高技术通讯 2012 年 第22 卷 第2 期:140 ~ 146

doi:10.3772/j.issn.1002-0470.2012.02.005

基于场景平移的网络安全态势预测①

李志东② 杨 武 王 巍 茆大鹏

(哈尔滨工程大学信息安全研究中心 哈尔滨 150001)

摘 要 针对网络安全态势序列复杂多变,蕴含各种各样的演化规律,传统网络安全态势预测方法难以处理的问题,提出了一种专用的预测算法,该算法从长程相关的视角辨识态势序列蕴含的规律,依据事发迹象推断延续效应,经相似度、普遍性、对比度和缩放比加权后,合成预测序列。继而引入进化算法,依据预测效果调节相关参数,通过在线反馈式学习强化泛例的作用、弱化特例的干扰,提升预测算法的适应性。实验表明,该预测算法从超长态势序列中辨识多种类远距离相关性的能力很强,能对复杂多变的趋势保持自适应,预测结果更为精准可信。

关键词 网络安全,安全态势,趋势预测,场景平移,长程相关

0 引 言

随着网络的日趋庞大、复杂和多变,网络的安全威胁也趋于多元化,使得传统的静态被动防御或全面系统加固愈发显得不堪重负(难以跟上变化节奏)或代价高昂(耗费大量投资、影响网络性能),由此催生出了动态的、主动的、有针对性的防御,而这种防御是以切实有效的态势预测为基础的。态势预测旨在依据历史规律及当前迹象推测未来网络安全状况的演化趋势,以辅助或指导动态防御,使管理员能提前采取应对措施,及时有效地对复杂多变的安全威胁做出快速响应。

态势评估是态势预测的基础,该领域的研究起步较早。Bass[1]基于入侵检测系统,借助多传感器数据融合评估安全态势。陈秀真等人[2]从网络、主机、服务、漏洞、攻击等层面评估了威胁态势。态势预测的专项研究很少,且大都沿用现有预测算法或模型,典型的有自回归移动平均模型(autoregressive moving average model, ARMA)、灰色模型(grey model, GM)、径向基函数神经网络(radial basis function neural network, RBFNN)和隐马尔可夫模型(hidden Markov model, HMM)。ARMA 辨识态势序列中蕴含的依存关系或自相关性[3,4],抽象出数学预测模型。它要求态势序列或其某级差分满足平稳性假设,这种前提条件过于苛刻,极大地限制了其适用范围。灰色预测中的 GM(1,1)模型[5]先借助累加削弱态势序列的随机性,再使用指数曲线拟合生成序列,预测后经累减还原,能体现单调缓变趋势,难以反映随机游走、周期波动等特征。Grey Verhulst 比较适合描述按“S”或反“S”形摆动发展的态势序列[6],将变迁划分为若干阶段也不乏合理性,但难点在于预测各阶段的发生时刻及持续时长。RBFNN 使用非线性映射[7]描述态势序列中蕴含的规律,然而演化规律是无限的、多变的,远非中小规模实用型神经网络借助某次训练就能应对的,需要但也很难做到持续增量学习,更为棘手的是消解样本间的冲突,特征相似而期望输出差异很大的现象在态势预测中极为常见。HMM 方法[8]使用双重随机过程描述态势序列中蕴含的规律,依据观察值序列确定初始状态概率分布矩阵、状态转移概率矩阵和输出概率分布矩阵,其状态数目和可能的观察值数目均难确定,牵涉到存储开销、训练速度与预测精度的折中。综合而言,安全态势序列蕴含了一系列复杂多变的随机趋势,突发性很强,不确定性很大,很难辨识和描述其规律,不是传统算法靠某个公式或函数就能表达及预测的,依据超长态势序列估计参数既很必要、也很困难,沿用现有的预测算法将难

① 863 计划(2007AA01Z473)和国家 242 信息安全计划(2007B17)资助项目。

② 男,1975 年生,博士生;研究方向:网络安全态势评估与预测;联系人,E-mail: zhdlee@126.com

(收稿日期:2010-06-21)

以避免与该领域特有问题的严重脱节。针对以上问题,本文提出了一种基于场景平移的预测方法,从录制的历史态势序列中查找相似迹象,衡量事发迹象与延续效应的联系强度,依据当前迹象推测某种效应重现的可能性,加权合成预测结果,另辅以进化算法,从形态及精度上计量预测偏差,通过逐步微调持续提升适应性。

1 概述

1.1 基础定义

按等时隔 τ 采样网络安全态势值,构成离散型时间序列。以 z_k 表示 t_k 时刻的态势值,将横纵坐标为 (t_k,z_k) 的点连成折线图,如第 5 节图 1 所示。

