DOI: 10.11992/tis.201706041
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1250.018.html

基于门禁日志挖掘的内部威胁异常行为分析

王培超, 周鋆, 朱承, 黄金才, 张维明
(国防科技大学 信息系统工程重点实验室, 湖南 长沙 410072)

摘 要: 门禁系统是保护重要场所安全的重要手段, 可以有效防止未授权用户的进入。然而, 近年来大量案例表明重要场所的威胁主要来自于具有合法权限的内部人员。针对这个问题, 提出基于门禁日志数据挖掘的内部威胁异常行为分析方法。该方法首先利用 PrefixSpan 算法对正常行为序列进行提取, 之后计算待检测序列的序列异常度分数, 并根据决策者设定的阈值来找出异常序列。通过真实门禁数据中的实验, 验证了本方法可以降低精确匹配在数据较少时带来的高误报率, 实现对内部人员异常行为的有效发现, 为加强重要场所安全保护提供了新的途径。

关键词: 门禁系统; 日志数据挖掘; 内部威胁检测; 异常行为分析
中图分类号: TP311 **文献标志码:** A **文章编号:** 1673−4785(2017)06−0781−09

中文引用格式: 王培超, 周鋆, 朱承, 等. 基于门禁日志挖掘的内部威胁异常行为分析[J]. 智能系统学报, 2017, 12(6): 781–789.
英文引用格式: WANG Peichao, ZHOU Yun, ZHU Cheng, et al. Analysis on abnormal behavior of insider threats based on access-log mining[J]. CAAI transactions on intelligent systems, 2017, 12(6): 781–789.

Analysis on abnormal behavior of insider threats based on accesslog mining

WANG Peichao, ZHOU Yun, ZHU Cheng, HUANG Jincai, ZHANG Weiming
(Key Laboratory of Information System Engineering, National University of Defense Technology, Changsha 410072, China)

Abstract: Using an access control system is an important method of guarding key places, and it can effectively prohibit the entry of unauthorized users. However, many recent cases indicate that threats to key places mostly come from insiders. To address this challenge, this paper proposes a method for analyzing the abnormal behavior of insider threats based on accesslog data mining. First, the PrefixSpan algorithm is used to extract normal behavior sequences; then, the anomaly scores of the access sequences are calculated. Finally, the abnormal sequences are identified according to a threshold determined by decision makers. Experiments on real access data show that this method can decrease high false alarm rates caused by an exact match when there is limited data and can also effectively reveal abnormal behavior by insiders. Therefore, this method provides a new approach for enhancing the protection of key places.
Keywords: access control system; accesslog mining; insider threat detection; analysis on abnormal behavior

重要场所的安全保护历来是人们关注的焦点, 其安保措施有钥匙专人携带、雇用安保人员、使用门禁系统等, 然而, 随着经济社会的发展, 门禁系统(access control system)在各重要场所的所占比重越来越大, 传统安保措施的应用越来越少。这一方面由于传统安保措施的弊端(如钥匙易丢失、易被复制, 人员被收买等), 另一方面得益于门禁系统日益完善的强大功能, 指纹识别[1]、虹膜识别[2]等新识别技术的应用使门禁系统已经成为了涉及诸多新技术的新型现代化安全管理系统, 在银行、宾馆、重要办公场所等地发挥着无可替代的作用。

门禁系统发挥作用的主要途径, 是通过对不同用户授予不同的权限, 从而规范不同地点的进出人员。但是, 近年来大量案例表明, 一个重要场所的最大威胁往往不是来自外部人员, 而是来自那些拥有合法权限的内部人员, 内部威胁(insider threat)随着“棱镜门”等事件的曝光而越来越受到人们的重

收稿日期: 2017−06−10. **网络出版日期:** 2017−11−09.
基金项目: 国家自然科学基金项目(71571186); 教育部在线教育研究基金项目(2017YB119).
通信作者: 周鋆. E-mail: zhouyun@nudt.edu.cn.

视。国外已有不少学者对于内部威胁进行了研究，这些研究主要存在于信息域，如 D.F.Ferraiolo[3]提出了基于角色的访问控制（role based access control，RBAC），Bishop Matt[4]在此基础上，提出了基于属性的组访问控制（attribute based group access control，ABGAC）等，这些研究对本文进行物理域异常检测提供了很好的指导作用。

门禁日志分析是发现内部威胁的重要途径，作为物理域信息的重要来源，国内外已有不少学者对此展开研究，通常借鉴网络空间异常检测的方法来刻画人员的行为。序列模式挖掘是利用门禁日志数据刻画人员行为的有效方法，本文在此基础上提出了一种计算序列异常度分数的方法，通过利用 PrefixSpan 算法[5]找出人员行为的频繁序列，并通过计算序列异常度分数对人员的行为序列的异常度进行了定量刻画表示，进而可根据阈值来找出异常行为，有效减少了因精确匹配造成的高误报率，适用于对各种门禁日志的分析处理。

