A Model of Intrusion Detection Based on Data Mining in Campus Network

QI Bei[1], DONG Yun-feng[2]

1. Network Center, Shandong Institute of Light Industry, 250353, Jinan, China

2. Computing Center, Shandong Institute of Light Industry, 250353, Jinan, China

1. qb@sdili.edu.cn, 2.dyf@sdili.edu.cn

Abstract: In campus network, it is the precondition of network's normal running that establishes the integrate network security system. Campus network mostly adopts the firewall technology to resist various network attack, but now, network security is threaten from inside and outside campus network, so the intrusion detection technology becomes the main defense measure that assists the firewall. Now, network intrusion detection has problem of high false alarm rate, we have adopted the method which combines the data mining and intrusion detection system together, incorporated the technology of network traffic monitoring and analysis. We propose a new model of intrusion detection based on data mining and network traffic analysis. The experiment result indicates this model can find many kinds of network intrusion behavior effectively and have higher intelligence and environment adaptability.

Keywords: campus network; intrusion detection; data mining; network traffic analysis

校园网中基于数据挖掘的入侵检测模型

亓　蓓[1]，董云峰[2]

1.山东轻工业学院现代网络中心，济南，中国，250353

2. 山东轻工业学院计算中心，济南，中国，250353

1.qb@sdili.edu.cn, 2.dyf@sdili.edu.cn

【摘要】在校园网中，建立完整的网络安全体系是其正常运行的前提条件。校园网大多采取防火墙技术来抵御各种网络攻击，但是现在的网络安全受到来自校园网内外的双层威胁，因此入侵检测技术成为辅助防火墙的主要防御手段。针对目前网络入侵检测中存在虚警率较高的问题，我们采取了将数据挖掘与入侵检测系统相结合的方法，融入网络流量监控分析技术，提出了一种新的基于数据挖掘和流量分析的网络入侵检测模型。实验结果表明该模型能够有效地发现多种网络入侵行为，具有更高的智能性和环境适应性。

【关键词】校园网；入侵检测；数据挖掘；流量监控分析　好

1 引言

近几年来，全国各所高校的规模都在不断的扩大，校园网网络的规格也在快速的增长，校园网的网络安全受到来自校园网内外的双层威胁。为了保证网络的正常运行，现在大多采取防火墙技术来抵御各种网络攻击。但是再好的防火墙一旦被绕过就会成为一个“马奇诺防线”，所以校园网多数采取入侵检测技术进行辅助。在对大量网络入侵检测技术进行研究的基础上，根据高校校园网入侵数据只占网络中传输的所有数据总量的很小一部分和入侵数据和正常数据是可以区分开来的特点，我们结合网络流量监控和异常分析技术，建立了一个基于数据挖掘的智能化入侵检测模型。

2 入侵检测模型的总体结构

该模型主要由四大模块组成：数据采集与分类模块，不确定数据流量异常检测模块，不确定数据流量异常分析模块和神经网络判决模块。总体框架如图 1 所示。在数据采集与分类模块中，使用聚类算法对网络数据分类，分为确定数据和不确定数据，确定数据用于训练第四个模块的神经网络，不确定数据转入流量分析；在不确定数据流量异常检测模块中，利用异常检测技术，对

978-1-935068-31-0

网络数据包流量进行检测，当发现异常后，启动异常分析模块；异常分析模块对出现流量异常的网络采用数据挖掘算法，挖掘出频繁序列，建立精简高效的序列规则，对不确定数据的流量异常做出判断，给神经网络判决模块提供流量数据参考；神经网络判决模块通过确定数据学习，同时根据流量数据和学习的规则判断不确定数据，实现入侵检测模型的智能化。

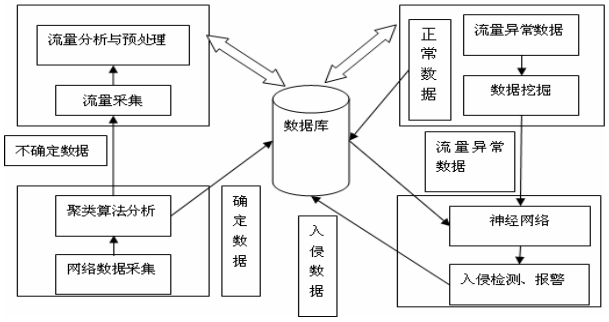

Figure 1. Frame of The Model

图 1. 模型的框架

3 数据预处理

我们用于入侵检测的实验数据来自 1999 KDDCUP 入侵检测数据集，数据集中总共包括 4898431 条记录，每一条记录都是由在模拟入侵的过程中搜集的网络数据的连接记录所提取的特征向量。所提取的特征包括独立的 TCP 连接包含的基本特征。这些实验数据集有以下特点：

