一种基于成本的入侵检测系统评估决策树分析方法*)

谢 亮

(浙江金融职业学院 杭州 310016)

摘 要 随着入侵检测系统的广泛应用,对入侵检测系统进行评估的要求也越来越迫切。本文首先对入侵检测评估的现状进行了深入的总结研究,然后在已有方法的基础上,提出了基于成本的基于成本的入侵检测系统评估决策树分析方法。该方法也是基于ROC曲线的,它通过决策树引入成本,利用最优预计成本表征入侵检测系统性能。最后通过实验表明了该方法的有效性。

关键词 入侵检测系统评估,成本,ROC曲线,决策树

A Decision Tree Analytical Method Based on Cost for IDS Evaluation

XIE Liang

(Zhejiang Financial Professional College, Hangzhou 310016)

Abstract With the broad application of intrusion detection systems (IDSs), it's more and more urgent to evaluate these IDSs. First, this paper summarizes the status of IDSs evaluation. Then based on existed methods, we bring forward an analytical method on performance: the decision tree analytical method based on cost for IDSs evaluation. This method based on ROC curve, but introduces cost by decision tree. Then it scales the performance of IDSs with corresponding cost. At last, we make a experiment to prove the validity of this method.

Keywords Intrusion detection systems evaluation, Cost, ROC curve, Decision tree

1 引言

目前入侵检测系统(Intrusion Detection System, IDS)[1]已经广泛应用于政府、企业和科研机构等组织,成为网络安全的重要防御工具之一[2,3]。尽管入侵检测已发展了20多年,但对其评估方面开展的工作还不是很多。大多数IDS的研制厂家出于各方面的原因,宣传时常常夸大其词,而IDS的用户对此往往又不是很清楚,所以迫切需要建立起一个科学的评估方法及可信的评估标准。对IDS进行评估,可以为购买者提供产品选购的参考依据,为安全分析员提供安全报警信息的相关性分析,更重要的是,还可为入侵检测技术的研发者提供IDS的最优配置、系统的优缺点以及性能改进的相关数据。

入侵检测必须考虑成本问题,IDS应该以最小的代价实现最大的安全目标。理想的IDS对所有检测到的入侵行为都做出响应,但是从资源限制等实际情况考虑,"不借一切代价"的响应显然是不合理的,响应成本不能超过预计入侵带来的损失。因此IDS的评估研究意义重大,是入侵检测的另一个重要的研究方向[4]。

但是,就目前而言,技术研究人员更多地考虑的是技术可行性或技术有效性,而忽视了成本可行性或成本有效性。这一方面是因为技术人员与企业用户考虑问题的侧重点不同,另外一个方面,是因为成本分析需要解决一个成本量化的问题,而目前量化是一个难点。

2 基本概念

在说明基于成本的IDS评估决策树分析方法之前,先对一些基本概念进行说明。

1)DS在某个时刻所处的状态:a)有入侵发生,用 I 表示;b)没有入侵发生,用 N 表示。

2)入侵的先验概率,用 p 表示,即 $p(I)=p$,根据不同的环境,p 会有不同的取值,用于描述环境的敌对性。

3)S 的检测报告:a)有入侵报警,用 A 表示;b)没有入侵报警,用 NA 表示。

4)由ROC曲线可知:a)检测率 β;b)误警率 α。

5)综合以上参量得到如下条件概率表达式:

a)$P(A|I)=\beta$,表示有入侵发生时系统报警的概率,即检测率;

b)$P(NA|I)=1-\beta$,表示有入侵发生时系统无报警的概率,即漏检率;

c)$P(A|NI)=\alpha$,表示无入侵发生时系统报警的概率,即误警率。

6)对于IDS提交的报告,操作者有两种选择:a)响应,该操作用 R 表示;b)不响应,用 NR 表示。

这是成本分析的重点,因为前面已经提到,响应与不响应具有不同的成本,我们采取最小化预计成本的策略。至于如何选取最小化预计成本值,将在后文详细介绍。

7)在入侵检测过程中,会有各种事件发生,各事件概率表示如下:

