带野值的单类分类器在安全审计中的应用

李佳桢,潘志松,倪桂强,王 琼
LI Jia-zhen,PAN Zhi-song,NI Gui-qiang,WANG Qiong

解放军理工大学 指挥自动化学院,南京 210007
Institute of Command Automation,PLA University of Science and Technology,Nanjing 210007,China

LI Jia-zhen,PAN Zhi-song,NI Gui-qiang,et al.Application of one-class classifier with negatives in security audit data analysis.Computer Engineering and Applications,2008,44(21):154-156.

Abstract: One-class classifier is currently a hot spot of pattern recognition field.One-class classifier with negatives is based on one-classifier,by leading into a few costful abnormal samples to reinforce the classification.This model applies to the problems handling the two kind data categories imbalances where positives more over than negatives.It is proposed in this paper that using support vector data description with negatives in security audit data analysis system.Through some experiments,it is proved to be more sensitive with exceptional samples,so it will be more valuable in practice.

Key words: one-class classifier;support vector data description;security audit

摘 要: 单类分类器是当前模式识别领域的一个研究热点。带野值的单类分类器是在单类分类器的基础上,通过引入少量珍贵的异常样本(野值),以加强分类器的性能。该模型适用于处理正类样本数目远多于反类样本的两类数据类别不平衡问题。提出了将带野值的支持向量描述方法应用于安全审计数据分析中,并通过实验证实了该方法对异常样本更为敏感,具有良好的应用潜力。

关键词: 单类分类器;支持向量数据描述;安全审计
DOI: 10.3778/j.issn.1002-8331.2008.21.043 **文章编号:** 1002-8331(2008)21-0154-03 **文献标识码:** A **中图分类号:** TP393

一般来说可以认为安全审计数据分析是一个标准的单类问题。首先,审计数据中大部分数据流都是正常的,正常应用行为的流量比异常应用行为的流量高出了若干个数量级; 其次,异常应用行为"可遇不可求",或者很难"求"。由于攻击行为具有不确定性,即使知道了攻击方法也很难模仿;并且很多特定网络受攻击的概率很小,可是一旦被攻破,损失将无法估量(如军用卫星网)。

单类分类器利用正常审计数据建立一个正常模式,然后对当前的系统或用户的行为进行比较,判断出与正常模式的偏离程度。用是否超过预先设定的边界,来判别当前入侵事件的归属。罗隽[1]提出了基于支持向量数据描述(SVDD)的安全审计数据分析模型,并使用国际标准数据集 MIT Lpr,通过大量的实验来测试基于 SVDD 的安全审计数据分析模型的性能,证明了其具有较好的效果。

但是在实际应用中,我们往往能够获取少量攻击样本(从已有的攻击手段中获取)。这就成为了模式分类中的数据类别不平衡问题[2,3],即样本实例中一些类的实例很多,而另一些类实例很少。事实上在这类数据类别不平衡问题中,把任何一个新样本预测为大类就已经可以获得很高的预测精度[4],而实际应用中更需要检测出那些少数(异常)类。所以一旦获得了可用的异常样本(野值),我们就希望能够充分利用,以加强分类器对少数类的检测效果。本文所使用的 SVDD-negative 模型,就是在 SVDD 单类分类器的基础上,引入对野值的利用,以得到更好的对异常的检测性能。

1 安全审计数据描述和预处理

审计跟踪作为一种安全机制,可以审查基于每个目标或每个用户的访问模式,并使用系统的保护机制。安全审计信息包含了很多方面的内容,文中主要关注的是网络安全审计中的系统调用执行迹。1996 年,墨西哥大学的 Forrest 等人提出了在异常检测中通过分析程序执行过程产生的系统调用序列(称为系统调用执行迹,Sequences System Calls,SSC)来构建特征轮廓的方法,并用实验证明了程序执行轨迹的局部模式(系统调用的短序列)可以完全刻画程序行为的特征[5]。

系统调用序列的数据采集方法比较简单,通过对审计系统进行配置, 就可使其根据用户的要求监控相关程序的执行过程。下面以安全审计的国际标准数据集 MIT lpr(http://www.cs.unm.edu/immsec/data/,新墨西哥大学提供的真实环境下采集的实验数据集,基于SunOS 4.1.4)为例,描述数据采集和与预处理方法。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60603029);江苏省自然科学基金(the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2005009)。

作者简介: 李佳桢(1981-),男,研究方向为网络安全、模式识别;潘志松(1973-),男,研究方向为机器学习、模式识别;倪桂强(1966-),男,研究方向为网络管理、网络安全、模式识别;王琼(1979-),女,研究方向为机器学习、模式识别。

