软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(5):1020−1030 [doi: 10.3724/SP.J.1001.2011.03823]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

多级安全中敏感标记的最优化挖掘*

杨 智[1,2,3], 金舒原[1,4+], 段淼毅[1], 方滨兴[1,4]

[1](中国科学院 计算技术研究所,北京 100190)

[2](中国科学院 研究生院,北京 100049)

[3](信息工程大学 电子技术学院,河南 郑州 450004)

[4](信息内容安全技术国家工程实验室,北京 100190)

Optimal Mining on Security Labels in Multilevel Security System

YANG Zhi[1,2,3], JIN Shu-Yuan[1,4+], DUAN Mi-Yi[1], FANG Bin-Xing[1,4]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

[3](Institute of Electronic Technology, Information Engineering University, Zhengzhou 450004, China)

[4](National Engineering Laboratory for Information Security Technologies, Beijing 100190, China)

+ Corresponding author: E-mail: jinshuyuan@software.ict.ac.cn

Yang Z, Jin SY, Duan MY, Fang BX. Optimal mining on security labels in multilevel security system. *Journal of Software*, 2011,22(5):1020−1030. http://www.jos.org.cn/1000-9825/3823.htm

Abstract: This paper presents a bottom-up approach to implement the automatic and scientific transform of access control policies in the migration. First, the problem of mining security labels optimally is described formally, and it is then proved to be NP-complete. Next, an approximate optimization algorithm based on hierarchical clustering and genetic algorithm is presented, which decomposes the problem into two parts: category partition and secret level assignation. Finally, experimental results show that the algorithm is effective in finding an optimal solution. The proposed approach can be applied to migration projects in hierarchy protection in information security.

Key words: multilevel security; security label; optimal mining; computational complexity; hierarchical clustering algorithm; genetic algorithm

摘 要: 提出了一种自底向上的方法来实现系统迁移过程中自动和科学的访问控制策略转换.首先对多级安全中敏感标记最优化挖掘问题作了形式化描述,证明了该问题是 NP 完全问题,不存在多项式时间算法.然后,在此基础上提出了基于层次聚类和遗传算法的近似最优化挖掘算法,将该问题分解为范畴划分和密级分配两个阶段.最后,实验结果表明,算法能够有效地挖掘出最优的敏感标记.该方法可以应用于等级保护工作中的系统迁移工程.

关键词: 多级安全;敏感标记;最优化挖掘;计算复杂度;层次聚类算法;遗传算法

中图法分类号: TP309 **文献标识码**: A

* 基金项目: 国家自然科学基金(60703021); 国家高技术研究发展计划(863)(2009AA01Z438, 2009AA01Z431, 2006AA01Z457); 国家重点基础研究发展计划(973)(2007CB311100)

收稿时间: 2009-08-18; 定稿时间: 2010-02-01

等级保护是我国信息安全保障工作的一项基本制度和根本方法.全国重要信息系统的评定等级工作已基本完成,目前的主要任务是系统安全整改建设,这标志着我国等级保护工作已进入全面实施阶段.然而现实的情况是,很多重要电子政务系统、电子商务系统,由于其本身的重要程度被要求达到 3 级或以上级别.且 3 级以上信息系统的应用安全、主机安全和网络安全要求必须实现强制访问控制策略,其核心是多级安全(multi-level security,简称 MLS)策略.然而,这些已存在系统的访问控制策略由于历史等原因并不是多级安全策略(常见的是自主访问控制).因此,如何进行策略转换并保证策略转换的可行性,是一个迫切需要解决的问题.

强制访问控制策略要求对主体、客体指定敏感标记,这些敏感标记是等级分类和非等级类别的组合.对于多级安全来说,敏感标记是密级和范畴集的组合.目前,对策略转换主要采用自顶向底(top-down)的方法.该方法由系统用户、系统开发者和安全专家一起对信息系统工作流程及场景、已有授权策略库进行分析,根据专家经验和分析的结果手动为主体和客体分配标记.这种人工分配标记的方式对于小型系统也许可行,但当面对有数以千计的用户、数以万计的资源、百万级的授权规模的大系统时,其工作量巨大,难以保证科学性,甚至可能由于严重破坏原有系统访问控制逻辑而根本无法实现策略的转换,使原有系统无法实现向多级安全的迁移.

本文提出了一种自底向上(bottom-up)的方法来实现系统迁移过程中访问控制策略的自动转换.该方法的基本思想是,利用数据挖掘的方法从原有的授权策略库中发现而不是重新定义符合多级安全规则的模式,找出最优化的实体标记,我们称该方法为多级安全中敏感标记的最优化挖掘方法.本文分析和证明了敏感标记最优化挖掘的复杂度,提出了有效的挖掘算法来求解敏感标记中的范畴和密级的优化问题,并通过模拟和真实数据的实验验证了该方法的良好的性能.另外,利用该方法可实现策略自动转换,从而避免了手工转换的诸多弊病,最小化了定级系统策略转化的代价.因此,该方法对我国的等级保护工作的推进具有一定的理论和实践意义.

1 相关工作

多级安全的经典模型是 BLP 保密性模型[1,2],它是许多信息系统安全评测标准的制定依据和理论基础.为了防止机密信息的泄漏,BLP 模型规定主体不能读访问安全级高于自己的客体,不能写访问安全低于自己的客体,从而实现信息向高安全级单向流动.为了提高系统的可用性,BLP 模型引入可信实体概念,使高安全级可信主体能够通过一定范围的安全级调整写访问低安全级客体.目前,有些研究对可信实体处理做了深入分析,试图解决可信主体安全隐患,例如,文献[3]提出了离散标记序列多级安全模型.与 BLP 的可信主体级别范围模型相比,该模型在多级安全的策略范围内实现了可信主体特权最小化.由于多级安全模型只是一个保密性模型,不具有完整性模型的优点,也有些文献研究保密性模型与完整性模型的融合问题,以增强 BLP 模型的安全性.例如,文献[4]的安全模型使敏感标记同时包含机密级和完整级,同时,可信访问控制检查员可动态调整客体机密级和完整级,以保证系统可用性.文献[5]研究了在主客体动态变化的访问控制系统中,敏感标记中级别升降对系统安全带来的影响.它使用了一种新的逻辑编程语言对这种系统进行建模分析,寻找信息流动脆弱性或证明其安全性.概括来说,已有的研究基本上偏重于对多级安全模型本身的研究和改进,很少见到如何从旧系统向多级安全系统迁移方法方面的研究.例如,很少见到如何从已有的授权策略库中发现符合多级安全策略模式的研究.