定义 1 令 $\boldsymbol{G}(i,m)$ 指代始于 t_i 时刻含 m 段邻接线的子图。线段斜率 q_k 按式

$$q_k = (z_{k+1} - z_k)/(t_{k+1} - t_k) \tag{1}$$

计算,斜率序列 $(q_i,q_{i+1},\cdots,q_{i+m-1})$ 用 $\boldsymbol{Q}(i,m)$ 向量表示。

定义 2 以 $\boldsymbol{L}(i,m)$ 表示 $\boldsymbol{Q}(i,m)$ 的特征谱,符号 $\boldsymbol{O}$ 指代零向量。当 $\boldsymbol{Q}(i,m) = \boldsymbol{O}$ 时,$\boldsymbol{L}(i,m) = \boldsymbol{O}$,否则按式

$$\boldsymbol{L}(i,m) = \boldsymbol{Q}(i,m)/\|\boldsymbol{Q}(i,m)\| \tag{2}$$

表述为 $\boldsymbol{Q}(i,m)$ 的单位化向量。针对 $\boldsymbol{L}(i,m)$ 的第 k 个分量 l_{i+k},定义其倾角 θ_{i+k} 等于 $\arctan l_{i+k}$,满足 $-\pi/2 < \theta_{i+k} < \pi/2$ 约束。

定义 3 函数 $f_\sigma(i,j,m)$ 计算从 $\boldsymbol{Q}(i,m)$ 到 $\boldsymbol{Q}(j,m)$ 的伸缩比。当 $\boldsymbol{Q}(i,m) = \boldsymbol{O}$ 且 $\boldsymbol{Q}(j,m) = \boldsymbol{O}$ 时 $f_\sigma(i,j,m) = 1$;当 $\boldsymbol{Q}(i,m) \neq \boldsymbol{O}$ 且 $\boldsymbol{Q}(j,m) \neq \boldsymbol{O}$ 时 $f_\sigma(i,j,m) = \|\boldsymbol{Q}(j,m)\| / \|\boldsymbol{Q}(i,m)\|$;其余的情况不会被用到,没有定义。

1.2 基本思想

尽管从逻辑上看,依据结果反推起因、依据历史预测未来均缺乏理论支撑,但从概率论和统计学上看,相似的态势曲线形状更可能源自相似的起因、机理及影响,继而引发相似的延续效应。

假设 $\boldsymbol{G}(i,m)$ 与 $\boldsymbol{G}(j,m)$ 是两个已知的历史子图,摘自同一总图,满足 $t_i < t_j$,而 $t > t_{j+m}$ 的未来趋势是未知的、有待预测的。

如果 $\boldsymbol{G}(j,m)$ 与 $\boldsymbol{G}(i,m)$ 相似,那么能较为可信地推断:在 $[t_j,t_{j+m})$ 时段内发生的情况、作用的机理、造成的影响与 $[t_i,t_{i+m})$ 时段的也比较相似,虑及延续效应,t_{i+m} 之后的历史很可能会在 t_{j+m} 之后重演,尽管幅度未必一致。

据此原理,可按 $\hat{q}_{j+m+k} = f_\sigma(i,j,m) \times q_{i+m+k}$ 预测后续的线段斜率。使用参数 ρ 控制预测步数,当 $k = 0,1,2,\cdots,\rho - 1$ 时,按步距 τ、斜率 $\hat{q}_{j+m+k}$ 逐段衔接递推,构成趋势预测曲线。

其本质是依据事发迹象 $\boldsymbol{G}(j,m)$ 与 $\boldsymbol{G}(i,m)$ 的相似、推断延续效应 $\hat{\boldsymbol{G}}(j+m,\rho)$ 与 $\boldsymbol{G}(i+m,\rho)$ 的相似,将 t_i 之后的场景平移至 t_j 之后,按预期规模重建。

2 支撑体系

2.1 拟合度

定义 4 函数 $\phi_\Theta(i,j,m)$ 计算 $\boldsymbol{Q}(i,m)$ 与 $\boldsymbol{Q}(j,m)$ 之间的夹角相似度:当两者均为零向量时 $\phi_\Theta(i,j,m) = 1$;当有且仅有一个是零向量时 $\phi_\Theta(i,j,m) = 0$;当两向量均非零时,按式

$$\phi_\Theta(i,j,m) = \frac{\boldsymbol{Q}(i,m) \times \boldsymbol{Q}^{\mathrm{T}}(j,m)}{\|\boldsymbol{Q}(i,m)\| \times \|\boldsymbol{Q}(j,m)\|} \tag{3}$$

定义为夹角余弦。

定义 5 函数 $\phi_1(x)$、$\phi_2(x)$ 分别计算质变、量变趋势,各自关注形态、精度差异。当 x 值大于、等于、小于零时,$\phi_1(x)$ 值分别等于 +1、0、-1,表示上升、平稳、下降。$\phi_2(x)$ 选用 $\sin x$。综合趋势 $\phi_\perp(x)$ 按式

$$\phi_\perp(x) = 0.2 \times \phi_1(x) + 0.8 \times \phi_2(x) \tag{4}$$

计算。

定义 6 以 ∇ 符号表示向后差分算子,设 $\nabla^0\theta_k = \theta_k$,针对正整数 α,按式

$$\nabla^\alpha\theta_k = 0.5 \times (\nabla^{\alpha-1}\theta_k - \nabla^{\alpha-1}\theta_{k-1}) \tag{5}$$

