基于入侵事件预测的网络安全预警方法*)

张 峰 秦志光 刘锦德
(电子科技大学计算机学院 成都610054)

摘 要 提出了一种基于入侵事件统计规律的安全预警方法,包括聚类分析、周期分析、趋势预测。依据某一攻击发生的历史分布特点,通过聚类分析,取得入侵频数序列;周期分析确定入侵事件发生的周期性;预测未来时间入侵发生趋势。讨论了时间粒度对预测效果的影响,以及算法对周期性攻击预测的适应性。实验结果表明:该方法对周期性攻击的预警误报率为19%和漏报率为27%。

关键词 入侵事件,预警,预测,网络安全

Intrusion Event Based Early Warning Method for Network Security

ZHANG Feng QIN Zhi-Guang LIU Jing-De
(College of Computer and Engineering, University of Electronic Science and Technology, Chengdu 610054)

Abstract Statistics based early warning method is proposed. It covers clustering, cycle analyzis and prediction. Clustering results in intrusion frequency according historical intrusion events. Cycle analysis testifies whether there is a cycle. Prediction gives future frequency of attacks. Relationship between clustering time and false positive rate and false negative rate is discussed and experimented. It shows periodical intrusion events gains better result than the non-periodical.

Keywords Intrusion event, Early warning, Prediction, Network security

1 引言

在网络安全领域,预防、检测和响应入侵已经得到人们的广泛关注,防火墙、入侵检测系统、蜜罐[1]、源回溯技术[2]被部署到重要的应用系统中。但是,这些技术和系统都是要等待攻击发生了,才能有所反应。安全预警试图在攻击发生之前,对其攻击发生的数量及时空特性进行预测。战略预警和监管技术是为防御突然的网络攻击,运用预警技术监视、识别大规模网络上入侵行为的综合性警戒手段,是国家战略防御系统的重要组成部分。战略预警的基本任务是从大规模网络中采集和分析网络信息,从中提取、过滤重要的安全信息,及时准确探测、识别网络中的入侵行为,对来袭目标、攻击类型等进行分析和判断,并将分析结果传送给监管系统,以便后者及早采取相应防御措施。

1999 年, IAAC (Information Assurance Advisory Council)启动了"信息安全保障的威胁评估与预警方法"项目,研究网络攻击的威胁评估的量化方法和预警方法[3]。他们分析了一个用于安全子域(sub-state)内的预警方法,但不能对未来的攻击进行预测。Jim Y.[4]描述了一种通过构造 attack profile(包括攻击历史活动、攻击工具、操作步骤、动机、目标、审记标识等内容)的方法来预测攻击。但是构造 attack profile 本身就是很难或是开销巨大的。Ming-Yuh H.[5]提出了一种概念模型,利用攻击树对攻击意图建模,预测攻击者可能的后续攻击。

基于入侵事件预测技术,我们提出了一种基于入侵事件预测的网络安全预警方法:对入侵事件进行聚类分析,根据特定攻击发生的历史规律,预测大规模网络的安全趋势,为监管系统提供防御参考。

2 基于入侵事件的预警方法

基于安全事件安全趋势预测方法,根据入侵事件发生的历史规律性预测将来一段时间的安全趋势,进行中短期预警和长期的安全形势预测。首先对获取的入侵事件进行聚类分析,得到特定攻击类型发生的历史数据。再按其历史规律的周期性和非周期性分别预测未来的发生趋势。

2.1 基本定义

入侵事件,由以下五元组表示:$E=\{D,S,R,C,T\}$。每个入侵事件具有五个属性。其中,D 为目标地址集合,S 为源地址集合,R 为请求服务类型集合,C 为攻击类型集合(文中以 Snort 定义的攻击类型为例),T 为时间标记集合。

