计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800056

基于 PCA 和随机树的数据库异常访问检测

冯安然[1,2]　王旭仁[1,2]　汪秋云[2]　熊梦博[1,2]
1 首都师范大学信息工程学院　北京 100048
2 中国科学院信息工程研究所,中国科学院网络测评技术重点实验室　北京 100093
(ann654175863@163.com)

摘　要　数据库作为数据存储与交互的平台,其中包含了机密与重要信息,是恶意人员攻击的对象。外部人员的攻击可通过基于角色的权限控制系统对未经授权的用户访问进行限制,而来自内部人员的伪装攻击往往不易被察觉。针对数据库的访问行为,提出一种基于主成分分析(Principal Component Analysis,PCA)和随机树(Random Tree,RT)的异常访问检测算法 PCA-RT。首先,根据用户提交的查询语句特征构造用户数据库访问行为轮廓向量;然后,利用 PCA 算法对用户行为轮廓进行降维,使用随机树算法训练异常检测器。基于事务处理性能委员会(Transaction Processing Performance Council,TPC)组织发布的新一代数据库性能评测标准 TPC-E 构造实验数据集,提取较为全面的用户数据库访问行为轮廓特征向量。仿真实验结果表明,使用 PCA 算法对数据的约简达到 35%以上,PCA-RT 算法的精确率和召回率分别提高了 1.78%和 9.76%,从而证明了用户行为轮廓向量构造方法和 PCA-RT 算法对 TPC-E 数据库用户访问行为的异常检测是有效的。

关键词:异常检测;数据库安全;TPC-E;用户行为轮廓;主成分分析;随机树算法
中图法分类号　TP309

Database Anomaly Access Detection Based on Principal Component Analysis and Random Tree

FENG An-ran[1,2],WANG Xu-ren[1,2],WANG Qiu-yun[2] and XIONG Meng-bo[1,2]
1 College of Information Engineering,Capital Normal University,Beijing 100048,China
2 Key Laboratory of Network Assessment Technology,Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100093,China

Abstract　As a platform for data storage and interaction,database contains confidential and important information,making it a target of malicious personnel attacks. To prevent attacks from outsiders,database administrators can limit unauthorized user access through role-based access control system,while masquerade attacks from insiders are often less noticeable. Therefore,the research on database anomaly detection based on user behavior have important practical application value. A user anomaly detection algorithm PCA-RT based on Principal Component Analysis (PCA) and Random Tree (RT) is proposed for the anomaly detection of database user access behavior. Firstly,users' profile is constructed according to the characteristics of the query submitted by the users,then the principal component analysis is applied to reducing the dimension of the users' profile and feature selection. Finally,random tree has trained anomaly detector. The experiments,based on dataset constructed according to TPC-E,which is a new generation of database performance evaluation standard issued by TPC (Transaction Processing Performance Council),show that the user profile and PCA-RT are fast and effective for anomaly detecting of database user access behavior. PCA algorithm reduces data during data preprocessing up to more than 35%. The accuracy and recall of PCA-RT algorithm are improved by 1.78% and 9.76% respectively. It is proved that the construction method of user profile vector and the PCA-RT algorithm are effective for anomaly detection of user access behavior in TPC-E database.

Keywords　Anomaly detection,Database security,TPC-E,User behavior profile,Principal component analysis,Random tree algorithm

1　引言

随着数据在现代生活中占据着越来越重要的地位,其安全问题逐渐凸显。2019 年全球数据泄露的平均成本为392 万美元[1]。数据泄露的原因是多样的。Verizon 通过对 65 个国家的 5.3 万起安全事件和 2216 起数据泄露事件进行分析,于 2018 年发布了第 11 次数据泄露调查报告[2]。报告显示约 28%的安全泄露来自企业或组织的内部,其中,医疗方面的内部泄露超过了外部攻击所引起的数据泄露,占总泄露事件的 56%。内部人员出于利益考虑或操作不当引发的数据泄露一

