DOI:10.19304/j.cnki.issn1000-7180.2008.05.016

基于聚类的网络异常检测

张西广[1], 郑秋生[1], 王虎祥[2], 陈国强[3]
(1 中原工学院 计算机学院, 河南 郑州 450007; 2 忻州师范学院 计算机系, 山西 忻州 034000;
3 河南大学 计算机与信息工程学院, 河南 开封 475001)

摘 要: 探索一种基于聚类来识别异常的方法, 这个方法不需要手动标示的训练数据集却可以探测到很多不同类型的入侵行为.实验结果表明该方法是可行的和有效的, 使用它来进行异常检测可以得到探测率和误报率的一个平衡, 从而为异常检测问题提供一个较好的解决办法.

关键词: 入侵检测; 聚类; 特征检测; 异常检测
中图分类号: TP393 文献标识码: A 文章编号: 1000-7180(2008)05-0062-04

Anomaly Detection by Clustering in the Network

ZHANG Xi-guang[1], ZHENG Qiu-sheng[1], WANG Hu-xiang[2], CHEN Guo-qiang[3]
(1 College of Computer, Zhongyuan Institute of Technology, Zhengzhou 450007, China;
2 Department of Computer, Xinzhou Teacher's University, Xinzhou 034000, China;
3 College of Computer and Information Engineering, Henan University, Kaifeng 475001, China)

Abstract: In this paper we explore a clustering algorithm to identify outliers. No manually classified data is necessary for training and it is able to detect many different types of intrusion. The experiment result shows that the algorithm is feasible and effective, and anomaly detection by using this algorithm could get a balance between false positive rate and detection rate, so it could be a better solution to anomaly detection.

Key words: intrusion detection; clustering; signature detection; anomaly detection

1 引言

随着互联网与我们日常生活联系的愈加紧密, 它也日益成为一个吸引人的攻击目标.有许多安全策略都集中在如何预防攻击, 如认证、过滤以及密码技术等; 而入侵检测是在安全预防措施被攻破以后如何检测到这些攻击; 只有将预防和检测结合起来才能提供更加安全可靠的网络环境.

为了知道一次网络攻击是否发生, 需要大量分析从网络上收集来的数据, 从中寻找可疑的活动.目前有两大类常规的做法: 特征检测和异常检测.特征检测利用已知的攻击模式来发现攻击, 它对已知攻击的探测非常有效, 但无法探测未知攻击.异常检测是寻找网络中与正常网络行为偏离的行为, 它解决了无法察觉未知攻击的问题, 不过它发现的只是和正常网络行为不相符的行为, 这样的行为未必一定是有害的, 因此可能产生误报情况.

特征检测是利用安全专家生成的一套规则来进行判定, 当新的漏洞或者攻击发生的时候, 判定规则需要进行更新.在调查新的未知攻击时, 安全专家通常利用自身经验形成的正常模型来区分异常事件.而异常检测的思想是通过对历史数据的学习, 赋予计算机同样的区分异常事件的能力.由于在不同的环境下, 对正常行为的定义也可能不同, 因此每种环境都需要建立一个相应的正常行为模型, 这样潜在的攻击行为想绕过异常检测系统就会比较困难, 文中的目标就是探索一种能够为单独的网络环境建立正常行为判别模型的方法.

但是异常检测也有其自身的难点和需要解决的问题.首先比起确定的攻击特征, 正常行为的特征定

收稿日期: 2007-07-11

义更加抽象和模糊. 其次, 常规的机器学习方式都是通过对标记为不同类别的数据来进行判定学习, 然而异常检测只能通过给定的正常数据这样一个种类来建立正常模型, 异常数据类别是缺失的. 为了建立一个正常行为模型, 还需要一个没有任何攻击行为的数据集来训练, 但实际上这样的数据集是很难获取的. 文中提出通过聚类的方法来解决第二个问题, 它不需要没有任何攻击行为的、干净的数据集, 最后的试验结果表明, 文中使用的方法可以检测到攻击而且误报率也相对比较低.

2　相关工作

异常检测有基于主机的异常检测和基于网络的异常检测. 因为有些攻击仅仅来自内部, 并不产生网络流量, 因此基于主机的检测是很有必要的, 但是基于网络的异常检测可以在攻击还只是发起却没有到达主机之前进行察觉, 从某种程度上说这是一个很有效的补充.