视图 V 是一组聚类条件,$V=\text{or}Expr \mid \text{and}Expr \mid group\text{-}Expr$, 其中,$\text{or}Expr=\text{OR}((D|S|R|C|T)=val)$,$\text{and}Expr=\text{AND}((D|S|R|C|T)=val)$。条件表达式由项属性、属性值表达式和关系符 OR、AND 的组合构成。视图可以产生具有相同属性值的一组入侵事件集。

预测某种类型的入侵事件的未来趋势,首先要取得该种事件的历史规律。为了计算特定视图下的入侵事件的发生频次,采用了基于统计信息网格(STING)的多分辨率聚类方法。聚类结果是将相似的记录分成若干组,得到相关目标聚类的入侵事件频次集。入侵事件的属性(目标地址,源地址,请求服务类型,攻击类型,时间)看作 n 维空间 S 的维,分别有一个有界定义域。输入的入侵事件为 n 维空间中的点集。聚类方法如下:

确定包含聚类的子空间

利用单调性引理(基于关联规则挖掘的先验性质 apriory property):频繁项集的所有非空子集也是频繁的。设 $k=1$,遍

*)基金项目:863资助项目"战略预警与监管体系结构研究"(2002AA142040)。张 峰 博士生,主研方向:网络安全主动防御技术。秦志光 教授,博导,主研方向:网络安全、办公自动化。刘锦德 教授,博导,主研方向:开放系统与其安全性技术、中间件技术。

