面向内部威胁检测的用户跨域行为模式挖掘

文 雨 王伟平 孟 丹
(中国科学院信息工程研究所 北京 100093)

摘 要 内部用户行为分析是系统安全领域中一个重要的研究问题.近期的工作主要集中在用户单域行为的单一模式分析技术,同时依赖于领域知识和用户背景,不适用于多检测域场景.文中提出一种新的用户跨域行为模式分析方法.该方法能够分析用户行为的多元模式.此外,该方法是完全数据驱动的方法,不需要依赖相关领域知识和用户背景属性.最后作者基于文中的用户行为模式分析方法设计了一种面向内部攻击的检测方法.在实验中,作者使用文中方法分析了真实场景中的5种用户审计日志,实验结果验证了文中分析方法在多检测域场景中分析用户行为多元模式的有效性,同时文中检测方法优于两种已有方法:单域检测方法和基于单一行为模式的检测方法.

关键词 内部威胁;多检测域;用户跨域行为分析;非负矩阵分解;高斯混合模型;机器学习
中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2016.01555

Mining User Cross-Domain Behavior Patterns for Insider Threat Detection

WEN Yu WANG Wei-Ping MENG Dan
(*Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

Abstract User behavior analysis is an important problem in the system security research filed. Recently existing work mainly focused on the single pattern analysis of user single-domain behavior, which needed to rely on expert's knowledge and user background knowledge. Thus, these work were not suitable for user behavior pattern analysis in the multi-domain scenarios. In this paper, we proposed a novel method for the user cross-domain behavior analysis. Our method could identify multi-pattern of user cross-domain behavior. Moreover, our method was a completely data driven resolution which did not need any expert's knowledge and user background knowledge. At last, we also designed an insider attack detection method based on our user behavior analysis approach. In our experiment, we used our methods to analyze and detect five user audit logs in real environment. The experimental results showed that our user behavior analysis method was effective on the multi-pattern analysis of the user cross-domain behavior in the multi-domain scenarios, and our insider attack detection method was better than two existing solutions: a single-domain detection method and a single patterns based detection method.

Keywords insider threat; multi-detection domain; user cross-domain behavior analysis; non-negative matrix factorization; Gaussian mixture model; machine learning

1 引 言

内部用户行为分析是系统安全领域中一个重要研究问题.近期许多安全事故中,内部用户攻击(Insider Attack)已经成为主要原因之一[1].内部用户通常指组织机构的内部人员,他们通常是组织机构中信息系统的用户,如政府雇员、企业员工等,或

收稿日期:2015-09-06;在线出版日期:2016-01-22.本课题得到国家"八六三"高技术研究发展计划项目基金(2013AA013204)资助.文 雨,男,1976年生,博士,主要研究方向为大数据、系统安全.E-mail:wenyu@iie.ac.cn.王伟平,男,1975年生,博士,研究员,博士生导师,中国计算机学会(CCF)会员,主要研究领域为数据库、大数据、系统安全.孟 丹,男,1965年生,博士,研究员,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为大数据、系统安全.

者公共服务的使用者,如数字图书馆的用户等.而用户或用户进程在计算机系统中的各种活动记录(又称为用户审计日志)是分析用户行为的重要依据,如用户的命令执行记录[2]、文件搜索记录[3]、数据库访问记录[1,4]、鼠标操作[5]等.本文将各种用于用户行为检测的审计日志产生环境统一称为检测"域",如用户命令域、数据库域、文件系统域等.此外,随着各界对计算机系统可问责性的重视[6-7],系统中检测域的种类呈现出越来越多样化的趋势.

已有许多工作提出面向内部威胁检测的用户行为分析方法,如文献[2,4-5,8].这些方法通常使用审计日志来分析用户在某个检测域的域内行为模式,然后基于这些被识别的用户行为模式检测用户的异常行为.因此这些工作主要关注用户单域行为分析.对于有技巧的攻击者,他们能够将攻击行为巧妙的分解为多个步骤,而且每步都被伪装成正常行为[9].这类攻击通常被称为复合攻击(Multi-step Attack).因此当恶意用户的攻击行为在不同检测域被独立分析时,极大可能被分别识别为无害的正常行为.例如,某个用户作为系统开发人员,需要每天登录一台计算机编写多个源代码文件.同时他还作为系统管理员,需要每天登录多台计算机进行系统文件查看和配置.因此对于系统登录域,该用户每天登录一台还是多台计算机都不算异常行为;而对于文件系统域,该用户读写源代码或系统文件也不算异常行为.因此,当该用户每天登录多台计算机并隐蔽的收集敏感数据时,各域独立检测将不易发现其异常行为.因此,检测系统需要具有集成不同检测域和分析用户跨域行为的能力.

然而,在复合攻击研究方面,大部分工作主要集中在基于主机和网络安全缺陷分析进行复合攻击路径计算[10-11],没有考虑用户行为分析和检测.最近,相关研究人员发现复合攻击具有一定的动态性,如攻击者可以利用窃取的信息和合法权限对无安全缺陷的主机和网络实施进一步的攻击.但是,这方面工作主要侧重于攻击路径的动态性分析,需要与用户行为检测技术结合起来实现攻击路径推断,如 Chen 等人[9]提出的基于概率攻击图的内部攻击意图推断方法.在面向多域的内部攻击检测工作方面,Maloof 等人[12]提出一种通过融合用户各单域行为检测结果进行内部攻击检测的方法.实际上,该工作仍然采用用户各单域行为分析的方式,同样没有考虑进行用户多域行为分析.

本文关注面向内部威胁检测的用户跨域行为模式分析问题.用户跨域行为模式包括用户域内行为模式以及用户各域域间行为的关联模式.用户跨域行为模式分析通过用户多域行为联合分析实现.用户跨域行为模式可用于识别和发现单域场景不能发现的用户异常行为.用户的行为模式通常由自身个性和习惯、所任职位、承担角色、所在部门、所从事的项目等众多因素所决定.同时,对于每个检测域,主导用户行为差异的主要因素可能不尽相同.例如,有些域可能由用户个性化特点或习惯所决定,因此不同用户具有不同的行为模式;而有些域可能由用户工作职责所决定,相同工作角色的人通常具有相同的行为模式.因此,用户跨域行为模式不会由单一的系统条件或用户背景属性所决定.此外,由于职位变动、角色变更、部门调动、项目变化等原因,用户行为模式也可能发生必要的演变.因此,用户跨域行为模式应该是多元的.然而,现有方法主要面向用户单域行为分析,通常基于领域知识[12]或用户背景属性,如用户标示[2,4-5,8]或用户角色[1,4,13]等,采用数据分析技术构建用户单一行为模式,如用户分类、典型行为等.显然这些方法不适用于用户多域行为联合分析的应用场景.

本文提出一种新的用户跨域行为模式分析方法.我们首先分别为各个检测域构建归一化的用户单域行为描述,并基于时间窗口通过集成各域域内行为特征构建用户多域行为描述.然后通过从用户多域行为描述中提取基模式,生成用户行为特征.基模式是那些未知的用户行为主导因素对用户各域行为影响的量化形式,而用户行为特征则量化了这些未知主导因素对每个用户行为的影响.最后,我们使用非监督学习技术挖掘多元的用户行为模式.在应用方面,我们设计一种基于用户跨域行为模式分析的内部攻击检测方法.该方法通过检测用户异常行为模式以及对用户正常行为模式产生负面影响的用户异常行为,来发现可能的内部攻击行为.本文的主要贡献有:

(1)基于多域异质审计日志融合的用户多域行为描述构建方法.我们在审计日志特征层面,通过用户单域行为描述(即单域日志特征)生成和集成,来刻画用户多域行为.

(2)结构化的用户行为特征生成方法.基模式使得用户行为潜在结构变得清晰,并使用户行为描述得到一定程度的约简,因此使得用户行为特征对用户行为的解释变得方便.同时,结构化的生成方式

使得用户行为特征粒度和表达倾向可以根据分析需要通过观察和计算进行调整.

(3) 完全基于数据驱动的用户行为多元模式分析. 不依赖任何领域知识和用户背景，完全以数据驱动的方式挖掘多元的用户行为模式.

(4) 基于本文用户行为模式分析的内部攻击检测方法. 该方法能够利用多元的用户跨域行为模式进行内部攻击检测.

本文第 2 节是问题描述；第 3 节介绍本文方法主要思路和设计；第 4 节是本文方法的实验部分；第 5 节是本文方法在内部威胁检测方面的应用；第 6 节是相关工作介绍；最后是本文的结论部分.