1　相关研究

内部威胁是异常检测所面临的巨大挑战，国外学者均提出了检测内部威胁的相关理论或实践方法[6-12]，并取得了较好的效果。无论是国内外的学者，在进行实际的内部威胁检测时，通常借鉴网络空间异常检测的方法来对用户行为进行分析，即构建用户的正常行为模型后查找离群点，常用方法包括有监督的异常检测、半监督的异常检测和无监督的异常检测[13]，通过建立正常的行为模式集，将实际行为模式与正常行为模式进行对比，看两者是否匹配，若不匹配，说明是异常，反之则为正常。

门禁日志分析是内部威胁检测的一个小分支，国外学者对其已有一定的研究，如 Bostjan 等[14]通过将刷卡数据与监控数据结合提出了多层框架模型对用户行为进行分析，M. Davis 等[15]采用图挖掘算法检测门禁数据中的结构异常（建筑物中不正常路径）和数值异常（不正常计时数据）等。与国外学者相比，国内学者对于门禁日志数据挖掘的研究较少，不少学者关注于门禁系统架构的设计[16-17]，而对于门禁数据仅进行了统计学分析[18]，较少对异常行为进行相应分析。郑伟平等[19]对社区管理数据利用 k-means 找出社区人流规律，从而加强对社区治安的管理；史殿习等[20]提出了可视为 Apriori 算法扩展的加权模式挖掘算法，利用此算法刻画用户的日常行为模式，取得了较好的效果；顾兆军等[21]对大型航站楼门禁日志进行了序列模式挖掘，对机场员工行为进行了有效刻画，并利用精确匹配找出内部人员的异常行为序列。

序列模式挖掘是查找序列集合中的频繁序列的重要方法，给定一个由不同序列组成的集合，其中，每个序列由不同的元素按顺序有序排列，同时给定一个用户指定的最小支持度阈值 min_sup，序列模式挖掘就是找出所有出现频率不低于 min_sup 的子序列[22]。常用的基本序列模式挖掘算法有类 Apriori 算法（AprioriAll、AprioriSome、DynamicSome）和基于数据投影的算法（FreeSpan[23]，PrefixSpan）等。Apriori 类算法的思想大致相同，首先遍历序列数据库生成候选序列，并利用先验性质进行剪枝得到频繁序列，每次遍历都是通过连接上次得到的频繁序列生成新的长度加 1 的候选序列，然后扫描每个候选序列验证其是否为频繁序列，要对数据库进行反复多次的扫描。FreeSpan 算法利用当前挖掘的频繁序列集将序列数据库递归地投影到一组更小的投影数据库上，分别在每个投影数据库上增长子序列，PrefixSpan 是 FreeSpan 的改进算法，其投影时不考虑所有可能出现的频繁子序列，只检查前缀序列，然后把相应的后缀投影成投影数据库，之后在其中只检查局部频繁模式，不需要生成候选子序列。PrefixSpan 算法在处理数据时有较高的效率，故本文在后续的实验中采用此算法。

序列模式挖掘是查找序列集合中的频繁序列的重要方法，给定一个由不同序列组成的集合，其中，每个序列由不同的元素按顺序有序排列，同时给定一个用户指定的最小支持度阈值 min_sup，序列模式挖掘就是找出所有出现频率不低于 min_sup 的子序列[22]。常用的基本序列模式挖掘算法有类 Apriori 算法（AprioriAll、AprioriSome、DynamicSome）和基于数据投影的算法（FreeSpan[23]，PrefixSpan）等。类 Apriori 算法的思想大致相同，首先遍历序列数据库生成候选序列并利用先验性质进行剪枝得到频繁序列，每次遍历都是通过连接上次得到的频繁序列生成新的长度加 1 的候选序列，然后扫描每个候选序列验证其是否为频繁序列，要对数据库进行反复多次的扫描。FreeSpan 算法利用当前挖掘的频繁序列集将序列数据库递归地投影到一组更小的投影数据库上，分别在每个投影数据库上增长子序列，PrefixSpan 是 FreeSpan 的改进算法，其投影时不考虑所有可能出现的频繁子序列，只检查前缀序列，然后把相应的后缀投影成投影数据库，之后在其中只检查局部频繁模式，不需要生成候选子序列。PrefixSpan 算法在处理数据时有较高的效率，故本文在后续的实验中采用此算法。

2 问题描述

2.1 路径序列数据

由于门禁系统的存在, 每个人的卡会由管理人员统一进行授权, 只被允许访问特定的区域。当一个人进入某个区域时, 需要预先刷卡, 门禁系统会将当前刷卡时间、刷卡人姓名、卡号、刷卡地点等重要信息进行记录。对于内部人员来说, 他们的行为路径是本文进行异常行为分析的重要对象, 将一个人每天的刷卡地点按顺序进行采集, 即可得到一个人每天的行为序列。

2.2 正常路径序列数据

对一个人来说, 每天工作的流程是基本确定的, 因此每天的行为路径序列应该有较大的相似性。例如, 对于一个老师来说, 每天来到办公室后, 在短暂准备后会去实验室和学生讨论问题, 之后再回到办公室备课或完成论文等, 这样就形成了“办公室—实验室—办公室”的行为序列。将人员访问的门禁点用 $p_i(i=1, 2, \cdots, n)$ 表示, 按采集顺序排列就可以得到人员的路径序列 $\langle p_1, p_2, \cdots, p_i, \cdots, p_n \rangle$, 之后由决策者人为设定 min_sup, 即可将这些数据进行频繁模式挖掘, 将支持度高于 min_sup 的行为序列视为正常行为序列, 从而得到正常行为序列库。