历报警数据库，找出所有的一维密集单元格(攻击)。

a 频次大于 min$f$(事件发生的最小频次)，其组成的集合记为 E_1；

b 若 $k<n$ 则由 k 维的密集单元格集合 E_k 生成 $k+1$维的候选密集单元格，否则转 d；

c 若 E_{k+1}不为空则，过滤掉非密集的单元格，$k=k+1$，转 b；

d 得到最高维的密集单元格构成的子空间。

回答查询的方式

a 确定与查询相关的聚类子空间的维数 k；

b 从 k 维聚类子空间集合中选择与查询最相关的聚类子空间；

c 只考虑第 k 层中满足查询条件的单元，$k+1$层的处理仅对这些单元进行；

d 重复 c 直到满足查询要求；

e 对最终结果的处理：过滤掉非密集单元格。

2.2 趋势预测算法

趋势预测主要包括：选取历史数据、选择拟合曲线、建立回归方程、计算趋势信息。

2.2.1 *算法步骤* 预测算法的处理流程如图1。

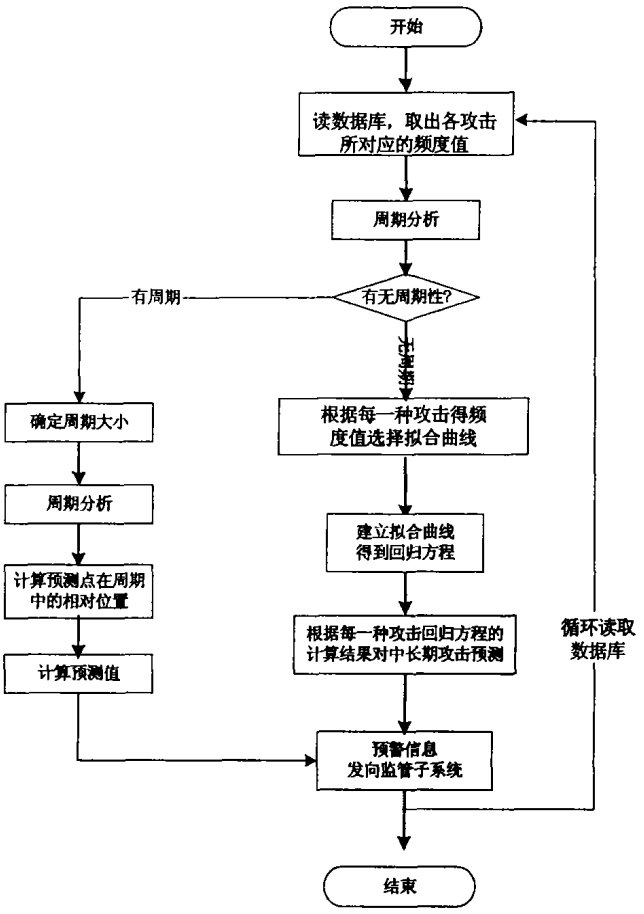

图1 趋势预测算法流程

算法步骤：

a. 读出历史数据(攻击频度值)，共取 N 个数据，给数据编号，0到$5N-1$。

b. 对所取的历史频度值进行周期分析：若无周期，则转 c，直接建立预测周期模型；若有周期，则转 d，确定周期大小，进入周期模型预测。

周期分析的方法如下：

限制条件：首先默认该数据的周期 $cycle=3$，至少从 $cycle=3$开始向后试探；历史数据中至少要有5组周期的数据，才认为该组数据可判周期。

1)从第四组数据($i=cycle$)开始，依次往后求其它各组与第一组数据的差的绝对值，直到该差的绝对值小于等于 ε(周期推测误差限)，记下当前数据序号 i ($3\leqslant i\leqslant N$)；

2)尝试以 i 为周期，从0～i 中随机选取3组数据，然后分别与下面四个周期中的对应点相减取绝对值；

3)如果相减取绝对值$<=\varepsilon$ 的数据的个数小于等于3，则认为周期判断成功，输出 i 作为周期，并退出所有循环，跳至程序出口处；

4)如果相减取绝对值$>\varepsilon$ 的数据的个数大于3，即有3个以上的奇异点，则尝试以 i 为周期失败，令 $i=i+1$，重做2)，3)，4)步，共重做2次；

5)如果 i 的上述候选值均不符合条件，则再从当前数据起，依次求各数据与第一个数据的差的绝对值，直到该差值小于等于 ε，记下当前数据序号 i；

6)重复第2)，3)，4)，5)步，直到 $i>N$；

7)如果 $i>N$，输出0，表示不存在周期。

c. 依据样本数据特点，选择回归模型。计算将来的攻击次数。

根据历史样本值选取拟合曲线，进行回归分析。数据库中记录着过去每一时间粒度 x_i 内该攻击所发生的频度值 x_i，这对应于一组数据(x_i，y_i)，根据该组数据样本，选择合适的回归方程。所用的回归分析的方法包括：(1)线性模型。线性模型是曲线模型中最简单的一种，其数学公式为 $y=a+bx$。(2)指数模型。也叫复比增长模型，其数学公式为 $y=k+ab^x$。(3) 修正指数曲线模型。其数学公式为 $y=k+ax^b$。(4) 逻辑斯谛曲线模型(皮尔曲线模型)，呈 S 形，是生长曲线的一种，其数学公式为 $y=1/(k+ab^x)$。(5) 非线性模型。是多项式回归模型中最常用的一种，其数学公式为 $y=a+bx+cx^2+dx^3\cdots$

在给定一个实际观察时序列 y_t($t=0,1,2,\cdots,n$) 的条件下，能建立的预测模型可以不同，但预测模型选择的正确与否直接关系到预测的准确程度。设给定的实际观察时序列为：$y_t(t=0,1,2,\cdots,n)$。

多项式曲线模型的选择：(1)若一阶差分 $\Delta y_t=y_t-y_{t-1}=c(t=0,1,2,\cdots,n)$，则可用线性模型进行预测；(2)若 k 阶差分 $\Delta^k y_t=\Delta^{k-1}y_t-\Delta^{k-1}y_{t-1}=c(t=k,k+1,\cdots,n)$，其中 c 为不等于零的常数，则可用 k 阶多项式曲线趋势模拟进行预测。

