一种基于文法压缩的日志异常检测算法

高赟[1),2)]　周薇[1)]　韩冀中[1)]　孟丹[1)]

[1)](中国科学院信息工程研究所信息内容安全技术国家工程实验室 北京 100093)

[2)](中国科学院大学计算机与控制学院 北京 100049)

摘　要　近年来日志挖掘是一种广泛使用的检测应用状态异常的方法．现有的异常检测算法需要大量计算，或者它们的有效性依赖于测试日志满足一些预先定义的日志事件概率分布．因此，它们无法用于在线检测并且在假设不成立时会失效，为了解决这些问题，该文提出了一种新的异常检测算法 CADM.CADM 使用正常日志和待检测日志之间的相对熵作为异常程度的标识，为了计算相对熵，CADM 充分利用了相对熵和文法压缩编码大小之间的关系而不是预先定义日志事件概率分布的族，通过这种方式，CADM 避免了对日志分布的预先假设，除此之外，CADM 的计算复杂度为O（n），因此在日志较大的情况下有较好的扩展性，通过在仿真的日志和公开日志集上的评测结果可以看出，CADM 不仅可以应用在更广泛的程序日志上，也有更高的检测精度，因此更适合在线日志挖掘异常检测发的工作．

关键词　异常检测 ； 文法压缩 ； 日志挖掘 ； 相对熵

中图法分类号　TP393　　**DOI 号**　10.3724/SP.J.1016.2014.00073

An Online Log Anomaly Detection Method Based on Grammar Compression

GAO　Yun[1),2)]　ZHOU　Wei[1)]　HAN Ji-Zhong[1)]　MENG　Dan[1)]

[1)]（National Engineering Laboratory for Information Security Technologies，Institute　of　Information　Engineering，Chinese Academy of Sciences，Beijing　100093）

[2)]（School of　Computer and Control Engineering，University of　Chinese Academy of Sciences，Beijing　100049）

Abstract　Nowadays,mining program logs is a widely used technique for detecting Anomalies in program states.Basically, existing anomaly detection methods require considerable computation efforts, or their effectiveness relies on some prior assumptions of the distribution holding on test logs.Therefore,they can hardly work online and cannot be used in all scenarios.To address the aforementioned problems, this paper proposed a new anomaly detection method called CADM.CADM exploited relative entropy between test logs and normal logs to measure the anomalous level.Instead of computing the　relative entropy directly based on predefined distribution family,our method took advantage of the relationship between relative entropy and compression size by an adapted grammar-based compression method and eliminated such kind of assumptions.In addition,our method has only an O(n) computation complexity and scales well on large logs.Experiments with both synthetic logs and real world logs show that our method is more suitable for online log mining tasks since it has higher detection accuracy on broader variety of program logs.

Keywords　anomaly detection；grammar compression；log mining；relative entropy

收稿日期：2013-04-16；最终修改稿收到日期：2013-11-22.本课题得到国家自然科学基金（61070028）、国家“八六三”高技术研究发展计划项目基金（2012AA01A401）及先导专项（XDA06030200）资助. 高 赟，男，1991 年生，博士研究生，主要研究方向为异常检测、大规模数据处理，E-mail：gaoyun@iie.ac.cn.周 薇（通信作者），女，1987 年生，硕士，研究实习生，主要研究方向为大规模数据处理和图计算.E-mail:zhouwei@iie.ac.cn.韩冀中，男，1972 年生，博士，正研级高工，主要研究领域为云平台和大规模数据处理. 孟 丹，男，1965 年生，博士，研究员，主要研究领域为大规模数据处理.

1 引 言

随着互联网的不断发展，各类网络应用在生产生活中发挥着越来越重要的作用，但与此同时，针对网络应用的攻击不断涌现，如各类病毒、蠕虫等①. 另一方面，随着网络应用规模的不断增大以及为节省成本使用廉价机器部署服务的趋势的发展，程序的执行路径难以预测，并且硬件环境不再可靠，因此难以在部署前发现所有可能的错误. 一旦针对网络应用的攻击得手或者网络应用自身出现异常，将给应用的所有者及用户带来不可估量的损失②.

