

基于人工蜂群的业务流异常状态检测方法

段谟意*

(南京铁道职业技术学院 软件学院, 南京 210031)

(*通信作者电子邮箱 my_duan1964@163.com)

摘要:针对日益严重的网络安全问题,基于人工蜂群与聚类方法提出一种新的状态检测算法——DASA。该算法首先根据 SKETCH 方法和 Hash 函数建立业务流异常状态模型,并且利用人工蜂群技术实现对异常状态的检测。最后,以实际数据进行仿真实验,对比分析了样本数据与 DASA 算法检测的结果,发现 DASA 具有较好的适应性,而且聚类个数、丢弃阈值和邻域半径等因素对状态检测产生较大影响。

关键词:业务流;异常状态;人工蜂群;聚类

中图分类号: TP393.08 **文献标志码:** A

Detection method of anomaly traffic state based on artificial bee colony

DUAN Moyi*

(School of Software Engineering, Nanjing Railway Vocational and Technical College, Nanjing Jiangsu 210031, China)

Abstract: In order to deal with the worsening network security problem, a new state detection algorithm, detection method of Anomaly traffic State based-Artificial bee colony (DASA), was proposed by Artificial Bee Colony (ABC) and clustering. In this algorithm, the anomaly traffic model was presented with SKETCH and Hash function at first, and the anomaly state was detected based on ABC. Then, a simulation with actual data is conducted to compare the results between Sample and DASA, which shows that DASA has better adaptability. And it has large impact on state detection with clustering number, dropping threshold and domain radius.

Key words: traffic; anomaly state; Artificial Bee Colony (ABC); clustering

0 引言

随着 Internet 规模越来越大,网络的安全问题日益突出,通过对实际业务流的状态检测来判断当前网络是否遭受了攻击或者病毒入侵^[1-3],也成为目前网络安全研究的热点和重点。为了实现网络的实时监控,业务流状态检测方法也得到了充分的研究。目前最常用的网络异常检测方法^[4]主要有时间序列模型、聚类分析、统计方法和小波变换技术等。文献[5]基于工业测控网络业务流矩阵提出了一种概率主成分分析的检测方法,用于解决业务流异常检测时存在误差率较高的问题。文献[6-7]将概率主成分分析的检测方法用于网络的异常检测,解决了实时检测拒绝服务和端口扫描等入侵。文献[8]通过计算网络业务流的特征量,并基于蝴蝶突变模型刻画业务流异常行为,从而提出了一种突变级数的异常状态检测算法。文献[9]针对单一算法并基于特征选择和支持向量机建立了一种异常检测技术。文献[10]利用大偏差理论以及决策理论来判断当前系统的行为是否存在异常,并对其进行实时检测。而小波变换技术常常和尺度特性相结合来对网络业务流异常状态进行检测^[11]。同时,其他很多方法也应用于网络异常检测中,SKETCH 方法^[12]就是最近提出的用于大规模数据处理和计算的一种数据结构,而基于 SKETCH 方法也存在大量关于网络测量变化、业务流状态估计和网络异常分布的技术手段。文献[13]基于交互式网络流模型,针对异常业务流的检测与定位问题,提出了一种不完整业务流

的有效识别方法,通过设计 SKETCH 方法的 Hash 函数来减缓业务流异常行为的扩散速度。

在上述工作基础上,本文结合人工蜂群与聚类方法^[14-15]提出了一种新的业务流状态检测方法。该方法通过 Hash 函数建立业务流异常状态模型,并且利用类间离散度来对业务流进行聚类,同时基于人工蜂群方法来对异常状态进行检测。最后以实际业务流进行仿真实验,来验证所提出的状态检测方法的有效性。

1 业务流异常状态模型

Internet 的快速发展,使得实际业务流呈现出各种复杂现象。为了深入研究业务流特征,并且实时判断是否因为病毒或者攻击等因素产生异常状态,首先建立如图 1 所示的网络结构图,其中, A_i 为内网主机 ($i = 1, 2, \dots, m$), Router1 为内网出口处的路由器, Router2 为外网入口处的路由器,这里将监控点置于 Router1 处。假设设有 m 台内网主机 A 与 n 台外网主机 B 进行通信,其状态采用矩阵 S 进行表示:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix} \quad (1)$$