增长型曲线模型的选择：(1)若 $y_t/y_{t-1}=b$ ($t=1,2,\cdots,n$)，其中 b 是不为零的常数。则用指数曲线模型 $y_t=ab^t(t=0,1,2,\cdots,n)$进行预测。(2)若 $\Delta y_t/\Delta y_{t-1}(t=2,3,\cdots,n)$，其中 b 是不为零的常数。则用修正指数曲线模型 $y_t=k+ab^t(t=0,1,2,\cdots,n)$进行预测。(3)若$(1/y_t-1/y_{t-1})(1/y_{t-1}-1/y_{t-2})=b$ ($t=2,3,\cdots,n$)，其中 b 是不为零的常数，则可用逻辑斯谛曲线模型 $y=1/(k+ab^x)$ ($t=1,2,\cdots,n$)进行预测[6]。

在确定了预测模型之后，为了通过回归分析求出曲线模型的系数，需要将一些曲线方程变为线性形式：$y=k+ab^x\rightarrow \ln(y-k)=\ln a+(\ln b)*x$；$y=1/(k+ab^x)\rightarrow\ln(1/y-k)=\ln a+(\ln b)*x$。设共取 m 组数据(所取数据量与所预测的未来时间有关，本文中取5倍时间段的历史数据量)。然后根据选定的曲线方程，建立函数关系 $y=f(x,a_1,\cdots,a_n)$，确定其中 n 个参数 $a_1,\cdots,a_n$ 通过最小二乘法确定，使得 $\sum_{k=1}^{m}[f(x_k,a_1,\cdots,$

$a_n)-y_t]^2$最小。根据确定的该模型和参数，计算未来 t，可能发生的该种攻击的次数。

d. 计算预测值在周期内的相对位置。

对所取的历史数据，以长度 T 顺序分组，统计各个周期内相对位置的平均值，得到预测向量$[y_1, y_2, \cdots, y_T]$。所求的预测点在周期内对应于坐标 $X=6N \bmod T$，如图2。则返回预测值 $y[X]$。

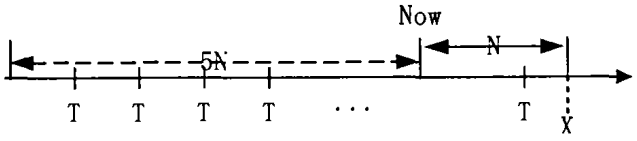

图2 预测点在周期内的相对位置

2.2.2 *效率分析* 趋势预测的复杂度取决于选择回归曲线的复杂度。设给定的实际观察时序列为：$y_t(t=0,1,2,\cdots,n)$。若为线性模型，计算复杂度为：$O(n)$；若为指数模型，也叫复比增长模型：$O(n)$；若为修正指数曲线模型，计算复杂度为：$O(n^2)$；若为逻辑斯谛曲线模型，计算复杂度为：$O(n)$；若为非线性(m 阶多项式)模型，计算复杂度为：$O(n^m)$。

2.2.3 *结果与讨论* 为了评价预测效果的好坏，定义预测算法中的基本指标误报率和漏报率。针对特定类型的网络攻击。设攻击发生的真实值为 R，预测值为 P。误报率(正错误)：$\frac{P-R}{R}\times 100\%$，即预测值与真实值之差和真实值的比率(预测值大于真实值)；漏报率(负错误)：$\frac{R-P}{R}\times 100\%$，即真实值与预测值之差和真实值的比率(预测值小于真实值)。