攻击和错误发现的越早，所能采用的补救措施就越多，造成的损失就会越少. 因此，在线异常检测受到了学术界和工业界的广泛重视. 常用的异常检测方法是分析应用的日志. 在应用运行的过程中，应用本身和监控程序都会产生各类日志来记录应用的状态、重要的运行事件和网络流量，因此日志包含应用运行的动态信息，适合用于异常检测[1].

传统日志分析通过人工检测或者使用事先定义好的异常规则来实现. 当日志大小有限以及异常类型可知时，这些方法十分有效并且也比较灵活，但是对于当前程序产生的百万行规模的日志，人工检测很难实现[2]. 提前获取所以可能的异常也很困难[3]. 并且难以实时在线感知异常的发生.

因此，业界使用异常检测来实现程序的自动化分析. 通过建模正常日志并对待测试日志与正常模式的偏离进行检查，异常检测算法可以自动地发现可能的问题，包括未知的异常，异常检测算法通常使用一类针对离散序列的异常检测方法来分析日志，这类算法将日志所蕴含的事件序列分为段，并赋予每一段一个异常得分. 之前的序列异常检测算法大都基于统计模型或者马尔可夫模型[4]. .但是，他们都依赖于特定的假设：

（1）正常的日志事件满足一个特定的分布.

（2）各个日志事件的出现相互独立.

（3）日志事件的出现满足马尔可夫性质，即一条日志事件只与它之前有限个日志事件相关.

但是这些假设并不是总是成立.如果日志事件之间有较负责和隐式的关系[3]. 这些方法就会忽略掉这些关联或者无法正确地描述这些关联[5]. 因此. 当预先的假设不成立时，基于统计模型和马尔可夫模型的方法就会有较高的误差.

为了解决上述问题，本文提出了一种基于压缩的异常检测方法（Compression-based Anomaly Detection Method ，CADM）,这种方法不依赖于日志事件分布的假设.CAMD 选择相对熵作为评分标准对正常日志和异常日志进行区分[6]. 相对熵是一种对概率分布相似程度的度量.相对熵越大，那么测试日志的事件分布和训练日志事件分布的区别也就越大，测试日志为异常日志的可能性越大.

但是，由于正常日志和异常日志的事件分布都是未知的，直接计算相对熵是不可行的. 因此 CADM 利用相对熵与压缩的关系来估计相对熵的大小.由于 CADM 中所需的操作仅有时间复杂度较低的压缩. 因此 CADM 可以实现实时异常检测.综上所得，本文的贡献包括：

（1）提出了一种基于压缩估计相对熵的在线异常检测方法 CADM.CADM 不需要提前假设日志项分布，因此，它可以应用到更广泛的场景.

(2)选择并改造了一种通用的压缩方法，即文法压缩来提供更好的相对熵估计.改造后的算法具有更低的计算复杂度并且更适合用于异常检测.

(3)提供了 CADM 的理论分析，证明了它的正确性、高效性以及可扩展性.

(4)设计和实现了一个原型系统并对 CADM 进行评估，实验结果表明 CADM 具有更高的适用性和检测正确率，因此更适合用于日志异常检测.

本文第 2 节讨论离散序列异常检测问题的相关研究;第 3 节描述 CADM 的背景;第 4 节详细描述 CADM 的设计与实现；第 5 节通过实验评估 CADM；第 6 节总结本文且提出未来的研究计划.

2 相关工作

异常检测在许多领域都有很重要的应用[7].其中，用于检测异常序列的方法被广泛地应用在日志挖掘中，文献[4]综述了现有的离散序列异常检测方法.

大部分异常检测算法基于统计模型或马尔科夫模型实现.其中，基于统计模型的检测方法又可分为两类：第 1 类方法假设正常日志满足含参分布 P(x)并使用参数估计的方法根据训练日志确定参数值，然后采用假设检验的方法去评估测试日志是否异常.例如,文献[8]提出了一种基于 X^2 统计量的

以上 3 种情况都能够在常量时间内处理.上述文法转换算法的结果是一种容许文法.在文法编码阶段，零阶自适用算数编码被用于对文法转换中每一步识别的最长子序列对应的符号进行编码.可以证明，顺序压缩算法对所有稳定遍历的信源都是通用的，即可以将该信源产生的序列压缩至熵率.