其中: s_{ij} 表示内网主机与外网主机之间的通信状态矢量,这里令 $s_{ij} = [w_1, w_2, w_3, \dots, w_k]$, w_1 表示数据包数, w_2 表示数据到达速率, w_3 表示时间戳,等等 (k 为最大状态指标数)。通过

上述定义,用来描述内外主机在通信过程中状态发生的动态变化,监控内网主机的异常行为,防止外网主机对内网的入侵。

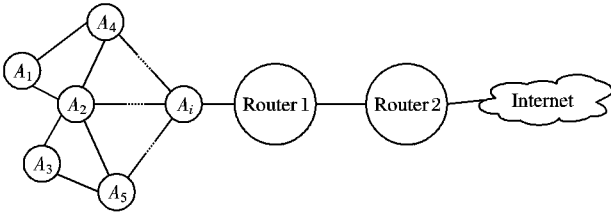


图1 网络拓扑结构示意图

本文基于 SKETCH 方法来对网络状态进行实时统计,并设计 Hash 函数来对异常主机状态进行获取和分析。结合文献[15],令 SKETCH 连接度 $C = [c_1, c_2, c_3, \dots, c_k]$,与状态矢量 s_{ij} 对应。其中, c_k 为 $a \times b$ 大小矩阵,并且每个 c_k 都关联某个 Hash 函数 $h_k: \{1, 2, \dots, a\} \rightarrow \{1, 2, \dots, b\}$ 。初始时刻,将 c_k 每一位清零,当有数据包到达时, c_k 按照式(2)对每一位的行 $\eta(w_k)$ 和列 $\xi(w_k)$ 进行更新:

$$c_k(\eta(w_{ij})\xi(w_{ij})) = 1 \quad (2)$$

这里设计如下 Hash 函数对状态进行分析:

$$h_k(z) = (\alpha z) \% k + \beta \quad (3)$$

其中 α 和 β 为正整数。通过分析 SKETCH 结构中业务流的异常位置,反向找到异常状态信息,从而确定异常源,即可转化为求解 Hash 表中 $c_1, c_2, c_3, \dots, c_k$ 对应为 $w_1, w_2, w_3, \dots, w_k$ 的业务流状态信息。这里建立如下模型:

$$\begin{cases} w_1 = (\alpha_1 z) \% k + \beta_1 \\ w_2 = (\alpha_2 z) \% k + \beta_2 \\ \vdots \\ w_n = (\alpha_n z) \% k + \beta_n \end{cases} \quad (4)$$

那么,在满足各约束条件的情况下对上述模型进行求解,通过获得 z 值来确定异常状态与异常源。但如何快速有效地获得 z 值,对于排除网络故障和异常有着非常重要作用。本文结合人工蜂群的聚类方法进行研究。

2 基于人工蜂群的异常状态检测方法

本文首先通过采用聚类方法检测业务流异常状态。在网络发生异常初期,异常状态与正常状态相比是较少的,所以聚类可以将相似度较高的正常状态聚合在一起,而将异常状态快速排查。对于状态 $s_{ij} = [w_1, w_2, w_3, \dots, w_k]$ 来说,可以找到某个划分 ε ,使得如下类间离散度 λ 总和达到最小:

$$\lambda = \sum_k \sum_{i \in k} d(w_i, \delta_k) \quad (5)$$

其中: δ_k 表示第 k 类指标的中心, d 表示当前指标 w_k 与中心 δ_k 的距离。

同时,本文基于人工蜂群算法与聚类方法来实现模型中 z 值的求解。人工蜂群算法是基于蜜蜂自组织和群体智能的计算优化算法,它将蜜蜂划分为侦察蜂、采蜜蜂和跟随蜂。其思想可以归纳为:首先采蜜蜂对邻域内蜜源进行搜索,选取花蜜较多的蜜源。然后跟随蜂根据蜜源大小以一定概率选择蜜源并在此邻域内重新进行搜索;如果跟随蜂所搜索到的蜜源优于当前采蜜蜂获得到的蜜源,则替换最优结果;否则保持不变,侦察蜂继续在此邻域内搜索,直至算法结束。以下给出具体的实现算法——DASA (Detection method of Anomaly traffic State based-Artificial bee colony):