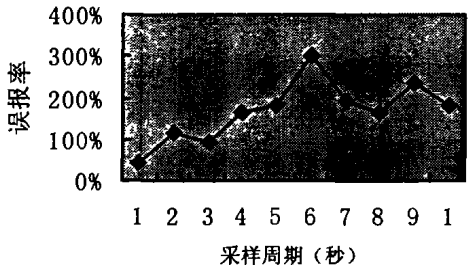

图3 非周期情况下误报率与采样周期关系

在上述预测算法中的时间粒度，也即统计攻击次数时的取样周期，是影响预测模型的重要参数。时间粒度的长短直接影响到数据的分布情况，将影响回归模型的选取。而数据趋势与回归模型近似的攻击类型能够得到更好的预测效果。对于周期性攻击，因其入侵事件本身具有周期分布的特点，只与周期内的相对位置有关，而与数据分布无关，所以，只须在非周期情况下，找到使误报率和漏报率较小的时间粒度 T_a 即可。我们利用专业攻击库软件 IDS Informer，定制了一组攻击，并对攻击类型为“misc-activity”的报警事件运用预测算法。这组攻击产生的报警事件经聚类后，频度值序列经周期分析程序判断，没有明显的周期性。在同一组攻击策略下，测试了时间粒度分别为1秒到10秒的算法效果。数据库中的数据分为两部分，5/6用于生成预测模型，1/6用于检验预测结果。误报率和漏报率随时间粒度(采样周期)的变化关系分别见图3、图4。由图3，可以清晰看到，随着时间粒度的增大，误报率有增大的趋势。时间粒度为1秒时，误报率为46%。图4说明，漏报率对时间粒度的变动不敏感，没有明显的变化趋势。漏报率随时间粒度变化上下浮动。考虑到过小的时间粒度可能造成每个时间粒度内的聚类频次过少，出现多个连续零的情况，因此取 $T_a=3$ 秒(误报率为95%，漏报率为49%)。

分析预测算法对于周期性攻击的预测效果。选定固定的时间粒度 T_a(3秒)。考察当攻击周期接近、大于、远大于时间粒度 T_a 时的预测效果。制定了五组不同的攻击策略，这些攻击策略使得产生的攻击周期分别为2秒、6秒、9秒、15秒、42秒。预测算法对不同的周期性数据的适应性情况如表1。在攻击周期 T_a 小于时间粒度的情况下($T_a=2$秒)，漏报率与误报率都很小(1%)，这是因为在一个时间粒度内，攻击可以完成一轮，每个时间粒度内的数据几乎是相等的。周期性数据的漏报率与误报率都比非周期性的预测效果要好。而且预测较近的未来比预测较远的未来效果好。

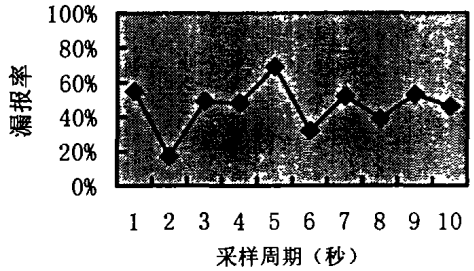

图4 非周期情况下漏报率与采样周期关系

表1 周期情况下误报率漏报率与攻击周期关系

攻击周期(秒)	$T_a=2$	$T_a=6$	$T_a=9$	$T_a=15$	$T_a=42$
误报率	1%	29%	34%	25%	5%
漏报率	1%	27%	35%	25%	46%

实验在100M 局域网中进行，一台攻击机(Windows 平台，IDS Informer V4)，报警事件产生器(Linux 平台，Snort V1.84)。

3 应用实例

我们将上述预警方法应用到战略预警项目。在实现的原型系统中，战略预警系统为用户或监管系统提供历史安全信息与未来安全趋势，是一个相对独立的系统。预警子系统逻辑结构划分为三层：入侵报警收集器(底层网段层)，聚类单元(中层聚类层)，高层规律表示与趋势预测(预警中心)。考虑系统要适用于大规模网络，对情报的收集和处理采用网络分区方式进行。网络分区是将整个互联网划分为多个区域，每个区域可以包含多个网段，划分的依据是便于情报的收集和处理。每个网络分区由一个低层聚类单元负责，每个低层聚类单元包括若干个入侵报警收集器和一个预警分析部件。每个网段部署一个入侵报警收集器，负责收集情报。低层聚类单元对报警信息采用聚类分析的方法，从中找出相关联的攻击事件，以及它们分布情况。分析结果发布功能由预警中心完成，主要完成全局安全状态管理，分析结果秘密、完整地发布。预警子系统的功能结构如图5所示。