到稿日期:2019-08-13　返修日期:2019-10-22　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家电网有限公司总部科技项目(5700-201972227A-0-0-00)
This work was supported by the science and technology project of State Grid Corporation of China (5700-201972227A-0-0-00).
通信作者:王旭仁(wangxuren@cnu.edu.cn)

般难以被发现，会造成严重的经济损失。报告显示，约有68%的数据泄露在发生后的几个月仍未被察觉。

数据库作为存储数据的“仓库”，往往是恶意人员攻击的重点对象。传统的数据库安全防护主要依赖于防火墙在网络层面对异常网络通信的阻断机制[3-5]。然而，内部人员拥有合理的访问权限，在对数据库进行操作的过程中不会引发异常通信告警，由此造成的数据泄露不易被人察觉。本文提出了一种基于主成分分析和随机树的数据库用户访问行为异常检测方法 PCA-RT，并采用事务处理性能测试委员会所制定的标准 TPC-E 测试数据库作为实验数据集对其进行性能评测。

本文第 2 节主要介绍数据库异常检测的相关工作；第 3 节介绍 PCA 算法和随机树算法；第 4 节讨论基于 PCA-RT 算法的异常检测系统的结构和实施步骤；第 5 节讨论符合 TPC-E 规范的数据集准备和用户访问数据库系统的行为轮廓提取方法；第 6 节比较分析 PCA-RT 算法的实验结果；最后总结全文。

2 相关工作

为了对访问数据库的用户进行管理，一些学者对数据库的访问权限控制方法进行了改进，使得权限控制策略能够解决更加复杂的实际应用问题[6-10]。但这种权限控制方法只能防止来自外部无授权的恶意攻击，对内部授权用户和伪装成内部用户的恶意攻击无效。自 Lee[11] 将数据挖掘技术引入入侵检测领域开始，研究者便利用数据挖掘技术构建针对数据库的异常检测系统。

Chung 等[12] 设计开发了一种数据库误用检测系统 DEMIDS，该系统根据数据库自身的结构特点与查询语句中包含的丰富语义信息，利用关联规则构造关系数据库中的用户轮廓来描述用户的正常工作行为；同时，该研究还提出了一种基本的异常检测模型。

Kamra 等[13] 利用审计日志中的 SQL 语句，分别采用 3 种粒度构造用户行为轮廓，使用朴素贝叶斯分类法构造异常检测模型；但其只考虑了用户提交查询语句的语法结构，不能检测出用户的伪装异常。

与文献[13]中的方法相比，Mathew[14] 将查询语句的返回结果作为用户行为特征，利用查询结果的统计特征构造用户行为轮廓；分别采用了朴素贝叶斯分类法和属性偏差计算法作为异常检测器。这种特征提取方法将检测的重点放在用户查询结果中，而不是用户提交的查询语句上，因此可以有效地检测出用户的伪装异常。但是，由于查询结果的数目往往比较庞大，利用统计特征来构造用户模型花费的时间较长，因此其训练效率较低。

Sallam 等[15] 将查询语句的语法结构和查询结果结合在一起构成新的用户轮廓，分别采用朴素贝叶斯分类法和多标签分类器构造异常检测模型。这种方法并没有计算查询结果的各项统计特征，仅将查询结果的数目占总查询表的比例作为一项特征加入用户行为轮廓中。

Ronao 等[16] 针对结构复杂的数据库查询语句提出一种细粒度的行为轮廓构建方法，首先提取出上百种属性，然后使用 PCA 对所提属性进行降维，最后使用随机树分类法构造异常检测器。这种方法在构造用户行为轮廓时，将各属性的数值类型作为一类特征，增加了计算的复杂度。

本文提出的基于主成分分析和随机树的数据库异常访问算法 PCA-RT，使用 PCA 对用户行为轮廓向量进行降维，缩短了训练时间；同时采用随机树算法训练异常检测模型，有效地提高了异常检测的准确率。