1) 在某时刻 t ,初始化网络参数以及 Hash 函数,并确定业务流状态指标聚类个数 r 。

2) 按照式(6)获得状态指标聚类中心 O_r :

$$O_r = i + \frac{U-D}{r+1}(i * \text{rand}() - 1) \quad (6)$$

其中: U 和 D 分别为该聚类区域内的上界和下界, $\text{rand}()$ 产生 0 到 1 之间的随机数。

3) 由深度优先遍历算法对状态指标进行遍历,当所有指标被遍历时完成聚类。

4) 按照上述聚类的状态指标结果初始化蜂群邻域,将状态指标看作蜂源,并设置蜂源位置、迭代次数 N 、丢弃阈值 M 。

5) 根据式(7)~(8),计算当前区域 r 的邻域半径内蜂源的初始解 $w(0, r)$ (将此解 $w(0, r)$ 作为最优解 OPT) 和适应度 $f(0, r)$,将蜜蜂与蜂源进行对应。

$$w(k, r) = w(k, r) + 2(1 - \text{rand}())(w(k, r) - w(j, r)) \quad (7)$$

$$f(k, r) = N \log\left(\frac{1}{N} \sum_{k,j=1}^N (w(k, r) - w(j, r))\right) \quad (8)$$

其中: $0 < j < N, 0 < k < N$ 。

6) 采蜜蜂按照式(6)进行邻域搜索,获得当前解 $w(k, r)$ 以及适应度 $f(k, r)$,如果 $f(k, r) > f(0, r)$,则将最优解 OPT 替换为 $w(k, r)$, $f(0, r)$ 替换为 $f(k, r)$;否则保持不变。

7) 跟随蜂根据式(9)计算的概率 $p(i, r)$ 来选择蜂源并进行邻域搜索,计算新解 $w(i, r)$ 以及适应度 $f(i, r)$,如果 $f(i, r) > f(0, r)$,则将最优解 OPT 替换为 $w(i, r)$, $f(0, r)$ 替换为 $f(i, r)$;否则保持不变。

$$p(i, r) = \frac{1 + (f(i, r))^\theta}{r + \sum_i (f(i, r))^\theta} \quad (9)$$

其中 θ 为正常数。

8) 令 $i = i + 1$,跳转到6),重复执行,直至 $i > N$ 。

9) 令 $r = r + 1$,如果 $r \leq M$ 时,侦察蜂根据式(7)~(8)计算新解 $w(k, r)$ 与适应度 $f(k, r)$,并根据 $f(0, r)$ 决定是否进行替换,跳转到6);否则跳转到10)。

10) 输出当前最优解 OPT,即为模型所求 z 值。

11) 算法结束。

3 仿真实验

首先,本文在 NS2 中建立如图2所示的网络拓扑结构,其中内网主机数为6,与内网通信的外网主机数为8;并且对网络参数进行初始化:发送数据包大小为1024 B,节点缓存区大小为100 KB,链路带宽为10 Mb/s,延时15 ms。令业务流状态 $s_{ij} = [w_1, w_2, w_3, w_4]$, w_1 表示数据包数, w_2 表示数据到达速率, w_3 表示时间戳, w_4 表示数据长度;SKETCH 结构中 c_k 为 10×12 大小矩阵,Hash 函数中 α 和 β 的值分别取6和8。同时,令蜂群的初始邻域半径为10,迭代次数50次。这里在 Router1 处进行监控(为方便研究,仅仅打开80端口和21端口),收集5000 s的业务流数据。为了能够保证数据的平稳性,取最后1000 s数据作为样本。而前4000 s数据作为先验信息,采用本文提出的 DASA 算法进行对比分析,结果如图3所示。图3给出了数据包数和数据到达速率这两种状态的比较情况,从图中可看出,DASA 算法检测的状态信息与样本数据比较吻合。对其进行误差分析,两种状态的误差分别为8.17%和7.39%。