4 CADM 算法的设计

本节将详细讨论 CADM 算法的设计.如图一所示，CADM 算法一共包括以下 3 步：

(1)使用背景中讨论的方式将整个日志转化为一个离散事件序列.

(2)使用训练日志所转化的序列来训练 CADM 的模型.CADM 所需要使用的训练日志为正常日志.

(3)将待测试日志分段，并使用异常评分公式为每一段的异常程度进行评分.

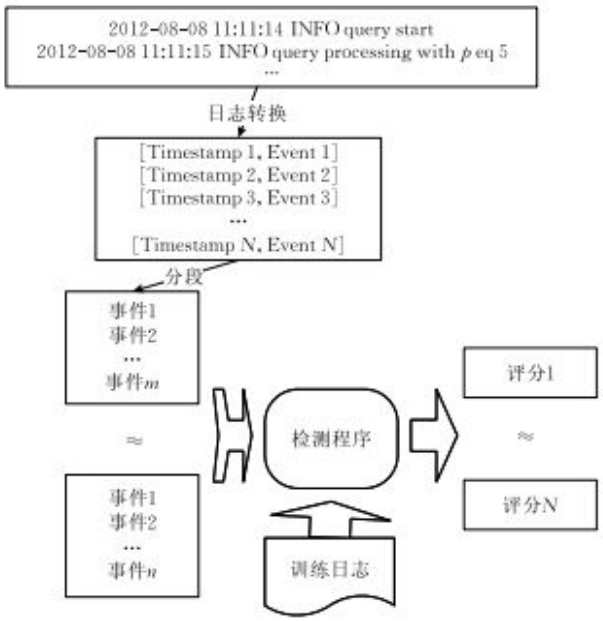

图 1 CADM 算法的主要步骤

图 1 CADM 算法的主要步骤

对于第 1 步中的日志转换，CADM 结合使用方法 2,3 和 4.大部分场景下启发式算法对异常检测任务来说已经足够精确，如果有需要，还可以通过对源码或二进制分析来精确找出所有的事件类型.最后，如果事件类型无法从消息模板中推断，那么还需要结合领域相关的知识来完成日志转换.下文将详细讨论将日志转化为事件序列后进行检测的过程,其中所用到的主要符号如表 1 所示.

表 1 CADM 算法用到的主要符号

$D(Q\|\|P)$	分布 Q 与 P 之间的相对熵
P_n或 P_n	在分布P与Q下长为 n 的事件序列的概率分布
E_n	所有长为 n 的事件序列,其中的事件取自事件集合 E
$uc(*)$	将事件看作字符并使用某种通用压缩算法压缩序列后得到的比特数
$uc(m)$或 $uc(Mm)$	单独压缩事件序列 m 或 Mm 所产生的编码的比特数
$uc(m\|M)$	$uc(Mm)-uc(m)$

4.1 异常评分公式

本节将描述 CADM 所使用的异常评分公式，并对该公式给出理论分析.CADM 首先将从日志转换中获取的事件序列分段.一个日志文件所对应的事件序列可以被分为若干段，每一段包含一定数量的事件.在 CADM 中可以使用任意的分段策略，一般情况下，CADM 使用等事件数目的方式进行段的划分，即事件缓冲区中的事件个数到达指定段长时，就将事件缓冲区中的所有事件作为一段并使用评分公式进行评分.如果有可以直接利用的段的语义，如将某个用户一次会话作为一段,CADM 也可以使用这种语义进行分段.段是异常检测的最小单位，每一段都会根据式（1)获得一个异常程度的评分.