在授权策略挖掘领域,近年来比较活跃的是对角色挖掘问题(role mining problem,简称 RMP)的研究[6−11].角色挖掘问题被定义为从已存在的授权策略中发现最优的角色集合,角色集合可以完备地、正确地、有效地表达已有授权关系[6,8].目前,已有一些有效的角色挖掘方法,如子集枚举方法[6]、图优化方法[7]、布尔矩阵分解方法[9]和概率统计方法[10]等等,也有些文献[11]从语义角度对角色挖掘质量进行了评价.这里的角色与多级安全中的范畴有一定相似之处,它们都体现了分组的概念,但是也有明显的区别:角色可反映某个职能岗位拥有的权限集合,而范畴大多表示标记应用的类别和领域.授权策略挖掘的另一个方向是噪声处理,以发现和纠正授权误配和漏配.文献[12]研究了用关联规则挖掘算法 Apriori 检测授权策略的错误,它通过分析授权的频繁模式发现异常的授权规则.文献[10]用极大关系模型 IRM 去检测错误的授权,它通过对多种集合同时聚类来发现关联关系.

2 问题定义及复杂度分析

2.1 BLP模型介绍

为了便于描述问题,我们先介绍BLP模型的基本概念.参照文献[1],设S是系统主体的集合,O是客体集合,P是权限集合$\{r,a,w,e\}$,分别表示读权限(r)、写权限(a)、读/写权限(w)和空权限(e).敏感标记集合 $L=\{(c,k)|c\in C, k\subseteq K\}$,密级 C 是大小可以比较的线性序列,范畴集 K 中的范畴是非等级的应用领域或类别.

定义 1. 标记(c,k)支配标记(c',k'),当且仅当 $c'\leqslant c$,且 $k'\subseteq k$,记作$(c,k)\succ(c',k')$.

用 $f_S:S\to L$ 和 $f_O:O\to L$ 表示主体和客体的标记函数.

规则 1. 简单安全条件:s 可以读 o,当且仅当$f_S(s)\succ f_O(o)$.

规则 2. *-属性:s 可以写 o,当且仅当$f_O(o)\succ f_S(s)$.

规则 3. 自主安全特性:状态的每一次存取操作都是由存取矩阵所限定的.

简单安全条件通常称为“不向上读”,*-属性通常称为“不向下写”.另外,主体的访问需要受到自主安全特性的制约.在实际应用中我们发现,敏感标记中范畴的层次包含关系较难获得语义支持.例如,一个保密系统中范畴集为{人事,财务,业务},密级为{公开,秘密,机密},标记人事部门普通用户 s_1 为[秘密,{人事}],标记财务部门普通用户 s_2 为[秘密,{财务}],标记主管 s_3 为[机密,{人事,财务,业务}].现有资源 o_1 是人事和财务部门普通用户均可读访问的共享资源,资源 o_2 是人事兼财务管理级别以上的用户才能读/写访问的保密资源.在这种情况下,难以恰当地标记 o_1 和 o_2 以满足需求.若标记 o_1 为[公开,{人事,财务}],则 s_1 和 s_2 均无法读访问 o_1,即无法实现“下读”;若标记 o_2 为[秘密,{人事,财务}],则 s_3 只能读访问 o_2,无法实现“上写”,而 s_1 和 s_2 却可写访问 o_2.

因此,在不改变多级安全规则的前提下,为了提高原有授权系统转换到多级安全系统时的可用性,本文对敏感标记做出约束,限制标记的范畴集只包括一个范畴,以简化标记之间的支配关系,从而更关注范畴内受密级支配的信息单向流动.对于公共资源,我们可以通过数据挖掘的方法将它们单独聚类成一个新范畴.同时,考虑主体的多角色多领域访问情况,允许主体有多个标记.主体能否访问某个客体取决于主体是否存在某个标记,该标记和客体的标记的关系是否满足多级安全规则要求的支配关系.

2.2 敏感标记的最优化挖掘问题

由于大多数授权最终都可以转化为访问控制矩阵形式,因此,本文主要研究如何从访问控制矩阵向多级安全策略转换的问题.访问控制矩阵的行表示主体,列表示客体,矩阵中的元素表示相应主体访问相应客体的权限.本文中,我们用访问控制矩阵距离作为授权策略间差异的指标.

定义 2(访问控制矩阵距离). 设 $P=\{r,a,w,e\}$代表权限集合;访问控制矩阵 $A=(a_{ij})_{m\times n}$,其中,$a_{ij}\in P$;访问控制矩阵 $B=(b_{ij})_{m\times n}$,其中,$b_{ij}\in P$. $\| A-B \|=\sum_{i=1}^{m}\sum_{j=1}^{n}\ell(a_{ij},b_{ij})$ 称为访问控制矩阵 A 和 B 之间的距离,其中,访问权限距离函数 $\ell(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases}$.

在定义2中,我们将所有的距离差异看作是一样的,即当$x\neq y$时距离差异都是1.我们也可以根据实际情况给出函数ℓ的其他定义.例如,可以根据实际需要定义权限 w 和 e 的距离大于 w 和 a 的距离.根据规则 1 和规则 2,可以推导出多级安全系统的访问控制矩阵.这里,我们约束每个标记中只有 1 个范畴,并允许主体多标记.

定义 3(多级安全系统导出的访问控制矩阵). 对于一个有 m 个主体和 n 个客体的多级安全系统,标记集合记为 $L=\{(c,k)|c\in C,k\in K\}$;所有主体的标记记为向量 $p=(ls_1,ls_2,\ldots,ls_m)^T$,其中,$ls_i\subset L$,它是为第 i 个主体分配的标记元组;所有客体的标记记为向量 $q=(lo_1,lo_2,\ldots,lo_n)^T$,其中,$lo_j\in L$,它是为第 j 个客体分配的标记.由多级安全系统

(L,p,q)导出的访问控制矩阵 M 定义为 $M=pq^T=(x_{ij})_{m\times n}$,其中, $x_{ij}=\begin{cases} w, \exists l\in ls_i, l=lo_j \\ a, \exists l\in ls_i, l\prec lo_j \\ r, \exists l\in ls_i, l\succ lo_j \\ e, \text{其他} \end{cases}$.

