doi:10.3969/j.issn.1673-9833.2013.06.011

基于因子分析的混合贝叶斯入侵检测算法

吴欣欣，文志诚，叶健健

（湖南工业大学 计算机与通信学院，湖南 株洲 412007）

摘 要：传统的基于贝叶斯网络的入侵检测技术中，未考虑到入侵检测数据量过多的问题，导致贝叶斯网络构造过程中计算量过大，从而使得检测效率偏低；还有其检测的数据仅来源于网络或者主机，使得数据来源单一，对检验的准确性造成了一定程度的影响，针对上述2个问题，提出了基于因子分析的混合贝叶斯入侵检测技术，利用因子分析对网络连接数据的属性特征进行选择，降低了数据相关性，同时将网络数据和主机数据综合起来分析评定网络当前安全状态，以提高入侵检测的准确度。试验结果表明：改进后的检测技术能降低数据维数，提高了计算效率和检测精度。

关键词：贝叶斯网络；因子分析；入侵检测

中图分类号：TP393 **文献标志码**：A **文章编号**：1673-9833(2013)06-0052-05

Study on Hybrid Bayesian Intrusion Detection Method Based on Factor Analysis

Wu Xinxin，Wen Zhicheng，Ye Jianjian

（School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：As the traditional Bayesian network intrusion detection technology exists some problems of low detection efficiency owing to too much calculation and the single data source resulting in a certain influence on the accuracy of the examination, puts forward a hybrid Bayesian intrusion detection technology based on factor analysis. The method applies factor analysis to the attribute features selection of network connecting data and reduces the data correlation, at the same time, integrates network data and the host data for analysis and evaluation of the current network security state to improve the detection accuracy. The experimental results show that the proposed detection technology greatly reduces the dimension of data and improves the computational efficiency and detection accuracy.

Keywords：Bayesian network；factor analysis；intrusion detection

0 引言

随着计算机网络技术的普及与发展，网络在人们日常生活、工作以及学习中所扮演的角色越来越重要，对社会的影响力也越来越大。然而计算机网络技术发展的同时，也带动了网络攻击手段的不断翻新，因此网络安全受到了前所未有的挑战。入侵检测作为确保网络安全的一种重要手段，已成为网络安全领域中研究的热点，但传统的入侵检测技术存在一些缺陷[1]：1）漏报率与误报率非常高；2）无法检测未知的攻击模式，只能对模式库中已有的攻击模式进行识别；3）试验数据来源单一，导致检测准确率偏低，因此，寻求高有效性与强自适应性的入侵检测模型成为了研究重点。近年来，国内外学者提出了许多入侵检测方法来解决上述问题，这些方法都能提高入侵检测率。常用的方法是将贝叶斯

收稿日期：2013-09-05

作者简介：吴欣欣（1989-），女，湖南张家界人，湖南工业大学硕士生，主要研究方向为网络安全态势感知，
E-mail：2533818144@qq.com

网络与入侵检测技术相结合。这种方法虽然提高了入侵检测的有效性与自适应性，并且可以主动防御外来的攻击，但是基于贝叶斯网络的入侵检测技术有着明显的不足：对数据进行贝叶斯分类时，未考虑冗余的数据属性会提高数据维度，这导致贝叶斯分类工作强度加大，且分类效果不好，还加大了贝叶斯网络构造过程的难度，并且数据之间存在的相关性也增加了分析问题的难度。

针对上述问题，本文提出基于因子分析的混合贝叶斯入侵检测方法。该方法从 2 个方面对已有的基于贝叶斯网络的入侵检测技术进行了改进：1）分析 2 种数据。本方法所分析的数据来源于主机和整个网络环境，这解决了基于贝叶斯网络的入侵检测技术的数据来源单一、不具代表性的问题。本方法通过综合评析主机系统和网络的数据流两部分数据来判断被保护的系统是否受到攻击，进而提高入侵检测的准确率。2）简化检测数据。将因子分析法应用于入侵检测系统，能有效地消除数据冗余，降低数据相关性，从而达到简化数据的目的，避免了因为数据量过大而导致贝叶斯分类的计算工作纷繁复杂，使后续工作无法进行等问题，因此，入侵检测率有明显提高。

