一种基于粗糙集理论的数据库入侵检测模型*

蔡 敏，汪世义，梁宝华
（巢湖学院 计算机科学与技术系，安徽 巢湖 238000）

摘 要：提出了一种可以检测数据库管理系统中异常事务入侵检测模型。该模型运用粗糙集理论从用户历史会话中提取用户正常行为轮廓，并利用散列算法来加速SQL模板的匹配，既可以有效检测异常事务，又可以避免因为一两次误用而把无辜的用户误认为是恶意攻击者。对模型的性能做了测试和分析。

关键词：数据库入侵检测；行为模式；SQL模板；粗糙集
中图分类号:TP309.2 文献标识码:A

Model of database intrusion detection based on rough set

CAI Min, WANG Shi Yi, LIANG Bao Hua
(Department of Computer Science and Technology, Chaohu College, Chaohu 238000, China)

Abstract: An intrusion detection model is proposed to detect anomalous transactions in DBMS. The model is based on rough set and capable of extracting uses´ normal behavior profile from user´ normal historical audit data. It can not only detect intrusions efficiently, but also can avoid mistaking an innocent user for a malicious attacker. Hash algorithm is applied to accelerate the match process of SQL templates. Finally, a performance analysis to the model is reported.

Key words: database intrusion detection; behavior profile; SQL template; rough set

数据库中往往保存着对公司或组织极为重要的数据，由于数据的重要性和价值，数据库经常成为攻击者的目标。网络化进一步延伸了数据库受到攻击的时间和空间。传统的数据库安全机制(如身份认证、访问控制等)认为未授权用户的所有行为都是违背系统安全策略的，应该禁止，而授权用户符合授权的行为则都不会造成系统破坏。这种以预防为主的被动安全机制已越来越不能满足日益增长的数据库安全的需要。一方面外部攻击者总能找到新方法闯入数据库，另一方面传统的数据库安全机制无法检测和阻止恶意授权用户的威胁。例如未授权用户通过窃取到合法的身份或权限等手段，伪装成合法用户，授权用户也可能会在情感或利益的驱使下执行一些恶意数据库事务。据统计，对数据库的攻击80%来自内部，内部滥用是数据库系统的主要威胁[1]。

入侵检测系统(IDS)作为一种积极主动的安全防护技术，能够检测到应用程序和用户的异常行为。可以通过有效地结合入侵检测技术来加强数据库的安全性。然而目前国内外对入侵检测技术的研究主要集中在网络层和操作系统层，对处于应用层的数据库入侵检测研究较少。DEMIDS机制[2]只从内容上分析了用户对数据库的访问，没有从时间上给予分析。此外，DEMIDS机制审计粒度过细，会严重降低DBMS性能。DBMTD机制[3]需要数据库管理员(DBA)手工创建授权事务轮廓。而事务轮廓只能反映事务内的SQL序列关系，不能反映事务间的SQL模板序列关系，也不能反映事务内路径的执行

*基金项目：安徽省高校省级自然科学基金（KJ2008B38ZC）

频度。极少执行的路径不应该作为正常路径，极少执行的事务也可能会存在异常。本文针对以上问题，提出了一种基于粗糙集的数据库异常检测机制。

1 基于粗糙集的数据库异常检测模型

1.1 基本原理与系统模型

定义 1 用通配符(%)替代用户递交的 SQL 语句中数据值，就得到了 SQL 模板[4]。

用户常常使用相同的 SQL 模板，但较少使用相同的查询，即 SQL 模板具有稳定性。用于训练的所有查询都是正常的，它们所操作的属性之间的距离[2]是相近的，即 SQL 模板的稳定性中实际已包含了属性结构的稳定性，因此不必再另外考虑属性间的距离。

典型数据库应用程序是客户/服务器系统(或三层系统)，用户通过客户端程序连接到 DBMS (当今的趋势是使用浏览器通过互联网访问数据库)。用户所能执行的 SQL 模板是事先定义好并被编写在数据库应用程序的代码里，用户只能按照 SQL 模板规定的形式来操作数据库。在基于角色的访问控制机制(RBAC)中，属同一角色的不同用户可以调用的 SQL 模板集是相同的，但每个用户使用这些 SQL 模板的规律却不尽相同。用户使用数据库的方式随着时间的推移会落入一定的模式中，因此，在事务或应用模式中，SQL 模板之间是存在一定执行顺序的。粗糙集可以从比较小的样本序列中提取预测规则集，得到的行为模型能有效逼近理想正常行为模型。本文的方法是：在训练阶段，根据用户正常历史审计数据，确定每个角色可以调用的 SQL 模板集，并通过粗糙集约简建立用户正常行为轮廓；在检测阶段，使用用户正常行为轮廓预测下一个 SQL 模板。若预测失败，认为发生异常。尽管异常并不一定就是入侵，但至少应该引起 DBA 的密切注意。系统模型如图 1 所示。

