

文章编号:1001-9081(2010)06-1480-03

网络安全态势预测及其在智能防护中的应用

王晋东, 沈柳青, 王 坤, 王 娜

(信息工程大学 电子技术学院, 郑州 450004)

(muyue1983@163.com)

摘 要:针对智能安全防护系统这一特殊应用,提出了一种以灰色 GM(1,1) 为原型,以 Markov 链为误差校正的新型网络安全态势预测算法。首先,对网络安全态势进行量化建模,通过 GM(1,1) 预测模型拟合网络安全态势预测曲线,然后通过真实值曲线与预测曲线之间对比建立 Markov 偏移概率矩阵,得到平均偏移百分比,从而对原预测值进行误差校正。实验证明:该方法预测结果与真实值较为接近,能较好地体现态势值的趋势性和波动性,且计算量小,适合智能防护中的应用。

关键词:网络安全态势预测;智能安全防护软件;加权平均;预测精度

中图分类号: TP393.08 **文献标志码:** A

Network security status forecasting and its application in intelligent defense

WANG Jin-dong, SHEN Liu-qing, WANG Kun, WANG Na

(Institute of Electronic Technology, Information Engineering University, Zhengzhou Henan 450004, China)

Abstract: Concerning the application of network security status prediction to Intelligent Security Defense Software (ISDS), this paper introduced a new prediction algorithm based on GM(1, 1) and Markov models. Firstly, the GM(1, 1) model was used to predict the original status value. Then, the Markov model was used for error compensation. And the simulation experiment proves that this prediction algorithm is good at status prediction and suitable for intelligent defense application.

Key words: network security status forecasting; Intelligent Security Defense Software (ISDS); weighted average; prediction precision

0 引言

随着 Internet 的发展,网络攻击的日益频繁,网络安全问题得到了广泛的重视,针对网络安全的防护措施不断得到提升。当前,应对网络攻击的对策主要是事前预防、事发反应和事后补救“三位一体”的防护方法。网络安全态势预测作为事前预防中不可缺少的一部分,有重要的研究意义。一方面,态势预测可以使管理员更好地了解网络安全状态,能在网络遭受攻击和损失之前,及时采取防御措施,加强网络安全设备的安全策略,真正达到网络安全主动防御;另一方面,依据态势预测建立网络安全预警机制和自主决策机制,根据不同的安全趋势或阈值做出不同的反应,也是实现网络安全防护自动化、智能化的重要基础。

目前网络安全态势预测,没有形成标准的方法体系,大部分网络安全态势预测方法都是基于时间序列分析领域的研究成果进行设计的,种类繁多且各有优缺点,主要有回归分析方法^[1]、灰色预测方法^[1-2]、人工智能方法^[3-5],以及综合预测方法^[3-4]等几类。本文分别对各类预测方法以及其典型模型进行分析,在此基础上结合智能安全防护软件这一特殊应用,设计了一个有效的网络安全态势预测方法。

1 各类预测方法及其典型模型分析

1.1 回归分析方法

实际问题中,通常几个变量之间存在相关关系但难以确

定。通过对历史数据的统计分析得到它们之间关系的统计规律称为回归关系。有关回归关系的理论、计算和分析称为回归分析。回归分析又可分为一元线性回归分析、多元线性回归分析,以及非线性回归分析。目前用于态势预测的典型回归分析方法是自回归移动平均模型 ARMA^[1]和自回归单整移动平均模型 ARIMA^[3]。ARIMA(p, d, q)模型是 ARMA(p, q)模型的一种扩展形式,它可以将一个非平稳时间序列转化为平稳时间序列。两种模型的步骤相同,差异就是 ARIMA 在估计之间要确定原时间序列的单整(差分)阶数 d 。

ARIMA 和 ARMA 模型预测方法的优点是:反映了时间序列的自相关性,预测结果能够体现时间序列的随机性和周期性等因素,对短期预测效果较好。缺点是:步骤较为复杂,需要较多人工操作,对中长期预测误差较大。

1.2 灰色预测方法

灰色系统理论是 20 世纪 80 年代,由邓聚龙教授提出的,并已得到广泛地应用。灰色系统即含有部分确定信息和部分不确定性信息的系统。灰色系统理论主要研究“小样本”、“贫信息”的不确定性系统的建模、预测、决策,以及控制等问题,通过对少量已知信息的分析,进行系统建模从而找出其规律。目前应用于态势预测的典型灰色模型为 GM(1,1)模型^[1-2]。