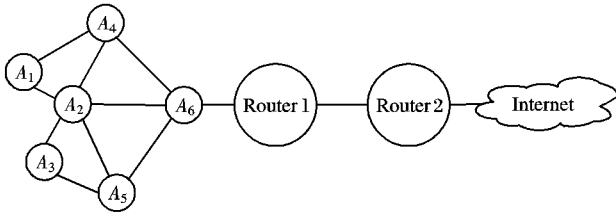
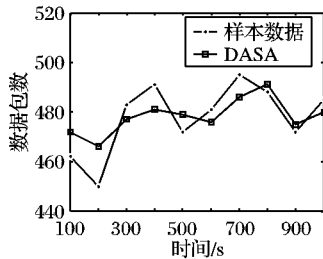
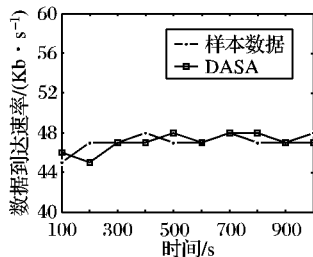


图2 网络仿真结构



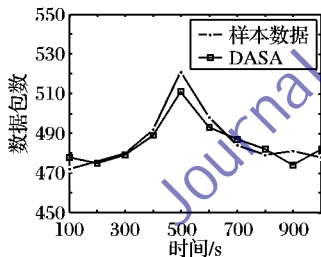
(a) 数据包数



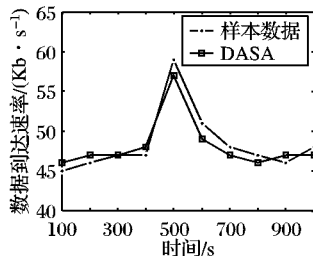
(b) 数据到达速率

图3 样本数据与DASA检测数据对比

进一步地,为了检验DASA算法的有效性,这里在主机 A_6 处采用分形ARIMA模型来产生一些突发业务流,设定相关性参数 $d = 0.45$ 。样本数据与DASA检测数据的状态对比情况如图4所示,从图中可看出,大约在500s样本业务流出现突发情况,此时DASA检测数据与样本数据相一致,有效地模拟了这个现象。



(a) 数据包数



(b) 数据到达速率

图4 突发业务流状态比较情况

其次,这里将影响检测误差的人工蜂群和聚类方法的参数进行深入讨论。图5给出了不同聚类个数 r 下,数据包数状态 w_1 的检测误差变化情况。从图5可看出,在仿真前期,聚类个数越多对应的检测误差越小,而在仿真后期情况正好相反,聚类个数越少对应的检测误差越小。由于前期聚集的业务流数量较少,划分的聚类个数越多越能清楚反映异常状态,其误

差率就越低;而到了仿真后期随着到达业务流的增多,划分的聚类个数越多可能将越多的一些正常业务流归纳为异常状态,则表现出误差率的增加。

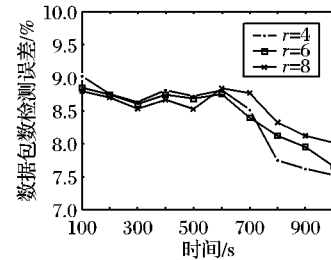


图5 不同聚类个数的数据包数状态检测误差

进一步地,图6给出了不同丢弃阈值 M 下,数据到达速率状态 w_2 的检测误差的变化情况。从图6可以看出,在仿真前期,丢弃阈值越大对应的检测误差越小,而在仿真后期,丢弃阈值越小对应的检测误差越小。造成这种状态的原因可能是由于前期业务流较少,发生异常状态的可能性偏低,丢弃阈值越大能够接受正常状态的业务流越多,其检测误差就越小;但是当业务流聚集增多达到一定程度,存在异常状态的可能性增加,丢弃阈值越大意味着接受异常状态的业务流越多,其检测误差就越大。

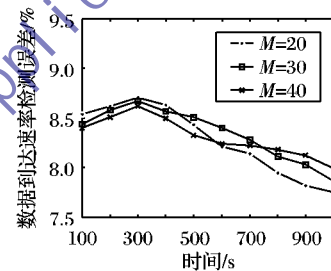


图6 不同丢弃阈值的数据到达速率状态检测误差比较

同时,图7给出了不同类间离散度 λ 下,数据长度状态 w_4 的检测误差与蜂群邻域半径之间的关系。从图7可以看出,随着蜂群邻域半径的增加,数据长度状态检测误差先呈现出下降趋势,随后呈现上升趋势。并且类间离散度不同,将导致检测误差的变化速率产生较大差异,图中 $\lambda = 20$ 的曲线较 $\lambda = 30$ 的曲线更加平滑,这说明类间离散度越小,使得数据长度状态检测误差更加稳定。