2　问题描述

本文的目标是提出一种面向多检测域场景的用户跨域行为模式挖掘方法. 用户跨域行为模式包含了用户域内行为模式以及各域内行为模式之间的关联模式. 本文方法应该能够：

(1) 构建用户多域行为描述. 由于各域用户审计日志相对独立的产生，它们通常在频率、数据均匀性等方面具有一定差异. 因此，分析方法应该能够融合异质的多域审计日志，并刻画用户多域行为.

(2) 分析用户行为多元模式. 由于用户各域域内行为受到不同因素的影响，用户多域行为不会由个别系统条件或用户属性所主导. 相对于单域场景，用户多域行为可能具有多种行为模式. 因此，分析方法应该能够不依赖任何已知的因素或领域知识，挖掘用户行为的多元模式.

3　用户跨域行为模式分析

本节主要介绍本文用户跨域行为模式分析方法的设计思路.

3.1　方法概述

本文方法包括 3 个主要步骤(如图 1 所示)：(1) 构建用户多域行为描述；(2) 生成用户行为特征；(3) 分析用户行为多元模式.

图 1　本文的用户跨域行为模式挖掘过程

我们首先构建用户多域行为描述. 最直观的构建方法是先将各域用户审计日志进行融合，然后通过数据特征提取生成用户多域行为描述. 然而，由于各域审计日志之间的异质性，直接融合多域审计日志并不现实. 因此，我们选择在审计日志的特征层面构建用户多域行为描述. 首先分别从各域审计日志中提取归一化的用户单域行为描述，然后再基于一个选定的时间窗口将这些单域行为进行集成来构建用户多域行为描述. 如有必要这里可以通过相关性分析进行日志特征筛选，如 Pearson 相关系数. 这方面的工作较多，本文不做赘述.

然后，我们设计了一种结构化的用户行为特征生成方法. 该方法将用户多域行为描述结构化为一组基模式和用户行为特征两个部分. 基模式是一种用户行为的底层抽象，也是刻画用户行为的基本单元. 用户行为特征则刻画了用户行为关于基模式的线性组成结构. 实际上，每个基模式可看作某种未知的用户行为主导因素对用户各域行为影响的一种量化形式，而用户行为特征则量化了这些未知主导因素对用户行为的影响. 由于用户多域行为描述、基模式和用户行为特征均为非负向量，本文使用非负矩阵分解(简称 NMF)方法计算基模式和用户行为特征.

最后，我们基于非监督学习技术进行用户行为多元模式分析. 正常情况下，当系统中用户行为主导因素稳定时，用户行为也相应地趋于稳定. 从统计学角度，我们假设用户行为特征符合多维正态分布，并在实验中验证了该假设. 此外，由于存在用户行为主导因素发生变化而带来的用户行为合理变化的情况，用户行为特征应该符合多个正态分布. 因此，本文使用高斯混和模型(简称 GMM)分析了用户行为的多元模式. GMM 模型能够从用户行为特征中构建多个高斯分布，其中每个高斯分布可作为一种用户行为模式.

3.2　构建用户多域行为描述

我们首先分别为各域用户审计日志提取归一化的数据特征，即用户单域行为描述. 用户审计日志各项属性的数据类型通常包括标签类型、数值类型和文本类型. 标签类型属性记录了名称、事件、操作等信息，如计算机名、用户登录事件、用户文件写操作等. 这类属性的取值之间没有任何顺序、大小等相关性. 对于这类属性，我们采用二进制向量表示它们的

特征值.每个可能取值对应特征向量的一个元素,当日志中该属性的此值出现时,相应的向量元素为"1",否则为"0".例如,假设用户文件操作属性包括读、写、新建、拷贝、删除5种操作,则它的特征可表示为"10000"(读)、"01000"(写)、"00100"(新建)、"00010"(拷贝)、"00001"(删除)5种向量.用户审计日志中的绝大多数属性属于标签类数据类型.

数值类型的属性由实数表示,如资源使用率、网络流量等.显然这种数值型属性的取值与标签型属性的二进制特征值在尺度上可能有很大差异.为了避免这种差异和统一特征形式,对于数值类型,我们依然采用二进制特征向量的表示方法.我们将数值型属性的取值范围划分成若干个长度相等的分段,每个分段对应特征向量的一个元素.当属性取值落入某个分段时,则相应的向量元素为"1",而其他元素为"0".例如某个属性的取值区间为[0.0,10.0),我们可将其分为5个等长分段,当属性取值为"6.5"时,落入第4分段,则属性特征表示为"00010".文本类型如邮件和文件内容,主要由词串组成.在某些场景中,由于有保护用户隐私和防止敏感信息泄露的要求,这些内容不宜被用于分析,因此本文对文本类型不作考虑.

在得到用户的单域行为特征后,我们基于一个时间窗口,统计合并同一时间窗口的用户所有单域行为描述,构建成用户多域行为描述.本文选择以天为单位的时间窗口.这里可以根据研究需要选择不同时间粒度的时间窗口,如以小时、周等为时间单位.

3.3 用户行为特征生成

这里需要首先介绍基模式的概念.基模式是一种用户行为的底层抽象,也是刻画用户行为的基本构成单元.每个基模式可看作某种未知的用户行为主导因素对用户各域行为影响的一种量化形式.而用户行为特征则刻画了用户行为关于基模式的线性组成结构,它量化了基模式代表的未知主导因素对用户行为的影响.

我们假设用户行为有 k 种基模式,则用户多域行为关于基模式线性组合可写为

$$u_a(i)=w_{a,0}(i)\boldsymbol{h}_0+w_{a,1}(i)\boldsymbol{h}_1+\cdots+w_{a,k}(i)\boldsymbol{h}_k \quad (1)$$

其中 $u_a(i)$ 表示用户 a 的第 i 个时间窗口的行为描述.$\boldsymbol{h}_j$ 是第 j 个基模式向量.$w_{a,j}(i)$ 表示相应基模式 $\boldsymbol{h}_j$ 的系数(权重).相应的,用户 a 的第 i 个时间窗口的行为特征可表示为

$$\boldsymbol{f}_a(i)=(w_{a,0}(i),w_{a,1}(i),\cdots,w_{a,k}(i)).$$

这里 $\boldsymbol{f}_a(i)$ 表示用户 a 在第 i 个时间窗口的行为特征.需要说明的是,基模式数量 k 的选取对用户行为分析具有一定的影响.k 值较小时,基模式数量相对较少,意味着考虑的用户行为主导因素较少.因此用户行为特征对用户行为的描述更宏观,利于不同用户之间的行为差异比较.而 k 值较大时,基模式数量相对较多,意味着考虑的用户行为主导因素较多.而此时用户行为特征对用户行为的描述更细致,利于用户行为个体变化的分析.

基模式规模可通过基模式的相关性分析来确定,如 Pearson 相关系数.我们首先设定一个相关性阈值.如果系统倾向于分析不同用户的行为差异,则基模式相关系数应小于该阈值.如果系统倾向于分析用户个体行为的变化,则基模式相关系数应大于该阈值.因此我们的方法可根据具体分析需要,通过观察和计算来确定基模式数量,并生成不同粒度和不同表达倾向的用户行为特征.

接下来我们根据已知的用户多域行为,求解基模式 $\boldsymbol{h}_0,\boldsymbol{h}_1,\cdots,\boldsymbol{h}_k$ 以及用户 a 的行为特征 $\boldsymbol{f}_a$.首先我们构建所有用户的多域行为模型为

$$\boldsymbol{U}_{n\times m}=\boldsymbol{W}_{n\times k}\boldsymbol{H}_{k\times m} \quad (2)$$

其中 $\boldsymbol{U}_{n\times m}$ 是所有用户的多域行为描述矩阵,每行代表一个用户多域行为描述向量,n 是用户行为特征向量集的规模,m 是用户多域行为描述长度.$\boldsymbol{W}_{n\times k}$ 是包含所有用户的用户行为特征矩阵,每行代表一个用户在某个时间窗口的行为特征,k 表示基模式集合的规模.$\boldsymbol{H}_{k\times m}$ 则是基模式矩阵,每行代表一个基模式向量.

基模式计算可通过求解优化问题实现.当给定基模式规模 k,计算可被表示为如下优化问题:

$$\min_{\boldsymbol{W},\boldsymbol{H}}\|\boldsymbol{U}-\boldsymbol{W}\boldsymbol{H}\|^2 \quad (3)$$

这里 $\|\cdot\|$ 表示 Frobenius-范数.该目标函数意义是求解最优的用户行为特征矩阵 $\boldsymbol{W}$ 和基模式矩阵 $\boldsymbol{H}$,它们的乘积与用户行为矩阵 $\boldsymbol{F}$ 的绝对误差最小.因为矩阵 $\boldsymbol{V}$ 是非负的,而且对于用户行为来说,纯加性(非负性)的矩阵 $\boldsymbol{W}$ 和矩阵 $\boldsymbol{H}$ 才有意义,则以上优化问题的求解实际上是经典的非负矩阵分解问题.