a)用 p_1 表示没有警报的概率;

b)用 p_2 表示没有报警时不是入侵的概率;

c)用 p_3 表示有报警时不是入侵的概率;

d)其中 $1-p_2$、p_3 都对应着检测器报告错误的概率。

8)成本定义

a) C_α 没有入侵发生却响应的成本,即误警所带来的成本;

b) $C_{1-\beta}$ 有入侵发生却没有响应的成本,即漏检所带来的成本。

一般来讲,漏检所带来的成本要大于误警所带来的成本。我们可以假定正确响应的成本为0,设 C_α 为单位1,相对成

*)基金:2006年度浙江省教育厅科研项目立项:"基于Web使用挖掘的个性化推荐系统的研发",项目编号20060441。谢 亮 硕士生,讲师,研究方向为计算机网络安全,数据挖掘,入侵检测。

本值 $C=C_{1-\beta}/C_{\alpha}$，多数情况下，$C>1$。

3 基于成本的 IDS 评估决策树分析方法

对于 IDS 的报告，操作者或自动响应模块采取的操作不同，带来的成本就不同，是否响应取决于操作所带来的成本，如果响应带来的成本小于不响应带来的成本，我们就应该响应，反之就不响应；另外，成本也是 IDS 评估的一个重要方面，我们可以通过系统的最优预计成本评估 IDS 的性能。那么，如何将成本引入 IDS 评估呢？我们采用决策树分析方法[5]，将成本引入 IDS，并在此基础上推导出计算 IDS 的最优预计成本的公式。

根据以上基本概念，使用如下决策树分析方法来计算 IDS 的预计成本，如图 1 所示。该方法是在被评估系统的 ROC 曲线已知的情况下进行的。

从形式上看，图 1 中的决策树深度为 4，从左到右的层次依次为：系统检测报告、响应操作和系统所处的状态。其中有 2 个全概率事件结点（用□表示）、5 个概率事件结点（用○表示），8 个成本结点（用⊗表示），总共 8 条分支。

从意义上看，该决策树表示了系统检测报告、操作员或自动响应模块的响应操作、IDS 所处的状态（是否有入侵发生）以及成本四个属性之间的复杂关系。全概率事件结点表示的是操作员或自动响应模块在接到系统报告时的操作：要么响应，要么不响应。概率事件结点表示某事件的发生具有不确定性，比如对于系统报警事件，系统不报警的概率为 p_1，报警的概率（$1-p_1$）；再比如系统没有报警时某个事件是否是入侵这也是一个不确定事件，是，入侵的概率为 p_2，不是，入侵的概率为（$1-p_2$）；相应，系统报警时某个事件是否是入侵这也是一个不确定事件，是，入侵的概率为 p_3，不是，入侵的概率为（$1-p_3$）。决策树描述了系统检测报告、操作员或自动响应模块的响应操作、IDS 所处的状态（是否有入侵发生）以及成本四个属性之间的所有可能关系。以分支 II 为例，它表示系统无报警（报警概率为 p_1），操作者或自动响应模块不进行响应，但此时系统有入侵发生（概率为（$1-p_2$）），同时漏检所带来的成本值为 C。