我们假设所有可能的事件来自一个有限但不能预知的集合 E.段的事件序列为 m.那么，CADM 对 m 的评分公式为

$$Score(m)=\frac{1}{|m|}[uc(Mm)-uc(M)-uc(m)]-\lambda \quad (1)$$

式(1)中 Mm 代表序列 M 和 m 按字符串拼接方式进行的连接.如果 Score（m)是正值，那么当前段 m 就会被识别为异常.Score(m)的值越大，就说明该段异常的可能性越高.在式（1)中，λ是用来控制是否异常的阈值.整个检测过程如算法 1 所示.

算法 1 在线异常检测算法.

输入：ctx:压缩过之前所有正常事件序列的压
segment:当前事件序列段

输出：当前段是否异常

过程：PROCEDURE DETECT(ctx，segment)

incre_size ← compress(ctx，segment)

初始化压缩器 tmp_ctx

score ← (incre_size - standalone _size)
standalone_size ← compress(tmp_ctx，segment)
/Len(segment)

IF score> LEARNING_THRES THEN

恢复 ctx 到压缩 segment 之前的状态

END IF

个两状态的一阶马尔可夫链生成的，该马尔可夫链的状态转移矩阵为

其中 p 是一个可配置的参数.对于正常事件序列和异常事件序列，我们使用不同的 p，因此它们符合不同的分布.异常检测算法负责将异常序列从正常序列中识别出来.

我们使用两类不同的序列对 CADM 算法进行评估.对于第 1 类序列，如果马尔可夫链在两类状态下分别输出字符“0”和“1”.而对于第 2 类序列，两种状态下的输出分别变为“30”和“31”，因此马尔可夫生成器的状态轮换无法直接从最终生成的事件序列中获取.

对于每一类序列，我们设计了两个测试用例，在每个测试用例中，马尔可夫生成器选取特定 p 值生成 10000 个事件作为训练序列.待测试的序列包含 16000 个事件.在待测试的序列中，前 8000 个序列为生成器使用与训练序列相同的 p 值，因此前 8000 个事件应被识别为正常；后 8000 个序列则使用不同的 p 值生成，因此应被识别为异常.我们将正常序列的 p 值记作 p_n，异常序列的 p 值记作 p_a.另外，我们采用等事件数分段的方式，每一段的事件数设为 80.在每个测试用例中，针对较易检测和较难检测两种场景，我们分别选取 P_n=0.2,P_a=0.8 以及 P_n=0.5,P_a=0.8 两组值作为代表.

作为比较，我们首先将文献[12]中提出的马尔可夫模型应用到这两个测试用例上.该方法首先使用训练数据来训练一个一阶马尔可夫模型，并使用该模型来判断每一段出现的概率，如果概率较小则视为异常.将马尔可夫模型应用到两个测试用例上的结果如图 2 所示.对于第 1 个测试用例,当 pn 和 pa 差距较大时，马尔可夫模型对正常序列和异常序列

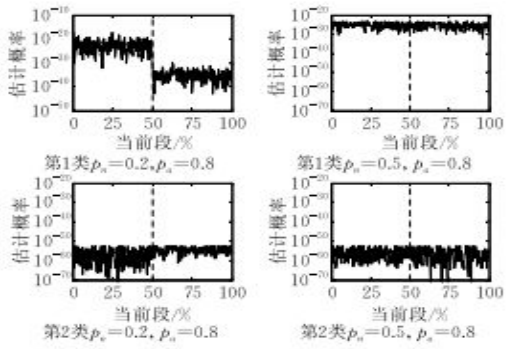

图 2　马尔可夫模型对人工生成数据进行异常检测的结果

列的概率估计有明显的差异，因此可以将这两类序列分辨开.但是，当 P_n=0.5,P_a=0.8 时，由于差距比较小,马尔可夫过程没能将这两类序列分辨开.对于第 2 种测试用例，无论 P_n 和 P_a 的值是什么，马尔可夫模型均无法将两类序列分辨开来.这说明马尔可夫的性质不仅与测试场景的难度有关，当序列不符合马尔可夫性质时，即使在测试场景比较简单的情况下，马尔可夫过程也无法将序列分辨开.