3 相关算法

3.1 主成分分析

PCA 的主要思想是从原始数据的 n 维向量空间中找出 k ($k<n$)个新的坐标轴。其中，新坐标轴之间彼此正交且与原始数据存在密切的关系：在原有 n 维数据的基础上，重新构造出包含绝大部分方差的 k 维特征[17]。PCA 在实际应用中的作用主要为：在数据预处理阶段对原始的多维向量进行降维处理，在保证信息损失最小的情况下得到一组低维向量。

设有 m 条 n 维数据，PCA 算法的目标为得到 k 维主成分，其步骤如下。

1)将原始数据构建为 m 行 n 列的矩阵 $\boldsymbol{X}$；

2)计算矩阵 $\boldsymbol{X}$ 每一列的均值，各列上的元素值分别减去该列均值，得到 $\boldsymbol{X}_{\text{new}}$；

3)求出协方差矩阵 $\boldsymbol{C}=\frac{1}{m}\boldsymbol{X}_{\text{new}}^{\mathrm{T}}\boldsymbol{X}_{\text{new}}$；

4)求出协方差矩阵 $\boldsymbol{C}$ 的特征值及每个特征值对应的特征向量；

5)将特征向量根据特征值从大到小的顺序按列排列，取前 k 列得到矩阵 $\boldsymbol{P}$；

6)$\boldsymbol{Y}=\boldsymbol{X}_{\text{new}}\boldsymbol{P}$ 即为矩阵 $\boldsymbol{X}$ 降到 k 维后的数据。

3.2 随机树

随机树是决策树算法的一种改进算法[18]。决策树算法模型是一种非参数型的分类器，是分类模型中应用最广泛的算法之一。决策树算法模型的构建一般包括 3 个步骤：属性选择、决策树生成和剪枝。决策树算法的关键在于如何在生成树的过程中选择最优的划分属性作为子节点。根据属性选择标准的不同，决策树算法分为 ID3，C4.5，CART 等。ID3，C4.5，CART 算法分别使用信息增益、信息增益比、基尼指数作为属性选择标准。

设训练数据集 D 有 n 个类 C_k ($k=1,2,\cdots,n$)，其中特征 A 有 m 个不同的取值$\{a_1,a_2,\cdots,a_m\}$，根据特征 A 的取值可将数据集 D 划分为 m 个子集$\{D_1,D_2,\cdots,D_m\}$。特征 A 的信息增益 $G(D,A)$ 定义为：

$$G(D,A)=H(D|A) \tag{1}$$

其中，$H(D)$ 表示数据集 D 的熵，用来衡量属性取值的不确定度，记为：

$$H(D)=-\sum_{k=1}^{n}\frac{|C_k|}{D}\log_2\frac{|C_k|}{D} \tag{2}$$

而 $H(D|A)$ 表示属性 A 对数据集 D 的条件熵，记为：

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i) \tag{3}$$

决策树算法计算速度快、准确率高，但在分类过程中容易出现过拟合现象。随机树算法改进了决策树构建过程中的属性选择方式，将遍历所有属性特征取最优改为随机选取 k 个属性计算其信息增益。相较于决策树算法，随机树算法运算时间短且不易出现过拟合现象。