在上述两个定义基础上,我们给出如下的敏感标记的最优化挖掘问题的形式化描述:

定义 4(敏感标记的最优化挖掘问题,security label optimal mining problem,简称 SLOMP). 给定一个有 m 个主体、n 个客体的授权系统,已存在访问控制矩阵为 $A\in P^{m\times n}$,正整数 $s\leqslant m+n$,正整数 $t\leqslant\min(m,n)$.求标记集合 $L=\{(c,k)|c\in C,k\in K,|C|\leqslant s,|K|\leqslant t\}$,所有主体标记 $p=(ls_1,ls_2,\ldots,ls_m)^T$,所有客体标记 $q=(lo_1,lo_2,\ldots,lo_n)^T$,使得多级安全系统$(L,p,q)$导出的访问控制矩阵与 A 之间的距离最小,即求 $\arg\min_{L,p,q}\|A-pq^T\|$.

2.3 SLOMP计算复杂度分析

为了研究 SLOMP 的计算复杂性,我们将求最优解的 SLOMP 表述为如下判定问题:

定义 5(SLOMP 的判定版本). 给定一个有 m 个主体、n 个客体的授权系统,已有访问控制矩阵是 $A\in P^{m\times n}$,正整数 $s\leqslant m+n$,正整数 $t\leqslant\min(m,n)$,$\delta\geqslant 0$.判定是否存在多级安全系统(L,p,q),$L=\{(c,k)|c\in C,k\in K,|C|\leqslant s,|K|\leqslant t\}$,所有主体标记 $p=(ls_1,ls_2,\ldots,ls_m)^T$,所有客体标记 $q=(lo_1,lo_2,\ldots,lo_n)^T$,使得$\|A-pq^T\|\leqslant\delta$,问题记作 $dSLOMP(m,n,s,t,A,\delta)$.

下面,我们用归约的方法分析 $dSLOMP(m,n,s,t,A,\delta)$的复杂性.先看一个 NP 完全问题[13,14],问题描述如下:

定义 6(离散基划分问题,discrete basis partition problem,简称 DBPP). 给定有限集合 U,U 的一个子集集合 H,正整数$\lambda<\min\{|H|,|U|\}$,$\eta\geqslant 0$,判定是否存在 U 的一个子集集合 B,B 是 U 的一个划分,且$|B|=\lambda$,使得

$$\ell_{\Delta}(H,B)=\sum_{h\in H}\min_{S\subseteq B}|h\Delta(\cup S)|\leqslant\eta,$$

其中,$\cup S=\bigcup_{s\in S}s$,$h\Delta(\cup S)=|h\cup(\cup S)|/|h\cap(\cup S)|$,记作 $DBPP(U,H,\lambda,\eta)$.

定理 1. dSLOMP 是 NP 完全问题.

证明:先证明 DBPP$\propto_P$SLOMP.设有 DBPP 的一个实例 $D(U,H,\lambda,\eta)$.构造 dSLOMP 实例 $S(m,n,s,t,A,\delta)$如下,令授权系统中客体个数 $n=|U|$,客体集合 $O=\{o_1,o_2,\ldots,o_n\}$对应 $D(U,H,\lambda,\eta)$中集合 $U=\{u_1,u_2,\ldots,u_n\}$,主体个数 $m=|C|$,$D(U,H,\lambda,\eta)$的 $H=\{h_1,h_2,\ldots,h_m\}$对应各个主体能读/写访问的客体集合,即访问控制 $A=(a_{ij})_{m\times n}=\begin{cases} w, u_j\in h_i \\ e, u_j\notin h_i \end{cases}$,构造 SLOMP 实例显然可在多项式时间内完成.

事实上,因为 A 中访问权限只有 w 和 e 两种,这意味着主体和客体要么不在同一范畴中(权限为 e),要么在同一范畴中且密级相同(权限为 w).不失一般性,我们可假设实例 $S(m,n,s,t,A,\delta)$中密级个数 s 为 1,即所有主客体密级相等.主体能否读写访问客体取决于主体是否存在一个标记,该标记与客体标记相等.实例 $S(m,n,A,s,t,\delta')$等价于 $S'(m,n,A,1,t,\delta)$,继续构造该实例,令 $t=\lambda$,$\delta=\eta$.若实例 $S'(m,n,1,t,A,\delta)$存在解,记为(L,p,q),则其范畴集 $K=\{k_1,k_2,\ldots,k_r\}$.若 $r<t$,则令 $K=\{k_1,k_2,\ldots,k_r,k_{r+1},\ldots,k_t\}$,$k_{r+1},\ldots,k_t$ 是虚设的范畴,并无实际意义;取实例 $D(U,H,\lambda,\eta)$中一个子集集合 $B'=\{b_1,b_2,\ldots,b_\lambda\}$与 K 对应,$u_j\in b_i$ 当且仅当 $q=(lo_1,lo_2,\ldots,lo_n)^T$ 中 lo_j 的范畴是 k_i;取实例 $D(U,H,\lambda,\eta)$中一个子集集合 $S'=\{s_1,s_2,\ldots,s_m\}$与 $p=(ls_1,ls_2,\ldots,ls_m)^T$ 对应,$b_j\in s_i$ 当且仅当 ls_i 中包含范畴 k_j.显然$\|A-pq^T\|\leqslant\delta$,则有 $\sum_{i=1}^{m}|h_i\Delta(\cup s_i)|\leqslant\delta=\eta$,则 B'即是要求的 B,S'是要求的 S;反之,若实例 $D(U,H,\lambda,\eta)$存在解,因为 B 和 K,S 和 p 的一一对应关系,而且 $\sum_{i=1}^{m}|h_i\Delta(\cup s_i)|\leqslant\delta=\eta$,则$\|A-pq^T\|\leqslant\delta$.即,实例 $D(U,H,\lambda,\eta)$求解转化为求解实例(m,n,A,s,t,δ').因此,DBPP$\propto_P$SLOMP,即 SLOMP 是 NP 难的.