定义 α 阶差分递推式,以便从更深层次辨析趋势差异,满足 $-\pi/2 < \nabla^\alpha\theta_k < \pi/2$ 约束。

定义 7 特征谱 $\boldsymbol{L}(i,m)$ 与 $\boldsymbol{L}(j,m)$ 之间的趋势差 $\phi_\nabla(i,j,m)$ 按式

$$\phi_\nabla(i,j,m) = (0.5 \times m \times (m+1))^{-1} \times \sum_{\alpha=0}^{m-1}\sum_{k=\alpha}^{m-1} |\phi_\perp(\nabla^\alpha\theta_{i+k}) - \phi_\perp(\nabla^\alpha\theta_{j+k})| \tag{6}$$

计算。

定义 8 拟合度函数 $\phi(i,j,m)$ 是按式

$$\phi(i,j,m) = \phi_\Theta(i,j,m) - \phi_\nabla(i,j,m) \tag{7}$$

定义的,鉴于 $-1 \leqslant \phi_\Theta(i,j,m) \leqslant 1$、$0 \leqslant \phi_\nabla(i,j,m) < 2$,故 $\phi(i,j,m)$ 满足 $-3 < \phi(i,j,m) \leqslant 1$ 约束,取值越大,拟合越佳。

统计而言，满足 $\phi_{\Theta}(i,j,m) > 0$ 的情况约占半数，过于宽泛，因此，有必要先减去 $\phi_{\nabla}(i,j,m)$ 惩罚项，再按阈值 ε_{ϕ} 筛选 $\phi(i,j,m)$，以达到窄化的目的。ε_{ϕ} 初值为 0.2，满足 $0 \leqslant \varepsilon_{\phi} < 1$ 约束。

2.2 普适度

定义 9 在历史态势图 $\boldsymbol{G}(0,n)$ 中，使用标量 $\chi[k,m,\rho]$ 表示 $\boldsymbol{Q}(k,m+\rho)$ 的普适量，$\chi_{\max}$ 值按式

$$\chi_{\max} = \max\{\chi[k,m,\rho] \mid 0 \leqslant k < n-m\} \quad (8)$$

求取，初值均为 0.0。

定义 10 函数 $f_{\chi}(k,m,\rho)$ 依据式

$$f_{\chi}(k,m,\rho) = (n-m) \times (1 + 2 \times \pi^{-1} \times \arctan(\chi[k,m,\rho] - \chi_{\max})) \quad (9)$$

将普适量映射为普适度，落入 $(0,n-m]$ 区间。

$f_{\chi}(k,m,\rho)$ 采用随变量程，以免因计算精度受限而泯灭极小值或差异值，函数曲线呈大值敏化小值拖尾形状。其中 $\chi_{\max}$ 是缓存值，不必现算。

普适度越大，表明 $\boldsymbol{Q}(k,m)$ 及其外延的代表性越好，依据 $\boldsymbol{Q}(k,m+\rho)$ 预测出的形态也更精准；否则，表明 $\boldsymbol{Q}(k,m+\rho)$ 只是较为罕见的特例，据其预测的效果欠佳。

2.3 对比度

假设有 n 个未归一权重 $\tilde{w}_1,\tilde{w}_2,\cdots,\tilde{w}_n$，经敏化指数 η 调整成 $\tilde{w}_1^{\eta},\tilde{w}_2^{\eta},\cdots,\tilde{w}_n^{\eta},\eta > 0$，标准化权重 w_k 按式

$$w_k = \tilde{w}_k^{\eta} / \sum_{k=1}^{n} \tilde{w}_k^{\eta} \quad (10)$$

计算，对比度 $\frac{w_i}{w_j}$ 等于 $\frac{\tilde{w}_i^{\eta}}{\tilde{w}_j^{\eta}}$。当 $\tilde{w}_i \neq \tilde{w}_j$ 时，不妨假设 $\tilde{w}_i < \tilde{w}_j$，按升序将界点排列为 $0,\frac{\tilde{w}_i}{\tilde{w}_j},1,\frac{\tilde{w}_j}{\tilde{w}_i},\infty$。

在钝化区（$0 < \eta < 1$），当 η 取值从 1 向 0 递减时，w_i/w_j 从 $\tilde{w}_i/\tilde{w}_j$ 向 1 递增，w_j/w_i 从 $\tilde{w}_j/\tilde{w}_i$ 向 1 递减，从而弱化 w_i 与 w_j 的差异。在锐化区（$\eta > 1$），当 η 从 1 向 ∞ 递增时，w_i/w_j 从 $\tilde{w}_i/\tilde{w}_j$ 向 0 递减，w_j/w_i 从 $\tilde{w}_j/\tilde{w}_i$ 向 ∞ 递增，从而突显 w_i 与 w_j 的差异。当 $\eta = 1$ 时无敏化。若以 $\tilde{w}_k^{\eta} = w_k \times \xi$ 表述式(10)，则可按式

$$\frac{\tilde{w}_k^{\eta \times x}}{\sum_{k=1}^{n} \tilde{w}_k^{\eta \times x}} = \frac{(w_k \times \xi)^x}{\sum_{k=1}^{n} (w_k \times \xi)^x} = \frac{w_k^x}{\sum_{k=1}^{n} w_k^x} \quad (11)$$

