基于无监督聚类的入侵检测方法

罗 敏,王丽娜,张焕国
(武汉大学计算机学院,软件工程国家重点实验室,湖北武汉 430072)

摘 要: 研究了基于无监督聚类的入侵检测算法. 算法的基本思想是首先通过比较无类标训练集样本间的距离来生成聚类,并根据正常类比例 N 来确定异常数据类别,然后再用于真实数据的检测. 该方法的优点在于不需要用人工的或其他的方法来对训练集进行分类. 实验采用了 KDD99 的测试数据,结果表明,该方法能够比较有效的检测真实网络数据中的未知入侵行为.

关键词: 入侵检测; 数据挖掘; 无监督聚类; 无类标数据
中图分类号: TP393. 8 **文献标识码**: A **文章编号**: 0372-2112 (2003) 11-1713-04

An Unsupervised Clustering-Based Intrusion Detection Method

LUO Min, WANG Li-na, ZHANG Huan-guo
(School of Computer Science and Technology, State Key Laboratory of Software Engineering, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: An unsupervised clustering-based intrusion detection algorithm is discussed. The basic idea of the algorithm is to produce the cluster by comparing the distances of unlabeled training data sets. With the classified data instances, anomaly data clusters can be easily identified by normal cluster ratio. And then the identified cluster can be used in real data detection. The benefit of the algorithm is that it needn't labeled training data sets. Using the data sets of KDD99, the experiment result shows that this approach can detect unknown intrusions efficiently in the real network connections.

Key words: intrusion detection; data mining; unsupervised clustering; unlabeled data

1 引言

随着网络技术和网络规模的不断发展,网络入侵的风险性和机会也越来越多,网络安全已经成为一个全球性的重要问题. 在网络安全问题日益突出的今天,如何迅速、有效地发现各类新的入侵行为,对于保证系统和网络资源的安全显得十分重要.

入侵检测技术主要分为两类,即:异常检测(abnormal detection)和误用检测(misuse detection). 异常检测是指利用定量的方式来描述可接受的行为特征,以区分和正常行为相违背的、非正常的行为特征来检测入侵. 误用检测是指利用已知系统和应用软件的弱点攻击模式来检测入侵[1].

对于异常检测,人们主要依赖于他们的直觉和经验选择统计特性以构造入侵检测系统. 而对于误用检测,则一般首先是由网络安全专家们对攻击模式和系统弱点来进行分析和分类,然后再手工的建立相应的检测规则和模式来构造入侵检测系统. 这样就造成了现有的许多入侵检测系统只能够对某一些特定的或已知的入侵行为取得比较好的结果. 而所有这些方法采用的模型的建立完全依赖于对训练数据集中数据样本的学习,所以保证该数据集的洁净性,对建立一个实用的入侵检测系统是至关重要的[2~6]. 而实际上,要为系统的学习收集一个洁净数据集往往是不太容易的,并且代价也很高. 因此研究无监督的入侵检测方法是非常必要的.

本文对采用无监督聚类方法的入侵检测算法作了一些研究,试验结果表明该方法在入侵检测中是可行的. 算法的优点是不需要用人工的或其他的方法来对训练集进行分类,并且能够比较有效地检测未知入侵攻击.

本文的组织如下:第二节介绍基于无监督聚类的入侵检测方法,第三节介绍实验采用的方法和结果,第四节总结实验结果并讨论算法的缺陷.

2 基于无监督聚类的入侵检测算法

2.1 聚类问题的产生和定义

聚类问题起源于许多学科并具有广泛的应用,例如数据压缩、信息检索、模式识别、公共设施选址和数据挖掘等. 众多的应用领域导致了许多不同的聚类问题的出现,不过,所有的聚类问题都需要在对给定的数据集进行划分的同时优化一个特定的目标函数[7].

令 S 为由 d 维度量空间的点代表的 n 个数据对象的集合. 将 S 分成 k 个子集 $C_1, C_2, \cdots, C_k$ 的一个划分称为 k-聚类

收稿日期:2002-11-05;修回日期:2003-06-10
基金项目:国家自然科学基金重点课题(No. 90104005, No. 90204011)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(k-clustering),其中每个 C_i 称为一个簇(cluster).

两个数据对象之间的距离通过一定的度量方法来确定.度量函数的选取与具体的应用息息相关,最广泛使用的是欧几里得距离(Euclidean distance).它的定义为:

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$$

其中 $i=(x_{i1},x_{i2},\cdots,x_{ip})$ 和 $j=(x_{j1},x_{j2},\cdots,x_{jp})$ 是两个 p 维的数据对象.