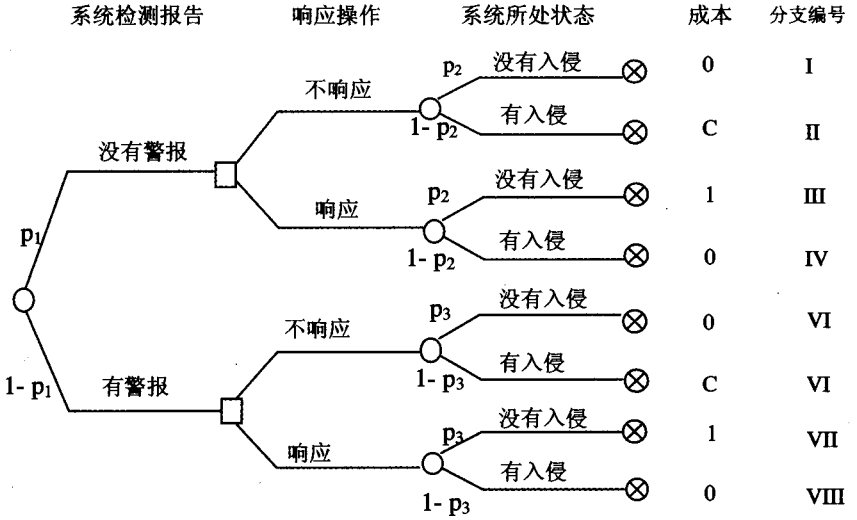

图 1 决策树成本分析

通过图 1 中的决策树可推导出计算预计成本值的公式。我们的分析基于已知 IDS 的 ROC 曲线这一假定之上。

对于一组已知的 α、β 值，假设当前环境的入侵先验概率为 p，相对成本为 C，为了计算 ROC 曲线上各个点的成本，首先需要计算 p_1、p_2、p_3 的值。由贝叶斯公式可知：

$$p_1=P(NA)=P(NA|NI)P(NI)+P(NA|I)P(I)=(1-\alpha)(1-p)-(1-\beta)p \quad (1)$$

$$1-p_1=P(A)=P(A|NI)P(NI)+P(A|I)P(I)=\alpha(1-p)+\beta p \quad (2)$$

$$p_2=P(NI|NA)=P(NA|NI)P(NI)/P(NA)=(1-\alpha)(1-p)/p_1=(1-\alpha)(1-p)/[(1-\alpha)(1-p)+(1-\beta)p]) \quad (3)$$

$$1-p_2=P(I|NA)=P(NA|I)P(I)/P(NA)=(1-\beta)p/p_1=(1-\beta)p/[1-\alpha)(1-p)+(1-\beta)p] \quad (4)$$

$$p_3=P(NI|A)=P(A|NI)P(NI)/P(A)=\alpha(1-p)/(1-p_1)=\alpha(1-p)/[\alpha(1-p)+\beta p] \quad (5)$$

$$1-p_3=P(I|A)=P(A|I)P(I)/P(A)=\beta p(1-p_1)=\beta p/[\alpha(1-p)+\beta p] \quad (6)$$

计算出 p_1、p_2、p_3 后，再结合决策树进行分析，可分别计算出 8 条分支下系统所花费的成本。由于我们假定正确操作情况下的成本为 0，正确操作分两种情况，一种是实际上没有入侵发生，操作员或自动响应模块没有响应，另一种情况是实际上有入侵发生，操作员或自动响应模块也有响应，因此所有正确操作的预计成本都为 0，如分支 I、IV、V、VIII。而对于错误操作，我们同样可计算其预计成本。错误操作也有两种情况，一种是实际上没有入侵发生，但操作员或自动响应模块有响应，此种情况下成本为 1，另一种情况是实际上有入侵发生，操作员或自动响应模块却没有响应，此种情况下成本为 C。根据公式（1）～（6），由决策树可以分别计算出没有警报也没有响应、没有警报却有响应、有警报也有响应和有警报却没有响应这四种情况下的预计成本，如表 1 所示。

表 1 错误操作的预计成本表

检测报告	不响应	响应
没有警报	$C(1-p_2)=C(1-\beta)p/p_1=(1-\beta)p/[(1-\alpha)(1-p)+(1-\beta)p]$	$p_2=(1-\alpha)(1-p)/p_1=(1-\alpha)(1-p)/[(1-\alpha)(1-p)+(1-\beta)p]$
有警报	$C(1-p_3)=C\beta p/(1-p_1)=C\beta p/[\alpha(1-p)+\beta p]$	$p_3=\alpha(1-p)/(1-p_1)=\alpha(1-p)/[\alpha(1-p)+\beta p]$