2.3 序列异常度分数

通过精确匹配直接判定异常在数据有一定缺失的情况下会导致极高的误报率, 为此, 本文引入了序列异常度分数(score of sequence's abnormal degree), 来定量刻画一个正常序列与一个待评判序列之间的差异。序列 $\langle p_1, p_2, p_3, p_4 \rangle$ 和 $\langle p_1, p_2, p_4, p_3 \rangle$ 以及序列 $\langle p_1, p_2, p_3, p_4 \rangle$ 和 $\langle p_3, p_5, p_1, p_6 \rangle$ 的差异显然是不同的, 传统异常检测方法通过进行精确匹配, 将与正常行为序列库中所有内容均不同的序列直接判定为异常, 不考虑两个序列之间的差异; 为了更好比较两个序列之间的差异, 可以采用编辑距离(edit distance, ED)对序列间的差异进行量化。由于不同序列长短和复杂程度各不相同, 用于比较的正常行为序列的支持度也不相同, 仅靠通过计算编辑距离会造成巨大误差, 因此, 本文计算相对编辑距离(relative edit distance, R_{ED})和相对支持度(relative support, R_{Sup}), 进而计算可以得到序列差异分数(score of sequence's difference degree), 之后根据时间规则计算时间异常分数(score of abnormal time), 通过将二者加权相加得到序列异常度分数, 根据决策者的阈值可以对异常序列进行发现。

3 异常序列挖掘

对于序列的异常程度, 本模型从两个方面来考虑: 一方面是当前行为序列与正常行为序列库中的序列的差异程度, 即序列差异分数, 这需要考虑相对编辑距离大小和相对支持度大小; 另一方面是刷卡的时间因素, 包括刷卡行为的发生时间和过于短暂的刷卡时间间隔两个方面。

3.1 序列差异分数

3.1.1 正常行为序列库建立

为了定量刻画异常序列的异常程度, 首先应进行正常序列库的建立。通过利用 PrefixSpan 算法, 设定合理的最小支持度 min_sup(通常为 20% 左右), 对预处理后得到的行为序列进行频繁模式挖掘, 可以得到行为序列中的高频序列。对于一个部门来说, 在常年的正常运行中已基本形成固定的行为模式, 每名员工在岗位不变的情况下均会形成自身固定的行为模式(例如先去实验室 a, 后去实验室 b), 因而通过对大量数据进行频繁模式挖掘得到的高频行为序列可以被认为是正常行为序列。

3.1.2 相对编辑距离计算

将一个序列变换成另一个序列, 其可能的最大编辑距离为正常行为序列长度和当前行为序列长度中较大的那个。为了更好比较不同序列进行变换时需要的编辑距离的相对大小, 计算相对编辑距离 R_{ED} 公式为

$$R_{\mathrm{ED}}\left(q_i^t, q_j^n\right) = \frac{\mathrm{ED}\left(q_i^t, q_j^n\right)}{\mathrm{Max}\left(\left|q_i^t\right|, \left|q_j^n\right|\right)} \tag{1}$$

式中: ED 为编辑距离函数, q_i^t 为测试序列中的第 i 条序列, q_j^n 为正常序列库中的第 j 条序列, $|q_i^t|$ 和 $|q_j^n|$ 为相应序列的序列长度。相对编辑距离可以有效比较在编辑距离相同时两序列之间的差距。例如, 将序列 $\langle p_1, p_2, p_3 \rangle$ 变换为 $\langle p_1, p_2, p_3, p_4 \rangle$ 所需的编辑距离为 2, 将序列 $\langle p_1, p_2 \rangle$ 变换为 $\langle p_1, p_2, p_3, p_4 \rangle$ 所需的编辑距离同样为 2, 然而, 两个正常行为序列的长度不同, 在编辑距离相同的情况下, 正常行为序列的长度越长, 当前行为序列和正常行为序列的相似度越高。相对编辑距离可以很好地刻画出两序列间的差异程度。

3.1.3 相对支持度计算

对于当前行为序列来说, 与其对比的正常行为序列的支持度对评价当前行为序列的差异程度有很大影响。为了定量刻画这种差异, 定义相对支持度 R_{sup} 为

$$R_{\sup}\left(q_j^n\right)=\frac{\log\left(C_{\sup}\left(q_j^n\right)\right)}{\log\left(\mathrm{Max}_{\sup}\right)} \tag{2}$$

式中：$C_{\sup}\left(q_j^n\right)$为正常序列库中第 i 条序列的支持度，$\mathrm{Max}_{\sup}$为正常行为序列库中最大的支持度；取对数可以减少因支持度间差距太大导致的分数过小。相对支持度越高，证明人员日常行为与当前正常行为序列存在差异时，日常行为序列的异常程度越大。