比较著名的网络异常监测系统有 ADAM[1]、Seurat[2]、NIDES 和 SPADE 等, ADAM 系统采用关联规则对集中发生的事件才进行处理, 但关联规则本身存在一定的缺陷: 假设一个异常事件对整体的状况影响并不大的时候, 为了检测到这个事件, 必须设置相应较低的阀值, 而这个过程会产生非常多不必要的关联规则, 从而在有效率和漏报率之间存在一个权衡. Seurat 系统从逆向进行判定, 如果用户的最终状态是正常的, 那么可以认为它在这一段时间内是属于正常的, 此时, 系统只需要做出一步的判断就能大致的判断用户在这段时间内的状态, Seurat 系统采取一个模糊的概念来进行判断, 它用一种抽象派的点画法来形容自己的系统. 由于大部分的攻击都会引起文件系统的更改, 并且局域网内主机系统、安装的软件和存在的漏洞都倾向于一致性, 因此一次发作的病毒倾向于更新同一个或者几个文件, Seurat 系统所针对的特征就是利用对文件系统共性更改的检测来探测病毒. 当该系统部署到更大的网络范围中时, 由于主机之间存在的差异性可能远远大于局域网内主机之间的差异性, 所以系统检测效果会大大减弱.

当前的网络异常检测机制大多采用的都是一个恒定的模型, 事件的发生概率判定取决于它在训练集中发生的概率, 这样的模型并不会随着时间而做出调整. 这样的做法实际上并不符合应用的真实情况, 网络上的数据流会存在一种自相似的行为[3], 它不是在任何时间段都很平均地发生, 相反它会在某个时刻产生正常的突发, 在一个给定的时间窗口里面, 很难预测它正常的发生概率.

3　通过聚类进行异常检测

3.1　特征向量和距离

每一个数据点 p 通过一个特征向量来表示, 而每一个簇 c 通过它的中心和半径来进行表示. 此处采用的距离度量是 Euclidean 距离: $\mathrm{distance}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \sqrt{\sum_{j=1}^{n}(p_{1j} - p_{2j})^2}$. 其中 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是特征向量, p_{ij} 表示对象 p_i 的第 j 个属性, n 是对象属性的维度.

为了获取每个数据点的特征向量, 需要把原始的输入属性向量 $\boldsymbol{X}_i$ 转化成特征向量 $\boldsymbol{Y}_i$, 根据属性特征是连续的变量还是离散变量而采取不同的转换方式. 对于离散型的变量, 那些越经常出现的离散值就越不可能成为异常, 因此可以利用这个离散值出现的频率来对属性进行描述, 那些发生频率越接近的属性越相关, 发生频率差异越大的越不相关, 这样就可以把离散值转化成连续性的变量.

为了适应各种可能的情况, 聚类算法必须能够适应任何数据分布的数据集, 算法将根据数据集情况把所有的数据转换到文中规定范围的数据空间中. 整个转换方式如下所示(avg _ v 表示属性向量的均值, std _ v 表示属性向量的标准差)

$$\mathrm{avg\_}v[j] = \frac{1}{N}\sum_{i=1}^{N} p_i[j],$$

$$\mathrm{std\_}v[j] = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p_i[j] - \mathrm{avg\_}v[j])^2}.$$

式中, $p[j]$ 就是点 p 的第 j 个属性, 最终的数据点会被转化为 $\mathrm{new\_}p[j] = \frac{p[j] - \mathrm{avg\_}v[j]}{\mathrm{std\_}v[j]}$. 换句话说, 对于每一个属性, 这里计算它偏离了平均值的距离相对均方差的比值来进行属性的标准化操作.

3.2　聚类

一旦完成了对象的空间转换和标准化, 聚类算法就可以针对对象向量之间的距离进行聚类操作. 聚类的操作有很多现成的方法, 例如 k - means[4]、BIRCH[5]、DBSCAN[6] 等经典算法. 此处采用的是一个较为简单的迭代算法, 用来对 k - means 进行初始化操作, 它一般是在人工无法确定如何给参数 k 赋值的时候使用的. 聚类算法假设每一个簇都有一个中心, 对象向量计算它与这些簇中心的距离并加入到距离最小的那个簇中. 算法在开始的时候随机

的选择一个对象作为簇的中心，然后迭代的选择距离自身簇中心最远的那个对象作为一个新簇的新中心点，根据这些新的中心点对所有的数据点进行一次重新的分配，这个过程一直持续到没有对象到它所在的簇的中心点距离超过簇与簇之间平均距离的一半结束，其伪代码流程如下：

输入：包含 n 个对象的数据库.