非负矩阵分解(简称 NMF)[14]是一种十分有效的数据处理方法,被广泛应用于文本挖掘、图像分析、推荐系统等领域.该方法将一个非负矩阵 $\boldsymbol{U}_{n\times m}$ 分解为两个低阶的非负矩阵 $\boldsymbol{W}_{n\times k}$ 和 $\boldsymbol{H}_{k\times m}$ 的乘积,

其中阶数 k 的选值远小于 m 和 n 的最小值. 虽然对于带有约束条件 $\boldsymbol{W},\boldsymbol{H}\geq 0$ 的式(3),同时求解矩阵 $\boldsymbol{W}$ 和矩阵 $\boldsymbol{H}$ 是一个非凸问题,但是通过分别求解矩阵 $\boldsymbol{W}$(设 $\boldsymbol{H}$ 固定)和矩阵 $\boldsymbol{H}$(设 $\boldsymbol{W}$ 固定),可将该问题转化为凸问题.

$$\boldsymbol{H}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}/\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} \tag{4}$$

$$\boldsymbol{W}=(\boldsymbol{H}\boldsymbol{U}^{\mathrm{T}}/\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{\mathrm{T}} \tag{5}$$

本文采用交替最小二乘法(简称 ALS)[15]计算上面的基模式矩阵和用户行为特征矩阵. ALS 方法具有简单和实用的优点. 我们首先使用文献[16]提出的方法初始化基模式矩阵 $\boldsymbol{W}$. 该初始化方法通过 k 次随机选取矩阵 $\boldsymbol{V}$ 的若干列向量并计算它们的均值,来初始化矩阵 $\boldsymbol{W}$. 假设 $\boldsymbol{W}$ 不变,使用式(4)求解矩阵 $\boldsymbol{H}$,然后将矩阵 $\boldsymbol{H}$ 的所有负值置为 0. 然后,假设矩阵 $\boldsymbol{H}$ 不变,使用式(5)求解矩阵 $\boldsymbol{W}$,同样将矩阵 $\boldsymbol{W}$ 中所有负值置为 0. 循环以上步骤直到计算收敛条件被满足为止,如矩阵误差小于某个阈值或分解计算达到一定的迭代阈值. 最后,分别对矩阵 $\boldsymbol{W}$ 和矩阵 $\boldsymbol{H}$ 执行标准化转换. 首先对矩阵 $\boldsymbol{W}$ 进行标准化转换,使得它的每个行向量模为 1. 然后根据矩阵 $\boldsymbol{W}$ 中标准化的行向量,调整矩阵 $\boldsymbol{H}$ 中对应的列向量.

3.4　用户行为多元模式分析

我们基于非监督学习技术进行用户行为多元模式分析. 本文中,非监督学习算法得到的用户行为特征聚簇即为用户行为模式. 非监督学习技术通常利用用户行为特征的相似度、聚集密度或者概率分布,对用户行为进行聚类分析. 特征的相似度通常采用距离指标进行度量,而聚集密度通常使用邻近特征规模进行刻画. 然而,从用户行为模式分析的角度,同模式用户行为特征应该具有一定的统计意义. 而基于用户行为相似度和聚集密度的分析方法,并不能保证得到的用户行为模式具有统计意义. 此外,系统中主导用户行为的各项因素稳定时,用户行为也应该趋于稳定. 因此,本文从统计学角度,假设同模式的用户行为特征符合正态分布. 在第 4.4 节,我们进行了该分布假设的检验实验,实验结果验证了用户行为特征符合正态分布.

由于用户行为具有多元模式,用户行为特征符合多个不同的正态分布. 因此,本文基于高斯混合模型(简称 GMM)[17]分析用户行为的多元模式. GMM 模型将用户行为特征的分布表示为多个高斯分布的线性组合. 其中,每个高斯分布代表一类用户行为特征,即一种用户行为模式. 同分布的用户行为特征的高斯分布函数可表示为

$$g(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{v})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{v})\right),$$

其中,$\boldsymbol{\Sigma}$ 为协方差矩阵,$\boldsymbol{v}$ 为均值向量,$\boldsymbol{x}$ 为同分布的用户行为特征向量,d 表示用户行为特征向量长度. 全体用户行为特征的分布密度表示为不同的高斯分布函数的加权线性组合:

$$G(x)=\sum_{i=1}^{m}\rho^{i}g^{i}(x),\ \sum_{i=1}^{m}\rho^{i}=1,$$

其中,x 表示任意的用户行为特征,$g(\cdot)$为高斯分布函数,ρ^{i}为第 i 个高斯分布的权值,m 表示高斯分布的数量. 实验中,我们采用交叉验证的方法来确定最佳的高斯分布数量.

GMM 模型的参数集 Θ 通常使用期望最大(EM)算法[17]估计得到. EM 算法基本思路是利用已知的用户行为特征 $\boldsymbol{X}=(f_a(0),f_a(1),\cdots,f_b(0),f_b(1),\cdots)$,从模型参数集的初始值 Θ_0 开始,估计一个新的参数集 Θ,使得在新的模型参数下样本的似然概率 $P(\boldsymbol{X}|\Theta)\geq P(\boldsymbol{X}|\Theta_0)$. 新的模型参数再作为当前参数进行训练,依次迭代运算直到模型收敛为止.

4　用户行为模式分析实验

本文的用户行为模式分析平台是一台有 4 颗 6 核 1 GHz AMD Opteron 处理器,64 GB 内存,1 TB 磁盘和 10 Gb 以太网卡的服务器. 服务器运行的软件包括 Red Hat Enterprise Linux 6.2 操作系统(2.6.32-220.el6.x86_64 SMP 内核),Python 2.7 运行时环境.

4.1　数据集

本文实验中,我们历时 3 个月收集了 21 名用户(编号 U1～U21)的 5 类审计日志,包括计算机登录、文件访问、USB 设备操作、CD/DVD 介质访问和打印机操作(见表 1). 系统登录日志记录了用户登入办公计算机的时间、计算机名(或地址)、用户名等信息. 文件访问日志记录了用户的新建、打开、修改、删除文件等操作. CD/DVD 介质访问日志记录了用户的 CD 和 DVD 介质使用信息. USB 设备操作日志记录了 USB 设备信息、用户插拔 USB 设备以及读写设备内容等操作. 打印机日志记录了打印机相关信息以及用户的文件打印等操作.

表 1 5 种审计日志数据集

编号	域	总数/条
1	计算机登录	3775
2	文件访问	43045
3	介质访问	414
4	USB 设备操作	1058
5	打印机操作	867

根据 5 类审计日志，我们生成 5 种用户域内行为特征(见表 2). 计算机登录特征包括用户登录操作和登出操作，向量长度为 2. 文件访问特征包括文件类型(共 5 类)、新建操作、打开操作、修改操作、拷贝操作、删除操作，向量长度为 10. 介质访问特征包括介质类型(共 2 类)、介质读、介质写，向量长度为 4. 设备操作特征包括设备类型、设备插入、设备拔下、设备读、设备写，向量长度为 5. 打印机操作特征包括文件类型(共 5 类)，向量长度为 5. 然后，我们以天为时间窗口，通过统计和合并 5 类域内行为特征，为每个用户生成跨域行为特征，特征向量长度为 26.

表 2 用户单域行为(日志特征)

编号	域	特征	长度
1	计算机登录	登录操作、登出操作	2
2	文件访问	文件类型、新建、打开、修改、拷贝、删除、修改权限、写入权限、读取权限、读取和执行权限	14
3	介质访问	介质类型、介质读、介质写	4
4	USB 设备操作	设备类型、插入、拔下、设备读、设备写	5
5	打印机操作	文件类型	5

我们使用 Pearson 系数对 30 个特征的相关性进行分析. 图 2 的计算结果显示绝大多数特征的相关系数集中在 0.1～0.3 区间，说明它们的相关性较小. 只有少量特征的相关系数达到了 0.9，如计算机登录域的用户登录操作和登出操作，以及 USB 设备域的设备插入操作和拔出操作. 最后我们删除了系统登录域的用户登出操作和 USB 设备域的设备拔出操作，保留其他 28 个特征.