3.1.4 序列差异分数计算

相对编辑距离和相对支持度两方面在计算序列差异分数时都需要考虑。当前行为序列应与正常行为序列库中每一条序列进行比较，从而得到序列差异分数 score_1：

1）当相对编辑距离计算结果中存在 0 时，意味着当前行为序列与正常行为序列库中的序列存在完全一致的情况，因此此时序列差异分数为 0；

2）当相对编辑距离计算结果中不存在 0 时，意味着当前行为序列与正常行为序列库中的序列不存在完全一致的情况，这时考虑当前行为序列与正常行为序列库中所有序列的整体差别，对计算出的多个得分求平均值，从而得到该行为序列偏离正常行为序列的总体程度：

$$\mathrm{score}_1\left(q_i^t,q^n\right)=\begin{cases}0,\exists R_{\mathrm{ED}}\left(q_i^t,q_j^n\right)=0\\ \mathrm{mean}\left(\sum_{j=1}^{M_{\min\_\sup}}100R_{\sup}\left(q_j^n\right)R_{\mathrm{ED}}\left(q_i^t,q_j^n\right)\right),\forall R_{\mathrm{ED}}\left(q_i^t,q_j^n\right)\neq 0\end{cases} \tag{3}$$

式中：$M_{\min\_\sup}$ 为在当前最小支持度下的正常行为序列库中行为序列的总数；mean 为求平均值，可得出该条测试序列与正常行为序列库中所有行为序列的整体差距。

3.2 时间异常分数

3.2.1 时间规则

通过序列差异分数只能对序列的次序异常进行刻画，由于没有考虑时间，对于异常的发现存在一定的缺陷。定义时间异常规则如下：

1）异常时间段进入：用户在非正常时间段进入某地。

2）刷卡间隔过短：两次刷卡时间间隔过于短暂，异于平常。

这两种异常利用序列差异分数的方法是无法发现的，例如，对于序列〈p_1,p_1,p_1,p_2〉，当对编号 p_1 的设备在 10 s 内刷卡 3 次时，这种行为显然是异常的；然而，这条序列可能出现在构建的正常行为序列库中，因为用户在一天内对设备刷卡 3 次的行为是正常的，此种情况使用异常度分数的方法无法发现其异常，而时间规则却可以很好地将其发现。

3.2.2 时间分数计算

对于两条时间规则，而对于异常时间的刷卡行为，不同部门的工作时间段有不同之处，部门的异常刷卡时间段阈值（threshold）应该根据对该部门的正常运行时间进行人为设定，当某天的刷卡时间出现在异常时间段内时，根据出现时间与阈值之间的相对差距来计算二者之间的分数；刷卡间隔为门禁系统中同一天内两条数据之间的时间差（min），对于刷卡间隔过短的异常，本文根据该部门的整体刷卡间隔情况来确定。通过绘制刷卡间隔与累积频率的曲线，并将此曲线进行拟合，可以得到刷卡时间间隔与累计频率的关系函数，刷卡时间间隔对应的累计频率越大，意味着该刷卡间隔过短的可能性越小，计算分数时使用式 (4)：

$$\mathrm{score}_2\left(q_i^t\right)=100\times\left(\frac{\sum_{k=1}^{N_i-1}\left(1-f\left(\Delta t_k\right)\right)}{N_i-1}+\frac{\left|\mathrm{Min}\left(0,t_i-\mathrm{threshold}\right)\right|}{\left|\mathrm{threshold}\right|}\right)\times\frac{1}{2} \tag{4}$$

式中：f 为拟合出的函数，Δt_k 为当天的第 k 时间间隔（min），threshold 为设定的异常时间阈值，t_i 为第 i 天最早的刷卡时间，N_i 为门禁测试序列中第 i 天的记录总数。

3.3 序列异常度分数

3.3.1 序列异常度分数计算

序列异常度分数由序列差异分数和时间异常分数两部分构成，计算第 i 条测试序列的序列异常度分数采用的方法如下：

$$\mathrm{score}\left(q_i^t\right)=\alpha\cdot\mathrm{score}_1\left(q_i^t,q^n\right)+\left(1-\alpha\right)\cdot\mathrm{score}_2\left(q_i^t\right) \tag{5}$$

式中：α 和 1-α 为两个子分数的权重，权重可根据决策者的偏好来决定，缺省值为 0.5，即简单平均。

3.3.2 异常路径行为发现

分数计算出来后，人员行为序列的异常程度大小就有了定量的刻画，而将哪些分数视为异常需要人为定性决定。不同部门的人员最后计算出的分数是不一样的，直接对所有人员划定统一的异常分数阈值会导致高误报率，同一部门的人只有在同一部门中进行比较才有较大的说服力；频繁序列的 min_sup 的设定也是一个问题，不同 min_sup 会对最后计算得到的分数产生一定的影响。为了给决策者进行决策提供更好的支持，应为决策者提供在不同支持度下设定不同阈值时产生的报警率，为此，本文通过绘制不同 min_sup 下的报警率曲线来为决