该预测方法的优点是:模型算法简单,易于实现,且消耗低,能够体现网络安全趋势,适用于小样本预测,此外不需要

收稿日期:2010-01-06;修回日期:2010-02-10。 基金项目:河南省科技厅项目(082102320010)。

作者简介:王晋东(1965-),男,山西洪桐人,副教授,主要研究方向:信息安全、软件工程; 沈柳青(1983-),男,浙江武义人,硕士研究生,主要研究方向:信息安全; 王坤(1976-),男,河南郑州人,副教授,博士,主要研究方向:信息安全、软件自适应; 王娜(1970-),女,河南郑州人,副教授,主要研究方向:软件工程、信息安全。

进行参数设定或其他人工干预,适用性强。缺点是:能够体现网络安全走势但预测误差较大,无法体现随机性、周期性等因素,对中长期预测效果不佳。

1.3 人工智能预测方法

人工智能的基本思想是通过建立机器的自动感知和自学习机制,使其具有思维能力和行为能力。人工智能应用于网络安全态势预测的主要有支持向量机^[3-4]和神经网络^[5]两类方法。支持向量机与神经网络虽然有所区别,但是其基本思路和步骤是相同的。

此类方法的优点是:具有自学习能力,中短期预测精度较高,需要较少的人为参与。缺点是:训练时间较长,消耗较大,长期预测效果不佳,需要及时增量训练,特别是神经网络还存在泛化能力弱,易陷入局部极小值等问题。

1.4 综合预测方法

综合预测方法是指集成两个或两个以上的预测方法来进行时间序列预测。大部分研究表明综合预测方法比单独的线性和非线性模型有更好的预测效果^[3]。此类方法的优点是:预测精度较高。缺点是:算法复杂度高,消耗较大,对于长期预测效果不佳。

此外,还有贝叶斯预测算法、Markov 预测算法,等等。可应用于态势预测的方法较多,但各种方法都有其优缺点,没有一种方法是普遍适用的。对于一定的应用,应当分析其特征和需求,来选择或者集成最为合适的预测方法。

2 应用及设计

本文将态势预测应用于智能安全防护软件(Intelligent Security Defense Software, ISDS)^[8]中,作为其高层自主决策层中的一部分,负责对整个局域网安全态势的动态预测。这样的设计使智能安全防护软件不仅在各防护节点上有自主感知、决策、动态重构等智能特性,而且在指控中心能对整个局域网的安全状态进行感知和预测,进而辅助管理者(或自主进行)更高层次的决策,从整体上达到智能化。

2.1 应用环境分析

设计智能安全防护软件的网络安全态势预测方法,应当先分析其特征和需求,从而得出相应预测算法应当具有的特点。具体如下:

1)较好地体现网络安全趋势。ISDS 的态势预测主要针对大型的局域网,这种网络一般情况下态势较为平滑,对预测算法的要求是能够根据历史数据较好地预测出未来一段网络安全的走势,为管理者提供参考或据此采取对应措施。

2)短期预测效果好。大部分预测算法对中长期预测效果都不是很好,且鉴于网络态势变化的随机性成分较大,应当根据最新的数据预测下一步的趋势,所以要求以滑动时间窗来选择输入样本,以获取较好的短期预测效果。

3)简单性。作为应用型软件中的一部分,不能设计过于复杂的算法,应当以简便为宜,消耗应当较小。

4)适用性。ISDS 作为一套软件系统,应当适应各种网络应用环境,因此需要预测方法具有通用性,即预测过程应当尽量避免人工参与并且对样本数据量要求不高。

2.2 预测方法设计

由上分析,最适合本软件的应当是 GM(1,1) 预测模型,它能够较好地体现网络安全态势走势,短期预测效果较好,且具有简单性和适用性。文献[2]用 GM(1,1) 对网络态势进行了预测,并证明预测精度不低于 96%。但是,GM(1,1) 预测

模型也存在较大的缺陷。文献[1]也用 GM(1,1) 进行了网络态势预测,提出其不足之处是预测精度不高,不能体现网络安全态势中的随机性和波动性。而 Markov 链模型适合于预测平稳过程的随机波动,恰好能够弥补 GM(1,1) 的不足。将 GM(1,1) 与 Markov 结合来进行预测的方法已在经济预测^[7]、土地预测^[8]中使用,但其结合方法各有不同。本文针对网络态势预测,以 GM(1,1) 为预测原型,以 Markov 链进行误差校正,设计了预测算法 GMM,其基本步骤如下。