审计部件：进行数据采集。训练阶段要求 DBA 对用户行为进行监视，以保证采集的训练数据都是正常的。训练样本量越大，用户行为越规范，建立的轮廓就越准确。

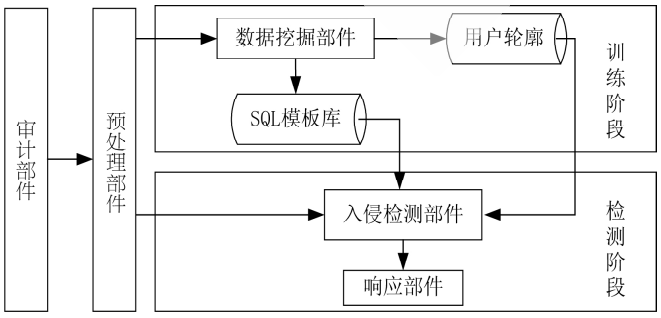

图 1 数据库入侵检测系统模型

预处理部件：将采集到的数据处理成数据挖掘部件和入侵检测部件需要的形式。

数据挖掘部件：构建 SQL 模板库；建立正常用户行为决策表，并运用粗糙集约简，建立正常用户行为轮廓。

SQL 模板库：规定各角色可用的 SQL 模板。

用户轮廓：存储用户正常行为轮廓。

入侵检测部件：根据入侵检测算法，判断新的用户行为是否为正常。

响应部件：如果发现异常行为则触发入侵警报，并按一定的策略采取相应的防范措施。

1.2 预处理过程

预处理过程包括以下 3 个步骤：

（1）集成用户会话连接。一个会话连接包含多个审计记录，需要通过登陆和退出命令以及会话连接 ID 把多个属于同一会话的审计记录归并到同一个会话连接中去。定义一个会话连接包含下面的属性：本次会话连接的 ID，本次会话的数据库用户名，该用户所属的角色名，本次会话用户所提交的所有 SQL 语句序列。

（2）提取 SQL 模板序列阶段。根据 SQL 模板的定义，依次将每个 SQL 语句转化成 SQL 模板。相应地，SQL 语句序列就转化成为 SQL 模板序列。

（3）计算 SQL 模板散列值。SQL 模板长度不一，特别是嵌套查询语句的模板一般较长，匹配效率较低。为了提高匹配效率，本文使用散列算法对每个 SQL 模板计算散列值，以后只要匹配散列值即可。这样 SQL 语句序列被进一步处理成散列值序列。散列函数[5]是一种把可变长度输入串(预映射)转换成固定长度输出串(散列值)的函数。散列函数具有单向性，从预映射的值很容易计算其散列值，但已知一个散列值，要找到预映射值，使其散列值等于已知散列值在计算上是不可行的。所以使用散列算法还可以增强系统安全性。

1.3 数据挖掘过程

1.3.1 粗糙集与知识约简

定义 2 决策系统是 1 个四元组 $S=(U,A,V,f)$，这里 U 是一个非空有限对象集，称为论域，A 是属性的非空有限集，$A=C\cup D$，C 为决策系统的条件属性集，D 为决策属性，且 $C\cap D=\phi$，$V=\bigcup_{a\in A}V_a$，V_a 是属性 $a\in A$ 的值域，$f:U\times A\to V$ 是一个信息函数。

定义 3 给定决策系统 $S=(U,A,V,f)$，对于属性集 $B\subset A$，称二元关系 $IND(B)=\{(x,y)\in U\times U\mid \forall a\in B, f(x,a)=f(y,a)\}$ 为 S 上的不可分辨关系。

定义4 令 $p \in R$，如果 $IND(R) = IND(R-\{p\})$，则称 p 是可缺的，否则为不可缺的。若 R 中每个关系都不可缺，则称 R 是独立的。

定义5 设 $Q \subseteq P$，如果 Q 是独立的，且 $IND(Q) = IND(P)$ 则称 Q 是 P 的一个约简。

1.3.2 SQL模板库的建立

一个好的散列算法，存在冲突的概率极小。可以认为散列值唯一代表一个SQL模板。由于散列值处理起来还是不太方便，本文还为每个散列值设置了连续的编号。每个编号唯一代表一个散列值，因此编号也唯一代表了一个SQL模板。编号与SQL模板库的建立过程是同时进行的，过程如下：