策者提供决策依据, 决策者可以自行决定需要设定的 min_sup 和异常分数阈值。

3.3.3 序列异常度分数计算示例

假设对某部门一个月的行为序列数据进行频繁模式挖掘, 得到表 1 结果。

表 1　测试用正常行为序列库

Table 1　Testing normal sequences library

行为序列	支持度计数
p_1, p_2, p_3, p_4, p_5	15
p_1, p_3, p_2, p_4	3
p_2, p_3, p_5	19

给定测试路径序列为〈p_1, p_2, p_5〉, 设定异常刷卡时间段阈值为早上 7 点, 刷卡时间间隔与累计频率的对应函数为 $f(\Delta t)=0.18\times(2.11\times\Delta t)^{0.37}$, 刷卡时间序列为〈2012-3-17:30:12, 2012-3-18:02:18, 2012-3-18:06:24〉, 测试阈值设定为 41, 权重 α 设定为 0.5。将测试路径分别与 3 条行为序列进行比较:

$$\text{score}_{1a}=\frac{\log(15)}{\log(19)}\times\frac{2}{5}\times100=36.79$$

$$\text{score}_{1b}=\frac{\log(3)}{\log(19)}\times\frac{2}{4}\times100=18.66$$

$$\text{score}_{1c}=\frac{\log(19)}{\log(19)}\times\frac{2}{3}\times100=66.67$$

$$\text{score}_1=\frac{(36.79+18.66+66.67)}{3}=40.71$$

接下来计算时间异常分数:

$$\text{scoer}_2=100\times\left[\frac{(1-f(32.1))+(1-f(4.1))}{2}+0\right]\times\frac{1}{2}=18.33$$

最后计算得到测试序列的序列异常度分数:

$$\text{score}=40.71\times0.5+18.33\times0.5=29.52$$

分数低于阈值, 因此该序列被标记为正常。

4 实验结果及分析

为了验证本文所提分数计算方法的有效性, 在 python2.7 和 MATLAB2014a 的环境下进行实验, 实验数据来自某涉密场所 2012—2016 年的门禁日志记录。

4.1 数据预处理

本文数据来自某单位的门禁系统, 其系统记录的门禁数据属性如表 2 所示。其中, 对本文的实验有帮助的属性有刷卡时间、设备名称、群组、用户姓名、卡片号码、用户编号, 6 个属性中, 最后两个是在找出异常序列后进行精确定位时使用, 其余属性的去除原因为重复或与题目无关。门禁系统无法没有对异常行为进行标记, 因此无法通过传统的分类方法对异常行为进行模型的构建。

表 2　门禁数据属性说明

Table 2　Instruction on attributes of access control system

属性	说明
刷卡时间	记录在门禁系统中的进入某地的时间
设备名称	门禁设备的安放地点
设备编号	与设备名称是一对多的关系
设备类型	有普通门禁和密码门禁两种
用户姓名	每名用户的真实姓名
用户编号	每个持卡用户的唯一编号
群组	用户所在的部门
开门方式	用户打开门的方式, 有按钮开门、刷卡开门等
开门结果	成功或失败
卡片号码	每张卡片对应的唯一号码
操作时间	由于系统延迟, 数据库相应群组数据更新的实际时间
操作人	数据库管理者, 均为SYSTEM

在门禁系统中, 总共有 103 个记录点, 但是由于系统存在较大问题, 很多门禁点在记录数据时存在不同程度的丢失现象, 而且不同群组的数据记录情况也各不相同, 为了建立可以反映用户正常行为的行为序列, 需要大量的数据, 因此需要寻找数据缺失最少的群组来进行正常路径行为模型建立。当使用有大量数据缺失的群组时, 会导致极高的误报率。导出数据后, 通过观察可以发现, 群组 *A*(部门 *A*)数据缺失较少, 因而接下来的实验中使用部门 *A* 的数据。

4.2 正常路径行为模型建立

本实验对部门 *A* 的数据进行了采集, 该部门是日常工作流程的关键部门之一, 应当进行高度关注。并且, 部门中的每个人在数据采集期间没有出现工作变动的情况, 长期工作性质未变, 最终得到的正常行为序列库可以较好地反映出部门人员的正常行为。通过对部门 *A* 在 2012—2014 年的数据进行处理, 并设定 min_sup 为 200 到 350 的非连续整数, 分别在不同支持度下对行为序列进行频繁序列挖掘。通过数据预处理, 得到每天的行为序列(共 1 049 条)后, 建立该部门在不同支持度下的正常行为序列库。如表 3 所示。

在 3 年时间里, 该部门的刷卡间隔数据共有 38 180 条, 而刷卡间隔在 1 h 内的记录有 32 585 条, 占总记录数的 85.34%, 因此认为超过 1 h 的刷卡时间间隔不存在刷卡时间间隔过短的情况, 统计 1 h 内的刷卡间隔与累积频率, 结果如图 1 所示。

表 3　正常路径序列模式集

Table 3　Normal set of road sequences

支持度	高频行为规则数	支持度	高频行为规则数
200	4 663	280	1 433
210	3 992	290	1 277
220	3 417	300	1 111
230	2 917	310	961
240	2 516	320	853
250	2 211	330	745
260	1 891	340	675
270	1 639	350	592