导出级联递推式。

利用 η 调节对比度，当 $0 < \eta < 1$ 时突出统计效果，在 $\eta > 1$ 时彰显个体优势，初值为 1.0。

3 预测算法

算法 1 给出了预测伪码描述，旨在依据事发迹象 $\boldsymbol{G}(n-m,m)$ 推断延续效应 $\hat{\boldsymbol{G}}(n,\rho)$，其中的 $\gamma[i,m,\rho]$ 是以 1.0 为初值的参数。

算法 1 依据历史图推出预测图。

输入：历史图 $\boldsymbol{G}(0,n)$ 及相关设置。

输出：预测图 $\hat{\boldsymbol{G}}(n,\rho)$。

```
01. Ψ←∅, ξ←0, j←n−m
02. FOR i = 0 TO j − 1 DO
03.   IF φ(i,j,m) > ε_φ THEN
04.     w←(f_χ(i,m,ρ) × φ(i,j,m))^η
05.     σ←f_σ(i,j,m) × γ[i,m,ρ]
06.     Ψ←Ψ ∪ {(i,w,σ)}
07.     ξ←ξ + w
08.   ENDIF
09. ENDFOR
10. (∀(i,w,σ) ∈ Ψ)(w←w/ξ)
11. IF Ψ ≠ ∅ THEN
12.   FOR k = 0 TO ρ − 1 DO
13.     q̂_{n+k} ← Σ_{(i,w,σ)∈Ψ} (w × σ × q_{i+m+k})
14.     ẑ_{n+k+1} ← z_{n+k} + τ × q̂_{n+k}
15.   ENDFOR
16. ENDIF
```

第 1 行初始化各个变量。第 2、9 行循环推进滑动窗口。第 3、8 行剔除拟合度太差的 $\boldsymbol{Q}(i,m)$ 向量。第 4 行先求普适度与拟合度的乘积，再将乘积的敏化值记入 w。第 5 行利用 $\gamma[i,m,\rho]$ 调节伸缩比。第 6 行将拟合位置、权重、伸缩比三元组记录到 Ψ 集合中。第 7 行把 w 值累加至 ξ 中。第 10 行执行权重归一化。

第 11、12 行检查 Ψ 中有无记录，若没有则拒绝预测。第 12、15 行循环推进预测位。第 13 行枚举 Ψ 集合，依据 (i,w,σ) 记录，将 q_{i+m+k} 缩放至预期尺度，按 w 加权后，叠加到预测斜率 $\hat{q}_{n+k}$ 上。第 14 行求出纵坐标值 $\hat{z}_{n+k+1}$。依此类推，预测出 $\{(t_{n+k},\hat{z}_{n+k}) \mid 1 \leqslant k \leqslant \rho\}$ 点集。

当 $\boldsymbol{Q}(i,m)$ 的普适度越小、或与 $\boldsymbol{Q}(j,m)$ 的拟合度越小时，其外延斜率 q_{i+m+k} 的权重 w_i 就越小，对预测值 $\hat{q}_{n+k}$ 的贡献也越小。

第 2－9 行的外循环不会超过 n 次，循环体呈 $O(m^2)$ 级；第 10 行呈 $O(n)$ 级；第 12－15 行的外循环为 ρ 次，循环体呈 $O(n)$ 级；故综合时间复杂度为 $O(n \times (m^2 + \rho))$，当 m 和 ρ 取常数时视为 $O(n)$。

4 进化算法

定义 11 函数 $f_\kappa(x)$ 基于当前权重集合,按式

$$f_\kappa(x)=\frac{\sum_{(i,w,\sigma)\in\Psi}(w^x\times\phi(i+m,n,\rho))}{\sum_{(i,w,\sigma)\in\Psi}w^x} \tag{12}$$

计算 η 调为 $\eta\times x$ 时的精准度,其正确性可依据式(11)证明,满足 $-3<f_\kappa(x)\leqslant 1$ 约束。继而定义 $f_\Lambda(x_1,x_2,\cdots)$ 函数,$f_\Lambda(x_1,x_2,\cdots)=x_i$ 当且仅当 $f_\kappa(x_i)$ 是在 $\{f_\kappa(x_1),f_\kappa(x_2),\cdots\}$ 集合中依据 x_k 的下标 k 按升序枚举时首次遇到的最大值。