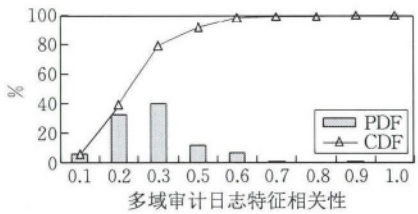

图 2 多域审计日志特征相关系数分布

4.2 用户行为基模式和用户行为特征计算

在生成用户行为特征过程中，关键是选取合适的基模式数量参数 k. 第 3.4 节我们提到，k 取值较小时，对用户行为的描述更宏观，利于不同用户之间的行为差异比较. 而 k 取值较大时，对用户行为的描述更细致，利于用户个体行为变化的分析. 因此，我们对比了 3 种规模(即 k 分别取值为 2、3 和 4)时的基模式计算结果.

图 3 显示了不同基模式数量的实验结果. 可以看出，$k=2$ 和 $k=3$ 时，基模式之间具有十分明显的差异. 而 $k=4$ 时，模式 1 与模式 3 相似的特征较多. 同时考虑到基模式数量较多时，便于用户个体行为变化的分析，最后我们在实验中选择 $k=3$ 的基模式计算结果. 从图 3 中还可看出，基模式能够很好地表达用户域内和域间行为的不同关系. 在我们选择的基模式中，基模式 0 突出了文件域中第 1、3 种文件类型、文件读操作、文件读取权限和设备域中设备插入操作等特征的关系；基模式 1 突出了文件域的第 2 种文件类型、文件写入权限、介质域的介质写操作等特征的关系；而基模式 2 突出了文件域的第 4 种文件类型、文件拷贝操作、文件修改权限和介质域的介质类型等特征的关系.

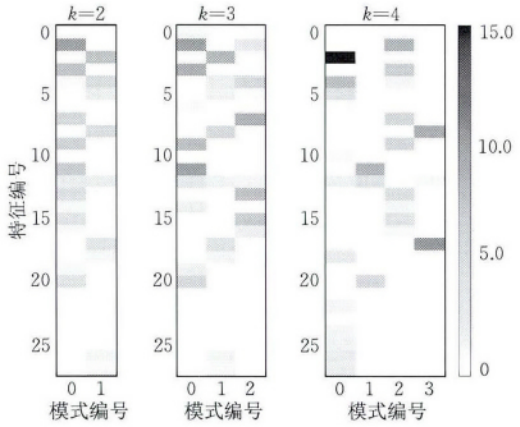

图 3 基模式规模系数 k 分别为 2、3、4 时的基模式计算结果比较(本文实验选择 $k=3$ 时的基模式)

我们进一步考察了用户行为特征的计算结果. 图 4 是所有用户行为特征计算结果. 可以看出，多数用户之间的行为差异较明显. 其次，大部分用户行为具有较明显的阶段性变化，如用户 U2 行为在第 54 天前后有较显著的不同，用户 U7 行为从第 36 天有明显的变化，用户 U20 行为在第 28 天前后也有显著差异等. 因此，基于基模式的用户行为特征，不仅能够表现不同用户之间行为的异同，同时能够较好地展现每个用户个体行为的变化.

由以上观察，我们知道用户行为具有明显的多元模式. 不可避免的，这种情况会给用户单一行为模

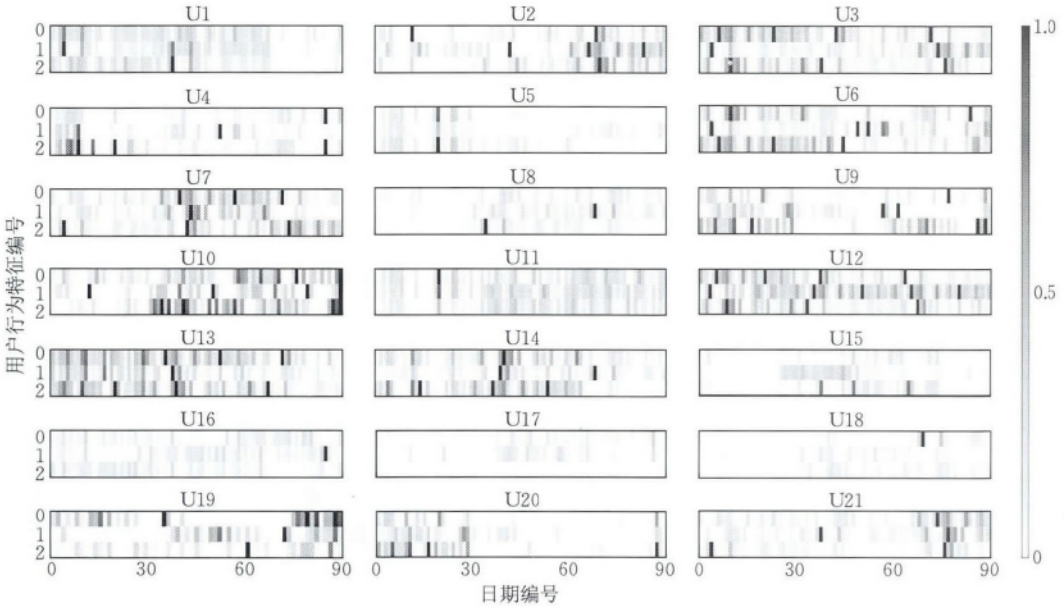

图 4　用户行为特征计算结果. 基于基模式的用户行为特征不仅能够表现不同用户之间行为的异同，同时能够较好地展现每个用户个体行为的变化

式分析带来困难. 因为单一行为模式分析方法很大可能仅能识别其中一种较强的行为模式，且得到的模式对于用户行为特点描述也较笼统. 而用户正常行为模式对于绝大多数攻击行为检测方法至关重要. 因此，单一行为模式分析会严重影响攻击检测方法的检测性能，这给识别出真正的攻击行为带来难度. 根据以上分析，本文强调用户行为的多元模式分析. 多元模式分析方法能够尽量全面刻画用户行为特点. 当应用于攻击检测时，相对于单一行为模式，能够较容易地发现用户异常行为.

4.3　用户行为多元模式分析

应用 GMM 模型分析用户行为模式，需要考虑选择合适的协方差约束方式和高斯分布数量. 实验中，我们采用交叉验证的方法寻找最佳的 GMM 模型. 我们使用常见的贝叶斯信息准则(BIC)[18]作为模型选择指标. 我们比较了分别采用 4 种协方差矩阵(spherical、diagonal、full 和 tied)和分别具有 1～20 个高斯分布的 GMM 模型的 BIC 指标. 通常情况下，模型的 BIC 指标越小说明模型的效果越好. 图 5 是不同模型 BIC 指标的对比. 其中使用 diagonal 协方差矩阵和具有 15 个高斯分布的 GMM 模型的 BIC 指标最小，故我们选择此模型进行用户行为模式分析.

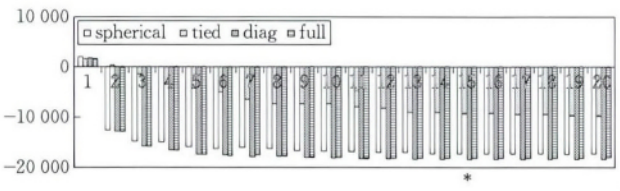

图 5　不同高斯分布规模和协方差的高斯混合模型(GMM)的贝叶斯信息准则(BIC)指标比较

图 6 是基于 GMM 模型的用户行为模式分析结果的可视化效果. 其中，每个 GMM 簇代表一种用户行为模式，共有 15 个模式.

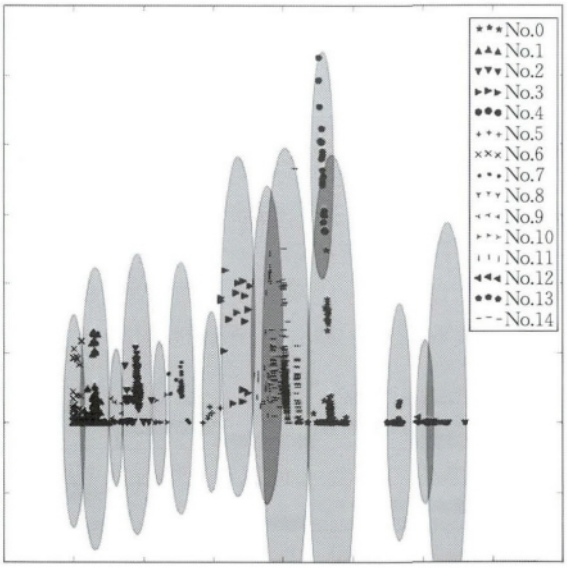

图 6　基于高斯混合模型(GMM)的用户行为多元模式分析结果

为了进一步了解这些用户行为模式的基本含义，我们求得每个模式中所有用户行为特征的平均值，并作为该模式的典型行为特征，如图 7 所示. 基于前面的基模式类型，我们可以知道模式 0 包含了大量的从不同介质拷贝第 4 种文件类型的操作，同