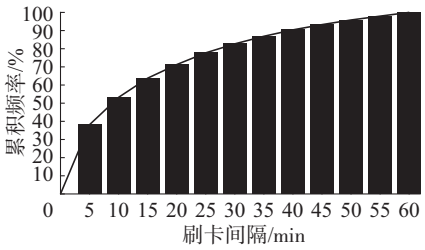


图 1　刷卡间隔与累积频率曲线

Fig. 1　Curve of swiping card interval and cumulative frequency

从图 1 中可以看出刷卡间隔和累积频率存在较明显的函数关系，利用 MATLAB 对其进行拟合，得到如下结果，图像如图 2 所示。

$$f(\Delta t) = 0.172\,6 \times (2.11 \times \Delta t)^{0.370\,8}$$

R^2 为 0.983，证明该函数对数据有较好的拟合效果。

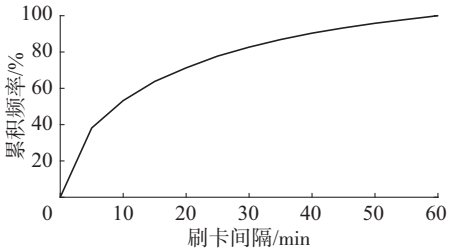


图 2　刷卡间隔与累积频率拟合曲线

Fig. 2　Fitted curve of swiping card interval and cumulative frequency

4.3　异常行为发现及对比

该部门的人员有不同程度的加班行为，但是早上的正常工作时间多为 8 点开始，不少工作人员 7 点多就会到达工作地点进行准备，因此设定 threshold 为 7，即异常时间段阈值为早上 7 点。采集该部门门禁记录中的剩余数据，处理后得到无标签的测试路径序列数据库（620 条），对得到的测试路径序列数据库计算序列异常度分数，得到在不同支持度下的结果，部分结果如图 3 所示。

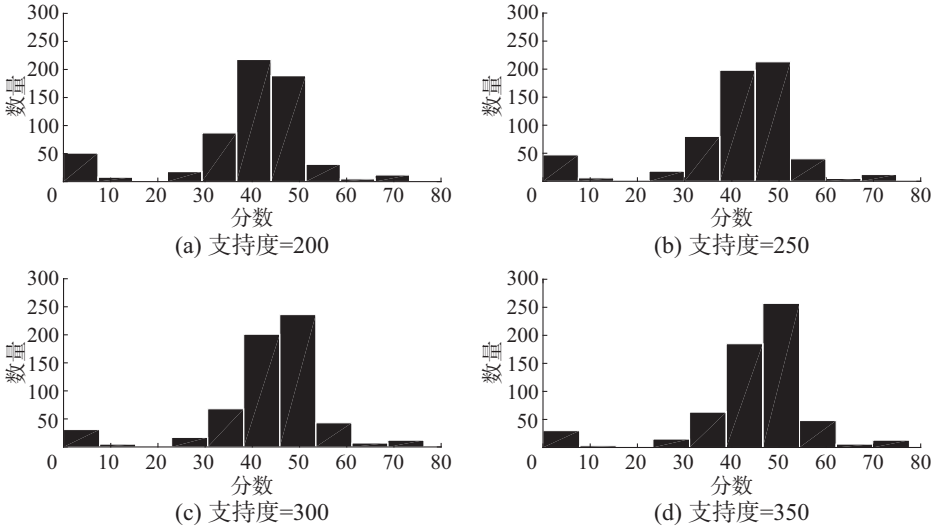


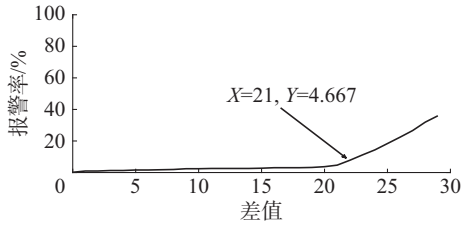
图 3　序列异常度分数分布

Fig. 3　Distribution of score of sequence's abnormal degree

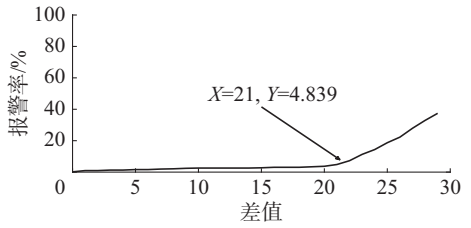
从图 3 中可以看出，随着支持度的增加，计算得到的序列异常度分数整体增大，高分段的集中区域在 x 轴上向右推进，计算得到的最大分数也逐渐增大。接下来，进行报警率曲线的绘制，在报警率曲线中，纵轴为报警率，即在当前阈值下报警的异常行为序列数占总序列数的百分比；横轴为人工设定的差值，从 0 开始递增，间隔为 1，在每个 min_sup 下通过将计算得到的待评价序列的序列异常度分数集合中的最高值减去当前差值后成为当前阈值。利用本文方法得到在不同支持度下的结果后，为了便于比较和传统方法得到结果的差别，在此一并绘制了利用精确匹配得到的结果图像，部分结果如图 4 所示。

从图 4 中可以看出，在本文方法得到的曲线

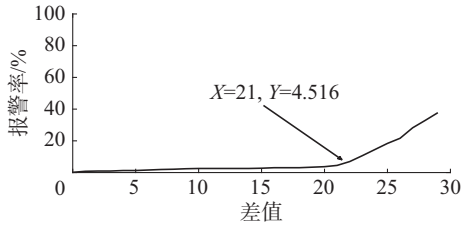
中, 随着当前阈值的逐渐下降(即差值的逐渐提升), 报警的异常行为序列越来越多, 决策者可根据图中结果来选定需要的阈值, 为今后的异常行为发现提供标准。在不同支持度下, 曲线的上升速度相似, 这是由于随着支持度的增加, 计算得到的分数整体增加的结果; 同时, 在差值为 21 左右的时候报警率曲线相对之前突然变陡, 决策者可根据此设定合理的分数阈值。在本实例中, 决策者可考虑将支持度 300、差值 21 时对应的阈值设定为今后的合理分数阈值(即序列异常度分数阈值=55.35)。这里对支持度为 250、差值为 4 的情况下查找出的序列进行人为观察, 部分结果如下所示。