然后，我们将 CADM 算法应用到这两个测试用例上，测试结果如图 3 所示.对于第 1 种情形，CADM 的结果不比马尔可夫模型差.但是，对于第 2 个测试用例，当序列上的马尔可夫性质不成立时，传统的马尔可夫模型无能为力，CADM 仍然有效.这是因为 CADM 不依赖于序列概率分布上的假设.因此，CADM 可以被应用到更广泛的场景下，并且在没有关于序列概率分布的先验知识时会有更优的检测结果.

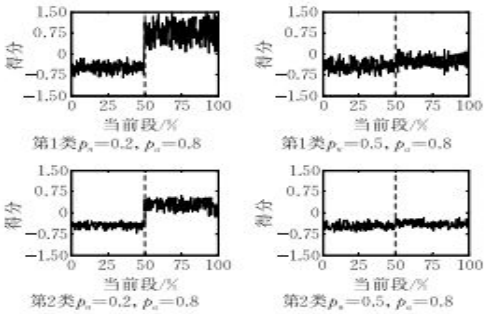

图 3　CADM 对人工生成数据进行异常检测的结果

5.2 参数调优与分析

在本实验中，我们将 CADM 应用到上一节中的马尔可夫生成器生成的更多测试用例上以对 CADM 进行评估和分析.我们采用等事件个数的分段方式并首先将每一段大小设为 80,并且将异常检测阈值和学习阈值均设为 0.由于在前一实验中，CADM 在 P_n=0.5，P_a=0.8 的测试用例中表现较差，本实验中采用相似的测试用例进行测试，并使用易测试的用例 P_n=0.2,P_a=0.9 作为对比.上述实验的结果如表 2 所示.在这个表格中，还列出了正常序列与异常序列概率分布的真实相对熵值.

从表 2 中，我们发现在当前参数下,CADM 在一些测试用例中召回率特别低.另外,召回率与正常事件序列及异常事件序列的相对熵存在联系.当正常序列与异常序列的相对熵差距较小时，算法将失效.为了找出算法失效的原因，我们在图 4 中描绘

表 2　参数调优之前的检测结果

	第 1 类				第 2 类			
	相对熵	真阳性数	精度	召回率	相对熵	真阳性数	精度	召回率
$p_m=0.2$ $p_a=0.4$	0.151	8	0.800	0.042	0.050	0	—	0
$p_m=0.2$ $p_a=0.5$	0.321	39	0.975	0.204	0.107	0	—	0
$p_m=0.2$ $p_a=0.6$	0.551	161	1.000	0.839	0.183	0	—	0
$p_m=0.5$ $p_a=0.7$	0.119	0	—	0	0.039	0	—	0
$p_m=0.5$ $p_a=0.8$	0.278	4	1.000	0.021	0.092	1	1	0.003
$p_m=0.5$ $p_a=0.9$	0.531	35	1.000	0.190	0.177	0	—	0
$p_m=0.2$ $p_a=0.9$	1.652	192	1.000	1.000	0.550	371	1	1

了

部分测试用例的详细结果.由于 CADM 算法的本质是使用增量压缩的大小 uc(m|M)和单独压缩的大小 uc(m)的差来估计相对熵，每一段的 uc(m|M)和 uc(m)的大小对估计和检测的准确度至关重要.因此，对于一些测试用例，我们在图 4 中描绘了每一段增量压缩大小和单独压缩大小.在图 4 中，左上角的图描绘了最理想的情况：对于正常的段,uc(m|M)和 uc(m)十分接近，因为 D(Q||P)等于 0;对于异常的段，uc(m|M)有一个明显的增量，因为异常的段是在错误概率估计的基础上进行压缩.利用这个差异，算法可以很容易地将正常序列和异常序列区分开来.但是,从其它几幅图可以看出，当相对熵较小时,uc(m)的冗余超过了使用错误概率压缩时 uc(m|M)的增量,从而导致了漏检.因此,算法失效主要是由于对 m 进行压缩时结果编码相对于信源熵率冗余过大.另外,从左下角图片还可以看出错误学习假阴性的序列会导致对正常事件序列的概率估计发生偏移，从而加重漏检的情况.

