014, 26(5): 576-583.

[11] 阮慧,党德鹏. 基于 RBF 模糊神经网络的信息安全风险评估[J]. 计算机工程与设计, 2011, 32(6): 2113-2115.

[12] 德鹏,孟真. 基于支持向量机的信息安全风险评估[J]. 华中科技大学学报(自然科学版), 2010, 38(3): 46-49.

[13] 楼俊钢,江建慧,沈张果,等. 软件可靠性估计的相关向量机模型[J]. 计算机研究与发展, 2013, 50(7): 1542-1550.

作者简介:

李欣蔚 男, (1973-), 硕士, 高级工程师. 研究方向为计算机网络. E-mail:lixingw908@163.com

陈 辉 男, (1974-), 博士, 副研究员. 研究方向为模式识别与智能控制、智能检测与非线性控制.

张亚红 女, (1977-), 硕士, 副教授. 研究方向为图像处理与识别、信息安全.