（1）依次根据会话中每个SQL模板的散列值搜索SQL模板库；

（2）如果模板库中没有该散列值，则为其设置一个编号。若有，则获取其编号；

（3）如果用户所属角色中没有该散列值，则将其加入，格式为角色、散列值和编号。

用编号代替相应的散列值，SQL语句序列就转化成编号序列。

1.3.3 建立决策系统

本文的目标是为每个用户建立一个正常行为轮廓，以用户作为独立的单元来进行训练。依次对每个用户执行这个过程，就得到所有用户的正常行为轮廓。

设 X 是某用户所有正常审计会话产生的全部SQL模板序列的集合，注意，此时的序列实际上是编号序列。用长度为 L+1的窗口依次沿每个SQL模板序列以步长为1滑动，得到一系列长度为L+1的序列段，所有这些序列段组成了序列段集 U。把 U 中每个序列段的前 L 个位置称作条件属性集，记作 A，把第 L+1个位置称作决策属性，记作 d。SQL模板序列段反映了用户会话中SQL模板之间的执行次序。如果选取序列段为1，就丢失SQL模板的次序信息，而长度太大，就会丢失会话的局部状态信息，无法正确反映正常情况下的局部SQL模板的调用状况。序列段长度的选取与具体数据库事务长度有关，约为2~5个事务所包含的模板数。

建立正常行为轮廓的基本思想是根据 U 中的前 L 个SQL模板预测第 L+1个SQL模板，即找到 U 中的条件属性集 A 与决策属性之间的关系。

1.3.4 提取用户正常行为轮廓

在决策系统中，不是所有的属性对知识表示都起作用。知识约简就是在决策系统保持分类能力不变的条件下，去掉那些在知识表示中不起作用的属性，以简化知识的表示。利用粗糙集理论约简知识，可得到最小决策规则分类模型。

在属性集 A 中找出约简是一个NP难题，但存在快速计算约简的启发性方法。假定 B 是 A 关于 d 的约简，可以提取到形如 $\alpha \rightarrow \beta$ 的规则集，其中 $\alpha = \bigwedge_{a \in B}(a = f(x,a))$，$\beta = (d = f(x,d))$，$x \in U$。

删除覆盖率小于指定覆盖率的规则。规则的覆盖率是指同该规则匹配的短序列段的数目占总序列段数目的百分比。用户的所有规则组成了用户轮廓。

1.4 入侵检测

定义6 会话的异常度定义为预测失败的序列段个数同会话中序列段总数之比。由安全管理员事先设定一个阈值，若会话异常度大于阈值，则认为本次会话异常。

借助训练阶段得到的用户行为轮廓可以检测出该用户会话是否异常。一条规则同一个长度为 L+1的序列段匹配，应满足如下条件：

（1）规则适用的长度为 L+1；

（2）规则条件中描述的序列段各个位置上的SQL模板应同序列段中的情况相符。

由于可能出现不一致规则的情况，因此本文引入了预测集概念。检测用户会话的算法如下：

（1）根据SQL模板库把用户会话转换成SQL模板编号序列；

（2）如果序列还没有检测完，用长度为 L+1的窗口在SQL模板序列上滑动，步长为1，每次截取一个长度为 L+1的序列段。否则结束检测；

（3）如果该序列段中包含不可用SQL模板，则会话异常度置为阈值+1，转（7）；

（4）在正常模型中，寻找与该序列段匹配的规则，以获得对该序列段预测集；

（5）若预测集为空，则以最常出现的几个预测结果组成对该序列段的预测集；

（6）如果该序列段的第 L+1个模板在预测集中，则预测成功。否则，将当前事务标记为异常事务，报告给DBA，并计算会话异常度；

（7）若会话异常度大于设定的阈值，则该会话连接为异常，报警并采取相应措施。否则转(2)。

2 实验结果与性能分析

本系统实验数据库为高校人事管理数据库，包含36个表，302个字段，使用SQL Server 2000的Profiler工具创建跟踪来采集审计数据。首先按第3节方法建立

SQL模板库，选取的序列段长度为9，并将用户会话连接处理成决策系统形式，然后运用波兰华沙大学与挪威科技大学联合开发的ROSSETA软件包对决策系统进行属性约简，并生成if-then形式的预测规则集，最后将预测规则集应用于检测。