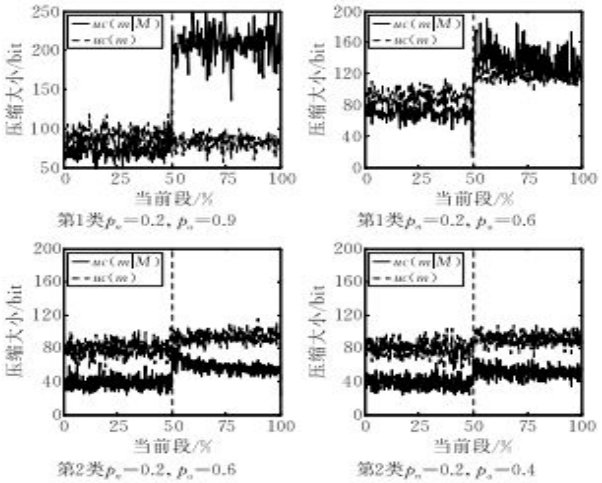

图 4　对于特定测试集单独压缩和增量压缩大小的关系

对于错误学习的问题，可以通过使用较高的学习阈值来解决.另一方面，可以通过增加每一段的事件数 W 来增加压缩率.通过使用较长的段，测试序列中的重复结构可以被充分利用，其单个事件压缩后的大小会更加接近信源的熵率，从而提高了相对熵估计的精确度.使用不同窗口大小的效果如表 3 左半部所示.对于每一个测试用例，我们列出了不同窗口大小下所有异常序列的相对熵估计 λ=1/n[uc(m|M)-uc(m)]的平均值.另外需要指出的是，在段的长度不同的情况下进行实验时，训练序列的长度应该随着段的长度线性增长，因为定理 1 中要求正常序列大小和段的大小满足条件 N≥Cn.从表中,我们可以看出相对熵估计值和理论值之间的差异随着段的长度增大而减小，因此使用较长的段有利于提高召回率.

表 3　不同窗口大小及是否采用重复对估计精度的影响

测试段长度	不重复				重重			
	W=80	W=160	W=320	W=640	W=80	W=160	W=320	W=640
$p_m=0.2, p_a=0.4$	−0.104	−0.069	−0.062	−0.049	0.0123	0.023	0.032	0.056
$p_m=0.2, p_a=0.5$	0.216	0.074	0.056	0.135	0.223	0.298	0.241	0.241
$p_m=0.2, p_a=0.6$	0.246	0.248	0.271	0.366	0.409	0.390	0.419	0.477
$p_m=0.5, p_a=0.7$	−0.192	−0.174	−0.124	−0.139	−0.030	−0.027	−0.006	−0.031
$p_m=0.5, p_a=0.8$	−0.063	0.022	0.059	0.061	0.082	0.130	0.152	0.162
$p_m=0.5, p_a=0.9$	0.133	0.180	0.255	0.332	0.294	0.347	0.406	0.424

但是，实际中段的长度无法设置得非常大.首先,段越长,每一个警告所能提供的信息就越小，因为管理人员需要通读整个段的日志来确定异常发生的原因.其次，在实际日志中，异常相关的信息在日志中可能不是连续的，如果使用较长的段可能导致异常信息引起的相对熵的差异被周围正常的日志所湮没.最后，如前所述，使用较长的段意味着必须使用更长的训练日志.

法分辨，则可以预期这两种方法无法有效工作.从图 6 中可以看出,正常序列和异常序列的这两种频率分布基本相似,这就解释了马尔可夫模型和统计模型失效的原因.