1　基于贝叶斯的入侵检测系统

贝叶斯网络是概率论与图论相结合的产物。它由一个有向无环图和条件概率表构成，图中的节点代表随机变量，有向弧代表变量之间的依赖关系，没有弧则表示变量之间相互独立，概率表中的概率值代表节点之间相互依赖关系的强度。节点变量是人们日常生活中问题的抽象，例如测试值、观测现象、意见征询等。贝叶斯网络的概率推理实际上是一个计算概率的过程。基于贝叶斯网络的入侵检测模型如图 1 所示。该检测模型是以网络连接数据作为研究对象，将网络连接数据集分为训练集和测试集。通过对训练集的学习，得到贝叶斯分类器模型，再利用模型对测试集进行测试，从而得到有效的检测模型。

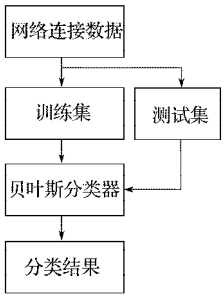

图 1　基于贝叶斯的入侵检测模型

Fig. 1　Detection model based on Bayesian intrusion

2　关键技术简介

因子分析法又称为因素分析法，是一种多元统计分析方法。该方法从一些信息重叠、具有错综复杂关系的变量中归纳出几个不相关的综合因子[2]，而这些综合因子可以解释原有信息中的大部分数据以及数据之间的联系。

因子分析法模型如下：

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + e_1, \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + e_2, \\ \vdots \\ x_n = a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{nm}F_m + e_n 。 \end{cases} \quad (1)$$

式中：$X=(x_1, x_2, \cdots, x_n)$是一组可观测的随机变量，其均值为 0，方差为 1；$F_i(i=1, 2, \cdots, m)$为 X 的一组公共因子；$e_i(i=1, 2, \cdots, n)$为特殊因子，是 X 的各个分量所特有的；a_{ij} 为因子载荷，a_{ij} 反应了第 i 个变量与第 j 个因子之间的相关性，即第 i 个变量在第 j 个因子上的权重，a_{ij} 的绝对值越大，则变量 x_i 与公共因子 F_j 的相关性越强。

因子载荷矩阵 $\boldsymbol{A}$ 可表示为

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}。$$

变量共同度和公共因子的方差贡献是因子分析法中 2 个非常重要的统计量。变量共同度用来反映样本 X 对公共因子 F 的每一个分量的依赖性大小。公共因子的方差贡献用来提取出重要的公共因子。

1）变量共同度。变量共同度是因子载荷矩阵 $\boldsymbol{A}$ 中的第 i 行的元素的平方和，其代表全部的公共因子对 x_i 的方差所做出的贡献，反映了全部公共因子对变量 x_i 的影响。变量共同度越大，则 X 的第 i 个分量 x_i 对 F 的每一个分量的依赖性就越大。

2）公共因子的方差贡献。公共因子的方差贡献是因子载荷矩阵 $\boldsymbol{A}$ 的第 j 列的元素的平方和，表示第 j 个公共因子对 X 的每个分量所提供的方差总和，是衡量每个公共因子重要性的指标。公共因子方差贡献越大，则第 j 个公共因子对 X 的影响就越大。将因子载荷矩阵 $\boldsymbol{A}$ 的每一列的公共因子方差贡献都计算出来，便可以将公共因子按照对 X 的影响力的大小进行排序，最终提取出最重要的公共因子。

提取最重要的公共因子之后，如果各公共因子的典型代表变量不是很突出，则需要进行因子旋转。因子旋转的目的就是使因子载荷矩阵中因子载荷的平均值尽量向 0 和 1 两个极值转化，大的载荷更大，小的载荷更小，让每个变量在尽可能少的因子上有

比较高的载荷，从而分辨出各个因子的重要性，最终得到比较满意的公共因子。

3 改进的入侵检测系统设计

3.1 系统模型

基于因子分析的混合贝叶斯入侵检测模型从两部分来检测当前网络是否安全：一部分是检测主机，从主机中提取出其日志、文件、系统调用等信息，通过判断这些信息是否发生异常来评判主机当前的安全情况；另一部分是检测网络系统，从网络中抓取数据包，将获取的大量数据首先运用因子分析法进行数据的简化与降维处理，从而去除了数据中多余的冗余属性，便于之后的数据分析与处理，结合贝叶斯网络对简化的数据进行概率推理，并计算出当前网络环境下攻击发生的概率，最后综合考虑主机和网络两部分的检测结果，得出当前网络是否受到攻击威胁以及安全程度，如图2所示。