算法 2 承袭了预测算法的变量及结果,在获知实测值之后执行,旨在提升适应性。

算法 2 依据实测值调节相关参数。

输入:$\varepsilon_\phi,\chi,\gamma,\eta,\boldsymbol{G}(0,n+\rho),\Psi,j$。

输出:$\varepsilon_\phi,\chi,\gamma,\eta$。

01. $\Delta\varepsilon_\phi\leftarrow((|\Psi|-\ln n)/(|\Psi|+\ln n))/n$
02. $\varepsilon_\phi\leftarrow\min\{\max\{\varepsilon_\phi+\Delta\varepsilon_\phi,0\},1-1/n\}$
03. FOR EACH $(i,w,\sigma)\in\Psi$ DO
04. $\Delta\chi\leftarrow\phi(i,j,m)\times\phi(i+m,n,\rho)$
05. $\chi[i,m,\rho]\leftarrow\chi[i,m,\rho]+\Delta\chi$
06. $\chi_{\max}\leftarrow\max\{\chi_{\max},\chi[i,m,\rho]\}$
07. IF $\boldsymbol{Q}(i+m,\rho)\neq\boldsymbol{O}$ AND $\boldsymbol{Q}(n,\rho)\neq\boldsymbol{O}$ THEN
08. $\kappa\leftarrow 0.20\times\phi(i,j,m)$
09. $\gamma[i,m,\rho]\leftarrow(1-\kappa)\times\gamma[i,m,\rho]+\kappa\times f_\sigma(i+m,n,\rho)/f_\sigma(i,j,m)$
10. ENDIF
11. ENDFOR
12. IF $\Psi\neq\emptyset$ THEN
13. $\eta\leftarrow\eta\times f_\Lambda(0.95,1.00,1.05)$
14. $\eta\leftarrow\min\{\max\{|\eta|,0.08\},32.00\}$
15. ENDIF

第 1 行求 ε_ϕ 的调节量 $\Delta\varepsilon_\phi$,在 $-1/n$ 与 $1/n$ 之间取值,当牵涉面 n 越大或 $|\Psi|$ 与 $\ln n$ 的差距越小时,调节幅度就越小,反之越大;若 $|\Psi|$ 小于 $\ln n$ 则下调阈值放宽条件,反之上调。第 2 行将 ε_ϕ 值限制在 $[0,1-1/n]$ 区间。

第 3 -11 行循环枚举 Ψ 集合,依据 (i,w,σ) 记录,调节 $\chi[i,m,\rho]$ 与 $\gamma[i,m,\rho]$ 的值。

第 4 行计算调节量 $\Delta\chi$ 值。预测算法筛选出的拟合度满足 $0<\phi(i,j,m)\leqslant 1$ 约束。当 $\boldsymbol{Q}(i,m)$ 与 $\boldsymbol{Q}(j,m)$ 的拟合越差时,对 $\chi[i,m,\rho]$ 的调节就越发谨小慎微,反之同理。当 $\phi(i+m,n,\rho)>0$ 时,上调 $\chi[i,m,\rho]$,外延值 $\boldsymbol{Q}(i+m,\rho)$ 与实测值 $\boldsymbol{Q}(n,\rho)$ 越接近,依据 $\boldsymbol{Q}(i+m,\rho)$ 预测的精准度就越高,上调幅度也就越大,反之同理下调。即使 $\chi[i,m,\rho]$ 不发生变化,也时常会因为相对退减而被逐步边缘化。当普适量增大、减小时,分别向普适度函数曲线的敏化区、拖尾区推进。

第 7 -10 行调节 $\gamma[i,m,\rho]$ 值,以便补偿外延规模与实测规模的差距。第 8 行中,比例值 κ 满足 $0<\kappa\leqslant 0.2$ 约束。按 $\gamma_{\text{new}}\leftarrow(1-\kappa)\times\gamma_{\text{old}}+\kappa\times\gamma_{\text{cpt}}$ 形式简写第 9 行。$\phi(i,j,m)$ 越大,κ 就越大,计算值 γ_{cpt} 对新值 γ_{new} 的贡献也越大,旧值 γ_{old} 中得以保留的成分就越少,反之同理。$\phi(i,j,m)$ 控制着衰减 γ_{old}、添加 γ_{cpt} 的谨慎程度,经多轮调节后,γ_{new} 将收敛于 γ_{cpt} 值。

第 13 行基于 η 的当前值,衡量下调 5%、维持现状、上调 5% 哪个最优,据此调节 η 值。第 14 行限制 η 的范围,以防过度钝化或锐化。

第 3 -11 行的外循环不超过 n 次,循环体呈 $O(m^2+\rho^2)$ 级;第 13 行呈 $O(n\times\rho^2)$ 级,故综合时间复杂度为 $O(n\times(m^2+\rho^2))$。

5 实例剖析

图 1 给出了依据历史图 $\boldsymbol{G}(0,13)$ 预测 $\hat{\boldsymbol{G}}(13,2)$ 的实例,参量组 (n,m,ρ) 设置为(13,3,2),图内上方的数字标示出所在列中的线段斜率,控制参数 $\varepsilon_\phi,\chi[i,m,\rho],\gamma[i,m,\rho],\eta$ 均取初始值。