SLOMP$\in$NP 是显然的,给定 SLOMP 实例(m,n,A,s,t,δ')和一个解(L,p,q),验证算法求$\|A-pq^T\|\leqslant\delta$的计算时间为 $O(mn)$,即在多项式时间内完成验证,即 SLOMP$\in$NP,因此 SLOMP$\in$NPC.定理 1 得证. □

由于 SLOMP 问题是 NP 完全问题,我们考虑该问题的近似最优解方法.我们分解该问题,首先划分范畴,并

确定主、客体所属范畴;然后在各个范畴内为主、客体分配密级,这些子问题的合解是 SLOMP 问题的解.范畴的最优划分和分配以及密级的最优分配算法将在第 3 节和第 4 节中探讨.

3 范畴最优划分和分配

3.1 范畴挖掘问题的数学模型

范畴反映了安全应用的类别或领域.若主体属于某个范畴,则通常他会对范畴内的大多数客体具有某种访问权限(r,a,w);若主体不属于某个范畴,则对该范畴内的客体无任何访问权限(e).因此,在范畴划分问题中,可以主要区分无权限和其他权限.此时,访问控制矩阵可以转换为布尔型矩阵,即对于权限 e,其对应的矩阵中元素取值为 0;对于权限 w,r,a,其对应的矩阵中元素取值为 1.为了便于描述和解决范畴划分问题,我们给出如下定义:

定义 7(布尔型矩阵乘法). 给定布尔型矩阵 $A\in\{0,1\}^{m\times k}$ 和 $B\in\{0,1\}^{k\times n}$,A 和 B 的乘法记作 $A\otimes B=C$,其中,$C\in\{0,1\}^{m\times k}$ 且 $c_{ij}=\bigvee_{l=1}^{k}(a_{il}\wedge b_{lj})$.

我们用布尔型矩阵 $M(SK)_{m\times k}$ 表示主体 S 与范畴 K 的关系 SK,矩阵中元素表示相应行的主体是否被分配相应列的范畴.同样地,我们用布尔型矩阵 $M(KO)_{k\times n}$ 表示范畴 K 与客体 O 的关系 KO,矩阵中元素表示相应行的范畴是否包含相应列的客体.在定义 7 下,$M(SK)\otimes M(KO)$反映了主体和客体是否共同属于某个范畴的情况.

定义 8(布尔型向量距离). 对于两个 d 维布尔型向量 $v,w\in\{0,1\}^d$,定义它们之间的距离为

$$\|v-w\|=\sum_{i=1}^{d}|v_i-w_i|.$$

定义 9(范畴的最优化挖掘问题). 给定主体(记为 S)对客体(记为 O)的布尔型访问控制矩阵 $M(SO),\delta\geqslant 0$,寻找范畴集 K、主体与范畴的关系 SK、范畴与客体的关系 KO,使得$\|M(SK)\otimes M(KO)-M(SO)\|\leqslant\delta$,并最小化$|K|$.

对于矩阵 $M(SO)_{m\times n},|K|=k$,如果搜索范畴分配的每一个状态,SK 状态数为 2^{mk},KO 状态数为 k^n,矩阵乘法时间为 $O(mk^2n)$,求解范畴的最优化挖掘问题的计算复杂度为 $O(2^{mk}k^nmkn)$,其计算量非常巨大.下面,我们提出一种基于层次聚类的范畴挖掘算法,使求解该问题可在多项式时间内完成.

3.2 基于层次聚类的范畴挖掘算法

在应用系统中,同部门或岗位的用户通常会访问相同的资源,紧密相关的资源通常会被一个用户以相同权限访问.基于这种分析,我们首先对访问控制矩阵进行 R-型聚类分析,将客体划分成类即范畴;然后再为用户分配范畴.聚类是无监督的,主要依据访问控制矩阵各元素之间的相似性或距离进行分类.在范畴划分问题中,用户根据经验或具备具体应用系统的背景知识,通常能够确定范畴数目的大致范围.所以,可在限定范畴个数范围条件下寻找质量最好的聚类.为了通过扫描一遍数据集一次性地构造出所有合理的划分组合,我们采用层次聚类算法[15],并通过在自底向上层次式的簇合并过程中评价聚类质量的方法来提高计算效率.

适合分类数据的类间距离计算方法有最短距离法和最长距离法.在层次聚类中,最短距离法由于在选择要合并的簇时对大簇有偏好,最终得到的分类的大小相差悬殊;而按最长距离法得到的各个类的大小相对均匀.因此,本文选择按最长距离法计算类间距离,即类 C_i,C_K 之间的距离 $d(C_i,C_K)=\max_{x\in C_i,y\in C_K}\|x-y\|$.

在聚类质量度量方面,评价指标有 Xiet-Beni 指标[16]、IGP 等[17].例如,Xiet-Beni 指标评价一个“好”的聚类原则是“簇内紧密,簇间尽可能分离”.在范畴挖掘中,我们主要考虑两方面:一是范畴划分与分配后与原访问控制矩阵的一致性;二是范畴个数对管理和安全带来的影响.通常范畴个数增加,分配关系就会复杂,管理代价也会增高,安全性就会降低.因此,本文给出基于二者的线性组合的评价指标,记为 $Q(C)$,其中,C 代表范畴.$Q(C)$的定义如下:

$$Q(C)=f(M(SK),M(KO),M(SO),|K|,\beta)=\|M(SK)\otimes M(KO)-M(SO)\|/mn+\beta|K|/m+n.$$

其中,$\beta(\beta>0)$为权重因子,用于平衡二者在取值范围上的差异.指标 $Q(C)$越小,聚类质量越高.

在定义 7~定义 9 以及聚类质量评价函数 $Q(C)$的基础上,我们给出如下基于层次聚类的范畴挖掘算法:

算法 1. 基于层次聚类的范畴挖掘算法.

输入:访问控制矩阵 $M(SO)\in\{0,1\}^{m\times n}$,$1<s\leqslant|K|\leqslant t<n$,$\beta$.

输出:最优的范畴个数 k^*,$M(SK^*)$,$M(KO^*)$.