时包含了较多向介质写第 2 种文件类型的操作. 模式 1 包含了大量向介质写第 2 种文件类型的操作，而其他操作几乎没有. 模式 2 包含了大量的从不同介质拷贝第 4 种文件类型的操作，同时有少量从设备中读取第 1、3 种文件类型的操作. 模式 3 包含了大量向介质写第 2 种文件类型的操作，同时还有少量从设备中读取第 1、3 种文件类型的操作. 模式 4 包含了大量从设备中读取第 1、3 种文件类型的操作，同时还有少量向介质写第 2 种文件类型的操作. 模式 5 包含了大量向介质写第 2 种文件类型的操作，同时有少量从不同介质拷贝第 4 种文件类型的操作. 模式 6 包含了少量向介质写第 2 种文件类型的操作，而其他操作较少. 模式 7 包含了大量的从不同介质拷贝第 4 种文件类型的操作，同时包含了少量向介质写第 2 种文件类型的操作. 模式 8 包含了极少的向介质写第 2 种文件类型的操作，同时几乎没有其他操作. 模式 9 包含了大量从设备中读取第 1、3 种文件类型的操作，同时还有极少的从不同介质拷贝第 4 种文件类型的操作. 模式 10 包含了少量的从不同介质拷贝第 4 种文件类型的操作，向介质写第 2 种文件类型的操作和从设备中读取第 1、3 种文件类型的操作. 模式 11 包含了大量的从不同介质拷贝第 4 种文件类型的操作，同时包含了较多的从设备中读取第 1、3 种文件类型的操作，和少量的向介质写第 2 种文件类型的操作. 模式 12 包含了较多的从设备中读取第 1、3 种文件类型的操作，同时还有少量向介质写第 2 种文件类型的操作. 模式 13 包含了少量的从不同介质拷贝第 4 种文件类型的操作和向介质写第 2 种文件类型的操作. 模式 14 包含了较多的从不同介质拷贝第 4 种文件类型的操作，同时包含了少量的向介质写第 2 种文件类型的操作和从设备中读取第 1、3 种文件类型的操作.

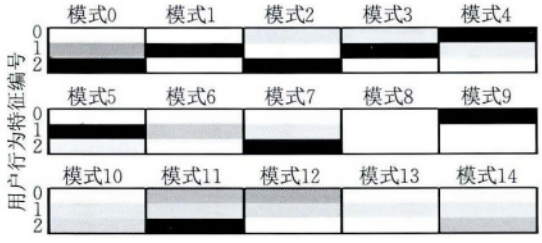

图 7 用户行为模式的典型行为特征

为了进一步验证用户行为的多元模式，我们分析了每个用户涉及的行为模式及其比重. 图 8 是每个用户关于已识别的 15 种模式的行为构成分析. 该图显示，所有用户的行为具有多元模式的特点. 图 9 是所有用户的模式占比分析结果. 其中，约 71.4% 用户的模式占比标准差小于 20%，其余用户的模式占比标准差也在 60% 左右. 结果说明，绝大多数用户行为不仅具有多元模式的特点，而且同时受多种模式主导. 因此单一模式不足以刻画用户的跨域行为特点.

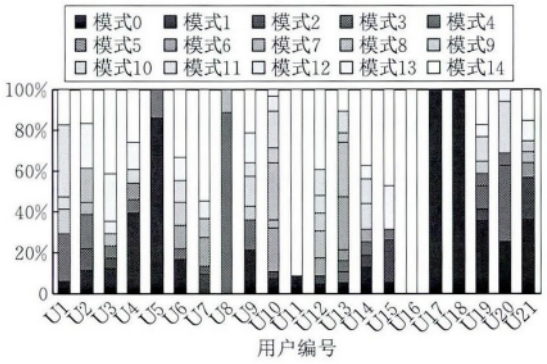

图 8 用户行为关于已知行为模式的构成成分分析

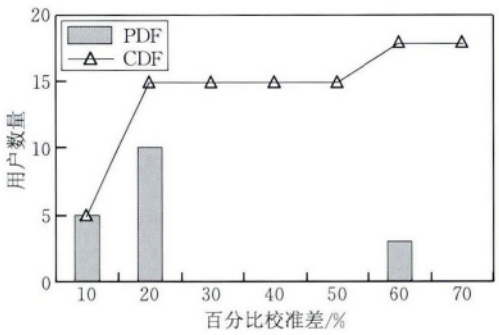

图 9 用户行为模式占比标准差分布

4.4 用户行为特征分布假设检验

实验中，我们对用户行为特征的高斯分布假设进行了假设检验. 对于上节实验得到的用户行为模式，我们采用正态概率图给出直观的检验结果，同时采用 Shapiro-Wilk 检验法进行了拟合性检验. 正态概率图描述了同模式的用户行为特征与标准正态分布的函数关系. 如果这些用户行为特征符合标准的正态分布，则它们的正态概率图将是一条直线. 而采用基于频率统计的 Shapiro-Wilk 检验法，我们可以量化的检验同模式用户行为特征分布的正态性. 该检验法给出称为 p-$value$ 的正态性检验指标. 当该指标高于某个阈值时(通常为 0.05)，则认为同模式用户行为特征的分布符合正态分布.

图 10 是假设检验结果，包括了每类模式中用户行为特征的正态概率图以及相应的 p-$value$ 检验指标. 不难看出，各模式的正态概率图基本上构成了一条直线. 同时，各模式的 p-$value$ 指标均高于 0.05. 实验说明用户行为模式符合高斯分布假设.

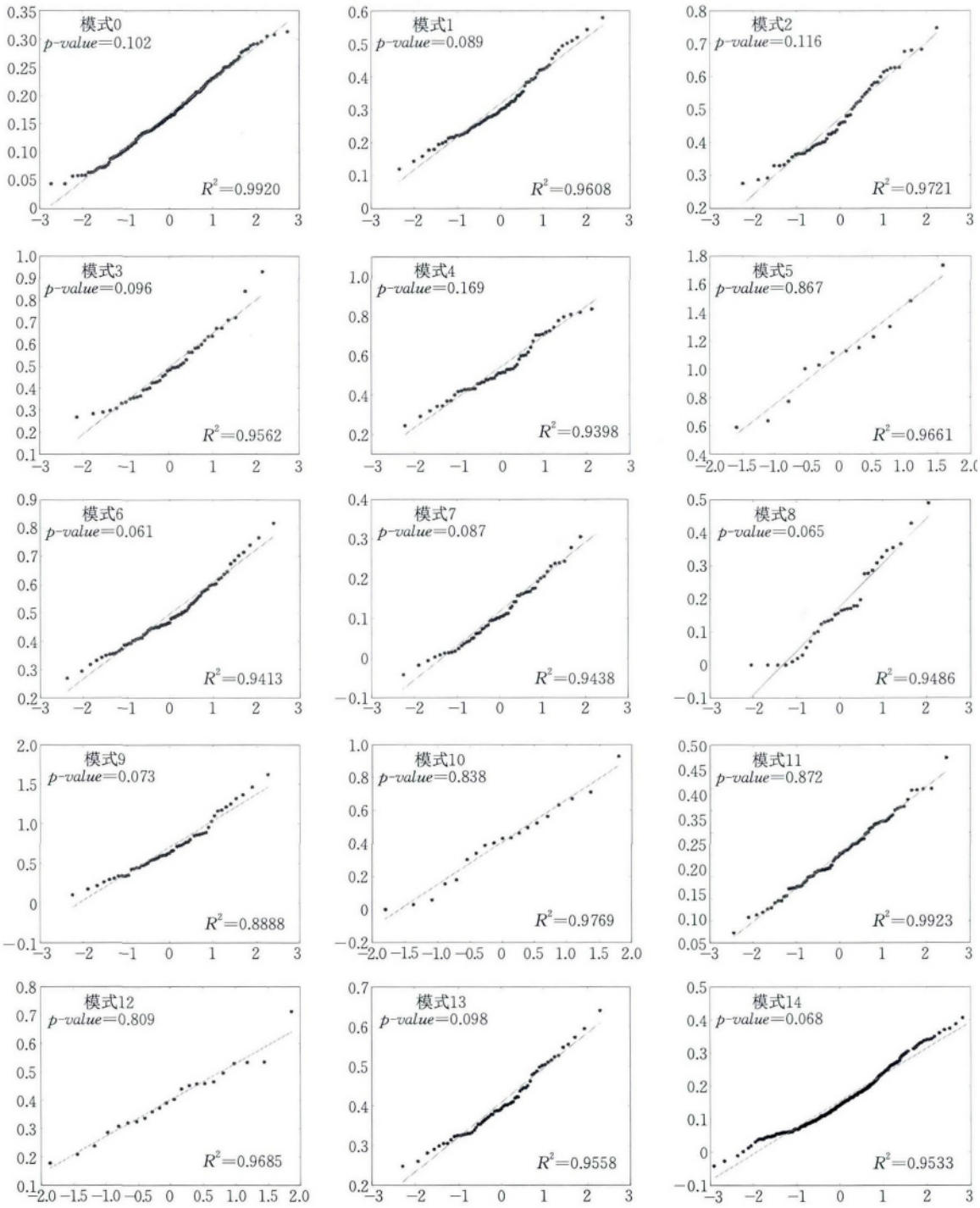

图 10　用户行为模式的高斯分布假设检验结果

5　基于用户行为模式的内部攻击检测

基于以上的用户行为模式分析结果，本文主要针对被较多关注的伪装攻击，设计了一种攻击检测方法. 伪装攻击是指恶意的内部用户通过某种手段（如猜密码、种木马、漏洞扫描等）非法窃取其他合法用户的账户和口令，并利用这些账户伪装成合法用户对系统实施的攻击.