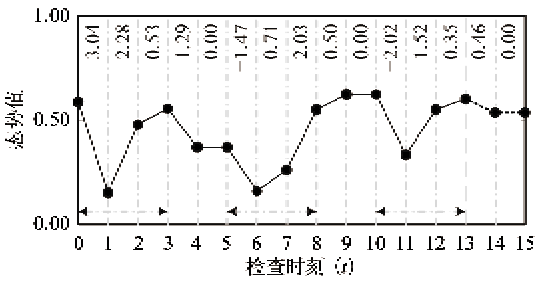

图 1 趋势预测示例

依据预测算法阐述。从第 2 -9 行 $\boldsymbol{Q}(i,3)$ 与 $\boldsymbol{Q}(10,3)$ 的逐个比对来看,当枚举至 $i=0$ 时,有 $\phi(0,10,3)=1.00$,故 $\boldsymbol{Q}(0,3)$ 入选;当 $i=5$ 时,有 $\phi(5,10,3)=0.42$,故 $\boldsymbol{Q}(5,3)$ 入选;其余的位置因 $\phi(i,10,3)\leqslant\varepsilon_\phi$ 而被排除。表 1 列出了一部分比对结果。当 $i=5$ 时,折点 $t=7$ 处的升势由缓转陡,与 $t=12$ 处的由陡转缓不同,从式(5)可知,一

阶差分值 $\nabla^1\theta_7 > 0$、$\nabla^1\theta_{12} < 0$，反映出了这种差异，相应的惩罚值计入 $\phi_\nabla(5,10,3)$ 中。经第 10 行归一化后，入选 Ψ 集合的三元组规整为(0,0.705,0.67)和(5,0.295,0.98)。从第 13 行可知，预测值 $\hat{q}_{13}$ 等于$0.705\times0.67\times(-1.29)+0.295\times0.98\times0.50$。依此类推，直至预测出 ρ 步趋势。不难看出，该预测机理能有效辨识同一序列中的多种类长程相关性。

表 1 相似度与惩罚值

i	0	2	3	5	6	8
ϕ_θ	1.00	−0.85	0.42	0.72	0.32	−0.32
$\varphi_\vee$	0.00	1.01	0.34	0.30	0.30	0.68

依据进化算法阐述。假设实测值 q_{13}、q_{14} 分别为 −0.86、0.00。从第 1 行来看，$|\Psi|=2$，是小于 ln13 的，故应下调 ε_ϕ 值，一旦 $|\Psi|>2$ 则会转入上调。从第 4、第 5 行可知，$\chi[0,3,2]$、$\chi[5,3,2]$ 的变化量分别为1.00×1.00、$0.42\times(-1.85)$，分别予以上调、下调之后，能强化 $Q(0,5)$ 的作用、弱化 $Q(5,5)$ 的干扰。从第 8、第 9 行可知，$\gamma[0,3,2]$、$\gamma[5,3,2]$ 分别被调节为$(1-0.20)\times1.00+0.20\times0.67/0.67$ 和$(1-0.08)\times1.00+0.08\times1.72/0.98$。从第 13 行可知 $(f_\kappa(0.95),f_\kappa(1.00),f_\kappa(1.05))=(0.13,0.16,0.19)$，故应上调 η 值。

表 2 给出了重复执行多轮时的效果。考虑到 $w_5=1-w_0$、$\hat{q}_{14}\equiv0.00$、$f_\chi(0,3,2)\equiv10.00$ 以及 $\gamma[0,3,2]\equiv1.00$，故这些参数未在表中列出。

表 2 预测及进化效果

轮数	1	3	5	7	9
w_0	0.71	0.95	0.98	0.99	1.00
$\hat{q}_{13}$	−0.46	−0.78	−0.83	−0.85	−0.86
ε_ϕ	0.19	0.17	0.15	0.13	0.11
$f_\chi(5,3,2)$	3.27	1.18	0.72	0.51	0.40
$\gamma[5,3,2]$	1.06	1.17	1.27	1.35	1.41
η	1.05	1.16	1.28	1.41	1.55

从表 2 可知，随着持续进化，$f_\chi(5,3,2)$ 逐渐减小，η 值逐渐增大，导致相对权重比 w_0/w_5 急剧攀升，使预测值逐轮向实测值逼近。

随着时间的推移，m 与 ρ 保持不变，n 呈线性增长，配套的辅助算法能淘汰过时数据，仅保留近期序列，同时修正拟合阈值、普适量等既支持自主进化，也支持人工修改或固化的参数。