初始时,每个客体 o_i 独自成一个簇,簇集合 $C^n=\{\{o_1\},\{o_2\},\ldots,\{o_n\}\}$有 n 个元素;令 $Q^*=\infty$;

根据 $M(SO)$计算出所有客体与客体距离矩阵 $D(O)$;

for $i=n,\ldots,s+1$ **do** //自底向上层次聚类

Select $(c'_p,c'_q)=\underset{c_p,c_q\in C^i,c_p\neq c_q}{\arg\min}\ d(c_p,c_q)$; //客体间距离等于 $M(SO)$相应列间距离

$C^{i-1}=C^i-\{c'_p\}-\{c'_q\};C^{i-1}=C^{i-1}\cup\{c'_p\cup c'_q\}$;

If $s\leqslant i-1\leqslant t$ //分析聚类质量

Construct $M(KO')$ from $C^{i-1}=\{c_1,c_2,\ldots,c_{i-1}\}$;

Construct $M(SK')$ from $M(KO')$, $M(SO)$; //依据是主体是否可访问范畴内半数以上客体

$Q(C^{i-1})=f(M(SK'),M(KO'),M(SO'),i-1,\beta)$;

If $Q(C^{i-1})<Q^*$ $M(SK^*)=M(SK')$; $M(KO^*)=M(KO')$; $k^*=i-1$; $Q^*=Q(C^{i-1})$; **End if**

End if

End for

算法复杂度分析.算法初始化时,计算所有客体与客体间距离,计算时间为 $O(n^2m)$.主循环有 $n-s$ 次迭代,第 i 次迭代包括从 $n-i+1$ 个簇中合并距离最近的两个簇和评价聚类质量两个环节.$n-s$ 次迭代中,前一环节计算时间为 $O(n(n-s))$,后一环节求解聚类质量计算时间不超过 $O(nm(s-t)t^2)$,算法可以在多项式时间内完成.

4 密级最优分配

4.1 密级挖掘的数学模型

对于 SLOMP 问题,给定 $A\in P^{m\times n}$,通过第 3 节中给出的范畴挖掘算法,确定 $M(SK)$和 $M(KO)$后,我们可将 A 分成一系列无交集的子访问控制矩阵,这些子访问控制矩阵是范畴内主体对客体的访问控制矩阵.此时,SLOMP 问题就转化为针对每个子访问控制矩阵,求相应的主客体密级的最佳分配问题.

定义 10(密级的最优化挖掘问题). 已知范畴 X 包含 m 个主体、n 个客体,X 内访问控制矩阵为 $A\in P^{m\times n}$,$\delta\geqslant 0$.求正整数 $k\leqslant m+n$,X 内主体的密级向量 $p=(cs_1,cs_2,\ldots,cs_m)^T$,其中,正整数 $cs_i\leqslant k$;X 内客体密级 $q=(co_1,co_2,\ldots,co_n)^T$,其中,正整数 $co_i\leqslant k$,使得$\|A-pq^T\|\leqslant\delta$,并最小化 k,其中,$pq^T=(x_{ij})_{m\times n}$.若 $cs_i=co_j$,则 $x_{ij}=w$;若 $cs_i\prec co_j$,则 $x_{ij}=a$;若 $cs_i\succ co_j$,则 $x_{ij}=r$.

现实应用中,从便于管理角度考虑会限定密级数目,若系统有 m 个主体、n 个客体,密级数目限定为 k,密级分配的搜索状态空间仍会达到 k^{mn}.因此,需要一种有效算法,能够在指定密级数目前提下实现最优的密级分配.

4.2 基于遗传算法的密级挖掘算法

我们使用遗传算法求解密级分配问题.遗传算法可用于很多复杂的搜索优化问题.遗传算法首先将产生候选解决方案的种群,然后通过自然选择使这些解决方案进化,从而使得不好的解决方案趋于被淘汰,好的解决方案存活并继续繁殖.不断重复这个过程,就得到了最优的解.基于遗传算法的密级挖掘算法的描述如下:

(1) 用染色体表示密级分配.遗传算法中对问题的解以编码形式呈现,一个解对应一条染色体.编码方式有二进制、整数、实数和非数值编码等.这里,我们将每个实体分配的密级看成染色体上的基因,这样,一个染色体将代表了对密级分配问题的一个解.为了更好地适应密级类型,算法中的编码方式采用整数表示.

(2) 种群初始化.算法需要一个初始种群作为初始解集合.初始种群通过随机方式产生,其产生的质量通常会对算法搜索效率和能否产生全局最优解产生大的影响.为了保证初始种群的多样性,定义第 i 个基因位上的基因熵[18] $Entropy(i)=-\sum_{j=0}^{k-1}p(j)\log_2 p(j)$,其中,$p(j)$是初始种群中在第 i 基因位取值为 j 的基因所占比例,

[0,k−1]是基因池,对应密级取值空间.设定阈值θ,若不能满足 $\log_2(k)-Entropy(i)<\theta$,则将所占比例最高的一个基因替换为所占比例最低的基因.重复该过程,直至该位基因熵满足上述不等式.这里,取值$\theta=0.3\log_2 k$.初始种群的规模可针对应用实例规模通过实验获得,若数目太小,则容易陷入局部最优解;若数目太大,则计算复杂度又较高.

(3) 适应度函数和选择方法.适应度函数反映了个体的适应能力,适应度函数值的大小决定某些个体是繁殖还是消亡.设范畴内访问控制矩阵 $A_{m\times n}$,染色体 b 表示对全部主、客体的一种密级分配方法,通过多级安全规则,我们可得到相应的访问控制矩阵 $A'_{m\times n}$.染色体 b 适应度函数定义为 $f(b)=1-\|A-A'\|/mn$.染色体选择方法采用轮盘赌选择(roulettete wheel selection)结合最优个体保存方法,文献[19]证明,二者结合的方法可使进化收敛到全局最优解.在密级分配问题中,从当代种群$\{b_1,b_2,\ldots,b_c\}$中轮盘赌选择当代个体 b_i 成为下一代成员父代的概率 $p(b_i)=f(b_i)\Big/\sum_{j=1}^{c}f(b_j)$. 当得到新一代种群后,将老一代中最优个体也直接加入其中,淘汰适应度值最小的个体.这里,实现轮盘赌的方法是计算各个 b_i 的轮盘刻度为 $s(b_i)=f(b_i)\Big/\sum_{j=1}^{i}f(b_j)$. 当随机产生(0,1)之间的一个刻度值 t 时,若 $b_{i-1}<t\leqslant b_i$,则 b_i 即为选中的个体.