收稿日期: 2008-04-30 **修回日期:** 2008-06-11

该数据集中的每个执行迹数据文件由两列数据构成,第1列为进程标识符;第2列为进程的系统调用命令在系统调用名称列表(mapping file)中的索引值,如'5'代表'Open',因此该数据已是数字序列。文件中,具有同一进程标识符的系统调用构成一个进程的执行迹,对应进程从开始执行到执行结束所使用系统调用序列。这些执行迹中既有在正常使用系统情况下收集的系统关键程序的执行迹,也有不同类型的入侵行为的执行迹(缓冲区溢出、特洛伊木马程序以及拒绝服务攻击等)。执行迹的长度各不相同。

预处理的主要目的就是要得到执行迹的系统调用短序列。由于执行迹中系统调用的次序关系是描述该程序行为的重要特征,分析这种次序关系的最简单方法就是利用长度为 K 的滑动窗口(Sliding Window)技术构造系统调用短序列。利用长度为 K 的窗口在程序执行迹上以步长为1从左到右滑动,以获得多个长度为 K 的短序列,作为输入数据。

系统调用短序列反映了进程执行过程中系统调用之间的次序关系,如果选取的短序列长度较短,就容易丢失系统调用的次序信息;长度太大,就丢掉了系统执行的局部信息状况。W.Lee 教授从信息论的角度研究了数据长度的选择情况[6],认为最合适的系统调用短序列长度为6~7。

2　基于 SVDD-negative 的安全审计数据分析模型

David Tax[7]建立了支持向量数据描述(SVDD)利用高斯核函数把样本空间映射到核空间,在核空间找到一个能够包含所有训练数据的一个球体。当判别时,如果测试样本位于这个高维球体中,那么就认为正常,否则就认为异常。在球面上的样本点即为 SVDD 所求得的支持向量。由于支持向量的个数是稀疏的,因此计算量得到相应的减少。

2.1　SVDD-negative 模型

David Tax[8]进一步提出了带野值的 SVDD 模型——SVDD-negative。该模型是对 SVDD 的改进,加入了对可用的异常样本(negative examples)即"野值"的利用,来加强 SVDD 的数据描述能力。其训练数据包括两个部分:正常样本(target)和少数异常样本(negative)。同不带野值的 SVDD 相比,SVDD-negative 还需要根据异常样本来调整球体的圆心和球半径。

假设正常样本的下标用 i,j 表示,异常样本的下标用 l,m 来表示;为正常样本贴标签 $y_i=1$,异常样本贴标签为 $y_l=1$。为所有训练样本都引入松弛变量 ξ,最小化问题为:

$$\varepsilon_{struct}(R,a,\xi)=R^2+C_1\sum_i\xi_i+C_2\sum_l\xi_l \tag{1}$$

约束条件为:

$$\|x_i-a\|^2\leqslant R^2+\xi_i,\ \|x_l-a\|^2\geqslant R^2+\xi_l,\xi_i\geqslant 0,\xi_l\geqslant 0,\forall i,l \tag{2}$$

结合约束(2)和等式(1),引入拉格朗日乘子 $\alpha_i,\alpha_l,\gamma_i,\gamma_l$,得到拉格朗日函数为:

$$L(R,a,\xi,\alpha,\gamma)=R^2+C_1\sum_i\xi_i+C_2\sum_l\xi_l-\sum_i\gamma\xi_i-\sum_l\gamma\xi_l-$$

$$\sum_i\alpha_i\lfloor R^2+\xi_i-(x_i-a)^2\rfloor-\sum_l\alpha_l\lfloor R^2+\xi_l-(x_l-a)^2\rfloor \tag{3}$$

对拉格朗日乘子有 $\alpha_i\geqslant 0,\alpha_l\geqslant 0,\gamma_i\geqslant 0,\gamma_l\geqslant 0$。

将拉格朗日函数分别对 R,a,ξ_i,ξ_l 求导并使求导后的函数为0,得到满足最优解的 KKT 条件为:

$$\sum_i\alpha_i-\sum_l\alpha_l=1 \tag{4}$$

$$a=\sum_i a_ix_i-\sum_l a_lx_l \tag{5}$$

$$0\leqslant\alpha_i\leqslant C_1,0\leqslant\alpha_l\leqslant C_2,\forall i,l \tag{6}$$