5.1　伪装攻击检测

伪装攻击的检测基础是，恶意用户的行为与被利用的伪装对象的行为存在一定的差异. 首先，恶意用户与被利用对象存在天然的个体和背景差异. 其次，恶意用户的目的是系统攻击，而被利用对象的行为则主要围绕日常的系统使用. 因此二者使用系统的方式和习惯会存在一定程度的不同.

基于以上思想，我们利用前面介绍的用户行为模式分析方法设计了一种新的面向多检测域的伪装攻击检测方法. 从用户行为模式的角度看，攻击行为可能带来两种结果，与用户正常行为差异较大，直接产生异常行为模式，或者与用户正常行为差异较小，间接对正常模式产生负面影响. 因此，对伪装攻击的检测主要包括两个方面：异常行为模式检测和正常行为模式的干扰检测. 通常攻击行为的频度远小于正常行为，因此异常模式是一些小的行为模式. 在 GMM 模型中，则是一些稀疏且规模较小的簇. 在检测过程中，我们通过设定一个异常行为模式阈值，用以区分用户的正常行为模式和异常行为模式. 在 GMM 模型中，该阈值是簇规模的下限，低于该阈值的簇则是异常行为模式. 异常模式所包含的用户行为则被认为是攻击行为.

在行为模式干扰检测方面，我们通过检验每个行为特征向量对其所属模式高斯分布的影响是否有利. 同样的，因为攻击行为的频度远小于正常行为，因此相对于正常行为更符合模式的高斯分布，而攻击行为则会减弱模式的高斯分布符合性. 基于第 4.4 节介绍的 $p\text{-}value$ 指标，我们设计了一个关于行为特征向量的模式支持度指标，用于计算每个行为特征向量对模式的高斯分布的支持程度：$s_{c,i}=p_{c,i}-p'_{c,i}$. 其中 $s_{c,i}$ 表示编号为 i 的行为特征向量对模式 c 的高斯分布支持度. $p_{c,i}$ 表示模式 c 包含了行为特征向量 i 的高斯分布假设检验 $p\text{-}value$ 指标，而 $p'_{c,i}$ 表示没有包含行为特征向量 i 的 $p\text{-}value$ 指标. 由于 $p\text{-}value$ 实际是一种概率指标，$s_{c,i}$ 的计算结果处于 $[-1,1]$ 区间. 如果 $s_{c,i}$ 指标越高，则行为特征向量 i 对模式 c 的高斯分布支持的越好，表示该行为越更大可能是正常行为，反之则越差，表示该行为更大可能是攻击行为. 通过设定一个模式支持度阈值，我们可以区分用户正常行为和攻击行为. 模式支持度指标高于该阈值的行为被认为是正常行为，而低于该阈值的行为则被认为是攻击行为.

检测算法如算法 1 所示. 首先将一个用户异常行为集 A 初始化为空集（第 1 行）. 接着使用 GMM 模型分析用户行为多元模式，并得到模式集合 C（第 2 行）. 然后分别对模式集 C 中各模式 c 进行攻击检测分析（第 3～15 行）. 首先进行异常模式检测（第 4～6 行）. 如果模式 c 的规模低于异常模式阈值 T_c，则将该模式识别为异常模式，并将该模式编号和其中所有行为特征向量编号记入异常行为集 A 中，然后转入对下一个模式的检测分析. 如果模式 c 规模高于异常模式阈值 T_c，则为正常模式，需要继续进行模式干扰检测分析（第 8～15 行）. 首先计算模式中每个行为特征向量的模式支持度指标. 如果该指标低于模式支持度阈值 T_s，则该特征向量被识别为异常行为，并记入异常行为集 A 中. 否则为正常行为，继续下一个行为特征向量检测. 最后返回异常行为集 A（第 16 行）.

算法 1. 伪装攻击检测算法.

输入：数据集 D，异常模式阈值 T_c，模式支持度阈值 T_s

输出：用户异常行为集 A

```
1.  A<-{}
2.  C<- GMM(D)
3.  FOR c IN C:
4.    IF c.points.size<T_c:
5.      FOR i IN range(c.points.size):
6.        A.append([c.name,i])
7.      CONTINUE
8.    FOR i IN range(c.points.size):
9.      points<-c.points
10.     points.delete(i)
11.     p<-p-value(c.model,c.points)
12.     p'<-p-value(c.model,points)
13.     s<-p-p'
14.     IF s<T_S:
15.       A.append([c.name,i])
16. RETURN A
```

6 内部攻击检测实验

6.1 数据集

在伪装攻击模拟方面，我们借鉴了 Schonlau 等人[19]在用户日志中注入攻击数据的方法. 该方法随机将用户分为被攻击对象和攻击者，然后将攻击者的部分日志作为攻击行为插入到被攻击用户的正常日志中. 实验中，我们随机选取了 6 名用户作为攻击者（U5/U8/U11/U16/U17/U18），并随机抽取每个攻击者 5 天的日志作为攻击数据（T1～T5）. 然后将其余 15 名用户作为被攻击对象，并将攻击数据随机、平均的插入到被攻击用户的日志数据中. 生成的攻击数据集见表 3 描述.

表 3　伪装攻击数据集

目标用户	攻击者	攻击日志(与攻击者对应)
U1	U17/U5	T3/T5
U2	U11/U8	T2/T3
U3	U17/U11	T5/T3
U4	U18/U16	T4/T2
U6	U17/U17	T4/T1
U7	U8/U18	T5/T1
U9	U5/U18	T2/T2
U10	U16/U11	T3/T4
U12	U5/U11	T1/T5
U13	U18/U5	T5/T4
U14	U16/U17	T4/T2
U15	U5/U8	T3/T4
U19	U18/U8	T3/T1
U20	U16/U16	T1/T5
U21	U8/U11	T2/T1

6.2　实验结果

我们对比了两种攻击检测方法与本文方法的实验结果. 第 1 种方法是基于用户单域行为多元模式的检测方法，称为 SDMP-GMM 方法. 基于第 5.1 节介绍的检测方法，SDMP-GMM 方法分别针对每个用户的各单域行为进行攻击检测，包括表 1 中的 5 类检测域. 该方法使用第 3.2 节介绍的用户单域行为描述作为用户单域行为特征. 第 2 种方法是基于用户跨域行为单一模式的检测方法，称为 CDSP-SVM 方法. 该方法类似于 Kamra 等人[4]和 Zheng 等人[5]提出的基于用户单一行为模式的攻击行为检测方法. CDSP-SVM 方法使用第 3.3 节介绍的用户跨域行为特征，通过标准的 SVM 分类器挖掘每个用户的单一行为模式，并使用这个分类器判断用户待检行为是否异常. 实验中，SVM 分类器采用了多阶的多项式核，并通过交叉验证的方式评价检测效果. SVM 分类器的训练数据由被攻击用户和其他用户的部分日志数据构成，而测试数据包括被攻击用户的其余日志数据和攻击者的日志数据. 最后，为了方便对比，本文方法称为 CDMP-GMM 方法. 实验中，CDMP-GMM 方法分别对每个用户的跨域行为进行了攻击检测.