输出：k 个簇，使平均误差准则最小.

方法：

（1）任意选择 1 个对象作为初始簇的中心；

（2）repeat

（3）选择距离自身簇中心最远的那个对象作为一个新簇的新中心点；

（4）更新数据点的分配情况；

（5）until 没有对象到它所在的簇的中心点距离超过簇与簇之间平均距离的一半.

3.3 标示簇和检测

由于这里处理的是没有任何标记的数据，也无法根据标记进行训练，为了发现哪些簇属于正常行为，哪些簇属于攻击，文中利用了如下一些假设：首先，属于同一个分类（属于正常或者攻击）的数据之间的特征向量距离应该是很近的，而属于不同分类的数据之间的特征向量距离应该是较远的. 同时，网络上正常行为的数据流条数也应该是远远超过异常行为的攻击数据流条数，因此算法假定那些包含 N %数据的簇所包含的数据都是正常的，而剩下的那些簇就标示为异常的.

一旦根据训练集生成了簇的中心点模型，系统就可以利用它来判定每一个新来的数据是否属于异常，其处理过程如下：

（1）根据训练集生成簇时计算的统计信息把新的对象 p 转化到新的特征空间 p'.

（2）找出距离 p' 最近的那个簇，也就是集合中的某个 C，对于所有的 C' 而言，$\mathrm{dist}(C, p') \leqslant \mathrm{dist}(C', p')$.

（3）根据簇 C 的标签（正常还是异常）来标示 p' 是正常还是异常.

换句话说，整个处理过程就是找到距离点 p 最近的那个簇，并且利用这个簇的标示对 p 进行赋值.

4 实验结果

实验使用的训练数据集和判定数据集是采用 KDD CUP 99[7] 数据集的不同子集. 为了测试性能，聚类算法将对假定的正常数据应该占有的比例进行调整，采用探测率和误报率进行综合的评价. 为了对这些指标进行统计，需要用到实验数据集的分类标签，它仅仅是为了算法计算效率的时候才需要使用，而在实际使用中，系统的运行不需要这些标签的.

当调整正常的数据应该占有的比重时，针对这个数据集评测的结果如表 1 所示. 相对来说，设定 80 %作为数据集中正常的数据比例可以得到探测率和误报率的一个平衡，它在一个可以接受的误报率范围之内有不错的检测效率. 图 1 显示的是误报率和探测率之间的 ROC 曲线图，这个曲线图是用来可视化误报率和探测率之间的一个权衡图形.

表 1 评测结果

正常数据比例/ %	探测率/ %	误报率/ %
90	27.42	0.52
85	31.43	0.94
80	66.51	2.32
75	80.89	6.45
70	85.98	10.42

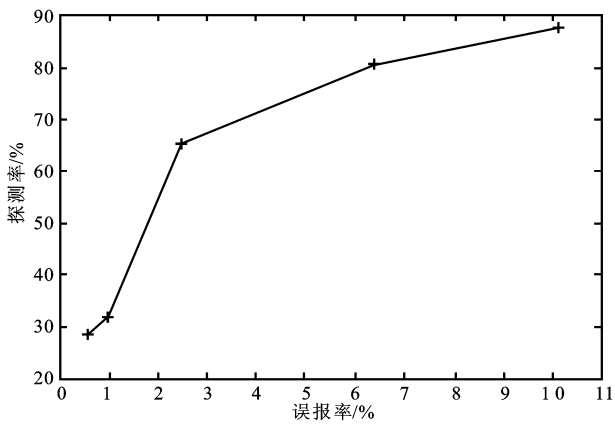

图 1 误报率和探测率的 ROC 曲线图

5 结束语

文中介绍了一种基于聚类来检测入侵的方法，它不需要在建立模型的时候提供训练数据集的标记信息，根据这个算法实现的检测系统可以检测到大量的入侵行为并保证了相对可以接受的误报率. 这个系统比起基于特征的、需要提供训练数据集的标记信息的方法而言有自身独特的优势. 首先它不需要手工地对训练数据集进行分类这项繁琐的工作，其次系统检测新的攻击类型的时候不需要详细地了解新攻击的特征，只需要新的攻击在训练数据集中出现即可，这样系统可以自动地检测数据究竟是正常还是异常[8].