上述的距离度量满足对距离函数的一些数学要求:

(1) $d(i,j)\geq 0$:距离值非负;

(2) $d(i,i)=0$:一个对象与自身的距离为 0;

(3) $d(i,j)=d(j,i)$:距离函数具有对称性;

(4) $d(i,j)\leq d(i,h)+d(h,j)$:从某一对象 i 到对象 j 的直接距离不会大于途经任何其他对象 h 总距离(三角不等式).

2.2 基于无监督聚类的入侵检测算法

基于无监督聚类的入侵检测算法建立在两个假设上[8].第一个假设是正常行为的数目远远大于入侵行为的数目.第二个假设是入侵的行为和正常的行为差异非常大.该方法的基本思想就是由于入侵行为是和正常行为不同的并且数目相对很少,因此它们在能够检测到的数据中呈现出比较特殊的特性.

基于无监督算法的入侵检测算法主要包括三部分,一是数据预处理部分,二是无监督聚类算法,三是检测算法.

在给定聚类半径 L 后,输入训练集合 $Z=\{x_1,x_2,\cdots,x_n\}$, $x_i\in R^n$,无监督聚类算法则可以实现对集合 Z 的无监督聚类.无监督聚类算法的描述如下:

给定一个固定常量 L.定义 $dist(c,x_i)$ 为欧几里得距离.聚类的中心点 O_i 由 x_i 的特征向量表示.

step 1 $C_1\leftarrow\{x_1\}$, $O_1\leftarrow x_1(feature)$, $num\_cluster\leftarrow 1$, $Z\leftarrow\{x_1,x_2,\cdots,x_n\}$;

step 2 若 $Z=\varnothing$,则 stop;

step 3 repeat;

step 4 选择 $x_i\in Z$, $i=2..n$.从已有中心找到一个离 x_i 最近的中心 O_j,即从 C_j 中找到一个聚类中心 O_j,使得对于所有的 $O_m\in C$,都有 $dist(O_j,x_i)\leq dist(O_m,x_i)$, $m=1..num\_cluster$;

step 5 若 $dist(c,x_i)\leq L$,则将 x_i 加入到聚类 C_j 中,即 $C_j\leftarrow C_j\cup\{x_i\}$,并调整 C_j 类的中心,即计算 C_j 类中所有数据的均值特征向量,并以该结果作为 C_j 类的中心;goto step8

step 6 否则生成一个新类, $num\_cluster\leftarrow num\_cluster+1$, $C_{num\_cluster}\leftarrow\{x_i\}$, $O_{num\_cluster}\leftarrow x_i(feature)$;

step 7 $Z=Z-\{x_i\}$;

step 8 until $Z=\varnothing$

其中 $num\_cluster$ 为当前产生的聚类数目, n 为整个训练集的数目, $C\cdots C_{num\_cluster}$ 为生成的聚类, O_j 为 C_j 的中心.显然整个算法在每次循环时需要遍历整个聚类,所以整个算法所花费的时间应该小于 $num\_cluster\times n$(在算法结束时 $num\_cluster$ 即为最终所生成的聚类数),因此该算法具有比较高的效率.

在聚类生成之后需要对其进行标类.由算法的第一个假设和第二个假设我们可以推测在最后生成的聚类中,如果某个聚类为正常数据所聚集成的,那么它包含的数据数目应该也远远大于那些由入侵数据聚集成的聚类所包含的数据数目.因此可以简单地将所有的类按其中包含的数据量大小排序,并设定一个比例数 N,那些位于 N 以上的包含最多数据量的类被判断为正常类,而其余的类则被认为异常.标类算法的描述如下:

假设 C_i, $i=1\cdots num\_cluster$ 为已经生成的聚类, N 为 0～1 之间的一个常数.

step 1 Sort(C_i),按照 C_i 中包含数据多少从大到小对 C_i 排序

step 2 $j=1$; $k=N*num\_cluster$;

step 3 repeat;

step 4 若 $j<k$,则将 C_i 标为正常类;

step 5 否则将 C_i 标为异常入侵类;

step 6 j++;

step 7 until j > num_cluster

这种方法[8]非常简单,且易于理解并实现,但其有效性与正常行为的子类数目有密切的关系.如果正常的行为类被划分过细,每个子类都在特征空间中有其独特的类中心,则必将导致单个子类中的数据量相对减少,甚至少到小于某些异常类包含的数据量.在这种情况下,就会错误地将正常的数据类划分为异常,或者是将异常的类划分为正常.为防止该问题的出现,应在生成训练数据集时尽量增大各类正常数据的容量,使得在任何子类中,都能包含足够多的数据量以从异常类中区分出来.