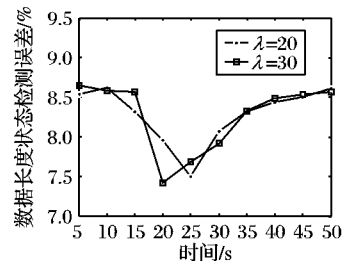


图7 数据长度状态检测误差与邻域半径之间的关系

4 结语

针对网络业务流的异常状态,本文结合人工蜂群与聚类方法提出了一种新的状态检测算法DASA。该算法首先基于Hash函数和SKETCH方法建立了业务流异常状态模型,并且采用深度优先遍历算法实现业务流聚类,同时通过人工蜂群算法完成对异常状态的检测。最后,将实际样本数据与DASA算法的检测数据进行对比仿真实验,结果发现DASA

(下转第738页)

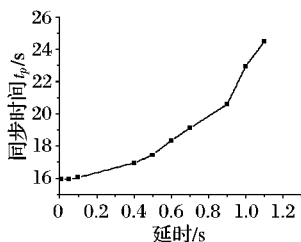


图9 延时对同步性能的影响

3 结语

实际应用中时滞是不可避免的,而这在以往的研究中常常被忽略,本文采用自适应控制方法研究了一类超混沌系统的广义函数投影滞后同步,重点分析了噪声和延时对系统同步质量的影响。研究结果还可以推广到混沌神经网络中。鉴于函数投影同步的特点,将其应用到混沌保密通信中,具有更强的保密性能。设计的同步控制器和参数更新规则中含有时间延迟参数,这是合理的,参数的大小与实际应用中驱动系统和响应系统之间的信号传输距离、信号传输的媒介等有关。另外,混沌的同步控制目前多停留在理论证明、实验仿真阶段,距离实际应用还有一段距离。

参考文献:

- [1] PECORA L M, CARROLL T L. Synchronization in chaotic system [J]. *Physical Review Letters*, 1990, 64(8): 821–824.
- [2] MAINIERI R, REHACEK J. Projective synchronization in three-dimensional chaotic systems[J]. *Physical Review Letters*, 1999, 82(15): 3042–3045.
- [3] WU X J, LU H T. Adaptive generalized function projective lag synchronization of different chaotic systems with fully uncertain parameters[J]. *Chaos Solitons Fractals*, 2011, 44(10): 802–810.
- [4] DU H Y, ZENG Q S, WANG C H, *et al.* Function projective synchronization in coupled chaotic systems[J]. *Nonlinear Analysis: Real World Applications*, 2010, 11(2): 705–712.
- [5] WU X J, LI S Z. Dynamics analysis and hybrid function projective synchronization of a new chaotic system[J]. *Nonlinear Dynamics*, 2012, 69(4): 1979–1994.
- [6] ZHANG Z Q, WANG Y X, DU Z B. Adaptive synchronization of single-degree-of-freedom oscillators with unknown parameters[J]. *Applied Mathematics and Computation*, 2012, 218(12): 6833–6840.
- [7] HU M F, YANG Y Q, XU Z Y, *et al.* Projective synchronization in drive-response dynamical networks[J]. *Physica A: Statistical Mechanics and its Applications*, 2007, 381: 457–466.
- [8] 王宇野, 许红珍. 异结构不确定混沌系统的广义投影同步[J]. *系统工程与电子技术*, 2010, 32(2): 355–358.
- [9] 王健安, 刘贺平. 不同超混沌系统的自适应修正函数投影[J]. *物理学报*, 2010, 59(4): 2265–2270.
- [10] PARK J H. Adaptive synchronization of Rossler system with uncertain parameters[J]. *Chaos Solitons Fractals*, 2005, 25(2): 333–338.
- [11] SEBASTIAN S K, SABIR M. Modified function projective synchronization of hyperchaotic systems through open-plus-closed-loop coupling[J]. *Physics Letters A*, 2010, 374(19/20): 2017–2023.
- [12] YU Y G, LI H X. Adaptive generalized function projective synchronization of uncertain chaotic systems[J]. *Nonlinear Analysis: Real World Applications*, 2010, 11(4): 2456–2464.
- [13] DU H Y, ZENG Q S, LU N. A general method for modified function projective lag synchronization in chaotic systems[J]. *Physics Letters A*, 2010, 374(13/14): 1493–1496.
- [14] 刘豹, 唐万生. 现代控制理论[M]. 北京: 机械工业出版社, 2006.
- [15] STENFLO L. Generalized Lorenz equations for acoustic gravity waves in the atmosphere[J]. *Physica Scripta*, 1996, 53(1): 83–84.
- [16] CHEN A, LU J, LU J L, *et al.* Generating hyperchaotic Lü attractor via state feedback control[J]. *Physica A: Statistical Mechanics and its Applications*, 2006, 364: 103–110.