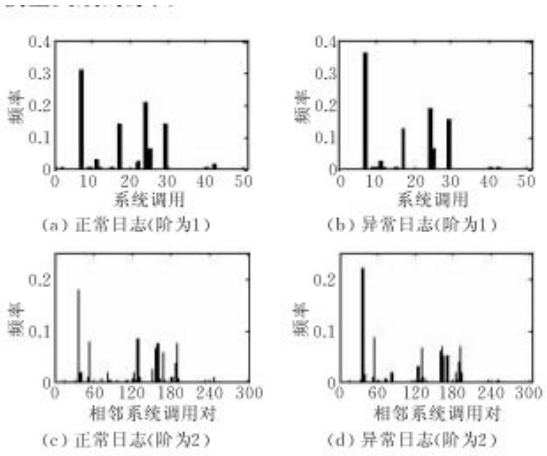

图 6　系统调用和相邻系统调用对的频率的统计

类似 BSM 审计日志，现实日志中事件的概率分布可能比较复杂，使用简单直观的假设将无法捕捉到异常日志的特征.从以上实验中，我们可以看出，通过去移除对这些假设的依赖，CADM 可以在更大范围的日志上有更好的效果.因此，CADM 更适合用于日志的在线异常检测.

5.4 可扩展性

在本实验中,我们对 CADM 的运行效率进行评测.我们使用 5.1 节中的马尔可夫生成器生成不同长度的日志，然后以一种离线的方式测量 CADM 在这些日志上的运行时间.

最终测试的结果如图 7 所示.从该图可以看出 CADM 是一种线性时间复杂度的算法,无论已经运行多长时间，CADM 检测到异常的时间延迟都不会增长.因此，CADM 有较好的可扩展性，更适合于在线异常检测.

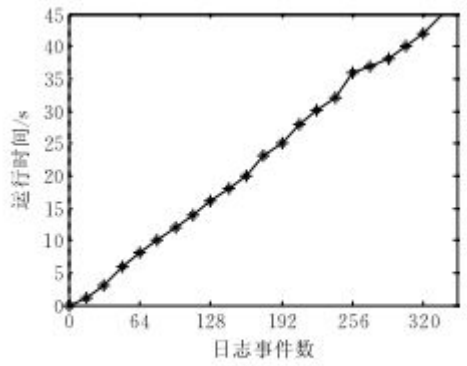

图 7　CADM 在不同大小日志上的运行时间

6　结论

本文提出了一种新的离散序列异常检测方法 CADM，并且给出了这种方法的理论分析和实验评估.CADM 选择相对熵作为测试日志与异常日志的度量并且使用压缩后的编码大小来估计相对熵.从实验中可以看出，与其它方法相比，CADM 不依赖于关于概率分布的先验知识并且具有线性时间复杂度，因此更适合于在线异常检测.

本文进行日志转换时使用了现有的方法.由于日志转换对整个异常检测过程的准确度非常重要,因此在未来我们将研究对专用的文本和高维数据的序列转换方法.

参　考　文　献

[1]Xu W，Huang L，Fox A，et al. Detecting large-scale system problems by mining console logs//Proceedings of the 22nd ACM SIGOPS Symposium on Operating Systems Principles. Montana，USA，2009: 117-132

[2] Jiang W，Hu C，Pasupathy S，et al. Understanding customer problem troubleshooting from storage system logs//Proceedings of the Conference on File and Storage Technologies. San Francisco，USA，2009：43-56

[3]Oliner A, Stearley J. What supercomputers say: A study of five system logs//Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Edinburgh, UK, 2007：575-584

[4]Chandola V. Anomaly detection for discrete sequences:A survey. IEEE Transactions on Knowledge and Data Engineering ，2012，24(5)：823-839

[5]Chandola V，Mithal V Kumar V. Comparative evaluation of anomaly detection techniques for sequence data//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa，Italy，2008: 743-748

[6] Cover T M，Thomas J A. Elements of information theory (Wiley-interscience，2006. 2006)

[7]Chandola V, Banerjee A，Kumar V. Anomaly detection:A survey. ACM Computing Surveys，2009，41 (3) : 1-58

[8]Ye N Chen Q. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. Quality and Reliability Engineering International，2001，17(2): 105-112

[9] Oliner A J，Kulkarni A V，Aiken A. Using correlated surprise to infer shared influence//Proceedings of the 2010 IEEE/IFIP International Conference on Dependable Systems

[10] Wang C,Viswanathan K,Choudur L,et al. Statistical techniques for online anomaly detection in data centers//Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management and Workshops. Dublin Ireland, 2011:385-392