检测算法如下所述:

假设 x 为要检测的一个网络数据包.

step 1 利用在预处理算法中得到的统计数据将 x 标准化,即 $x\rightarrow x'$;

step 2 j = 1;

step 3 repeat;

step 4 计算 C_j 的中心 O_j 与 x 的距离,即 $dist(O_j,x')$;

step 5 j++;

step 6 until j > num_cluster;

step 7 找到最小的 $dist(O_{min},x)$,并得到 O_{min} 所属类的类标 label;

step 8 若 label 是正常,则 x 是正常数据包,否则是入侵数据包.

检测算法非常简单快速,因此效率很高.

3 实验

3.1 样本集描述

选用的样本数据是目前入侵检测领域比较权威的测试数据,KDD CUP1999[9]数据,来源于 1998DARPA 入侵检测评估程序.该数据一共提供了 4,900,000 条数据,对于提供的每一个 TCP/IP 连接,除了一些基本属性(例如协议类型、传送的字节数等)外,还利用领域知识扩展了一些属性(例如登录失败的次数、文件生成操作的数目等),某些属性是在计算过去 2 秒

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

钟之内信息的基础上得到的,例如在过去 2 秒钟连接到同一个服务的连接数目.每个连接共有 41 种定性和定量的特征,其中有 8 个属性是离散型的变量,其余是连续型的数字变量.

入侵数据有 4 大类,24 小类.四大类是:DOS(Denial of Service)攻击,例如 ping of death;U2R(User to Root)攻击,例如 eject;R2U(Remote to User)攻击,例如 guest;PROBING 攻击,例如 port scanning.

为了满足检测算法的两个假设的需要,需要对测试集作一些过滤.所以我们从整个测试集中提取了 60638 条记录作为训练样本集.其中 60032 条记录为正常数据包,剩余 606 条记录为入侵数据包,所有正常数据的比例达到了 99 %,符合检测算法第一个假设的要求.在整个训练集中的入侵类型及数目见表 1:

表 1　训练集中入侵类型及数目

大类	小类及数目
DOS	共 284 条,其中 neptune (141),smurf (143).
U2R	共 68 条,其中 buffer_overflow (22),loadmodule (2),perl (2),ps(16),rootkit(13),xterm(13).
R2U	共 131 条,其中 ftp_write(3),guess_passwd(31),imap(1),multihop(18),named(17),sendmail(17),phf(2),warezmaster (29),xlock. (9),xsnoop(4).
PROBING	共 123 条,其中 ipsweep(30),nmap(19),portsweep(32),satan(42).

在测试样本集的选取中,我们一共选取了 4 组数据,每组各 3 万条记录.其中第一组和第二组测试集选自训练集,另外两组则选自 KDDCUP99 数据中除去训练集的部分(选择时特别选取了一些没有包含在训练集中的入侵,即未知的入侵).

3.2　预处理

因为整个原始的测试数据中包括离散的和连续的属性特征变量,所以需要对它们进行分别的处理.

对离散型的属性特征变量来说,我们举例说明处理方法:

假设在数据集中 service 属性中重复出现 http,ftp,telnet,smtp 等 4 项属性,形如:http,http,telnet,ftp,smtp,ftp,telnet,smtp,…,然后我们将这 4 项属性编码为 0001,0010,0100,1000,再将 service 属性分割为 4 个属性 service1,service2,service3,service4,凡是出现 http 的记录令 service1 = 0,service2 = 0,service3 = 0,service4 = 1;凡是出现 ftp 的记录令 service1 = 0,service2 = 0,service3 = 1,service4 = 0;依此类推可将离散型的 service 属性变量转换为连续性属性变量.

按照上述方法可同样处理其他离散型属性变量,这样处理的好处是可以确保每条记录之间的同一离散型属性之间距离相等,在计算中彼此不会产生偏差.

对于连续型的属性特征变量来说,不同的属性特征有不同的度量标准,因此如果对原始数据不进行预处理的话就有可能产生大数吃小数的问题.例如给定两个特征向量 $x_i = \{1000,1,2\}$, $x_j = \{2000,2,1\}$,则:

$$d(x_i,x_j) = \sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+|x_{i3}-x_{j3}|^2} = \sqrt{|1000-2000|^2+|1-2|^2+|2-1|^2}.$$

显然整个特征完全被第一项属性特征数据掩盖.