(4) 交叉.交叉是指两个父代个体的部分结构加以替换重组而生成新个体的操作.常用的交叉算子包括单点交叉、二点交叉、均匀交叉等.在密级分配问题中,主客体对应到染色体的位置并无前后次序要求,因此我们采用均匀交叉方法,即两个相同配对个体的每个基因都以相同的概率进行交换,从而形成两个新个体.

(5) 变异.变异用于对个体的编码串产生随机的小变化,即以很小概率选择从群体中选出一些染色体,随机选择某些基因位,改变其值,取值范围[0,k−1],对应密级取值空间.若变异概率太大,则会导致搜索产生振荡;若变异概率太小,则容易得到局部最优解.变异概率的选择可针对应用实例规模,通过实验获得.

(6) 终止条件.在指定遗传代数后中止遗传算法,并检查种群中的最优的染色体,如果没有得到满意的解决方案,则遗传算法重新启动.

算法复杂度分析.假设范畴内访问控制矩阵是 $A_{m\times n}$,群体规模为 l,迭代次数为 t,则上述过程中,初始化种群的时间为 $O(l(m+n))$;每一轮迭代中,计算个体适应度的时间为 $O(lmn)$,计算交叉和变异的时间为 $O(l(m+n))$.因此,总的计算适应度时间 $O(ltmn)$.

5 实验与性能分析

5.1 实验方法和评价指标

我们通过实验评测来分析和验证本文提出的算法性能.实验采用人工合成数据和实际数据两种数据源,前者能够较为全面地反映算法性能,后者能够反映算法解决实际场景中问题的应变能力.敏感标记算法实现形式是 Vsual C++6.0 控制台程序,包括数据预处理、范畴划分、密级分配、性能评估几个过程,时间计量精度为 ms.实验硬件环境是 Intel® Pentium® processor 1.73G,2G 内存,软件环境是 Microsoft Windows Server 2003 sp2.

我们给出的算法评价指标包括:范畴挖掘时间 KMT、密级挖掘时间 CMT、算法总的执行时间 TT、范畴挖掘准确率 KAR、密级挖掘准确率 CAR、算法总的准确率 TAR.评价指标的定义如下:

假设访问控制矩阵 $A=(a_{ij})_{m\times n}$ 将 A 转化成的有权限(读、写、读/写)和无权限(空权限)的布尔型矩阵记为 A'.设范畴挖掘得到的主体与范畴关系的布尔矩阵为 B,得到的范畴与客体关系矩阵为 C,则范畴挖掘准确率 $KAR=\|B\otimes C-A'\|/mn$.设共挖掘到 k 个范畴,范畴 i 内有 s_i 个主体、t_i 个客体,通过执行密级分配算法确定主、客体在该范畴内的密级,基于多级安全规则得到范畴 i 内访问控制矩阵 D_i,将它与 A 中相应主体对客体的权限比较,统计不同的个数记为 Ni,则密级挖掘准确率 $CAR=(1/k)\sum_{i=1}^{k}Ni/(s_i\times t_i)$.根据敏感标记挖掘算法得到的主、客体标记,可推出其对应的访问控制 E,则算法总的准确率 $TAR=\|E-A\|/mn$.

5.2 人工合成数据评测

这里,我们首先随机产生完全符合多级安全策略的授权策略;然后逐步加入噪音,直至授权策略数据完全随

机化,以此来探讨算法的挖掘能力.实际授权系统中的授权策略情况大部分应处于两者之间.

(1) 无噪音数据.无噪音数据产生办法如下:首先指定主体个数 m、客体个数 n、范畴个数 k、密级个数 c;然后基于集合{0,1}随机生成主体与范畴的关系矩阵 $M(SK)_{m\times k}$ 和范畴与客体的关系矩阵 $M(KO)_{k\times n}$;接着对这 k 个范畴内的主客体基于集合{0,1,2,…,c}随机标记出密级,这样就得到了多级安全策略;最后,将其导出的访问控制矩阵作为算法求解的问题.在不同 m,n,k 和 c 情况下的算法性能见表 1.算法估计的范畴范围记为 k_{zone},估计的密级记为 c_e,得到的最佳范畴个数记为 k'.出于分析准确率考虑,我们将层次聚类权重因子β设定的值较低,以降低管理复杂性对挖掘质量影响的权重.表 1 的结果是在β=1、染色体个数 l=100、遗传代数 t=1500、交叉概率 p_c=0.8、突变概率 p_m=0.05 下获得的.

Table 1 Results of algorithm performance for data without noise

表 1 数据无噪音情况下算法性能结果

m	n	k	c	k_{zone}	c_e	k'	KMT (s)	CMT (s)	TT (s)	KAR (%)	CAR (%)	TAR (%)
50	100	4	3	[2,4]	3	4	0.157	13.850	14.007	100	100	100
100	200	6	5	[6,6]	5	6	2.568	48.905	51.473	100	99.01	99.98
				[6,6]	4	6	2.561	49.522	52.113	100	94.59	97.21
				[3,5]	5	5	2.595	53.843	56.438	95.74	91.92	93.03
				[3,5]	4	5	2.637	54.432	57.069	93.57	87.91	90.09
200	100	6	5	[6,6]	5	6	0.730	49.601	50.331	100	98.82	99.25
				[6,6]	4	6	0.730	49.336	50.066	100	94.53	97.23
				[3,5]	5	5	0.728	50.259	50.987	95.20	90.47	92.78
				[3,5]	4	5	0.724	51.427	52.151	95.22	89.85	91.86
400	400	10	5	[10_10]	5	10	104	389	493	100	98.22	99.24
400	400	20	5	[20_20]	5	20	117	525	642	100	98.01	99.01
500	600	15	5	[15_15]	5	15	420	782	1 202	100	97.32	98.56

表 1 表明,对无噪音数据进行挖掘,如果范畴个数和密级个数已知,则当算法正确地设置了范畴范围和密级个数时,算法可基本正确地逆向求解出原来的主、客体标记.在上述实验中,算法均找到了完全正确的范畴划分.而且,如果指定足够的染色体个数和遗传代数,密级分配也会收敛到完全正确的解.若算法设置的范畴个数范围未能包含正确的范畴个数或设置的密级个数小于正确的密级个数,则算法不会收敛到正确解;然而,如果差距不大,从表 1 可以看出,正确率仍达到 90%以上.由于范畴划分的结果对随后的密级分配有直接影响,范畴个数范围设置是否合适,较密级个数设置是否合适对 TAR 的影响更大.