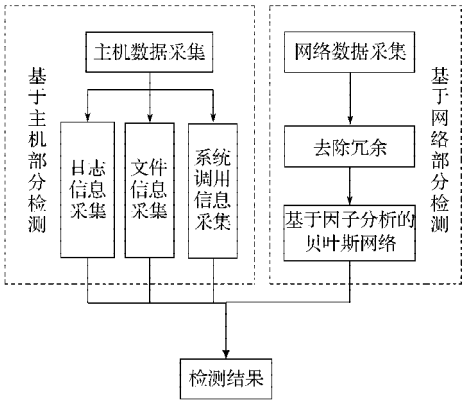

图2 改进的入侵检测系统模型

Fig. 2 The improved system model of intrusion detection

3.2 网络部分检测

基于贝叶斯的入侵检测系统通过从网络上抓取大量的数据包得到网络连接数据[5]，而这些数据间的相关性产生了大量冗余的数据属性，这就增加了贝叶斯分类的计算量，进而使后面的工作加大了难度，导致入侵检测率下降。而改进后的入侵检测系统运用因子分析法对网络连接数据进行分类，将相关性较高、联系比较紧密的数据变量分在同一类中，而不同类中的数据变量之间的相关性则较低。因子分析法可以在不丢失主要信息的情况下对数据进行简化处理，用少数不相关的因子变量代替原有信息中的大量数据，避免了数据分析过程中数据量过多，以及数据信息重叠等问题，且能够以最小的信息损失来解释变量之间的结构，从而较大降低了之后数据分析的难度以及数据的相关性，为之后的计算工作做好了充足的准备。

从网络获取的网络连接数据中，取出一部分作为研究样本。先用因子分析法对样本中的数据进行去除冗余处理，之后将其分成2部分，即取研究样本的3/4作为训练集，剩下的作为测试集，通过对训练集的学习得到了朴素贝叶斯分类模型，再通过此模型对测试集的数据进行检测[6-8]。网络部分检测模型如图3所示。

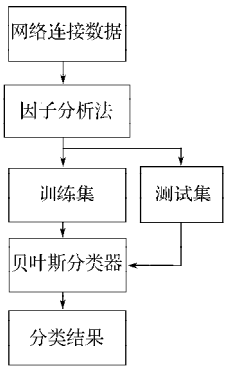

图3 网络部分检测模型图

Fig. 3 Network detection model

朴素贝叶斯分类器的工作原理如下。给定一个数据样本X，将每一条网络连接数据用一个特征向量$(x_1, x_2, \cdots, x_n)$来表示，x_i为它的第i个属性。假定有m个类$C_1, C_2, \cdots, C_m$，给定一个数据样本X，$P(X)$为样本X的全概率，$P(C_j|X)$表示在X发生的情况下C_j发生的概率。

由贝叶斯公式可得样本X属于C_j的概率为

$$P(C_j \mid X)=\frac{P(C_j)P(X \mid C_j)}{P(X)}, \qquad (2)$$

式中：$P(X)$为常数；$P(C_j)$与$P(X|C_j)$可以通过先验知识或者训练集学习得到。通过此式可以计算出样本事例X属于各个类的概率，将X归于后验概率最大的那个类，如果$P(C_j|X)$值超过了先验知识中规定的阈值，则判断有入侵发生。

3.3 主机部分检测

通过搜集主机的系统安全日志、审计数据、目录以及文件中的异常改变、程序执行中的异常行为等，获得系统活动信息、系统资源利用率、日志信息、系统调用信息和主机上的重要文件读写操作信息等，再根据这些信息来检查系统中是否存在违反安全策略的行为和被攻击的迹象。本文对主机进行入侵检测主要是检测3个部分：日志信息、系统调用和文件。

1）系统调用

系统调用是应用程序与操作系统之间的接口，可以使用户空间和内核空间之间进行信息的交互，用户可以通过系统调用请求操作系统来完成某些需要在内核状态下执行的操作，比如输入输出、进程管理、文件系统、存储管理等。同时，内核程序被计算机系统很好地保护起来，用户进程不可以直接访问，这样使得整个计算机系统更加稳定和安全，大部分的网络攻击都是通过非法改变系统调用来完成的，因此可以通过阻止网络攻击的相应进程系统调

用来达到预先避免攻击的目的。如在 Linux 系统中，系统调用在内核中实际上是一个数组列表指针对应的函数列表，通过替换需要阻止的函数的指针，就可以截获相应的系统调用。