为了解决这个问题,我们必须将数据的特征属性值进行标准化,可以进行如下的变换:

(1) 计算平均的绝对偏差(mean absolute deviation) S_f

$$S_f = \frac{1}{N}\sum_{i=1}^{n}(X_{if} - m_f) \tag{1}$$

其中 x_{1f}, …, x_{nf} 是 f 的 n 个属性特征值, m_f 是 f 的平均值,即

$$m_f = \frac{1}{N}\sum_{i=1}^{n}X_{if} \tag{2}$$

(2) 计算标准化的特征属性值

$$Z_{if} = \frac{x_{if} - m_f}{s_f} \tag{3}$$

这个平均的绝对偏差 S_f 比标准差 σ_f 对于孤立点具有更好的鲁棒性[10].

对于每个数据包都可以按照上面的三个公式来对它的特征属性进行计算并得到新的数据.这相当于利用统计特性将原始实例的特征属性映射到一个标准的属性空间上,有利于减少上面所述的问题.具体算法如下:

假设 x_i 为训练集中一个原始网络数据包, n 为整个训练集的数目.

step 1　i = 1;

step 2　repeat

step 3　选择 $x_i \in Z$.根据式(1),(2)计算 S_f 和 m_f;

step 4　i + + ;

step 5　until i = n

step 6　i = 1;

step 7　repeat;

step 8　根据式(3)计算 Z_{if},将 x_i 的特征属性值转换成标准化的值,记为 $x_i \rightarrow x_i'$;

step 9　i + + ;

step 10　until i = n.

3.3　实验结果

在实验时,有两个要用到的参数值,一是聚类半径 L,在聚类过程中决定在什么样的距离下,两个连接数据必须被分配到同一个类,即系统类型数据在特征空间中构成的类的平均半径;二是正常类比 N,在区分正常与异常类时决定正常类在所有类中所占的比例.实验中检测率定义为:算法正确检测到的入侵个数除以测试集的所有入侵数据个数的百分比;误警率定义为:算法将正常数据错误检测为异常数据的个数除以测试集的所有正常数据个数的百分比.

表 2 显示了几种 L 和 N 不同取值时的详细实验结果.

由于聚类半径 L 和正常类比 N 的选取没有一个比较好的方法,因此只能采取试探法进行实验.实验首先从固定 L = 20,改变 N 值做起.从表 2 可见,当聚类半径 L 固定时,各组的检测率和误警率基本是随着正常类比 N 的增加而降低的,这与我们估计的情况是一致的.因为将生成的聚类按照其中包含数据量大小进行排序之后,大体上是正常数据在前,异常数据集中在后.这样, N 取值越小,也即判断为异常的数据越多,检测率显然要增加.在理想的情况下,聚类结果形成的每个类都将只包括同种类型的数据,要么是正常数据,要么是入

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

侵数据,但实际情况下,这是不可能的,各个类都不可避免地会含有被错误地分配来的数据. 对入侵检测环境而言,也就是在正常类中含有异常数据,异常类中也含有正常数据. 因此,随着 N 的变小,检测率的增加,误警率也就自然增加了. 接着我们选取 N 固定为 20 %,改变 L 值实验. 由表 2 可见,当 L = 40, N = 20 %时算法的整体性能较好. 因此我们采用这组取值进行下一步的实验.

表 2 L 和 N 取不同值时各组的实验结果

聚类半径 L	正常类比 N(%)	第一组		第二组		第三组		第四组	
		检测率%	误警率%	检测率%	误警率%	检测率%	误警率%	检测率%	误警率%
20	10	71.7	9.1	89.2	9.8	63.3	11.3	67.2	10.8
20	15	65.6	5.2	72.5	6.3	56.1	8.5	60.2	7.9
20	20	57.3	1.1	62.1	1.2	48.8	3.2	51.6	2.7
20	25	46.1	0.6	54.2	0.9	39.9	2.1	28.3	1.8
10	20	50.9	1.7	60.7	1.5	46.2	9.7	36.8	7.1
20	20	57.3	1.1	62.1	1.2	48.8	3.2	51.6	2.7
30	20	64.8	1.6	53.8	1.1	45.3	2.9	51.3	2.2
40	20	62.7	0.7	49.3	0.8	51.2	1.5	53.1	0.9
50	20	50.6	1.1	46.2	0.9	40.8	1.6	50.6	1.4

表 3 显示了当 L = 40, N = 20 %时算法在第三组和第四组测试集上对已知和未知攻击入侵的检测情况. 已知入侵指的是在测试集和训练集中都包含的入侵(如表 1 中列出的入侵类型),未知入侵指的是测试集包含但训练集中没有的入侵(如 DOS 入侵中的 udpstorm 攻击, R2U 入侵中的 spy 攻击等).