在检测结果评价方面，我们采用了多种评价指标，除了基本的查全率 TP(True Positive)指标和错检率 FP(False Positive)指标，还包括常见的 ROC 曲线及其 AUC 面积. TP 指标是被检测出的攻击行为与攻击行为总量的比例，而 FP 指标是被误断的用户正常行为与用户正常行为总量的比例. AUC 指标能够综合反映检测效果的 FP 指标和 TP 指标. 检测方法的 AUC 值越趋近于 1，则检测效果越好. 对于 SDMP-GMM 方法和 CDMP-GMM 方法，可以通过同时对所有用户统一调节异常模式阈值和模式支持度阈值来得到两种方法的所有实验结果，因此我们可以使用 AUC 指标对它们的实验结果进行对比. 而 CDSP-SVM 方法的检测效果依赖于训练数据的选取，因此我们使用 FP 指标和 TP 指标对该方法多次实验后最理想的检测结果进行评价. 同时，与 SDMP-GMM 方法进行对比时不同，CDMP-GMM 方法分别调整对每个用户的异常模式阈值和模式支持度阈值.

图 11 对比了用户跨域行为检测(CDMP-GMM 方法)和单域行为检测(SDMP-GMM 方法)的实验结果. 可以看出，在 AUC 指标方面 CDMP-GMM 方法(0.95)优于 SDMP-GMM 方法分别在计算机登录域、文件域、USB 设备域、介质域和打印机域的检测效果(分别为 0.78、0.86、0.81、0.71 和 0.68). 实验说明，相比单域行为模式，用户跨域行为模式更有助于内部攻击检测分析.

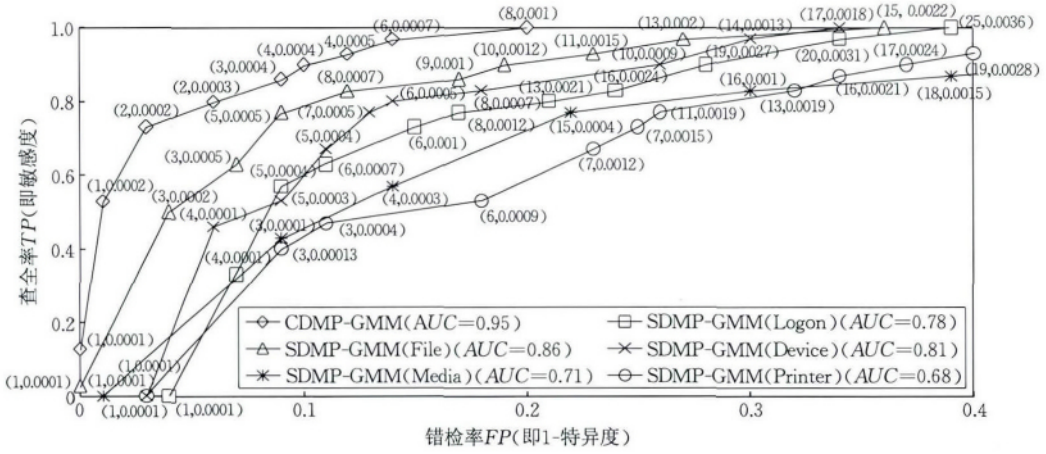

图 11　用户跨域行为检测(CDMP-GMM)与用户单域行为检测(SDMP-GMM)的实验结果对比

图 12 对比了基于用户行为多元模式的检测(CDMP-GMM 方法)与典型的基于用户行为单一模式的检测(CDSP-SVM 方法)分别对各用户的实验结果. 在 TP 指标方面,CDSP-SVM 方法仅检出用户 U1、U12、U15、U19 和 U20 的所有攻击行为,其余 10 名用户仅检出 50%的攻击行为($TP=50\%$),而 CDMP-GMM 方法能够检出所有用户的所有攻击行为($TP=100\%$). 在 FP 指标方面,CDSP-SVM 方法对用户 U3、U9、U10、U13 和 U21 的检测结果在 20%之内,对用户 U7 的检测结果在 30%之内,对用户 U1、U4、U6、U12、U19 和 U20 的检测结果在 40%以内,对用户 U2 的检测结果在 50%以内,

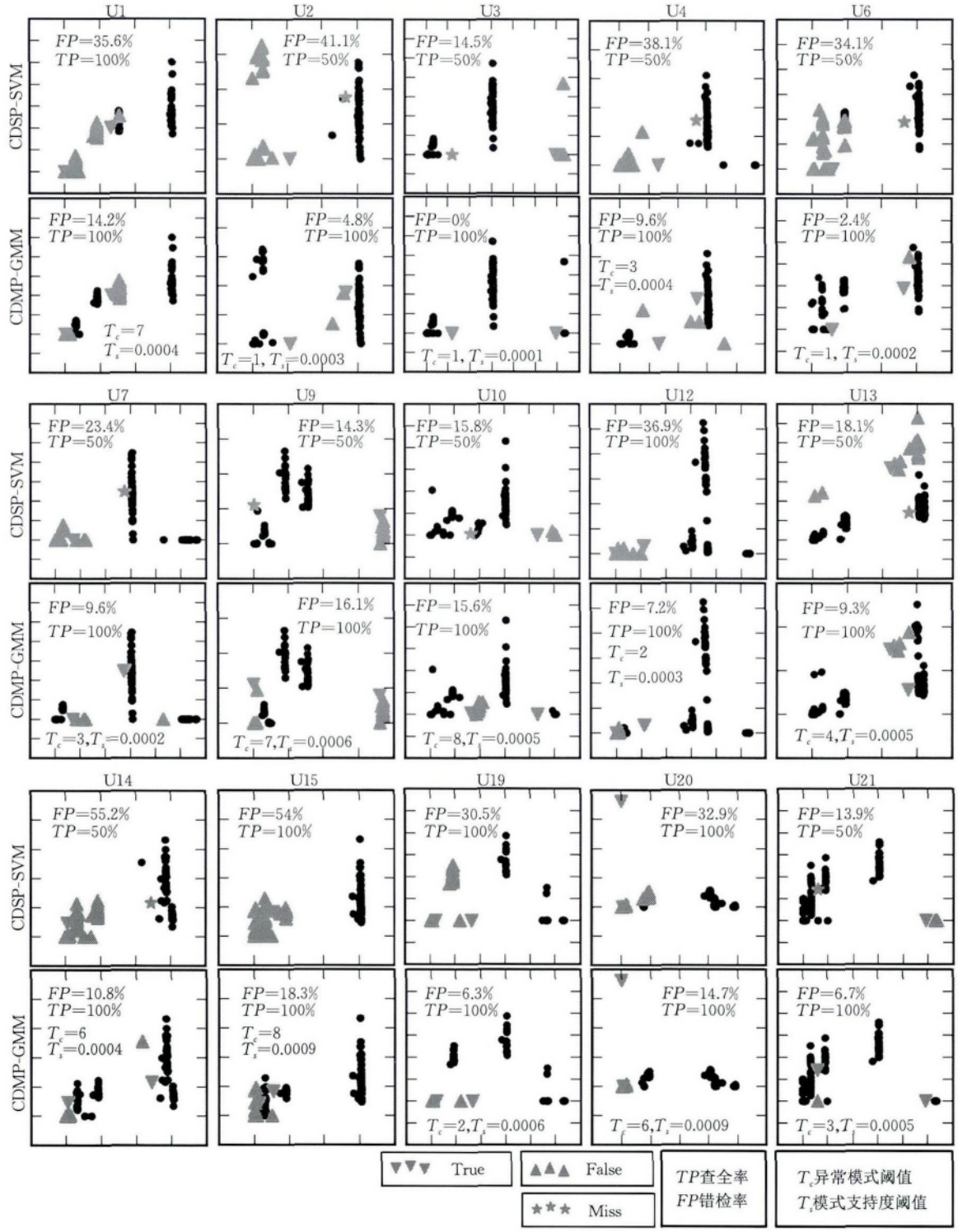

图 12 基于用户行为多元模式的检测(CDMP-GMM 方法)与典型的基于用户行为单一模式的检测(SDMP-GMM 方法)的实验结果对比

对用户 U14 和 U15 的检测结果超过 50%. 而 CDMP-GMM 方法仅对用户 U1、U9、U10、U14、U15 和 U20 的检测 *FP* 指标在 20%以内，对其余 9 名用户的检测结果均在 10%以内. 需要注意的是，对于用户 U9 的检测，虽然 CDSP-SVM 方法的 *FP* 指标(14.3%)优于 CDMP-GMM 方法(16.1%)，但是后者的 *TP* 指标(100%)却显著优于前者(50%). 对于用户 U10，尽管两种的 *FP* 指标比较接近，但是在 *TP* 指标上，CDMP-GMM 方法(100%)却显著优于 CDSP-SVM 方法(50%). 由此可知，基于用户行为多元模式的检测方法优于基于单一模式的检测方法.