2）日志信息

日志用来记录计算机系统日常所发生的重要事件，即记载了计算机系统的相关活动，包括用户登录、应用程序异常、系统事件、安全相关事件等。如果发生了上述事件，这些事件将会被记录到日志中，系统管理员可以通过事件查看器进行查看[7]。这样便于跟踪非法入侵或者查看发生问题的原因，如果日志信息被入侵者修改或者清除，则无法追踪到入侵发生的相关信息，为了阻止日志信息被修改，创建了一个线程来监控日志文件。如果日志文件被修改，则可获取被修改的日志文件名，并且对此日志文件进行分析。

3）文件

对文件的检测主要 2 个方面：1）利用病毒扫描软件对文件进行扫描，防止文件中毒；2）通过事先定义好的完整性规则对文件进行完整性检测。

3.4　改进的入侵检测系统流程图

改进的入侵检测系统的工作流程如图 4 所示。

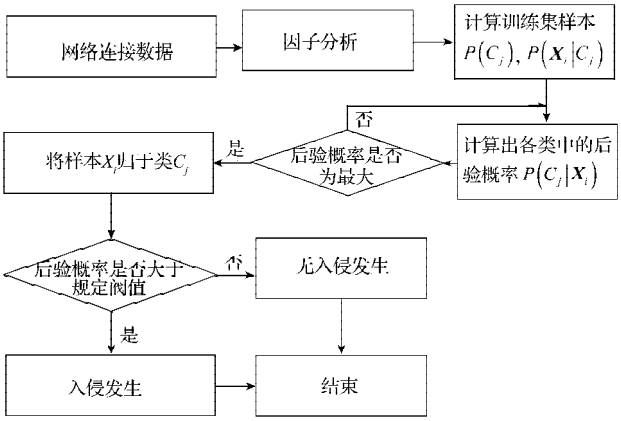

图 4　入侵检测流程图

Fig. 4　Flow chart of intrusion detection

步骤 1　从网络抓取的数据集中的每一条记录由 42 个属性构成，其中有 34 个连续属性值，为了便于计算，首先将数据集中的这些连续值进行离散化处理和数值化处理。

步骤 2　运用因子分析法对离散化之后的数据进行降维，去除冗余、减少相关性，得到具有较少属性的新数据集。

步骤 3　从新的数据集中提取出一部分数据。

步骤 4　对训练集进行学习。采用统计的方法或者专家知识得到每个类在训练集中出现的概率，即 $P(C_j)$，再计算样本 X_i 在 C_j 中出现的概率为 $P(X_i|C_j)$。

步骤 5　通过式（2）计算出 X_i 在每个类中的后验概率，再将 X_i 归于后验概率最大的那个类，

步骤 6　判断 X_i 的后验概率值是否大于该类规定的阈值，如果大于，则表示有入侵发生。

根据先验知识得到先验概率，再根据之后得到的新证据或者结果去修改概率（后验概率），不断地重复这个过程得到后验概率分布，而这个后验分布也可作为其他预测过程的先验分布，这样为最后的预测做好准备。

程序部分代码如下：

输入：网络数据样本，计数器的值。

输出：贝叶斯分类器，计数器的值。

```
BEGIN
    Let M represent the net data;
    ND=Net(M);   //获得网络数据
    DR=Reduce(ND); //对网络数据进行降维
  // 处理，将进行降维后的数据作为样本
    SV=Sample(DR); //取出样本中的每一条
  // 记录
    Let L represent the counter data;
    CV=counter(L);  //取出计数器中的数据
    If(SV.equals(CV)) //如果计数器中已经存 // 在此条记录的数据类型
    Counter++; //计数器自增
    Else  Counter 1++; //用新的计数器记录
  // 其类型的个数。
    return counter;
    return counter 1;
    CP=Compute(Counter ,Counter 1);
  // 根据两个计数器的值计算各类型的概率，及各类
  // 型中各属性的概率。
    return CP;
    END。
```

4　试验结果分析

试验所采用的网络数据来源于知识发现与数据挖掘国际会议。此数据集中的网络连接数据包含的攻击类型有 Land, Pod, Nmap, Satan, Mscan。通过对 20 000 条数据进行试验，分析本方法与传统的贝叶斯方法、神经网络方法的检测率，即