结论 提出的基于入侵事件预测的预警方法，依据特定攻击发生的历史规律，预测网络中的攻击在未来一段时间的发生次数。分析讨论了时间粒度对预测结果的影响，以及对周期性攻击和非周期性攻击的预测效果。该预测算法对于非周期性攻击误报率为95%，漏报率为49%，对于周期性攻击预测误报率为19%和漏报率为27%。可以采用时序列分析的方法替换文中的回归分析部分，并通过组合预测的方法降低单一方法的预测误差。

(下转第129页)

dio Frequency Identification, June 2002
4 Manes A. Introduction to Web Services. White Paper of Systinet, Inc., 2002
5 Bonsor K. How Smart Labels Will Work. http://www.howstuffworks.com/smart-label.htm, Feb. 2003
6 Ye X, Qiu R. Web Services Oriented Approach to High Availability of Product Information. 2003 International Conference of Industrial Engineering & Engineering Management, Shanghai, Dec. 2003
7 MIT. Auto-ID Center Technology Guide. http://www.autoidcenter.org, Feb. 2003
8 Alexander K, Gilliam T, Gramling K, et al. Focus on the Supply Chain: Applying Auto-ID within the Distribution Center. White Paper of MIT Auto-ID Center, June 2002
9 Kambil A, Brooks J. Auto-ID Across the Value Chain: From Dramatic Potential to Greater Efficiency & Profit. White Paper of MIT Auto-ID Center, June 2002
10 Alexander K, Birkhofer G, Gramling K, et al. Focus on Retail: Applying Auto-ID to Improve Product Availability at the Retail Shelf. White Paper of MIT Auto-ID Center, June 2002

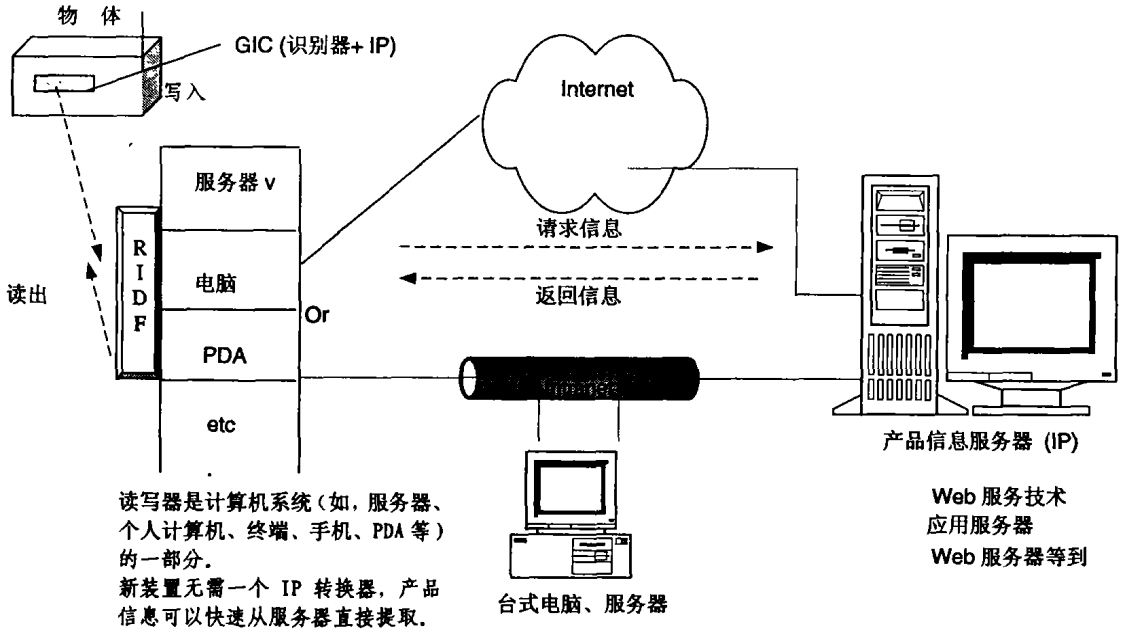

图5 全球识别码获取信息的新方法

(上接第79页)