将来还可以对算法进行一些新的扩展，因为网

络是一个不断进化的模型, 这就要求每隔一段时间手工的触发模型更新, 对新的数据重新计算一次判定模型. 如果系统可以不断的根据实际情况对判定模型进行增量更新, 那么就不需要每隔一段时间进行人工触发整个模型的更新了.

参考文献:

[1] Barbara D, Wu N, Jajodia S. Detecting novel network intrusion using bayes estimators[C]// First SIAM Int'l Conf. On Data Mining (SDM'01). USA: Chicago, 2001.

[2] Xie Y, Kim H, O'Hallaron D, et al. Seurat: a pointillist approach to anomaly detection[C]// Proc. of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID2004). France: French Riviera, 2004, Springer Verlag: 238—257.

[3] 赵佳宁, 李忠诚. 基于模拟的网络流量自相似现象分析[J]. 计算机科学, 2001, 28(11): 57—61.

[4] Mac Queen J. Some methods for classification and analysis of multivariate observations[C]// Proc. of the 5th Berkeley Symp. Bernoulli, Mathematical Statistics and Probability, 1967: 281—297.

[5] Zhang Tian, Raghu Ramakrishnan, Miron Livny. BIRCH: an efficient data clustering method for very large databases[C]// 1996 ACM—SIGMOD Int. Conf. Management of Data (SIGMOD'96). Canada, Montreal, 1996: 103—114.

[6] Ester M, Kriegel H P, Sander J, et al. A density—based algorithm for discovering clusters in large spatial databases with noises[C]// Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226—231.

[7] KDD99. Kdd99 cup dataset, 1999[EB/OL]. [2006—12—20]. http://www.ll.mit.edu/IST/ideval/data/dataindex.html.

[8] 马晓春, 高翔, 高德远. 聚类分析在入侵检测系统中的应用研究[J]. 微电子学与计算机, 2005, 22(4): 134—136.

作者简介:

张西广　男, (1977—), 讲师, 博士研究生. 研究方向为数据网格与网格中间件、计算机网络等.

郑秋生　男, (1965—), 硕士, 副教授. 研究方向为网络安全、软件复用技术.

王虎祥　男, (1974—), 硕士. 研究方向为计算机科学与技术领域.

陈国强　男, (1977—), 硕士. 研究方向为软件工程、网络安全.

(上接第 61 页)

4　结束语

文中描述了在 DSP 芯片 DM642 上实现新一代视频编码标准 H.264 算法的方法, 结合硬件芯片的特点和专有指令对算法进行优化, 实现了高性能的实时编码要求, 并提出一种基于 USB 2.0 的具有可扩展性强、传输速率高、支持"热插拔"的采集和存储系统的设计方案. 这一研究具有很大的实用价值.

参考文献:

[1] 钟玉琢, 向哲, 沈洪. 流媒体和视频服务器[M]. 北京: 清华大学出版社, 2003.

[2] 喻莉, 汪恒晶, 戴声奎, 等. 粒度区分和可调的无线流媒体可靠传输策略[J]. 微电子学与计算机, 2005, 22(8): 121—123.

[3] 张晓燕, 谢瑶堂. 最新视频编码标准 H.264 及其核心技术[J]. 现代电视技术, 2004(11): 62—65.

[4] 齐兵, 王群生, 杨春玲. H.264 解码芯片的比较与研究[J]. 电视技术, 2006(9): 34—36.

[5] Huang Y W, Hsieh B Y, Chen T C, et al. Analysis, fast algorithm, and VLSI architecture design for H.264/AVC intra frame coder[J]. IEEE Transactions on Circuits and System for Video Technology, 2005, 15(3): 378—401.

[6] 刘凌志, 路奇, 戎蒙恬, 等. 一种并行结构的 H.264 帧内预测器[J]. 上海交通大学学报, 2006, 40(1): 54—58.

[7] 杨晨, 李树国. 一种高并行度的 H.264 帧内预测器的 VLSI 设计[J]. 微电子学与计算机, 2006, 23(12): 111—117.

[8] 王争, 刘佩林. AVS 帧内预测算法及其解码器的硬件实现[J]. 计算机工程与应用, 2006, 42(19): 80—83.

[9] 陈其松. 基于 DSP 和 USB 的信号处理系统[J]. 化工自动化及仪表, 2005, 32(5): 37—39.

作者简介:

陈其松　男, (1974—), 博士研究生, 副教授. 研究方向为信号处理、自动控制.

陈孝威　男, (1945—), 教授, 博士生导师. 研究方向为多媒体技术、网络与通讯.

王国美　女, (1975—), 博士研究生.