(上接第729页)

检测数据与样本数据比较接近,而且聚类个数、丢弃阈值和邻域半径等因素对检测结果产生较大影响。在以后的研究中,可以考虑结合实际业务流的分形特性来建立一套完整的状态检测模型。

参考文献:

- [1] ASHFAQ A. A comparative evaluation of anomaly detectors under portscan attacks [C]// *Proceedings of the 11th International Symposium on Recent Advances in Intrusion Detection*. Berlin: Springer-Verlag, 2008: 351–371.
- [2] FEI R. ADIC: an anomaly detection algorithm using incremental clustering[J]. *Journal of Information and Computational Science*, 2009, 6(2): 1051–1057.
- [3] PASCHALIDIS L, CHEN Y. Anomaly detection in sensor networks based on large deviations of Markov chain model[C]// *Proceedings of the 47th IEEE Conference on Decision and Control*. Piscataway, NJ: IEEE Press, 2008: 2338–2343.
- [4] PATCHA A, PARK J. An overview of anomaly detection techniques: existing solutions and latest technological trends[J]. *Computer Networks*, 2007, 51(12): 3448–3470.
- [5] 侯重远, 江汉红, 芮万智, 等. 工业网络流量异常检测的概率主成分分析法[J]. *西安交通大学学报*, 2012, 46(2): 70–75.
- [6] ZAI Z, HAKAMI S, MOORS T, *et al.* Detection and identification of anomalies in wireless mesh networks using principal component analysis[J]. *Journal of Interconnection Networks*, 2009, 10(4): 517–534.
- [7] 张文铸, 刘佳, 袁坚, 等. 基于PCA的对等网络流量时空特性监测[J]. *清华大学学报: 自然科学版*, 2010, 50(4): 561–564.
- [8] 熊伟, 胡汉平, 王祖喜, 等. 基于突变级数的网络流量异常检测[J]. *华中科技大学学报: 自然科学版*, 2011, 39(1): 28–31.
- [9] 肖海军, 王小非, 洪帆, 等. 基于特征选择和支持向量机的异常检测[J]. *华中科技大学学报: 自然科学版*, 2008, 36(4): 99–102.
- [10] PASCHALIDIS I, SMARAGDAKIS G. Spatio-temporal network anomaly detection by assessing deviations of empirical measures[J]. *IEEE/ACM Transactions on Networking*, 2009, 17(3): 685–697.
- [11] SEONG S, REDDY A. Statistical techniques for detecting traffic anomalies through packet header data[J]. *IEEE Transactions on Networking*, 2008, 16(3): 562–575.
- [12] SCHWELLER R, LI Z C, CHEN Y, *et al.* Reversible sketches: enabling monitoring and analysis over high-speed data streams[J]. *Transactions on Networking*, 2007, 15(5): 1059–1072.
- [13] 杨柳静, 秦涛, 王晨旭. 应用交互式网络流模型的高速网络异常行为检测与控制方法[J]. *西安交通大学学报*, 2012, 46(6): 1–7.
- [14] 雷秀娟, 黄旭, 吴爽, 等. 基于连接强度的PPI网络蚁群优化聚类算法[J]. *电子学报*, 2012, 40(4): 695–702.
- [15] 田建芳, 雷秀娟. 基于蜂群和广度优先遍历的PPI网络聚类[J]. *模式识别与人工智能*, 2012, 25(3): 481–490.