1、每条记录包含 41 个特征，其中 34 个连续特征，7 个离散特征。

2、每条记录都有标签(正常的或是一种入侵行为)，共有 4 种攻击方式[1]。

4 改进聚类算法分析网络数据

现在越来越多的研究项目将数据挖掘技术应用到入侵检测中，各种数据挖掘算法，如关联规则挖掘算法、频繁情节规则挖掘算法、分类算法、聚类算法[2][3]都被提出来。在我们的这个模型中，我们选择聚类算法，聚类算法是一种“无指导的学习算法”，它摆脱了对带标识的训练数据的依赖，对于我们这个模型的第一部分——网络数据分类非常适合。我们采用的聚类算法是一种改进的聚类算法，是将 K-MEANS 算法和 DBSCAN 算法相结合的一种算法。

4.1 K-MEANS 算法

K-MEANS算法接受输入量k；然后将n个数据对象划分为k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。其工作过程说明如下：首先从n个数据对象任意选择 k 个对象作为初始聚类中心；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。k个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

K-MEANS 算法流程如下：算法 K-MEANS 中 k 是簇中对象平均值输入簇的数目。

（0）输入簇的数目值到 k，同时将一个较大值赋给 MIN；

（1）对所有对象数据度量值进行标准化处理；

（2）任意选择 K 个对象作为初始簇中心；

（3）Repeat；//3-6 循环

（4）计算剩余每个对象到每个簇中心的欧几里德距离；

（5）根据与各簇中心对象的距离，找出与之距离最小簇中心，并将该对象分配给以该簇中心为聚类核心的簇；

（6）Until 对象的分配不再变化；

（7）计算当前聚类情况的准则函数（1）的收敛值 VALUE；如果 VALUE>MIN，则 MIN=VALUE；

（8）转到（2）继续执行，直到准则函数（1）的收敛值最小 MIN 被找到。

准则函数：

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} \left| x - m_i \right|^2 \qquad (1)$$

结果是收敛值取最小 MIN 时的 K 个聚类即为所求结果。

基于上述算法计算结束后，得到 Cl，C2，…，Ci 个簇的集合 D，其中 i=l…k。及每个簇的对象集合 C1={obi，…}，…，Ci={obm…}；其中 i，…，m 属于 1～N。然后计算 DBSCAN 算法中需要的半径 E 和最少数目 MinPts 这两个条件，得到这两个数据后，利用 DBSCAN(E，MinPts)算法对 k 个聚类中的异常记录

集合进行再次分析，得到更精确的分析结果。

计算半径 E 的大概过程是先计算每一个集合 D 中的每一个对象数目在两个以上的簇的平均距离，然后将这些距离求和，用距离和除以集合 D 的簇数目得到半径 E。其中簇的平均距离是指簇中对象的欧几里德距离之和除以簇中对象的组合数。MinPts 是集合 D 的簇中相邻对象距离小于半径 E 的对象总数在集合 D 中所有对象总数所占的比例。

4.2 DBSCAN 算法

得到了半径 E 和最少数目 MinPts 后，我们使用 DBSCAN 算法再次分析，得到网络中的不确定数据聚类。DBSCAN 算法如下：

输入：包含 n 个对象的数据库，半径 e，最少数目 MinPts；

输出：所有生成的簇，达到密度要求。

（1）Repeat

（2）从数据库中抽出一个未处理的点；

（3）If 抽出的点是核心点 Then 找出所有从该点密度可达的对象，形成一个簇；

（4）Else 抽出的点是边缘点（非核心对象），跳出本次循环，寻找下一个点；

（5）Until 所有的点都被处理。

这样，模型的第一部分将网络数据分为确定的数据和不确定的数据两部分，给将已经判定为正常的数据用来训练神经网络，使用训练后的神经网络对不确定点再次进行判定。不确定的数据进行流量分析，先通过模型的第二部分采集流量，再使用数据挖掘算法分析异常流量的不确定数据，得到更精确的数据分析结果。

5 流量异常分析

5.1 流量收集及预处理

流量收集及预处理的主要工作是收集发生异常的流量数据，根据端口进行流量统计，记录端口流量在前 N 位（本文选择前 10 位）的端口及流量等相关信息，将提取的信息存入数据库中。

（a）流量收集。流量收集部分的数据来自数据库，当接到流量异常的通知后，在一定的时间内，记录发生异常的所有子网端口的数据信息：源地址，目的地址，源端口，目的端口，状态，端口号，流量值以及时间，把它们写入数据库中，并标记为异常数据文件。

（b）概要信息提取。因为异常检测系统所关心的问题就是找到这种流量剧烈变化的情况。设时间 T1，T2(T2>T1)，某端口 C。

一种情况是，T1 时刻端口 C 流量排名没有在前 N 位，而在 T2 时刻端口 C 的流量排名出现在前 N 位，则很可能是出现了异常情况。另一种情况是，T1 时刻端口 C 流量排名在前 N 位，而在 T2 时刻端口 C 的流量排名提高很多，则很可能是出现了异常情况。因此提取某时刻流量排名前 N 位的端口及其流量，挖掘这些端口及其流量的频繁模式来检测流量异常情况。