6 实验分析

定义 12 令 E_r 表示预测图与实测图的相对误差，按式

$$E_r=\rho^{-1}\times\sum_{k=n+1}^{n+\rho}|(\hat{z}_k-z_k)/z_k| \qquad (13)$$

计算，ρ 个相对误差分量的标准差记为 E_{std}。

通过 3 种实验评价预测算法。令 ATS 指代实测态势序列，SFM 指代本文提出的算法；m 设置为 5，n 和 ρ 按需设置，除个别情况外其余参数均取初值；关闭进化算法，以未经进化的首次预测结果为准。对比算法或模型选择 ARMA、GM(1,1)与 RBFNN。ARMA 依据 Akaike 信息准则(AIC)准则选择模型，以自相关和偏自相关的计算结果作为参考；GM(1,1)带有残差修正；RBFNN 训练样本的输入、输出向量长度分别为 m、ρ，为避免其规模过于庞大或发生过拟合，限制神经元数量不超过训练样本数的一半。

6.1 实验 1

图 2 给出的历史图包含了上升、饱和、下降等趋势以及周期波动、随机扰动等特征，比较有代表性，据此预测未来态势。

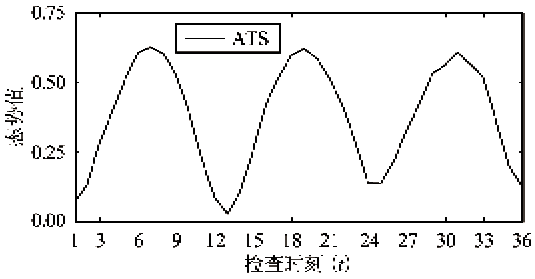

图 2 网络安全态势历史记录

从实验可知 SFM、ARMA、GM(1,1)、RBFNN 的相对误差分别为3.28%、5.89%、7.18%、16.11%。在图 3 中绘制预测结果时，将 SFM 上移 0.2 以利于分辨，辅以 ATS 的上移线以便对比，ARMA 和 RBFNN 同理。

针对 ARMA、GM(1,1)与 RBFNN，需人工识别、应对态势的周期性波动，在本实验中是以 12 为间距做差分变换，预测后再复原，否则预测效果太差，GM(1,1)、RBFNN 的相对误差分别高达59.67%、73.99%。然而，数据预处理无通用规则可循，常需一事一议，强烈依赖人工介入，很难用一套自动化程式应对多种趋势。

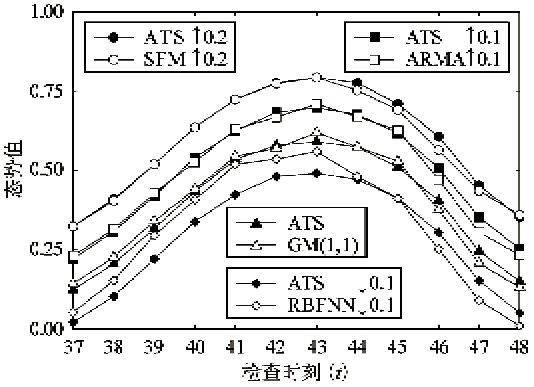

图3 网络安全态势预测结果

SFM 无需预处理,没有上述缺陷,能对复杂多变的趋势保持自适应。

6.2 实验 2

针对局域网,任选一段起伏较大且包含相似子片段的安全态势序列,做 20 组实验。得知 SFM、ARMA、GM(1,1)、RBFNN 的相对误差平均值分别为8.09%、20.89%、44.89%、34.75%;若选用几乎没有什么规律的序列做实验,则相对误差平均值依次为23.96%、21.72%、37.47%、53.54%,其中 ARMA 的稳定性略好一些。

任选一组实验,绘出图 4,以时间点 $t = 42$ 分隔历史图与预测图。从形态上看,ARMA 的预测尚可,只是精度稍差,而 GM(1,1)、RBFNN 却偏差甚远。从 GM(1,1)预测效果的反常波动及其在图 4 中输出的平坦曲线来看,它未能辨识出序列中蕴含的规律,做出针对性预测。

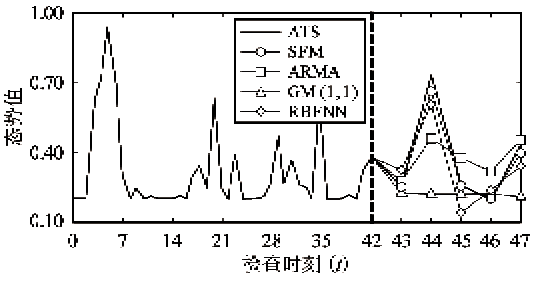

图4 网络安全态势预测结果

每组实验中,ARMA 与 GM(1,1)均需重新选择模型,如果将各组的历史图拼接成一个长序列再做上述预测,两者的性能均急剧恶化,而 SFM 不会,较长的序列更可能蕴含预测所需的相关性。

6.3 实验 3

用随机数据模拟态势序列能更全面地辨析算法差异。先从 Vista 操作系统的熵池中提取长度为 *Scale* 的随机序列;再从该序列中随机选取长度为 16 的子串,前 8 段作为事发迹象,后 8 段作为延续效应;将事发迹象拼接到随机序列的尾部构成历史图,把延续效应当作待预测图(实测图)。做 100 组实验,旨在考察各算法抵御随机干扰、辨识远距离相关性的能力,平均效果见表 3。