4 基于 PCA-RT 算法的异常检测系统

基于 PCA-RT 算法的异常检测系统的整体结构如图 1 所示。系统的工作流程分为训练和测试两个阶段。

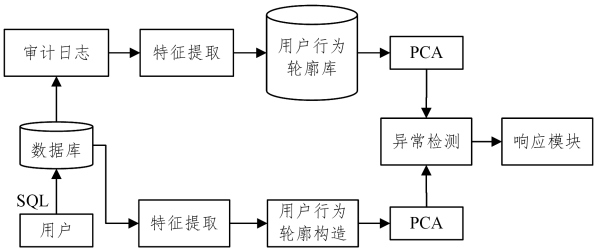

图 1 基于 PCA-RT 算法的系统结构

Fig. 1 PCA-RT system architecture

训练阶段的步骤如下：

1)对历史审计日志进行预处理，去除系统日志后得到用户查询数据；

2)提取查询数据的特征，构建用户的行为轮廓；

3)利用 PCA 算法对用户的行为轮廓向量进行降维，得到低维的用户行为轮廓向量；

4)使用随机树算法对训练数据进行训练，得到异常检测模型。

测试阶段的步骤如下：

1)对用户提交的查询进行预处理；

2)提取查询数据的特征，得到特征向量；

3)使用 PCA 算法对用户的行为轮廓向量进行降维，得到低维的用户行为轮廓向量；

4)将降维后的用户行为轮廓向量输入异常检测模型中，得到检测结果；

5)将检测结果输入响应器中，根据预先设定的响应策略发出响应。

5 基于 TPC-E 规范的数据库用户行为轮廓构造

由于真实数据库的后台审计日志不易获取，实验采用 TPC-E 数据库作为实验数据集。TPC 事务处理性能协会是一个评价大型数据库系统软硬件性能的非盈利性组织，其制定的规范在数据库异常检测领域已有多次应用[15-16,19]。TPC-E[20]是 TPC 协会制定的专门针对联机交易处理系统(OLTP 系统)的最新规范，其以美国纽约证券交易所为模型，模拟了客户和股票交易所与证券公司之间的交易往来，如账户查询、在线交易和市场调研等；证券公司还会根据市场变化执行指令并更新相关的账户和市场信息。TPC-E 包含 12 种交易事务，每种交易事务都有明确的执行者，即可以将各事务所属类别作为 TPC-E 数据库的查询语句的分类标签。

5.1 基于 TPC-E 规范的数据集构造

构造数据集的设备是一台处理器为四核 Intel Core i7 CPU 3.60 GHz、内存为 8 GB 的 64 位台式机，操作系统为 CentOS 6.5。实验所用的数据库为 Mysql5.6，辅助工具为 TPC 提供的 TPC-E 测试工具——基于 Mysql 的 EGen-v1.5.1。

利用 TPC 提供的脚本构建 TPC-E 数据库，模拟出 12 种交易事务。为了全面验证模型与方法的有效性，分别收集了 5 种不同规模的 TPC-E 数据集作为实验数据集。

TPC-E 原始数据集中包含大量系统命令等与用户交易无关的数据(如数据维护事务与交易清理事务)，去除这些数据，仅留下与查询相关的信息；由于交易事务都有明确的执行者(客户/证券公司/股票交易所)，因此每条数据使用该交易事务的执行者作为用户组别标签。分别将 5 个数据集的 80%作为训练数据集，剩余的 20%作为测试集。数据集的数据分布情况如表 1 所列。

表 1 TPC-E 产生的实验数据集

Table 1 Data distribution of five datasets of TPC-E

	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5
Training Data (Records)	86574	164290	246184	325165	401363
Testing Data (Records)	21643	41072	61546	81291	100341
Total	108217	205362	307730	406456	501704