第 1 步 根据历史网络态势值,通过 GM(1,1) 建立预测模型为 $\hat{x}^{(1)}(t+1) = [x^{(0)}(1) - b/a]e^{-at} + b/a$,还原得到预测值 $\hat{x}^{(0)}(t+1)$ 。

第 2 步 将真实态势值(用 $x^{(0)}(t+1)$ 表示)与第 1 步中得到的预测值(用 $\hat{x}^{(0)}(t+1)$ 表示)进行对比,由偏离量来划分 n 个状态,并将状态划分与态势序列相对应。

任一状态 $E_i = [\otimes_{li}, \otimes_{2i}] (i = 1, 2, 3, \dots, n)$ 。其中: \otimes_{li} 和 \otimes_{2i} 为上下状态边界, $\otimes_{li} = x^{(0)}(t+1)(1 + a_i\%)$, $\otimes_{2i} = x^{(0)}(t+1)(1 + b_i\%)$ 。

相当于以真实态势曲线为中心,划分 n 个区域。可根据历史数据的多少和精度要求确定状态个数,状态越多预测精度越高(可根据具体需要预先设定状态个数)。状态 E_i 的意思为预测值偏离实际值的范围为 $[a_i\%, b_i\%]$ 。

第 3 步 由状态划分计算状态转移概率 $p_{ij}(k) = \frac{M_{ij}}{M_i}$,得到 m 步的概率转移矩阵 $P^{(k)}$ 。

其中: $p_{ij}(k)$ 表示从状态 E_i 经过 k 步到达状态 E_j 的概率, M_{ij} 表示样本中状态 E_i 经过 k 步到达状态 E_j 的次数, M_i 表示样本中状态 E_i 出现的次数, m 根据需要预先设定。因此,概率转移矩阵 $P^{(k)}$ 中的第 i 行数据形成要经历以下步骤:首先从状态划分表中搜索状态 E_i 的个数,然后从 E_i 状态出发搜索 k 步转移到达的状态,最后统计,计算各转移概率。

第 4 步 选定最近的 c 个时刻,将其转移步数分别定义为 $1, 2, \dots, c$,编制预测表,得到总概率矩阵 P_c ,将其列向量求和,取最大和所在列状态为 GM(1,1) 预测偏移状态,计算该状态的平均偏移百分比 $\delta\% = \frac{(a_i + b_i)}{2}\%$ 。

第 5 步 通过 Markov 状态预测校正 GM(1,1) 确定预测值,最终预测值为 $\hat{x}(t+1) = \hat{x}^{(0)}(t+1) \cdot (1 - \delta\%)$ 。

2.3 GMM 预测算法改进

为提高 GMM 的预测精度和降低其复杂度,对 2.2 节中的步骤做如下改进。

1) 预测精度改进。

为了提高预测精度主要做两点改变:

一是,设定时间窗口,只用最近一段时间的态势值来预测下一段时间的态势值,适时更新状态划分表和概率转移矩阵。因为,网络态势值有其规律性,但其随机性往往也较大,下一时刻的态势值与最近的态势关系较大。这一点上该方法比神经网络预测更有优势,神经网络定时以新样本进行训练消耗太大。

二是,第 4 步中 δ 的计算方法不以列向量最大和为准,而是采用加权平均值来确定 δ 。因为,当两个状态的概率很相近时,直接以稍大者来判断其状态从而得到其偏移率误差较大,而应当以概率为权进行加权平均以确定该预测点所处的位置,而且离预测时刻最近的态势值对态势趋势影响应当较大,相应的转移概率赋较大的权值。

$$\delta\% = A \times P_{en} \times B$$

其中权重向量 A 为:

$$A = \left(\frac{c}{1+2+\dots+c}, \frac{c-1}{1+2+\dots+c}, \dots, \frac{1}{1+2+\dots+c} \right)$$

P_{en} 为总概率转移矩阵。

权重向量 B 由各状态平均偏移率组成:

$$B = \left(\frac{a_1 + b_1}{2}\%, \frac{a_2 + b_2}{2}\%, \dots, \frac{a_n + b_n}{2}\% \right)$$