检测率 = 检测到的攻击数目 / 总攻击数目。

试验结果如表 1 所示。由表可知，基于因子分析

的混合贝叶斯入侵检测方法与传统的贝叶斯和神经网络相比，检测率有所提高。因子分析法可以在信息丢失最少的情况下化简数据，找出数据集的主要特征属性，删除冗余的数据属性，这样减少了数据之间的相关性，降低了分类难度，提高了检测率。

表1 检测率对比表

Table 1 Comparison of detection rate %

攻击	检测率		
	神经网络方法	传统贝叶斯方法	本文所用方法
Land	94.11	95.02	95.53
Pod	80.26	80.97	82.39
Nmap	87.85	88.18	89.32
Satan	78.95	79.63	80.86
Mscan	78.62	79.74	80.92

5 结语

已有的基于贝叶斯的入侵检测方法所检测的数据单一，用网络数据或主机数据。本文所研究的入侵检测方法既检测了主机数据，也分析了网络数据，从主机和网络2部分同时对网络安全进行检测，提高了入侵检测的准确率，同时运用因子分析法对数据进行化简，减少入侵检测系统中贝叶斯分类的数据计算时间，提高入侵检测率。本文的方法能更好地检测出已知的入侵攻击，在一定程度上提高了入侵检测系统对网络攻击的检测能力，降低了入侵检测的误报率与漏报率，并且在适应性、有效性以及智能性方面都有所提高。

参考文献：

[1] 卿斯汉，蒋建春，马恒太，等. 入侵检测技术研究综述[J]. 通信学报，2004，25(7)：19-29.
Qing Sihan，Jiang Jianchun，Ma Hengtai，et al. Research on Intrusion Detection Technology：A Survey[J]. Journal of China Institute of Communications，2004，25(7)：19-29.

[2] 王 娟，慈林林，姚康泽. 特征选择方法综述[J]. 计算机工程与科学，2005，27(12)：68-71.
Wang Juan，Ci Linlin，Yao Kangze. A Survey of Feature Selection[J]. Computer Engineering & Science，2005，27(12)：68-71.

[3] 朱慧明，孙雄志. 基于混合先验分布的贝叶斯因子分析模型[J]. 湖南大学学报：自然科学版，2007，34(9)：82-85.
Zhu Huiming，Sun Xiongzhi. Bayesian Model for Factor Analysis Using Mixing Prior Distribution[J]. Journal of Hunan University：Natural Sciences，2007，34(9)：82-85.

[4] 胡春玲. 贝叶斯网络研究综述[J]. 合肥学院学报：自然科学版，2013，23(1)：33-40.
Hu Chunling. Research Overview on Bayesian Network[J]. Journal of Hefei University：Natural Sciences，2013，23(1)：33-40.

[5] 高 俊，吕述望. 入侵检测系统：IDS[J]. 现代电信科技，2001，24(9)：34-52.
Gao Jun，L ü Shuwang. Intrusion Detection System：IDS[J]. Modern Telecommunications Technology，2001，24(9)：34-52.

[6] 周颜军，王双成，王 辉. 基于贝叶斯网络的分类器研究[J]. 东北师大学报：自然科学版，2003，35(2)：21-26.
Zhou Yanjun，Wang Shuangcheng，Wang Hui. Research for the Classifiers Based on Bayesian Networks[J]. Journal of Northeast Normal University：Natural Science，2003，35(2)：21-26.

[7] 蒋建春，马恒太，任党恩，等. 网络安全入侵检测：研究综述[J]. 软件学报，2000，11(11)：1460-1466.
Jiang Jianchun，Ma Hengtai，Ren Dang'en，et al. A Survey of Intrusion Detection Research on Network Security[J]. Journal of Software，2000，11(11)：1460-1466.

[8] Denning D E. An Intrusion-Detection Model[C]//IEEE Transaction on Software Engineering. [S. l.]：IEEE，1987：222-232.

[9] Li E. Intrusion Detection Systems[EB/OL]. [2013-7-10]. http://uwcisa.uwaterloo.ca/Biblio2/Year/2010/ACC626%20Intrusion%20Detection%20Systems%20E%20Li.pdf. 2010.

[10] 马恒太，蒋建春，陈伟锋，等. 基于Agent的分布式入侵检测系统模型[J]. 软件学报，2000，11(10)：1312-1319.
Ma Hengtai，Jiang Jianchun，Chen Weifeng，et al. Distributed Model of Intrusion Detection System Based on Agent[J]. Journal of Software，2000，11(10)：1312-1319.

（责任编辑：邓 彬）