表3 网络安全态势预测效果

	Scale	ε_ϕ	E_r (%)	E_{sd}
SFM	1.2×10^2	0.930	2.49	0.09
SFM	1.2×10^3	0.980	1.75	0.09
SFM	1.0×10^6	0.998	1.88	0.09
ARMA	1.2×10^2	—	32.96	0.15
GM(1,1)	1.2×10^2	—	49.32	0.20
GM(1,1)	1.2×10^3	—	51.06	0.18
RBFNN	1.2×10^2	—	27.87	0.23

若使用毫无迹象或规律的真随机序列,则完全不可预测。尽管每次从熵池中取出的序列不尽相同,但当实验规模较大时,均能较好地重现表 3 给出的统计效果。因选择、训练模型较困难,ARMA 与 RBFNN 均难以应对较长的随机序列。

7 结 论

本文提出的预测算法依据事发迹象推断延续效应,无需先决条件,支持非平稳序列,能充分利用当代的计算能力,从超长序列中辨识长程相关性,得出更为精准可信的预测。进而引入了进化算法,依据预测效果调节相关参数,支持在线增量学习,能有效提升适应性。

ARMA 与 GM(1,1)每次选择的模型仅能处理较为单一的相关性,而态势序列中通常蕴含多种相关性,不是用某个公式就能描述的。随着时间的流逝,态势序列不断延长,而一旦态势序列发生了变化,ARMA、GM(1,1)与 RBFNN 必须重新选择或训练模型,针对 τ 较小的细粒度预测而言,这很难跟得上态势序列的增长速度,也就难以持续、及时地跟踪、适应复杂多变的趋势。更长的态势序列通常蕴含更加多样化的相关性,为预测所需,但 ARMA 和 RBFNN 并不适合处理长序列。RBFNN 不易消解训练样本冲突,规模可扩展性较差,预测精度欠佳。当评测输入与训练样本中的输入稍有不同时,RBFNN 输出的弥合值与期望值相差甚远。为网络安全态势预测专门定制的 SFM 算法没有上述欠缺,能以一套

自动化程式对复杂多变的趋势保持自适应。尽管如此,更深入地挖掘该领域的独特性,使之与预测算法更紧密地结合,仍是一项棘手的难题,有待继续研究。

参考文献

[1] Bass T. Intrusion detection systems and multisensor data fusion: creating cyberspace situational awareness. *Communications of the ACM*, 2000, 43(4): 99-105

[2] 陈秀真,郑庆华,管晓宏等. 层次化网络安全威胁态势量化评估方法. 软件学报,2006,17(4):885-897

[3] 韦勇,连一峰. 基于日志审计与性能修正算法的网络安全态势评估模型. 计算机学报,2009,32(4):763-772

[4] 韦勇,连一峰,冯登国. 基于信息融合的网络安全态势评估模型. 计算机研究与发展,2009,46(3):353-362

[5] Lai J B, Wang H Q, Zhu L. Study of network security situation awareness model based on simple additive weight and grey theory. In: Proceedings of the International Conference on Computational Intelligence and Security, Piscataway, USA, 2006. 1545-1548

[6] 赵国生,王慧强,王健. 基于灰色 Verhulst 的网络安全态势感知模型. 哈尔滨工业大学学报,2008,40(5):798-801

[7] 任伟,蒋兴浩,孙锬锋. 基于 RBF 神经网络的网络安全态势预测方法. 计算机工程与应用,2006,42(31):136-144

[8] Kim D H, Lee T, Jung S O D, *et al*. Cyber threat trend analysis model using HMM. In: Proceedings of the 3rd International Symposium on Information Assurance and Security, Piscataway, USA, 2007. 177-182

Network security situation prediction based on scene shift

Li Zhidong, Yang Wu, Wang Wei, Man Dapeng

(Information Security Research Center, Harbin Engineering University, Harbin 150001)

Abstract

Based on the view that the traditional methods for prediction of network security situation are unable to deal with the complex and inconstant network security situation sequence and its various evolution rules, the paper presents a special prediction algorithm. The algorithm identifies the rules in the situation sequence from the perspective of long range correlation, infers the subsequent effect according to occurred indication, and synthesizes the prediction sequence with the weighting by the indicators of similarity, universality, contrast ratio and scaling ratio. Afterwards, an evolution algorithm is introduced to adjust related parameters according to the prediction effect, strengthen the significance of universal cases and weaken the interference of special ones via online feedback learning, and improve the adaptability of the prediction algorithm. The experimenal results show that the prediction algorithm can perform excellently in identifying various long distance correlations from the super long situation sequence, keep self-adaptive towards complex and inconstant tendencies, and is more accurate.

Key words: network security, security situation, tendency prediction, scene shift, long range correlation