5.2 用户行为轮廓构造

数据集中包含大量的交叉查询语句、属性索引和分组结构，通过从数据集中提取完整的用户行为特征，构造了用户行为轮廓向量。

使用向量 $\boldsymbol{V}=(C,\boldsymbol{T},\boldsymbol{A}_S,\boldsymbol{A}_P,\boldsymbol{A}_O,\boldsymbol{A}_G)$ 表示从查询中提取出的属性特征。C 表示语句的命令类型。$\boldsymbol{T}$ 为查询检索的数据表，用向量表示，其长度为数据库中数据表的数目。向量 $\boldsymbol{T}$ 与数据表之间的映射方式为：当查询中包含某个表时，将向量中该表所在位设为 1，否则设为 0。$\boldsymbol{A}_S$，$\boldsymbol{A}_P$，$\boldsymbol{A}_O$ 和 $\boldsymbol{A}_G$ 分别表示查询语句中包含的检索属性向量、执行属性向量、排序属性向量和分组属性向量。检索属性指查询语句中 WHERE 引导的属性；执行属性指命令词引导的属性；排序属性为 ORDER BY 引导的属性；分组属性为 GROUP BY 引导的属性。这 4 种属性向量的构造方法与数据表向量 $\boldsymbol{T}$ 类似，向量长度为语句中所有数据表的属性数目之和。这些向量与属性的映射方式为：按照查询中数据表出现的顺序对向量进行分块，分块的依据是表的属性数目；当引导词引导的子语句中包含第 i 个表中的第 j 个属性时，将该向量的第 i 部分的第 j 个位置设为 1，否则设为 0。

根据 TPC-E 构造数据库用户轮廓举例，表 2 为用户表(Client)，表 3 为产品表(Products)，表 4 列出了 $\boldsymbol{V}$ 向量举例。

表 2 用户表

Table 2 Clients' table

c_ID	*c_name*
1	c_1
2	c_2
3	c_3
4	c_4

表 3 产品表

Table 3 Products' table

p_ID	*p_price*
1	1
2	2
3	5
4	8

表 4 向量表示

Table 4 Vector representation

Query	$\boldsymbol{V}=(C,\boldsymbol{T},\boldsymbol{A}_S,\boldsymbol{A}_P,\boldsymbol{A}_O,\boldsymbol{A}_G)$
SELECT *c_ID*, *p_ID* FROM Clients, Products WHERE *c_name*=c_1 and *p_price*=2 ORDER BY *c_ID* GROUP BY *p_ID*;	['SELECT',[1,1],[0,1,0,1],[1,0,1,0],[1,0,0,0],[0,0,1,0]]

表 4 中，$\boldsymbol{V}$ 向量的含义为：该语句的命令 C 为 SELECT；检索的数据表 $\boldsymbol{T}$ 为[1,1]，即用户表和产品表两个表；检索属性向量 $\boldsymbol{A}_S$ 为[0,1,0,1]，即 WHERE 引导的属性为第一个表（用户表）的第二个属性 *c_name* 和第二个表（产品表）的第二个属性 *p_price*；执行属性向量 $\boldsymbol{A}_P$ 为[1,0,1,0]，即 SELECT 引导的属性为用户表的第一个属性 *c_ID* 和产品表的第一个属性 *p_ID*；排序属性向量 $\boldsymbol{A}_O$ 为[1,0,0,0]，即 ORDER BY 引导的属性为用户表的第一个属性 *c_ID*；分组属性向量 $\boldsymbol{A}_G$ 为[0,0,1,0]，即 GROUP BY 引导的属性为产品表的第一个属性 *p_ID*。

6 实验与结果

6.1 评估方法

使用以下 3 个评价指标对所提系统的异常检测效果进行评估。

精确率（Precision）：反映了被分类器判定的正常样本中真正的正常样本的比重。其定义如下：

$$P=\frac{TP}{TP+FP} \tag{4}$$

召回率（Recall）：也称为 True Positive Rate，反映了被正确判定的正常样本占总的正常样本的比重。其定义如下：

$$R=\frac{TP}{TP+FN} \tag{5}$$

F1 值：模型精确率和召回率的一种加权平均。其定义如下：

$$F1=\frac{2PR}{P+R} \tag{6}$$

其中，TP 表示将正常样本预测为正常类的数量；FN 表示将正常样本预测为异常类的数量；FP 表示将异常样本预测为正常类的数量；TN 表示将异常样本预测为异常类的数量。

6.2 PCA-RT 算法实验和讨论

通过第 5 节的数据准备和预处理，得到了 198 维的 $\boldsymbol{V}$ 向量集。利用 PCA 对此高维用户行为向量进行降维，得到 28 维的用户向量。PCA 算法有效降低了数据的存储容量，5 个数据集降维前后的文件大小对比如表 5 所列。PCA 降维完成后，使用随机树算法训练异常检测器。