在计算时间方面,表 1 的结果说明,当主、客体数目增大时,计算时间迅速增长.KMT 受客体影响较大,这主要是因为层次聚类算法计算时间和客体数目成平方关系而造成的.CMT 与主、客体总的数目成正比例关系.同时,CMT 受范畴个数影响也较大,这是由于每个范畴内都要独立进行密级分配的原因.

(2) 有噪音数据.在无噪音数据中加入噪音的办法如下:针对多级安全系统导出的访问控制矩阵 $A_{m\times n}$,若噪音比例为 P_{noise},则在 $A_{m\times n}$ 中随机选择 $P_{noise}\times m\times n$ 个元素,其值随机修改为 4 种权限之一.图 1 显示了在 m=100,n=200,k=6,c=5,β=1,l=150,t=2000,p_c=0.8,p_m=0.05 情况下,噪音对挖掘准确率的影响.

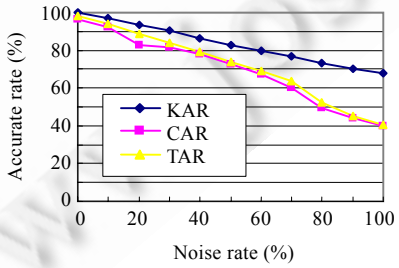

Fig.1 Algorithm performance for data with noise

图 1 数据有噪音情况下算法性能

图 1 表明,随着噪音比例 P_{noise} 的增大,准确率逐渐降低.当 P_{noise}=0.5 时,准确率仍达到 70%左右;当 P_{noise}=1.0 时,准确率降到 40%左右.考虑到授权策略的随机性以及策略冲突等原因,结果表明,算法还是有效地挖掘到了尽可能多的符合多级安全规则的策略.图 2 显示了在上述其他参数相同的情况下,分别连续增加 5 次 k 值和 c 值对算法的影响,其中,k 每次增加 25,c 每次增加 5.图 2 表明,k 值的增加提高了范畴划分的灵活性,可显著提高准确率.而 c 的增加对准确率没有显著影响,甚至产生振荡.原因可能是 c 已达到了最优值,或者遗传算法还没有充分进化.

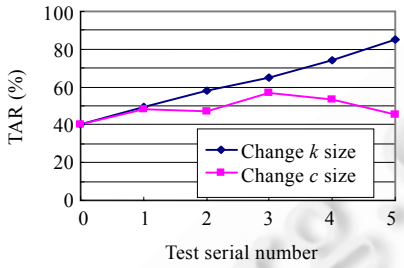

Fig.2 Parameters adjustment influence on algorithm performance for data with noise

图 2 数据有噪音情况下,参数调整对算法性能的影响

如果一个授权系统通过本算法分析评估,发现迁移到多级安全系统出错率较高,则说明其包含的符合严格 BLP 策略的模式较少.随着改进的多级安全模型的提出和标准化,相信这种情况下系统迁移仍将会可行.例如,扩展 BLP 模型[1−3]允许可信主体可以违背*-属性,这是通过一定范围内安全级调整来进行的.图 3 显示了在限定不同密级合法范围情况下,解符合扩展 BLP 模型的情况.可以看出,随着密级合法范围增大,准确率显著提高.

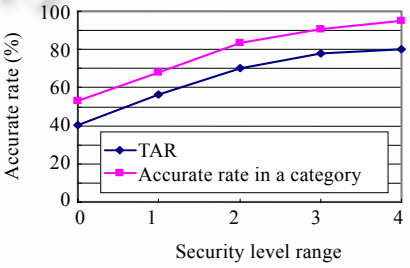

Fig.3 Influence of dynamic secret level adjustment on algorithm performance

图 3 密级可动态升降对算法性能的影响

5.3 实际授权数据评测

这里以信息工程大学的 OA 系统为例,挖掘其中的多级安全策略.系统面向学生、教师和行管人员,提供了教学管理、科研管理和行政管理等服务,采用了基于角色的授权访问控制(RBAC).学生虽然人数很多,但学生群体的权限相似性很大;不少教师是多角色,他们之间也有一定的相似性;行政人员人数较少,但权限个性化明显.主体权限以读/写和读为主,写权限极少.算法运行前需要将 RBAC 授权策略转化为主体对客体的访问控制矩阵.

在范畴分配阶段,通过分析,我们抽取有代表性的少部分学生和角色不同的部分老师,将他们和全部行管人员作为算法的输入主体.在该系统中,输入主体共有 48 个,客体有 162 个.令β=3,得到范畴挖掘准确率 KAR=83%,范畴划分结果见表 2,包括门户、本科教学、研究生教学等 8 个范畴.其中,staff 和 teacher 加下划线分别表示行政人员一部分和教师一部分.

Table 2 Results of category partition of an OA system

表 2 某 OA 系统的范畴划分结果

Category description	Portal	Undergraduate teaching	Graduate teaching	Research management	Administrative office	Employee services
User	All	Undergraduate/ Staff/Teacher	Graduate/Staff/ Teacher	Teacher/ Staff	Staff/ Executive	Teacher/Staff/ Executive
Resource number	15	35	30	26	34	22

在密级分配阶段,我们用 5,8 和 20 作为密级数目进行测试,算法运行参数为 l=150,t=10000,p_c=0.8,p_m=0.05,结果如图 4 所示.可以看出,密级数目 5 的挖掘在遗传到 6 000 代时,已经收敛到 64%附近;而密级数目为 8 和 20 的挖掘在遗传到 8 000 代时,均向 76%附近收敛.这说明 8 和 20 作为挖掘出来的密级数目,很有可能是全局最优解,最优的密级数目可能就在 8 左右.所以,我们选择密级数目为 8 的密级分配解作为最终的解,最后得到算法总的准确率 TAR=78%,实现了该 OA 系统向多级安全系统的迁移.