2) 算法复杂度改进。

第3步中的概率转移矩阵,要经过多项繁杂的操作才能形成,而一次预测过程中,第4步编制预测表只用到了最近 c 个时刻所处状态出发的 $1 \sim c$ 步转移概率,即每个转移矩阵 $P^{(k)}$ 只用到了其中一行。而且,设定时间窗使各步概率转移矩阵只被使用一次预测过程,需要实时更新。因此,应当除去不必要的计算,生成概率转移矩阵时,先定位最近 c 个时刻的状态,只计算特定的概率转移行来组成总概率矩阵。假设有 n 个状态,则改进后第3步的计算量是原来的 $1/n$,大大降低了算法的复杂度。

3 仿真实验

3.1 仿真环境与数据采集

为了力求实验数据的真实性和可信性,本实验采用 Honey Net 上公布的网络攻击数据,以 Matlab 为工具进行预测及分析。Honey Net 是为了研究网络攻击而建立的网站,其公布的网络攻击数据被很多研究者作为参考。本实验具体数据取自 Honey Project^[9] 获取的 2000 年 4 月到 2001 年 2 月期间攻击信息。Honey Project 建立了一个含 8 个 IP 地址高度控制和完全监视的网络,采用本地 ISP 提供的单线 ADSN 连接,包含了 Solaris Sparc、Windows NT、Windows 98 和 Linux Red Hat 等 3 个操作系统。这种实验环境与 IDS 的应用环境较为相似,因此其数据也比较适合用来进行仿真实验。

3.2 网络安全态势值生成

网络安全态势量化方法较多,大体思想都是以攻击或报警统计数据为基础,经过权值定义与加权平均计算得到风险值,以风险值来表示网络安全态势。因此,不同的指标选择和权值定义得到不同的网络态势值。为求精简,本实验仅取时间、攻击威胁度、攻击发生频率这 3 个主要指标来进行定义,得到网络安全态势值公式为:

$$SA(t) = \sum_{i=1}^m th_i f_{it}$$

其中: m 代表攻击威胁等级数, th_i 代表相应的威胁等级值, f_{it} 表示 t 时段相应威胁等级的攻击次数。攻击威胁等级划分参照 Sonrt 手册^[10] 分为 3 个等级,权值分别赋值 1、2、3。时间以天为单位。

由于 2000 年 9 月份攻击数据较完整,取该月 30 天的数据进行实验。9 月 1 日该网络发生的攻击总次数为 4 次,其中有 1 次为中危攻击,3 次低危攻击,从该日起算,因此 $SA(1) = 5$ 。同理可得该月的网络安全态势如图 1 所示。

3.3 态势预测及分析

用 Matlab 编写改进的 GMM 算法,以前 10 个态势值作为初始输入样本来预测后 3 个态势值,而后,时间窗口向前滑动来选择样本进行预测。

算法参数设置如下:时间窗口定为 10,即以离预测点最近 10 个时段数据形成概率转移矩阵;时间窗口滑动步长为 3;转移步长 c 设为 5,即取与预测点最近的 5 个时段态势值来进行预测转移状态;转移状态数定为 8,分别是

$$E_1: [x^{(0)}(t+1)(1+15\%), x^{(0)}(t+1)(1+30\%)]$$

$$E_2: [x^{(0)}(t+1)(1+5\%), x^{(0)}(t+1)(1+15\%)]$$

$$E_3: [x^{(0)}(t+1), x^{(0)}(t+1)(1+5\%)]$$

$$E_4: [x^{(0)}(t+1), x^{(0)}(t+1)(1-5\%)]$$

$$E_5: [x^{(0)}(t+1)(1-5\%), x^{(0)}(t+1)(1-15\%)]$$

$$E_6: [x^{(0)}(t+1)(1-15\%), x^{(0)}(t+1)(1-30\%)]$$

$$E_7: \leq x^{(0)}(t+1)(1-30\%)$$

$$E_8: \geq x^{(0)}(t+1)(1+30\%)$$

预测结果如图 2 所示,由此可以得出以下结论:

1) 采用了滑动时间窗口,来提高预测精度, GMM 和 GM(1,1) 预测曲线趋势与实际曲线基本上吻合,证明了两种预测方法对于短期预测效果都不错,能把握网络态势的趋势性。