[11] Xu W, Huang L, Fox A, et al. Mining console logs for large-scale system problem detection//Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles. Montana，USA，2009:117-132

[12]Ye N A markov chain model of temporal behavior for anomaly detection//Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop. New York，USA，2000:169

[13]Michael C C, Ghosh A. Two state-based approaches to program-based anomaly detection//Proceedings of the 16th Annual Conference on Computer Security Applications. New Orleans, USA, 2000:21-30

[14] Sun P. Mining for outliers in sequential databases//Proceedings of the 6th SIAM International Conference on Data Mining. Bethesda，USA，2006:94

[15]Zhang X,Fan P, Zhu Z. A new anomaly detection method based on hierarchical hmm//Proceedings of the 4th International Conference on Parallel and Distributed Computing, Applications and Technologies. Chengdu, China,2003：249-252

[16]Florez-Larrahondo G, Bridges S， Vaughn R. Efficient modeling of discrete events for anomaly detection using hidden markov models. Information Security, 2005，3650：506-514

[17]Yamanishi K, Maruyama Y. Dynamic syslog mining for network failure monitoring//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005：499-508

[18]Budalakoti S，Srivastava A，Akella R, Turkov E. Anomaly detection in large sets of high-dimensional symbol sequences. NASA Ames Research Center, Technical Report NASA TM-2006-214553, 2006

[19]Budalakoti S， Srivastava A N ,Otey M E. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. IEEE Transactions on Systems, Man， and Cybernetics, Part C:Applications and Reviews,2009, 39(1) : 101-113

[20]Keogh E, Lonardi S,Ratanamahatana C A. Towards parameter-free data mining//Proceedings of the 2004 ACM SIGK-DD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2004：206

[21]Kieffer J C, Yang E. Grammar-based codes:A new class of universal lossless source codes. IEEE Transactions on Information Theory, 2000, 46(3);737-754

[22]Yang E-H， Kieffer J C. Efficient universal lossless data compression algorithms based on a greedy sequential grammar

[23]Ziv J. On classification with empirically observed statistics and universal data compression. IEEE Transactions on Information Theory, 1988, 34(2)：278-286

[24]Rached Z, Alajaji F, Campbell L L. The kullback-leibler divergence rate between Markov sources. IEEE Transactions on Information Theory, 2004, 50(5)：917-921

[25]He D, Yang E. The universality of grammar-based codes for sources with countably infinite alphabets. IEEE Transactions on Information Theory, 2005，51(11)：3753-3765

GAO Yun. born in 1991，Ph. D. candidate. His research interests focus on anomaly detection and large-scale data processing.

ZHOU Wei, born in 1987, M.S. Her major research interests include large-scale data processing and graph computing.

HAN Ji-Zhong, born in 1972, Ph.D, professor, Ph. D. supervisor. His major research interests include cloud platform and large-scale data processing.

MENG Dan, born in 1965, Ph. D，professor, Ph. D. supervisor. His major research interests focus on large-scale data processing.

Background

Anomaly detection problem is a common problem emerging

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61070028) , National High Technology Research and Development Program (863 Program) of China (Grant No. 2012AA01A401) and “Strategic Priority Research Program ” of the Chinese Academy of Sciences (Grant No. XDA06030200). Any options, findings, conclusions,or recommendations expressed are the authors’ and do not necessarily reflect those of the sponsors. We'd like to thank National Engineering Laboratory for Information Security Technologies for preparing the platform in this study.

总结：在完成作业的过程中，知道把截图的版式改为四周型，便于修改。然后通过插入公式，更熟练了一些公式的构成。百度学会了设置上下标，而且改段落之间的距离。但是还存在以下问题：

1)插入页眉页脚插的不好，所以就没插入

2）明明格式和模板要求的一样，但是出来的效果却是比 PDF 上多了一行，减小了行距但是还是不行，为了保持格式的一样，就把那行删除了

3)有些英文单词下面有红线，添加到字典后还是不行，不知道如何清除。

弄这个文件真的很需要耐心，耐心百度学习，耐心修改。