（c）离散化方法。为减少连续属性值个数，我们对流量数据进行离散化处理。现采取自然划分分段的方法对流量数据进行划分。

5.2 网络流量频繁模式挖掘算法

频繁模式挖掘采用的算法是 CLOSET 算法。系统定期收集各端口的流量信息，定期计算频繁模式。为了适应流量处理的特点，在获得的频繁模式中加入了时间标记，记录频繁模式生成的时间，通过频繁模式的时间标记来反映频繁模式的存在期限和优先程度，挖掘出异常流量的频繁序列，得到不确定数据中流量异常的数据序列，建立精简高效的序列规则，交由人工神经网络进行训练、检测。

CLOSET 算法的实现：

输入：概要数据库，最小支持度阈值 min_support；

输出：频繁闭模式集合 FCI；

方法：

BEGIN

（1）建立 flist：

扫描概要数据库，提取信息加入数据库；

计算 item(port,stream)相同的项目数量；

将频繁项列表 flist 置为空；

对数据库中每个 transaction{

if(item.support>=min_support)

将 item 加入频繁项列表 flist；}

（2）将频繁项列表 flist 按照支持数降序排列：Descend_flist_item(flist)；

（3）根据 flist 在数据库中获得数据中的频繁项列表，并排序：transaction_item_descend()；

（4）FCI=φ；//初始化频繁闭项集 FCI

（5）使用 CLOSET 算法，获得频繁闭项集 FCI；

（6）CLOSETφ；

END

其中将 flist 设计为数组类型，每个数据项的数据

类型为：

type fist_node as
{port//端口值
stream//流量离散值
support//支持数
point//指向树节点的指针}

6 BP 神经网络

选择神经网络的目的就是要利用它的自学习性，使得系统在数据不完整和易变的情况下，仍然能得到一个全面的特征轮廓。通过以上聚类算法和流量分析算法，我们已经将确定数据的数据集划分成大小不同的聚类，并将流量异常的不确定数据，也就是我们认为的入侵数据提取了出来。根据聚类的规模，我们将子聚类从大到小排列，并把较大聚类中的数据看作是正常数据，将其作为训练数据用于训练神经网络。同时，使用检测到流量异常的数据也用于训练，最后训练过的神经网络用来对较小聚类中的数据做出判别。

训练采用 BP 算法是一种有教师的学习算法，该算法利用了均方误差和梯度下降法来实现对网络连接权的修正，修正的目标是使网络实际输出与规定输出之间的均方误差最小。BP 网络模型由三层组成，层与层之间全连接。由于 BP 网络对大数据集训练速度较慢，我们使用了自适应修改学习率的算法用来训练。在 BP 网络训练的负梯度算法中，学习率是一个固定的常数，而且它的值将直接影响到训练性能。如果选择得太大，会降低网络的稳定性；如果选择得过小，会导致过长的训练时间。本算法首先计算出网络的输出误差，然后在每次训练结束后，利用此时的学习率计算出网络的权值和阈值，并且计算出网络此时的输出误差。如果此时的输出误差与前一时刻的输出误差的比值大于预先定义的参数，那么就减小学习率，反之，就增加学习率。再重新计算网络的权值和阈值以及输出误差，直到前后输出误差的比值小于参数为止。

7 实验结果及结论

我们利用工作环境拥有的成型的校园网入侵检测系统，以校园网一年中收集的网络数据为基础，构造了数据集，总共 721,068 条数据，其中已经确定的入侵数据 5,423 条（占 0. 76% ）。经过聚类算法后，将各个子聚类按规模从大到小排列，将确定数据集中的数据（占全部数据总量的 72% ）作为正常数据，和检测到的流量异常的不确定数据作为入侵数据一起用于 BP 网络的训练。训练后的网络用于识别剩余 27%的未知数据，这部分的数据分类后再次用于 BP 网络训练，这时的神经网络就具有较高的准确率来检测入侵。最终实验表明检测率达到 82%，虚警率小于 6%。

在当前校园网网络安全形势日趋复杂、校园网安全状况日益严峻的情况下，作为一种主动防御技术的入侵检测技术越来越得到人们重视。针对当前入侵检测系统存在误报率和漏报率较高的问题，我们的这个基于数据挖掘的校园网入侵检测模型将聚类算法、网络流量监控分析和神经网络结合，取得了很好的效果。该模型的优势在于可自学习，自适应性好，对常规的入侵手段有着很高的检测率，同时也提高了检测未知入侵的能力。

References (参考文献)

[1] Pei J,Han J,Mao R.CLOSET:An EfficientAlgorithm for Mining Frequent Closed Itemsets [J].Proc.2000 ACM-SIGMOD Int.Workshop onData Mining and Knowledge Discovery(DMKD'00), 2000: 132-196.
[2] Sdkant. Fast Algorithms for Mining Association Rules and Sequential Patterns[C]. Madison: University of Wisconsin, 2003. 24(5): 324~355.
[3] 1999 KDD Cup Competition [DB/OL]. http: //kdd. ics.uc.i edu/databases/kddcup99/kddcup99. htm.l 2004. 8.22.