2) 改进的 GMM 预测精度高于 GM(1,1) 预测精度。图中大部分 GMM 预测曲线均比 GM(1,1) 预测曲线更接近真实值曲线,对波动性的把握较好。

3) 对态势值有较大突变的情况,如在时间段 10~15,两种预测方法误差都较大。而产生突变主要有两种情况,一种是攻击次数突增,另一种是攻击次数未增多,而高危攻击所占比率增大。

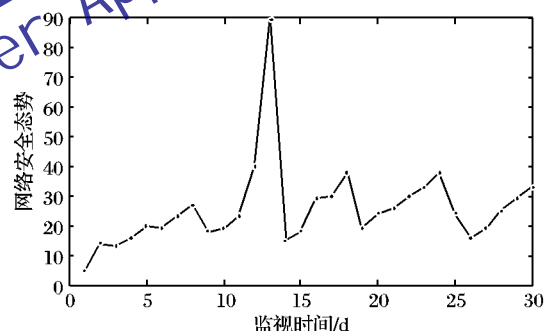


图1 9月份网络安全态势值曲线

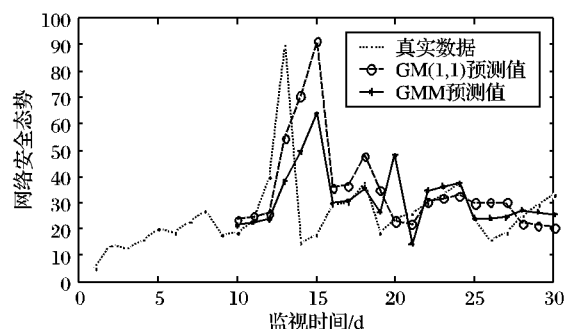


图2 网络安全态势预测曲线

4 结语

本文分析了各类网络安全态势预测方法及其典型模型,提出了各自的特点,为更好地利用和设计预测算法提供了参考。此外,与智能防护这一主题相结合,设计了符合智能安全防护软件的预测算法 GMM,并从算法复杂度和预测精度两个方面进行了改进。仿真实验表明该方法切实有效,是将网络安全态势预测应用于智能防护的一次有益探索。

(下转第 1488 页)

直接的关系,迭代次数越多,运行时间越长,遗传算法与粒子群算法的运行时间基本持平;而在适应值曲线的表现上,粒子群算法的曲线较陡峭,遗传算法的曲线较平缓,正确地反应出了遗传算法和粒子群算法的特点。粒子群算法由于粒子运动在大体上是向着粒子本身的历史最优位置和群体最优位置前进,但小范围内粒子的运动是离散随机的,某个粒子偶然找到一个更好的位置,其他粒子便会向着这个位置靠拢。在适应值曲线上表现出来就显得比较陡峭,突变点较多。而遗传算法每一代都选取了适应值较好的父代进行遗传,逐步逼近最优适应值,由于每次交叉和变异对种群的改变不是很大,因此在适应值曲线上表现出来就显得平缓。

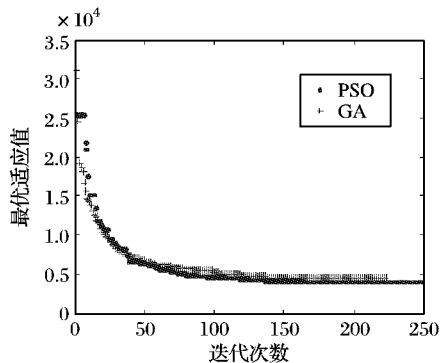


图1 种群规模 $popsiz = 20$ 时算法的收敛过程

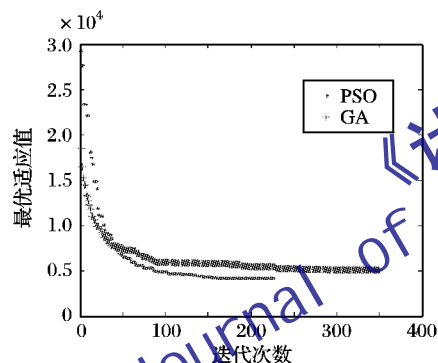


图2 种群规模 $popsiz = 10$ 时算法的收敛过程

4 结语

文中对遗传和粒子群算法在入侵检测中的优化特性进行对比分析。在遗传算法方面,引入了基因的概念,将染色体分成一个一个的基因进行交叉和变异运算,简化了算法的复杂度,且提高了交叉的范围。在粒子群优化算法方面采用了二