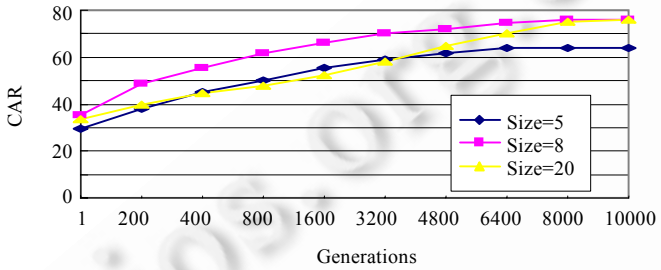

Fig.4 Results of secret level assignation for real-world data

图 4 对实际数据的密级分配结果

6 结 论

策略转换是目前我国等级保护推进工作的一个重要内容.本文形式化地描述了多级安全中敏感标记最优化挖掘问题,证明了它是 NP 完全问题.我们将问题分解为范畴划分和范畴内密级分配两个过程,提出了基于层次聚类和遗传算法的方法来求解该问题.算法的复杂度分析和实验结果表明,该算法能够有效地分析出已有授权策略中存在的符合多级安全规则的模式.该方法可应用于非多级安全系统向多级安全系统迁移的自动实现,也同样适用于其他策略向基于标记的完整性模型(如 Biba 模型)的转换.

下一步将研究在允许可信主体的安全级动态调整的扩展多级安全模型下敏感标记的最优化挖掘问题,如何获得更高准确率的策略转换算法也将会是进一步需要研究的问题.

References:

[1] Bell D, LaPadual LJ. Secure computer system: Unified exposition and MULTICS interpretation. Technical Report, MTR-2997 Rev.1, Bedford: The MITRE Corporation, 1976.

[2] Bell D, LaPadual LJ. Secure computer systems: Mathematical foundations. Technical Report, MTR-2547 Vol I, Bedford: The MITRE Corporation, 1973.

[3] Wu YJ, Liang HL, Zhao C. A multi-level security model with least privilege support for trusted subject. Journal of Software, 2007,18(3):730−738 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/730.htm [doi: 10.1360/jos180730]

[4] Li YF, Shen CX. A new security model for operating system. Science in China (Series E): Information Sciences, 2006,36(4): 347−356 (in Chinese with English abstract).

[5] Chaudhuri A, Naldurg P, Rajamani SK. EON: Modeling and analyzing dynamic access control systems with logic programs. In: Proc. of the 15th ACM Conf. on Computer and Communications Security (CCS 2008). New York: ACM Press, 2008. 181−390.

[6] Vaidya J, Atluri V, Warner J. Roleminer: Mining roles using subset enumeration. In: Proc. of the 13th ACM Conf. on Computer and Communications Security (CCS 2006). New York: ACM Press, 2006. 144−153. [doi: 10.1145/1180405.1180424]

[7] Zhang D, Ramamohanrao K, Ebringer T. Role engineering using graph optimisation. In: Proc. of the 12th ACM Symp. on Access Control Models and Technologies (SACMAT 2007). New York: ACM Press, 2007. 139−144. [doi: 10.1145/1266840.1266862]

[8] Vaidya J, Atluri V, Guo Q. The role mining problem: Finding a minimal descriptive set of roles. In: Proc. of the 12th ACM Symp. on Access Control Models and Technologies. New York: ACM Press, 2007. 175−184. [doi: 10.1145/1266840.1266870]

[9] Lu H, Vaidya J, Atluri V. Optimal Boolean matrix decomposition: Application to role engineering. In: Proc. of the 24th Int'l Conf. on Data Engineering (ICDE 2008). Cancun: IEEE Computer Society Press, 2008. 297−306. [doi: 10.1109/ICDE.2008.4497438]

[10] Frank M, Basin D, Buhmann JM. A class of probabilistic models for role engineering. In: Proc. of the 15th ACM Conf. on Computer and Communications Security (CCS 2008). New York: ACM Press, 2008. 299−309. [doi: 10.1145/1455770.1455809]

[11] Molloy I, Chen H, Li TC, Calo S, Lobo J, Wang QH, Li NH, Bertino E. Mining roles with semantic meanings. In: Proc. of the 13th ACM Symp. on Access Control Models and Technologies (SACMAT 2008). New York: ACM Press, 2008. 21−30. [doi: 10.1145/1377836.1377840]

[12] Bauer L, Garriss S, Reiter MK. Detecting and resolving policy misconfigurations in access-control systems. In: Proc. of the 13th ACM Symp. on Access Control Models and Technologies. New York: ACM Press, 2008. 185−194. [doi: 10.1145/1377836.1377866]

[13] Miettinen P. The discrete basis problem [MS. Thesis]. Helsinki: University of Helsinki, 2006.

[14] Miettinen P, Mielikainen T, Gionis A, Das G, Mannila H. The discrete basis problem. IEEE Trans. on Knowledge and Data Engineering, 2008,20(10):1348−1362. [doi: 10.1109/TKDE.2008.53]

[15] Berkhin P. Survey of clustering data mining techniques. Technical Report, EE242, San Jose: Accrue Software, 2002.

[16] Xie X, Bcni G. A validity measure for fuzzy clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1991,13(8): 84l−847. [doi: 10.1109/34.85677]

[17] Kapp AV, Tibshirani R. Are clusters found in one datasetpresent in another dataset? Biostatistics, 2007,8(1):9−31.

[18] Maekawa K, Mori N, Tamaki H, Kita H, Nishikawa Y. A genetic solution for the traveling salesman problem by means of a thermodynamical selection rule. In: Fukuda T, Furuhashi T, eds. Proc. of the IEEE Conf. on Evolutionary Computation. New York: IEEE Press, 1996. 529−534. [doi: 10.1109/ICEC.1996.542655]

[19] Rudolph G. Convergence analysis of canonical genetic algorithms. IEEE Trans. on Neural Networks, 1994,5(1):96−101. [doi: 10.1109/72.265964]

附中文参考文献:

[3] 武延军,梁洪亮,赵琛.一个支持可信主体特权最小化的多级安全模型.软件学报,2007,18(3):730−738. http://www.jos.org.cn/1000-9825/18/730.htm [doi: 10.1360/jos180730]

[4] 李益发,沈昌祥.一种新的操作系统安全模型.中国科学(E 辑,信息科学),2006,36(4):347−356.

杨智(1975－),男,河南开封人,博士生,讲师,主要研究领域为网络安全.

段淇毅(1953－),男,博士,研究员,博士生导师,主要研究领域为网络安全,人工智能.

金舒原(1974－),女,博士,副研究员,主要研究领域为网络安全.

方滨兴(1960－),男,博士,教授,博士生导师,中国工程院院士,主要研究领域为网络与信息安全,并行处理.