7　相关工作

7.1　用户单域行为分析

用于异常行为检测的用户单域行为模式分析方法通常依赖于某个用户属性来构建用户在某个检测域的域内正常行为模式，如用户标示、用户角色等. Kamra 等人[4]提出分别基于监督式学习和非监督式学习的用户数据库访问模式分析方法. 基于监督式学习的方法根据用户角色，使用朴素贝叶斯模型为每类角色构建正常的数据库访问模式. 而基于非监督式学习的方法在用户角色未知的情况下，根据用户标识，使用标准的聚类技术分别为每个用户构建典型的数据库访问行为，例如将规模最大的聚类作为用户的正常行为模式. Mathew 等人[1]和 Islam 等人[13]同样基于用户角色，分别使用 k-means 聚类技术和隐马尔可夫模型构建用户数据库查询行为模式. 而 Maxion 等人[2]和 Kholidy 等人[8]根据用户标示，分别使用朴素贝叶斯文本分类方法和半全局排列算法构建用户个性化的命令执行序列模式. Zheng 等人[5]根据用户标示，使用支持向量机分类器构建用户特征的鼠标移动模式. 与以上工作不同，本文方法不依赖于任何领域知识和用户背景，完全基于数据驱动方式实现用户多域行为模式分析. 其次，本文考虑用户多域行为的多元模式分析，而以上方法通常只是分析用户单一的正常行为模式.

7.2　基于多域的系统威胁检测技术

基于多域的系统威胁检测工作主要集中在各域检测结果的融合技术. Maloof 等人[12]提出一种通过融合多域用户行为检测结果的内网威胁检测方法. 该方法针对的内部恶意用户，能够利用自身和他人的合法权限，从系统的多个方面实施攻击. 该工作在内网中收集了用户在多种检测域中的行为事件，包括文件共享、http 访问、邮件和文件传输等多种用户日志. 然后基于领域知识分别为每类事件的每个属性设计一个检测器. 最后，设计了一个贝叶斯推理网络模型，用于综合所有检测器的报警信息来给用户行为评分. 虽然该方法涉及多个检测域，但是它需要分别检测用户事件各个属性，因此在用户行为检测方面仍然采用各域分离的用户行为分析方式. 其次，该方法需要依赖一定领域知识和已知异常模式来辅助设计各个检测器，因此不能检测未知的用户异常行为. Maggi 等人[20]提出一种面向网络入侵检测的多域报警关联技术. 该工作设计了一个将报警事件流和时间戳刻画为随机变量的报警产生统计模型，并使用统计测试方法构建了用于区分相关报警和无关报警的准则. 虽然该方法不依赖领域知识和已知异常模式，但是仍然需要基于各域独立的报警结果来检测网络异常. 与以上工作不同，本文方法在用户行为特征层面融合用户多域行为，能够集中分析用户多域行为，并且不依赖于领域知识和已知模式，完全采用数据驱动的方式进行用户行为模式分析.

8　结　论

内部用户行为分析是系统安全领域中一个重要的研究问题. 近期的工作主要集中在用户单域行为分析技术，不适用于多检测域场景. 本文提出一种新的用户跨域行为模式分析方法. 首先基于多域异质审计日志融合构建了用户多域行为描述，然后设计了一种结构化的用户行为特征提取方法，最后完全基于数据驱动的方式分析用户行为多元模式. 应用方面，我们设计了一个基于本文用户行为模式分析的内部攻击检测方法. 在实验中，我们使用本文方法分析了真实场景中的 5 种用户审计日志，实验结果验证了本文分析方法在多检测域场景分析用户行为多元模式的有效性，本文检测方法优于单域检测方法和典型的基于单一用户行为模式的检测方法.

参　考　文　献

[1] Mathew S, Petropoulos M, Ngo H Q, Upadhyaya S J. A data-centric approach to insider attack detection in database systems//Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection (RAID). Ottawa, Canada, 2010: 382-401

[2] Maxion R A, Townsend T N. Masquerade detection using truncated command lines//Proceedings of the International Conference on Dependable Systems and Networks (DSN). Bethesda, USA, 2002: 219-228

[3] Salem M B, Stolfo S J. Modeling user search behavior for masquerade detection//Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection (RAID). Menlo Park, USA, 2011: 181-200

[4] Kamra A, Terzi E, Bertino E. Detecting anomalous access patterns in relational databases. The VLDB Journal, 2008, 17(5): 1063-1077

[5] Zheng Nan, Paloski A, Wang Haining. An efficient user verification system via mouse movements//Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS). Chicago, USA, 2011: 139-150

[6] Sundareswaran S, Squicciarini A C, Lin D. Ensuring distributed accountability for data sharing in the cloud. IEEE Transactions on Dependable and Secure Computing, 2012, 9(4): 556-568

[7] Weitzner D J, Abelson H, Berners-Lee T, et al. Information accountability. Communications of the ACM, 2008, 51(6): 82-87

[8] Kholidy H A, Baiardi F, Hariri S. DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks. IEEE Transactions on Dependable and Secure Computing, 2015, 12(2): 164-178

[9] Chen Xiao-Jun, Fang Bin-Xing, Tan Qing-Feng, Zhang Hao-Liang. Inferring attack intent of malicious insider based on probabilistic attack graph model. Chinese Journal of Computers, 2014, 37(1): 62-72(in Chinese)
(**陈小军, 方滨兴, 谭庆丰, 张浩亮. 基于概率攻击图的内部攻击意图推断算法研究. 计算机学报**, 2014, 37(1): 62-72)

[10] Sheyner O, Haines J W, Jha S, et al. Automated generation and analysis of attack graphs//Proceedings of the IEEE Symposium on Security and Privacy. Berkeley, USA, 2002: 273-284

[11] Ou Xinming, Boyer W F, McQueen M A. A scalable approach to attack graph generation//Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS). Alexandria, USA, 2006: 336-345

[12] Maloof M A, Stephens G D. ELICIT: A system for detecting insiders who violate need-to-know//Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection (RAID). Gold Goast, Australia, 2007: 146-166

[13] Islam M S, Kuzu M, Kantarcioglu M. A dynamic approach to detect anomalous queries on relational databases//Proceedings of the 5th ACM Conference on Data and Application Security and Privacy (CODASPY). San Antonio, USA, 2015: 245-252

[14] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788-791

[15] Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 1994, 5(2): 111-126

[16] Berry M W, Browne M, Langville A N, et al. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, 2007, 52(1): 155-173

[17] Reynolds D. Gaussian mixture models//Li S Z, Jain A eds. Encyclopedia of Biometrics. USA: Springer, 2009: 659-663

[18] Schwarz G E. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2): 461-464

[19] Schonlau M, DuMouchel W, Ju W-H, et al. Computer intrusion: Detecting masquerades. Statistical Science, 2001, 16(1): 58-74

[20] Maggi F, Zanero S. On the use of different statistical tests for alert correlation//Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection (RAID). Gold Goast, Australia, 2007: 167-177

WEN Yu, born in 1976, Ph. D. His research interests include big data and system security.

WANG Wei-Ping, born in 1975, Ph. D., professor, Ph. D. supervisor. His research interests include database, big data and system security.

MENG Dan, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include big data and system security.

Background

User behavior analysis is an important tool for insider threat detection. Environment that generates user audit logs for such analysis are called detection domains in this paper. However, recently existing work mainly focused on single-pattern analysis of user single-domain behavior and relied on expert's knowledge and user background knowledge, which were not suitable for multi-domain scenarios. Smart attackers can cheat such single-domain detection systems by multi-step

attacks. Activities of the attackers in each single-domain are easily recognized as the normal behavior by separated detection analysis. Although a few work were proposed to detect anomalous user based on integrating anomaly alerts from each single-domain, they still separately detected user single-domain behavior.

In this paper, we presented a novel method for multi-pattern analysis of user cross-domain behavior which could be used to build multi-domain detection systems. Our method was a completely data driven resolution which did not rely on any expert knowledge and user background. Instead of direct multi-domain audit logs fusion at raw data level, we built user multi-domain behavior description through integrating features of each single-domain logs at feature generation level. Moreover, we designed a structured feature extraction method based on Non-negative Matrix Factorization for cross-domain feature generation of user multi-domain behavior. At last, we used Gaussian Mixture Model to analyze multi-pattern of user cross-domain behavior. In experiment, we used the proposed method to analyze five user audit logs in real environment. The experimental results showed that our method was effective on user cross-domain behavior pattern mining in multi-domain scenarios. Furthermore, we designed an insider attack detection method based on our user behavior pattern analysis approach. The experimental results also showed that our detection method was better than two existing solutions: a single-domain detection method and a single patterns based detection method.

This work is partially supported by the National High Technology Research and Development Program (863 Program) of China under Grant No. 2013AA013204.