进一步分析，在没有警报的情况下，操作员或自动响应模

块不响应的预计成本为 $C(1-\beta)p/p_1$，响应的预计成本$(1-\alpha)(1-p)/p_1$。为了使得成本最低，在计算最小预计成本时，理应取这两个值的最小值：$\min\{C(1-\beta)p/p_1,(1-\alpha)(1-p)/p_1\}$。同理，在有警报的情况下，预计成本为 $\min\{C\beta p/(1-p_1),\alpha(1-p)/(1-p_1)\}$。

再由决策树可知，系统有否警报是一个概率事件，没有警报的概率为 p_1，有警报的概率$(1-p_1)$，由此可以得到在给定 p、C 值时，ROC 曲线图上的每个点(α,β)的预计成本公式：

$$Cost(\alpha,\beta)=p_1\min\{C(1-\beta)p/p_1,(1-\alpha)(1-p)/p_1\}+(1-p_1)\min\{C\beta p/(1-p_1),\alpha(1-p)/(1-p_1)\}=\min\{C(1-\beta)p,(1-\alpha)(1-p)\}+\min\{C\beta p,\alpha(1-p)\} \quad (7)$$

设 ROC 曲线为 A，则 A 的最优预计成本公式为：

$$\text{Min } Cost(\alpha,\beta)=\min\{Cost(\alpha,\beta),\forall(\alpha,\beta)\in A\} \quad (8)$$

由公式(8)可以看出，IDS 的最优预计成本不仅与 IDS 的检测率及误警率有关，而且与相对成本 C 及 IDS 的运行环境的敌对性(入侵的先验概率 p 表示)有关。

4 基于成本的 IDS 评估决策树分析方法举例

为了进一步对基于成本的 IDS 评估决策树分析方法进行说明，并证明该方法的有效性，我们针对两个真实的 IDS，具体分析情况如下：

下面运用基于成本的 IDS 评估决策树分析方法对两个真实的 IDS 进行了评估，由于被评估的 IDS 的名称不宜公开，在这里分别为称这两个 IDS 为 IDS1 和 IDS2。在实验环境中，设定每天在 830,000 个网络会话中有 106 个入侵，也就是说环境敌对性为 106/830000 攻击/会话，即入侵检测的先验概率 $p=106/830000=1.277\times10^{-4}$。假设相对成本 $C=500$，也就是说漏检所造成的成本是误警所造成的成本的 500 倍。由于只是想说明基于成本的 IDS 评估决策树分析方法，因此对得到 ROC 曲线图的过程不予详述，在实验中可以得到 2 个 IDS 的 ROC 曲线图如图 2 所示。

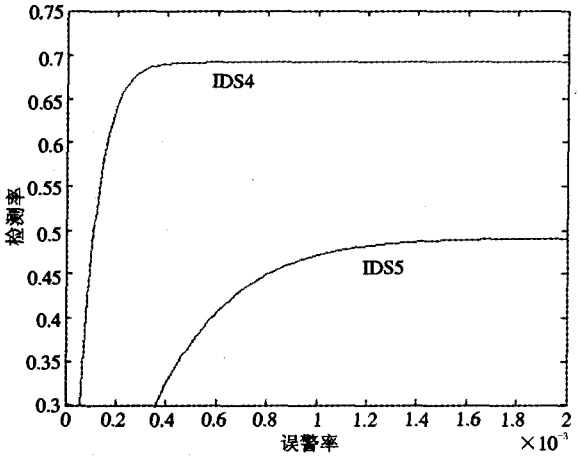

图 2 IDS1 与 IDS2 的 ROC 图

根据上面公式可以分别得到这两个 IDS 的最优点及其最优预计成本。如表 2 所示。

从表 2 中可以看出，在环境条件为 $p=1.277\times10^{-4}$，$C=500$ 的情况下。IDS1 比 IDS2 的预计成本低。

上面分析了通过最优预计成本来评估两个 IDS：IDS1 和 IDS2 孰好孰坏的问题。那么其性能是否能达到所要求的检测目标呢？假设 IDS 的检测目标为：检测率 $\beta=0.98$，误警率 $\alpha=0.01$。我们通过计算 IDS1 与 IDS2 在各个 p、C 值下的最优预计成本及目标点$(\alpha=0.01,\beta=0.98)$的预计成本，来判断能否达到目标要求。结果如表 3 所示。