表 3 L = 40, N = 20 %时,算法对已知和未知入侵的检测率情况

类别	第三组		第四组	
	已知入侵(%)	未知入侵(%)	已知入侵(%)	未知入侵(%)
DOS	59.5	26.1	48.6	18.9
U2R	69.1	62.9	71.7	58.4
R2U	40.2	10.3	61.3	9.6
PROBING	78.6	71.2	75.4	77.3
合计	61.9	45.9	64.3	41.1

从表 3 可见,算法对 U2R 和 PROBING 入侵攻击的检测都取得了比较高的检测率,但是对 DOS 和 R2U 入侵的检测效果不理想. 这与我们的估计也基本一致. 因为有很多 R2U 入侵是伪装合法用户身份进行攻击,这就使得其特征与正常数据包比较类似,造成了算法检测的困难. 另外由于在训练集中包含 DOS 攻击数目较多,在聚类标类时就有可能将入侵类标为正常类,造成检测率的降低. 算法在第三组和第四组测试集中对未知入侵的检测率都超过了 40 %,说明其能够比较有效的检测未知入侵行为.

4 结论及展望

实验结果表明,尽管无监督聚类入侵检测算法在检测性能上与传统的检测方法有所差距,但是算法不需要对训练集进行标类和严格的过滤,而且在对未知入侵的检测上也有比较好的效果. 因此算法在入侵检测领域应该有广泛的应用前景. 另外由于算法比较简单快速,也可用于其他无监督入侵检测算法的预处理工作中以提高检测效率.

但是算法对如何选取两个参数(聚类半径 L 和正常类比 N)没有合适的方法,只能采用试探的方法,还需要人工的干预. 下一步的工作就是继续研究如何确定两个参数的优化问题,并且希望能脱离人工的干预,和通过演化计算的方法,算法能够自己寻找合适的参数.

参考文献:

[1] 蒋建春,马恒太,任党恩,卿斯汉. 网络安全入侵检测:研究综述[J]. 软件学报,2000,11(11):1460 - 1466.

[2] Eskin E, Arnold A, Prerau M, et al. A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data [A]. Applications of Data Mining in Computer Security [C]. Kluwer Academic Publisher, Boston, 2002. 77 - 102.

[3] Honig A, Howard A, Eskin E, et al. Adaptive model generation: an architecture for the deployment of data minig-based intrusion detection systems [A]. Applications of Data Mining in Computer Security [C]. Kluwer Academic Publisher, Boston, 2002. 153 - 194.

[4] Schultz M, Eskin E, Zadok E, et al. Data mining methods for detection of new malicious executables [A]. In Proceedings of IEEE Symposium on Security and Privacy (IEEE S&P - 2001) [C]. Oakland, CA, May 2001. 38 - 49.

[5] Eskin E. Anomaly detection over noisy data using learned probability distributions [A]. In Proceeding International Conference on Machine Learning[C]. Morgan Kaufmann Press, Palo Alto, CA, 2000. 255 - 262.

[6] 刘海峰,卿斯汉,等. 一种基于审计的入侵检测模型及其实现机制[J]. 电子学报,2002,30(8):1167 - 1171.

[7] Cecilia M Procopiuc. Clustering problems and their applications (a survey) [DB/OL]. Department of Computer Science, Duke University. 1997. http://www.cs.duke.edu/~magda.

[8] Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering [A]. In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001) [C]. Philadelphia, PA, 2001.

[9] KDD99. KDD99 cup dataset [DB/OL]. http://kdd.ics.uci.edu/databases/kddcup99, 1999.

[10] Han Jiawei, Micheline Kambe. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.

作者简介:

罗 敏 男,1974 年出生于湖北蕲春,现为武汉大学计算机学院博士生,主要研究领域为信息系统安全理论与技术.

王丽娜 女,1964 年生于辽宁营口,博士,现为武汉大学计算机学院副教授,主要研究领域为密码学与信息安全理论与技术.

张焕国 男,1945 年生于河北,现为武汉大学计算机学院教授,博士生导师,主要研究领域为密码学与信息安全理论与技术.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net