进制粒子群优化算法,这对于组合优化问题具有更好的适用性。在两种算法的终止条件判定方面,既使用了最大迭代次数作为终止条件,同时又使用了收敛度判断。实验结果证明:两种进化优化算法在训练集上聚出的类与原始类相差很小,进化优化算法的高效性和优异性得到了充分的体现,但是粒子群算法在适应度的收敛值和收敛速度上均优于遗传算法。

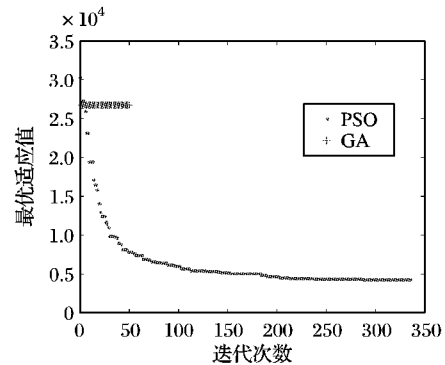


图3 种群规模 $popsiz = 5$ 时算法的收敛过程

参考文献:

- [1] 唐少先, 蔡文君. 基于无监督聚类混合遗传算法的入侵检测方法[J]. 计算机应用, 2008, 28(2): 400-411.
- [2] 徐东升, 艾晓燕, 阎世梁. 基于遗传优化与模糊规则挖掘的异常入侵检测[J]. 计算机应用, 2009, 29(8): 2227-2229.
- [3] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory [C]// Proceedings of the Sixth International Symposium on Micromachine and Human Science. Washington, DC: IEEE, 1995: 39-43.
- [4] KENNEDY J, EBERHART R. Particle swarm optimization [C]// Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Service Center, 1995: 1942-1948.
- [5] 谷保平, 许孝元, 郭红艳. 基于粒子群优化的k均值算法在网络入侵检测中的应用[J]. 计算机应用, 2007, 27(6): 1368-1370.
- [6] 郭惠玲, 唐勇, 张冬丽. 遗传算法在入侵检测规则提取中的应用[J]. 哈尔滨工业大学学报, 2009, 41(1): 348-350.
- [7] ABADEH M S, HABIBI J, BARZEGAR Z, et al. A parallel genetic local search algorithm for intrusion detection in computer networks [J]. Engineering Applications of Artificial Intelligence, 2007, 20(8): 1058-1069.
- [8] 姜永森, 王军霞, 杨慧中. 基于二进制粒子群优化的决策系统属性离散化[J]. 控制工程, 2008, 15(4): 360-363.
- [9] KDD Cup 1999 data [DB/OL]. [2008-10-10]. <http://kdd.ics.uci.edu/data-bases/kddcup.html>.

(上接第1482页)

参考文献:

- [1] 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[J]. 计算机学报, 2009, 32(4): 763-772.
- [2] 赖积保, 王慧强, 朱亮. 网络安全态势感知模型研究[J]. 计算机研究与发展, 2006, 43(增刊): 456-460.
- [3] 朱帮助, 林健. 基于ARIMA和LSSVM的非线性集成预测模型[J]. 数学的实践与认识, 2009, 39(12): 34-40.
- [4] 张翔, 胡昌振, 刘胜航, 等. 基于支持向量机的网络攻击态势预测技术研究[J]. 计算机工程, 2007, 33(11): 10-12.
- [5] 任伟, 蒋兴浩, 孙敏峰. 基于RBF神经网络的网络安全态势预测方法[J]. 计算机工程与应用, 2006, 42(3): 136-144.
- [6] SHEN LIU-QING, WANG JIN-DONG, WANG KUN, et al. The

design of intelligent security defensive software based on autonomic computing [C]// The Second International Conference on Intelligent Computation Technology and Automation. Washington, DC: IEEE Computer Society, 2009: 489-491.

- [7] 储小俊, 刘思峰. 基于新陈代谢GM-Markov模型的股价预测[J]. 山东财政学院学报, 2007, 3(1): 43-45.
- [8] 居玲华, 石培基. 基于Markov和GM(1,1)模型的土地利用结构预测[J]. 农业系统科学与综合研究, 2009, 25(2): 138-146.
- [9] PROJECT H. Know your enemy: statistics [EB/OL]. [2009-10-20]. <http://old.honeynet.org/papers/stats>.
- [10] ROESCH M, GREEN C. SNORT users manual [EB/OL]. [2009-11-02]. http://www.snort.org/assets/82/snort_manual-2_8_5_1.pdf.