表 2 IDS1 和 IDS2 的最优点及其最优预计成本
($p=1.277\times10^{-4}$，$C=500$)

	IDS4	IDS5
α	0.00038	0.0018
β	0.5876	0.4256
预计成本	0.0415	0.4517

表 3 C=500 下 IDS1、IDS2 与目标的预计成本

p		α	β	预计成本
1.277×10^{-4}	IDS4	0.0003	0.684	0.0020
	IDS5	0.0008	0.435	0.0034
	目标	0.01	0.98	0.0049
1×10^{-3}	IDS4	0.0003	0.675	0.0158
	IDS5	0.0018	0.479	0.270
	目标	0.01	0.98	0.0103
1×10^{-5}	IDS4	0.0004	0.691	0.0108
	IDS5	0.0017	0.492	0.0187
	目标	0.01	0.98	0.0105
1×10^{-6}	IDS4	0.00024	0.664	0.0013
	IDS5	0.0004	0.374	0.0022
	目标	0.01	0.98	0.0025

由表 3 可知，当 $p=1\times10^{-5}$ 时，目标预计成本是 0.0105，IDS1 的预计成本是 0.0108，和目标成本很接近。在 $p=1.277\times10^{-4}$ 和 $p=1\times10^{-6}$ 情形下，尽管 IDS1 比 IDS2 预计成本低不少，但两个系统都比目标系统成本低，因此任意一个都是有效的。而当 $p=1\times10^{-3}$ 情形下，两个系统都达不到目标要求。

同样，在固定 p 值情况下，改变 C 的值，可以得到如下结论：如果 C 的值很高，那么无论 p 值多小，IDS1 和 IDS2 都达不到目标。例如当 $C=5000$，$p=1\times10^{-6}$，IDS2 不能达到目标需求，而 IDS1 可以。而当 $C=7000$，$p=1\times10^{-6}$ 时，IDS1 或 IDS2 都没有达到目标需求。

结论 本文提出了一种新的 IDS 评估分析方法。该方法应用决策树分析方法，将已有的 ROC 曲线分析方法与成本分析相结合，引出一个新的评估量—预计成本，计算 IDS 的 ROC 曲线上各个点的预计成本，取最小预计成本为最优预计成本，具有最优预计成本的点叫做最优预计成本点。用最优预计成本作为评估的性能指标，评估 IDS 的性能目标，以及决定一个 IDS 在给定环境中的最优配置。比较各个不同的 IDS 的最优预计成本，就可以评估出这些系统的优劣。本文最后使用了两个真实的 IDS，验证了本方法的有效性。在该方法的分析过程中我们得到了一个附加的结论：IDS 的最优配置不仅与该 IDS 的 ROC 曲线有关，还与成本和运行环境的敌对性(可用入侵发生的先验概率表示)有关。

参 考 文 献

1 Kumar S. Classification and detection of computer intrusions: [Ph. D. Thsis]. Purdue University, 1995

2 Jiang Jian-chun, Ma Heng-tai, Ren Dang-en, et al. A survey of intrusion detection research on network security Journal of Software, 2000, 11(11): 1460～1466

3 Debar H, Dacier M, Wespi A. Towards a taxonomy of intrusion detection systems. Computer Networks, 1999, 31 (8): 805～822

4 Mell P, Hu V, Lippmann R, et al. An overview of issues in testing intrusion detection systems: [Interagency Report]. National Institute of Standards and Technology, 2003

5 Quinlan J R. Induction of Decision Trees [J]. Machine Learning, 1986, 1(1): 81～106