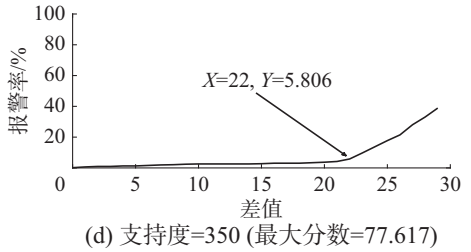
(a) 支持度=200 (最大分数=73.349)



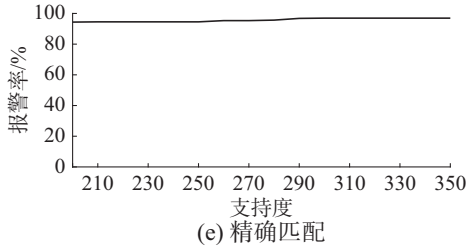
(b) 支持度=250 (最大分数=74.890)



(c) 支持度=300 (最大分数=76.349)



(d) 支持度=350 (最大分数=77.617)



(e) 精确匹配

图 4　部门 *A* 报警率曲线

Fig. 4　Department *A*'s curve of alarm rate

〈$p_0, p_{32}, p_{30}, p_{48}, p_{32}$〉: 当天的刷卡记录中只有从办公室之间的刷卡记录, 没有出现正门的刷卡记录。

〈 $p_{32}, p_{32}, p_4, \cdots, p_{48}, p_{32}, p_{32}$〉: 平日刷卡记录极少的 p_{32} 在当天出现了大量的刷卡记录。

传统的异常序列判别是通过精确匹配的方式, 若当前行为序列与正常行为序列库中的所有记录均无法完全匹配, 则当前行为序列被判定为异常。对于本数据来说, 若通过传统方式来找到异常序列, 直接将序列差异分数 score_1 的最终得分不为 0 的序列判定为异常序列即可。从图 4(e)中可以看出, 在不同支持度下, 利用传统方法查找得到的报警率均较高, 对于一个正常运行的单位来说, 这显然有悖于常识, 有较高的误报率。因此, 通过本文的方法, 可以对因精确匹配查找报警率高的序列集合提供有效的异常序列判断依据。

5　结束语

针对门禁日志, 本文没有像传统文章那样通过精确匹配找出异常, 而是提出了一种计算内部人员异常度分数的方法, 并通过补充的时间规则对异常数据进行良好判别, 适用于对内部人员的行为序列异常度进行有效的定量刻画。实验表明, 本文提出的序列异常度计算方法能够很好地对员工的路径行为异常度进行刻画, 并可以根据依图像设定的阈值将异常行为进行查找, 面对有一定缺失的数据可以有效减少由于精确匹配查找异常而造成的过高误报率, 这是对门禁日志数据的有效利用, 也是防范内部威胁的重要手段。在后续工作中, 考虑对人员行为进行更细粒度的分析, 异常时间段的设置将个人的日常行为习惯考虑进去, 并进一步利用贝叶斯网络对异常进行推断。

参考文献:

[1] 杨荣秀. 基于指纹识别技术的智能小区门禁系统的设计[J]. 科技与企业, 2016(5): 88–90.
YANG Xiurong. Design of intelligent community access control system based on fingerprint identification technique[J]. Technology and enterprise, 2016(5): 88–90.

[2] 李海青, 孙哲南, 谭铁牛, 等. 虹膜识别技术进展与发展趋势[J]. 信息安全研究, 2016, 2(1): 40–43.
LI Haiqing, SUN Zhenan, TAN Tieniu, et al. Progress and trends in iris recognition[J]. Journal of information security research, 2016, 2(1): 40–43.

[3] FERRAIOLO D F, KUHN R. Role based access control [C]//Proceedings of the 15th NIST-NCSC National Com-

puter Security Conference. Baltimore, Maryland, 1992: 554-563.

[4] MATT B, SOPHIE E, SEAN P, et al. We have met the enemy and he is us[C]//New Security Paradigms Workshop. Lake Tahoe, USA, 2008: 1-11.

[5] JIAN Pei, HAN Jiawei, BEHZAD M, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth[C]//20th International Council for Open and Distance Education World Conference on Open Learning and Distance Education. Heidelberg, Germany, 2001: 215-224.

[6] ANTONIO L, SIMON F, ZHUNAG Yan. A logical model for detecting irregular actions in physical access[C]// IEEE conference on database and expert systems applications. [S.l.], 2007: 560-564.

[7] DAVIS M, LIU W, MILLER P, et al. Detecting anomalise in graphs with numeric labels[C]//ACM Conference on Information and Knowledge Management. Glasgow, United Kingdom, 2011: 1197-1202.