为了验证方法的有效性，实验时，将阈值设成很大的值，以保证会话序列能被检测完。首先将50个正常会话用于检测，误报率为5.8%，序列平均异常度为2.52。然后将生成的500个异常事务平均分配到这50个正常会话中，只有4个异常事务没有被检测出来，漏报率为0.8%，序列平均异常度为30.66。由误报率和漏报率可以看出，本文建立的模型是比较精确的。由序列平均异常度可以看出，异常序列的平均异常度显著高于正常序列的平均异常度，所以只要选取适当阈值，尽管存在少量的误报，也能够准确地将正常会话同异常会话区分开来。

本系统能够检测的攻击类型如下：SQL注入攻击、合法用户的权限滥用、伪装攻击。

SQL注入攻击是最常见的一种数据库入侵方法，据了解，国内60%的论坛都存在SQL漏洞，它的危害程度应该引起足够的重视。攻击者利用数据库应用程序的漏洞，精心伪造恶意SQL语句，改变SQL查询结构以获取用户特权或对数据库产生影响。这种攻击是操作系统和网络层入侵检测系统所难以检测的。下面以SQL注入攻击为例说明本模型的实际应用。

例如用户让应用程序按指定类型列出自己的所有注册信用卡。这段功能的伪代码如下：

```
uname = getAuthenticatedUser()
cctype = getUserInput()
result = sql("SELECT nb FROM creditcards WHERE user='"
+ uname + "' AND type='" + cctype + "';")
print(result)
```

如果用户Bob搜索他的所有VISA卡，将执行以下查询：SELECT nb FROM creditcards WHERE user='Bob' AND type='VISA';。这个例子包含了一个SQL注入漏洞。若Bob想看用户Alice所有信用卡，他可以请求卡类型'OR user='Alice。这将引起如下查询执行：SELECT nb FROM creditcards WHERE user='Bob' AND type=''OR user='Alice';这个查询将Alice所有信用卡列表返回给Bob。

在本模型中，上述攻击语句将被提取为模板SELECT nb FROM creditcards WHERE user=% AND type=% OR user=%；然后对该模板进行散列，根据散列值搜索SQL模板库，结果没有与之匹配的散列值，因此可以断定该模板异常。同样地，该模型可以拒绝任何不符合事先定义的SQL模板的访问。

对于第二种攻击，授权用户绕过或回避安全控制，一般会访问那些原来不允许访问的数据库或数据库对象。对它们的检测类似第一种。

对于第三种攻击，攻击者通过合法的登录过程进入系统(可能是窃取到账号和口令)。不能单凭一两次异常就断定该用户非法，检测这种攻击的前提是入侵者会作出较多与正常用户不同的行为。如果一次会话中用户反常行为过多，就认为是一次入侵。

本文对已有的数据库入侵检测机制进行了分析研究，并提出了一种基于粗糙集的数据库入侵检测模型。为了提高检测的效率和精确度，使用了散列算法和预测集概念。该模型可以拒绝任何不符合事先定义的SQL模板的访问，还可以发现伪装成合法用户、但与合法用户正常使用方式不同的攻击者以及合法用户的滥用行为。研究可以有效检测数据库入侵的方法是数据库入侵容忍技术亟待解决的首要问题，下一步工作是结合其他检测方法进一步提高模型精确度，并将该模型应用到数据库入侵容忍技术中。

参考文献

[1] 文俊浩,徐玲,李立新，等. 安全增强的数据库系统的模型构建[J]. 计算机应用, 2005, 25(8): 1734-1736.

[2] CHUNG C Y, GERTZ M, LIVIN K. DEMID S. A misuse detection system for database systems[C]. In: Third Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems. Amsterdam, Netherlands: Kluwer Academic Publishers, 1999：159-178.

[3] MARCO V，HENIPUE M. Detection of malicious transactions in DBMS[C]. In: Proceedings of the 11th IEEE International Symposium Pacific Rim Dependable Computing, Washington. DC, USA: IEEE Computer Society, 2005：350-357.

[4] YAO Q，AN A，HUANG X.Finding and analyzing database user sessions[C]. In: Proceedings of the 10th International Conference on Database Systems for Advanced Applications, Heidelberg, Berlin: Springer, 2005：851-862.

[5] BRUCE S. Applied cryptography protocols, algorithms, and source code in C (Second Edition) [M]. New York: Willey, 1999：21-22.

（收稿日期：2008-12-26）