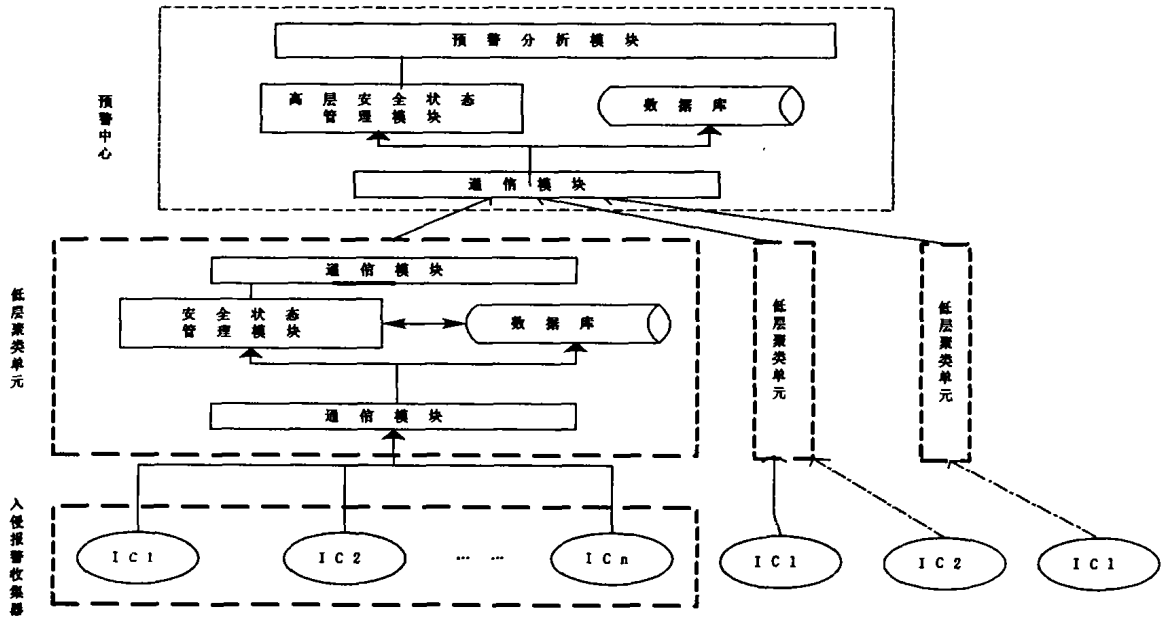

图5 预警系统处理流程

参考文献

1 Baumann R, Plattner C. Honeypots, Diploma thesis. http://security.rbaumann.net/download/diplomathesis.pdf. 2002
2 Buchholz F, Thomas E D, Kuperman B, et al. Packet Tracker Final Report, CERIAS Technical Report. Purdue University. http://www.cerias.purdue.edu/infosec/bibtex — archive//archive/2000-23.pdf. 2000
3 Rathmell A, Dorschner J, Knights M. Project: Threat Assessment and Early Warning Methodologies for Information Assurance, Http://www.icsa,ac.uk/Projects/ropa.html IAAC, Summary of Research Results: Early Warning & Threat Assessment Methodologies For Information Assurance. http://www.iaac.org.uk/Publications/ROPA/Website%20summary.pdf. May, 2001
4 Shyhtsun J Y, Felix W, Fengmin G, Ming-Yuh H. Intrusion Detection for an On-Going Attack. http://www.mnlab.cs.depaul.edu/seminar/ fall2002/IDSongoing.pdf. 1999
5 Ming-Yuh H, Jasper R J, Wicks T M. A Large-scale Distributed Intrusion Detection Framework Based on Attack Strategy Analysis. Computer Networks (Amsterdam, Netherlands), 1999, 31(23-24): 2465～2475
6 李正龙. 时序列特征与预测模型选择. 预测, 2001, 20(5): 70～73