[8] GOKHAN K, DUC L, TING X, et al. Ettu: analyzing query intents in corporate databases[C]//Proceedings of the 25th International Conference Companion on World Wide Web. Montreal, Canada, 2016: 463-466.

[9] TABISH R, IOANNIS A, JASON R. A new take on detecting insider threats: exploring the use of hidden markov models[C]//Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion. Limassol, Cyprus, 2016: 47-56.

[10] TED E S, DAVID A B, THOMAS G D, et al. Detecting insider threats in a real corporate database of computer usage activity[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 1393-1401.

[11] 王怀宝, 郭江利. 基于跟踪轨迹的徘徊行为分析[J]. 计算机与数字工程, 2016, 44(5): 843–846.
WANG Huaibao, GUO Jiangli. Wandering behavior analysis based on trajectory[J]. Computer and digital engineering, 2016, 44(5): 843–846.

[12] 邹一波, 陈一民. 基于运动标签的异常行为检测算法[J]. 计算机应用与软件, 2015, 5: 238–240, 266.
ZOU Yibo, CHEN Yimin. Anomalous behaviors detection algorithm based on motion label[J]. Computer applications and software, 2015, 5: 238–240, 266.

[13] HAN Jiawei, MICHELINE K, PEI Jian. Data mining concepts and techniques[M]. 3 版. 北京: 机械工业出版社: 2016: 355-356.

[14] BOSTJAN K, ERIK D, TEA T, et al. A probabilistic risk analysis for multimodal entry control[J]. Expert systems with applications, 2011, 38(6): 6696–6704.

[15] MICHAEL D, WEIRU L, PAUL M. Detecting anomalies in graphs with numeric labels[J]. ACM conference on information and knowledge management, 2011(10): 1197–1202.

[16] 胡向东, 韩恺敏, 许宏如. 智能家居物联网的安全性设计与验证[J]. 重庆邮电大学学报:自然科学版, 2016, 26(2): 171–176.
HU Xiangdong, Han Kaimin, XU Hongru. Design and implementation of security-focused intelligent household Internet of things[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 26(2): 171–176.

[17] 胡向东, 唐飞. 智能家居门禁系统的安全控制方法[J]. 重庆邮电大学学报：自然科学版, 2016, 28(6): 863–869.
HU Xiangdong, TANG Fei. Secure control methods of the entrance guard system for smart home[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 28(6): 863–869.

[18] 王菲. 数据挖掘在图书馆用户行为分析上的应用研究[D]. 上海: 上海交通大学, 2013: 26-49.
WANG Fei. Data mining applied in the library user behavior analysis[D]. Shanghai: Shanghai Jiao Tong University, 2013: 26-49.

[19] 郑伟平, 言专艺, 唐晓红. 电子门禁数据挖掘与应用方法[J]. 警察技术, 2015, 6: 47–50.
ZHENG Weiping, YAN Zhuanyi, TANG Xiaohong. Access control data mining and application methods[J]. Police technology, 2015, 6: 47–50.

[20] 史殿习, 李寒, 杨若松, 等. 用户日常频繁行为模式挖掘[J]. 国防科技大学学报, 2017, 39(1): 74–80.
SHI Dianxi, LI Han, YANG Ruosong, et al. Mining user frequent behavior patterns in daily life[J]. Journal of national university of defense technology, 2017, 39(1): 74–80.

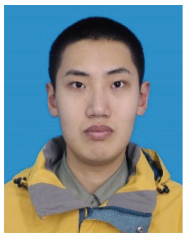
[21] 顾兆军, 安一然, 刘飞. 基于航站楼门禁日志挖掘的物理入侵检测技术[J]. 计算机应用与软件, 2015, 32(11): 317–320, 324.
GU Zhaojun, AN Yiran, LIU Fei. Physical intrusion detection technology based on terminal buildings access log mining[J]. Computer applications and software, 2015, 32(11): 317–320, 324.

[22] 陈卓, 杨炳儒, 宋威, 等. 序列模式挖掘综述[J]. 计算机应用研究, 2008, 25(7): 1960–1964.
CHEN Zhuo, YANG Bingru, SONG Wei, et al. Survey of

sequential pattern mining[J]. Application research of computers, 2008, 25(7): 1960–1964.

[23] HAN Jiawei, PEI Jian, BEHZAD M, et al. FreeSpan: frequent pattern-projected sequential pattern mining[C]// Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2000: 355-359.

作者简介:



王培超, 男, 1993 年生, 硕士研究生, 主要研究方向为网络空间数据挖掘, 参与国家自然科学基金面上项目 1 项, 教育部在线教育研究基金项目 1 项。

周鋆, 男, 1987 年生, 讲师, 博士, 主要研究方向为机器学习、贝叶斯网络学习及应用、网络空间的安全行为分析。发表学术论文 10 篇。

朱承, 男, 1976 年生, 研究员, 博士生导师, 博士, 中国指挥与控制学会 C4ISR 技术专委会总干事。主要研究方向为指挥控制、智能决策。主持国家自然科学基金项目 3 项、国家“863”计划项目 2 项, 担任多个国防重点型号项目的技术副总师, 获军队科研奖励 3 项。发表学术论文 30 余篇, 编著教材 3 部。