表 5 PCA 降维前后文件大小的对比

Table 5 Comparison of file size before and after PCA dimension reduction

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Original Size /MB	41.1	78.1	117	154	190
After Applying PCA/MB	27.4	51.8	77.8	103	127

为了验证 PCA-RT 算法的检测效果，使用以下算法进行实验对比。

1）常见的 5 种机器学习分类算法：反向传播神经网络（Back Propagation Neural Network，BP）、C4.5 决策树、随机森林（Random Forest）、序列最小最优化算法（Sequential Minimal Optimization，SMO）和支持向量机（Support Vector Machine，SVM）。在相同数据集上使用 PCA 进行特征选择和降维后，使用这 5 种机器学习算法进行模型训练。

2）文献[16]的工作。文献[16]使用 PCA 和随机森林算法对相同的数据集进行异常检测模型训练。

不同分类方法的检测结果对比如表 6 所列。可以看出，与文献[16]的工作相比，PCA-RT 算法的精确率和召回率分别提高了 1.78%和 9.76%，$F1$ 提高了 9.87%。这说明相比文献[16]采用的交易事务类型作为标签的方法，提交查询所属的用户来源作为用户分组的方法更加符合对用户行为模式的分组规范，说明基于 PCA-RT 算法构建的分类模型是有效的。

表 6 不同分类方法的检测结果对比

Table 6 Comparison of detection results of different classification algorithms

Classifier	P/%	R/%	$F1$/%	Average Time of Training/ (ms/Record)	Average Time of Testing/ (ms/ Record)
Random Forest[16]	98.22	90.24	90.13	—	—
BP	99.24	99.24	99.24	3.485	0.005
C4.5	100	100	100	0.090	0.001
Random Forest	100	100	100	0.597	0.009
PCA-RT (ours')	**100**	**100**	**100**	**0.011**	**0.001**
SMO	99.48	99.46	99.46	4.852	0.002
SVM	100	100	100	3.994	0.276

另外 5 种机器学习分类算法（BP，C4.5，Random Forest，SMO，SVM）的实验结果表明，检测的精确率和召回率均优于文献[16]的工作。这几种机器学习分类算法中，PCA-RT 算法训练所用平均时间和测试平均时间均最短，每条记录的平均训练时间为 0.011 ms，测试时间为 1 μs，且 C4.5 的训练时间约为 PCA-RT 算法的 8 倍。Random Forest 和 SVM 的检测效果与 PCA-RT 算法、C4.5 相同，但模型构造所用时间较长，分别为 PCA-RT 算法训练时间的 54 倍、363 倍，测试时间分别为 PCA-RT 算法的 9 倍、276 倍。BP 神经网络与 SMO 算法的检测效果良好但没有达到 100%，且两种算法的训练时间过长，不利于实时检测。实验对比结果说明，PCA-RT 算法使用随机树算法，在特征选取阶段的随机性极大地提高了模型构造与测试速度，且在不同规模的数据集上表现稳定。

结束语 本文提出了一种基于主成分分析和随机树算法的数据库用户行为异常检测算法 PCA-RT。本文的主要贡献如下：

1）在 TPC-E 数据集上提取了完整的数据库用户行为轮廓向量，提高了异常检测精确率；

2）提出了一种基于主成分分析和随机树的数据库用户行为异常检测方法 PCA-RT，PCA 算法较好地实现了对特征的降维，提高了检测效率，RT 算法极大地减少了模型构造与测试时间；

3)实验结果表明,本文提出的用户行为异常检测方法可快速有效地检测出数据库异常用户行为。

在仿真实验中,PCA-RT 算法的检测效果最好,检测率高、速度快且具有稳定性,可以应对不同类型数据库的异常检测需求。

本文实验主要采用了 TPC-E 规范构造实验数据集,而没有在真实环境的数据库上进行测试,今后将尝试利用现实数据库访问数据进行模型训练与测试。

参 考 文 献

[1] IBM PONEMON INSTITUTE. 2019 Cost of a data breach[EB/OL]. (2019-7-22) [2019-08-01]. https://www.ibm.com/security/data-breach.