将约束式(4)(5)(6)代入式(3)可得

$$L=\sum\alpha_i(x_i\cdot x_i)-\sum\alpha_l(x_l\cdot x_l)-\sum_{i,j}\alpha_i\alpha_jK(x_i\cdot x_j)+$$

$$2\sum_{l,j}\alpha_l\alpha_jK(x_l\cdot x_j)-\sum_{l,m}\alpha_l\alpha_mK(x_l\cdot x_m) \tag{7}$$

实际应用中仍然需要将样本空间映射到核空间,引入核函数得:

$$L=\sum\alpha_iK(x_i,x_i)-\sum\alpha_lK(x_l,x_l)-\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)+$$

$$2\sum_{l,j}\alpha_l\alpha_jK(x_l,x_j)-\sum_{l,m}\alpha_l\alpha_mK(x_l,x_m) \tag{8}$$

定义新的变量 α_i' 后,SVDD-negative 与不带野值的 SVDD 是一致的,可以用二次规划算法求得最优 α_i' 解。

2.2　实验对比

为了检验 SVDD-negative 模型在分类性能上比标准 SVDD 模型是否有提高,特别是我们期待的对异常样本检测能力的有无改进,下面通过在 Banana 数据集和 MIT lpr 数据集上进行的两个实验比较来说明。

实验1　Banana 数据集

香蕉型数据是二维空间典型的线性不可分的人工数据集。图1中的4张子图是 SVDD-negative 与 SVDD 的分类面的比较。SVDD-negative 与 SVDD 使用相同的参数,核函数为 RBF 核函数,同时调整 σ 从6,8,10到12。图中"*"号代表正常数据,"+"号代表异常样本。点线边界代表 SVDD-negative 训练形成的分类面,实线边界代表 SVDD 训练形成的分类面。

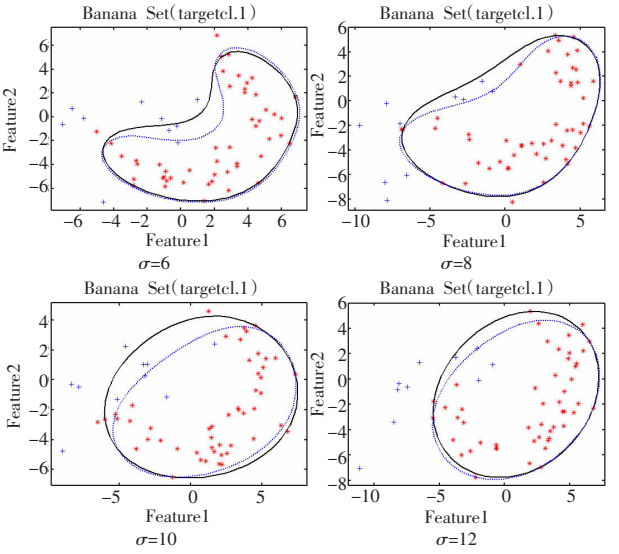

图1　SVDD-negative 与 SVDD 在 Banana 数据集上分类面的对比

分析4张子图可以看出,SVDD-negative 形成的分类面更为紧凑,对异常样本更为敏感。其中 σ=8,10和12的三张子图中,均有异常样本点在分类面上,表明该异常样本成为了支持向量。四张子图都表现出了 SVDD-negative 对异常样本的拒绝好于 SVDD。

实验2　安全审计中的系统调用执行序列数据

采用新墨西哥大学系统调用实验数据库的 MIT lpr 程序

的相关数据集进行实验来评估模型性能。该数据集包含 2 704 条正常序列和 1 001 条含有异常的序列。采用随机无放回抽样的方式选取 1 704 条正常执行序列和 11 条异常执行序列作为训练数据集,测试数据集包含与训练集独立的 1 000 条正常序列和 990 条异常序列。

选择切分长度 K=6,适用于实时检测。全部 2 704 条正常执行迹的数据量为 2 928 357 个系统调用,提取的特征序列共有 512 个;全部 1 001 条异常执行迹的数据量为 169 252 个系统调用,提取的特征序列共有 353 个。训练集中 1 704 条正常执行迹含有 1 895 341 个系统调用,生成的特征序列共 479 个;11 条异常执行迹含有 1 894 个系统调用,生成的特征序列共 37 个。可见数据预处理使数据量大为降低,同时也反映出难以获得特征序列集的完备集。

使用 RBF 核函数,σ=10。对测试数据集中的系统调用执行迹,选定一个阈值 α,当该执行迹中的短序列被判断为异常的数目超过 α,则判定该系统调用执行迹为异常。用对异常样本的检测率和对正常样本的误报率(误警率)来评判模型的检测效果,SVDD-negative 与 SVDD 检测结果对比如表 1 所示。