[2] VERIZON R T. Data breach investigations report [EB/OL]. (2019-02-15). https://enterprise.verizon.com/resources/reports/dbir/.

[3] WEI N. Anomaly detection and assessment of user behavior for database access [D]. Nanjing: Southeast University, 2017.

[4] CHEN D P. Intrusion detection system of database based on user behavior of analysis and identification [D]. Chengdu: University of Electronic Science and Technology of China, 2015.

[5] DUAN X Q. Research on database intrusion detection based on data mining [D]. Zhenjiang: Jiangsu University, 2009.

[6] LI N, TRIPUNITARA M V. Security analysis in role-based access control[C]// 9th ACM Symposium on Access Control Models and Technologies. New York: ACM, 2004: 126-135.

[7] NI Q, TROMBETTA A, BERTINO E, et al. Privacy-aware role-based access control [C] // 12th ACM Symposium on Access Control Models and Technologies. New York: ACM, 2007: 41-50.

[8] HADDAD M, STEVOVIC J, CHIASERA A, et al. Access control for data integration in presence of data dependencies[C]// 19th International Conference on Database Systems for Advanced Applications. Switzerland: Springer, 2014: 203-217.

[9] ABITEBOUL S, BOURHIS P, VIANU V. A formal study of collaborative access control in distributed datalog[C]// 19th International Conference on Database Theory. 2016: 1-17.

[10] BOSSI L, BERTINO E, HUSSAIN S R. A system for profiling and monitoring database access patterns by application programs for anomaly detection [J]. IEEE Transactions on Software Engineering, 2017, 43(5): 415-431.

[11] LEE W, STOLFO S J. Data Mining Approaches for Intrusion Detection[C] // Conference on USENIX Security Symposium. Berkeley: USENIX Association, 1998: 79-94.

[12] CHUNG C Y, GERTZ M, LEVITT K. DEMIDS: a misuse detection system for database systems[C]// Conference on Integrity and Internal Control Information Systems. Boston: Springer, 1999: 159-178.

[13] KAMRA A, TERZI E, BERTINO E. Detecting anomalous access patterns in relational databases [J]. Vldb Journal, 2008, 17(5): 1063-1077.

[14] MATHEW S, PETROPOULOS M, NGO H, et al. A data-centric approach to insider attack detection in database systems [C]// Conference on Recent advances in Intrusion Detection. Berlin: Springer, 2010: 382-401.

[15] SALLAM A, FADOLALKARIM D, BERTINO E, et al. Data and syntax centric anomaly detection for relational databases [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2016, 6(6): 231-239.

[16] RONAO C A, CHO S B. Mining SQL queries to detect anomalous database access using random forest and PCA[C]// Conference on Current Approaches in Applied Artificial Intelligence. Berlin: Springer, 2015: 151-160.

[17] HARRINGTON P. Machine learning in action [M]. New York: Manning Publications, 2012: 269-272.

[18] Java Code Examples for weka.classifiers.trees.RandomTree [EB/OL]. https://www.programcreek.com/java-api-examples/index.php?api=weka.classifiers.trees.RandomTree.

[19] ISLAM S M, KUZU M, KANTARCIOGLU M. A dynamic approach to detect anomalous queries on relational databases[C]// 5th ACM Conference on Data and Application Security and Privacy. New York: ACM, 2015: 245-252.

[20] TPC Benchmark E Standard Specification Revision 1.14.0[EB/OL]. [2019-02-23]. http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-e_v1.14.0.pdf.

FENG An-ran, born in 1993, master. Her main research interests include data mining, cyber security and database security.

WANG Xu-ren, born in 1972, postgraduate, Ph.D, vice professor. Her main research interests include data mining, cyber security and database security.