表 1 SVDD-negative 与 SVDD 在 MIT lpr 数据集上的检测效果对比

阈值	检测率		误报率	
	SVDD	SVDD-negative	SVDD	SVDD-negative
9	100%	100%	61.28%	62.25%
10	100%	100%	61.04%	62.03%
11	100%	100%	58.63%	58.90%
12	100%	100%	57.25%	57.55%
13	100%	100%	46.74%	48.45%
14	100%	100%	21.89%	22.11%
15	100%	100%	20.11%	20.54%
16	100%	100%	17.15%	18.41%
17	100%	100%	17.15%	17.99%
18	97.98%	100%	16.97%	17.54%
19	97.98%	100%	16.97%	17.05%
20	97.98%	100%	3.70%	3.94%
21	96.26%	100%	3.70%	3.84%
22	92.82%	100%	1.45%	1.59%
23	92.82%	100%	1.45%	1.59%
24	92.82%	100%	0.79%	0.77%
25	89.91%	100%	0.63%	0.77%
26	84.33%	100%	0.63%	0.77%
27	71.97%	100%	0.63%	0.63%
28	54.52%	100%	0.63%	0.63%
29	37.75%	100%	0.63%	0.63%
30	22.46%	100%	0.58%	0.63%
31	3.71%	99.29%	0.58%	0.63%
32	3.71%	98.99%	0.58%	0.63%
33	0.0%	98.99%	0.40%	0.42%
34	0.0%	96.26%	0.40%	0.42%
35	0.0%	96.26%	0.40%	0.42%

从表 1 中可以看出,SVDD-negative 模型的检测率始终保持着大于 SVDD 模型的检测率。当阈值大于 20 时,SVDD 的检测率开始显著下降,当阈值大于 28 时,检测率降低至 50%以下,实用性大大降低;而 SVDD-negative 一直到阈值增大到 30 仍然保持了 100%的检测率,在阈值大于 30 时也保持了较好的检测率。通过对比,带野值的 SVDD 模型 SVDD-negative 比不带野值的 SVDD 模型显著地提高了对异常样本的检测率。虽然相对来说带来了稍高一些的误报率,在相同的阈值条件下,SVDD-negative 的误报率比 SVDD 稍大,但是没有超过 2 个百分点。综合来看,SVDD-negative 对 SVDD 的改进是有效的。

实验 2 进一步验证了 SVDD-negative 对异常样本更加敏感这个特性。从安全审计的核心功能来说,SVDD-negative 对 SVDD 的改进是很有价值的。因为对于安全审计系统来说,检测到入侵最为关键,对入侵的漏报将导致系统面临巨大的威胁,甚至关系到系统的生死。

3 小结

带野值的 SVDD 单类分类器模型保留了纯单类分类器的优点,不需要为系统提供异常信息;通过数据的预处理,可以从较少的正常执行迹中学习正常的模式,并能取得比较理想的检测效果;检测部分只需要简单的计算,基本能够满足安全审计实时性的要求。同时,该模型又充分利用了少量珍贵的异常样本的信息,对异常样本比纯单类分类器更为敏感。基于带野值单类分类器的安全审计数据分析系统,可以在入侵检测应用中具有更好的适用性。其缺点是稍提高了系统的误警率,可能会更频繁地干扰用户的正常使用。如何进一步降低误警率是下一步的工作方向。

参考文献:

[1] 罗隽.基于单类分类器的网络应用行为入侵检测模型研究[D].南京:解放军理工大学,2007.

[2] AAAI-2000 Workshop on "Learning from Imbalanced Data Sets"[C/OL].http://www.site.uottawa.ca/~nat/Workshop2000/workshop2000.html.

[3] ICML'2003 Workshop on "Learning from Imbalanced Data Sets I-I"[C/OL].http://www.site.uottawa.ca/~nat/Workshop2003/workshop2003.html.

[4] Chawla N V,Japkowicz N,Kolcz A.Editorial:Special issue on learning from imbalanced data sets[J].ACM SIGKDD Explorations,2004,6(1):1-6.

[5] SVDDayaji A,Hofmeyr S,Forrest S.Principles of a computer immune system[C]//Proceeding of New Security Paradigms Workshop,1997:75-82.

[6] Lee W,Stolfo S J,Mok K W.A data mining framework for building intrusion detection models[C]//Proc the 1999 IEEE Symposium on Security and Privacy,Berkely,California,1999:120-132.

[7] David M J T.One-class Classification[D].1999.

[8] David M J T.Support Vector data description[J].Machine Learning,2004,54:45-66.