

# 一种基于粗糙集的多级安全数据库推理问题的量化分析方法

杨 武<sup>1</sup> 黄志真<sup>1</sup> 李立新<sup>2</sup>

(重庆工学院计算机学院 重庆 400050)<sup>1</sup> (解放军信息工程大学电子技术学院 郑州 450004)<sup>2</sup>

**摘 要** 信息大都存放在数据库中,信息安全很大程度上依赖数据库的安全。推理是从已知的信息推出新的信息,它是对数据库安全的一个重要威胁,特别数据挖掘的发展,这个威胁变得更加严重。简单描述了多级安全数据库系统的推理问题;介绍了粗糙集理论的相关概念;利用粗糙集理论提出了一种对推理风险进行量化的方法,它不依赖系统安全员的知识,能够处理确定性和非确定性的推理通道;并给出了一个分析例子。

**关键词** 安全数据库系统,数据挖掘,粗糙集,特征函数

## An Inference Quantized Analysis Scheme of MLS DBMS Using Rough Set

YANG Wu<sup>1</sup> HUANG Zhi-Zhen<sup>1</sup> LI Li-Xin<sup>2</sup>

(College of Computer Science & Technology, Chongqing Institute Technology, Chongqing 400050)<sup>1</sup>

(Electronic technology College of PLA's Information Engineering University, Zhengzhou 450004)<sup>2</sup>

**Abstract** Since information is almost stored in database system, information security depends on database system security. Inference, which is to get new information from known information, threatens database system security. Especially with the progress of data mining, the problem is more serious. The inference for multi-level security database system and the concepts about Rough set are introduced. Moreover, an inference quantized analysis scheme of multi-level security database system based on rough set and an example are presented.

**Keywords** Security database system, Data mining, Rough set, Eigenfunction

## 1 引言

推理,即从已知信息推出新的信息。在多级安全数据库系统中,推理不同于信息泄露,不是由安全机制被突破而引起,而是与应用相关的信息和语义。因此,推理问题不可能通过 DBMS 内部的机制得到一个完整的解决方法。推理是数据库安全的一个重要威胁,通过推理,用户能够从数据库中获得未授权访问的敏感数据或信息。多级安全数据库系统的推理问题是指如何阻止用户推理出未授权访问的信息<sup>[1,2]</sup>。

研究者们对多级安全数据库系统的推理问题进行了大量研究。目前,对多级安全数据库推理问题的研究基于系统安全员的 SSO 的知识,由 SSO 输入的知识代表了应用的语义,由于 SSO 不可能知道全部的数据依赖,因此只提供了一个部分的解决方法<sup>[3]</sup>。随着数据仓库和数据挖掘技术(又称数据库中的知识发现 KDD,指从大型数据库或数据仓库中提取隐含的、未知的、有潜在应用价值的信息或模式)的发展,利用数据挖掘方法,从数据库中的数据自身来研究推理问题成为一个很有希望的研究方向<sup>[4]</sup>。

粗糙集(Rough Set)理论是一种研究模糊、不确定性知识的数学方法,由波兰科学家 Pawlak 在 1982 年提出。该方法主要通过数据分析、近似分类、推理数据间的关系,来发现隐含的知识,揭示潜在的规律,因而粗糙集理论成为数据挖掘、模式识别、机器学习等领域的一种重要研究方法。粗糙集和 KDD 关系尤其密切,它为 KDD 提供了一种新的方法和工具<sup>[5,6]</sup>。

本文第 2 节介绍了粗糙集理论,第 3 节给出一种新的基于粗糙集的推理问题的量化分析方法,第 4 节给出一个实例,最后是本文的总结。

## 2 粗糙集理论简介

粗糙集的研究主要基于分类,它对不精确概念的描述是

通过上近似和下近似概念来描述的。下面首先介绍粗糙集理论的相关概念<sup>[7]</sup>。

### 2.1 知识表示系统

粗糙集理论把客观世界或对象世界抽象为一个知识表示系统  $S = \langle U, A, V, F \rangle$ 。其中,  $U$  是一组对象的有限集合,称论域,设有  $n$  个对象,则  $U = \{x_1, x_2, \dots, x_n\}$ 。  $A$  是有限个属性的集合,设有  $m$  个属性,则  $A = \{a_1, a_2, \dots, a_m\}$ 。  $V$  是属性的值域集:  $V = \{V_1, V_2, \dots, V_m\}$ , 其中  $V_i$  是属性  $a_i$  的值域。  $A$  又可进一步划分为两个不相交的集合: 条件属性集  $C$  和决策属性集  $D$  ( $D$  一般只有一个属性)。  $f$  是函数:  $f: U \times A \rightarrow V$ ,  $f(x_i, a_j) \in V_j$ <sup>[7]</sup>。

知识表示系统也称信息表,其中表列是属性,共有  $m$  列;表行是对象,共有  $n$  行。第  $i$  行第  $j$  列上的表项内容是  $f(x_i, a_j)$ 。表中的一行表示了知识表示系统中该对象的所有信息。

### 2.2 近似空间

近似空间是一个二元组:  $\langle U, IND(B) \rangle$ 。其中,  $U$  的含义同上,  $B$  是  $A$  的属性子集,  $IND(B)$  称为无差别关系,是  $U$  上的二元等价关系:  $IND(B) = \{(x_1, x_2) | f(x_1, b) = f(x_2, b)\}$ ,  $\forall b \in B$ 。

$IND(B)$  把  $U$  划分为若干个等价类:  $X_1, X_2, \dots, X_l$ 。对于归纳或分类学习,要学习的概念一般根据决策属性集  $D$  来划分,每个概念是  $IND(D)$  上的一个等价类,用  $U/B$  代表  $IND(B)$  上的等价类族<sup>[7]</sup>。

### 2.3 下近似和上近似

对任意一个概念(或集合)  $o$ ,  $u$  是对象集合,  $A$  是属性集合,  $B$  是  $A$  的一个属性子集,  $O$  的下近似定义为:  $\underline{BO} = \{x | x \in U : [x]_{IND(B)} \subseteq O\}$ ;  $O$  的上近似定义为:  $\overline{BO} = \{x | x \in U : [x]_{IND(B)} \cap O \neq \emptyset\}$ 。其中  $[x]_{IND(B)}$  表示  $x$  在  $IND(B)$  上的等价类。

## 2.4 约简和核

设有两个属性集  $B_1$  和  $B_2$ ,  $B_2 \subset B_1$ 。如果  $IND(B_1) = IND(B_2)$ , 且  $B_2$  不存在真子集  $B_3$  满足:  $IND(B_3) = IND(B_2)$ , 则称  $B_1$  可约简为  $B_2$ 。一个属性集的约简可能有多个,  $A$  的所有约简的交集称为核, 核可能为空。

## 2.5 属性依赖度

设有两个属性集  $P$  和  $Q$ , 则  $P$  对  $Q$  的依赖度定义为:  $\gamma_P(Q) = \#POS_P(Q) / \#U$ , 其中  $POS_P(Q) = \bigcup_{x \in U/Q} P - X$ ,  $P - X$  表示集合  $X$  在属性集上的下近似,  $U/Q$  表示论域  $U$  根据属性集  $Q$  所得到的划分类,  $\#U$  表示集合  $U$  的元素个数。

## 2.6 属性重要度

设属性  $a \in C$ ,  $C$  是条件属性集,  $D$  是决策属性集, 则  $a$  的属性重要度定义为:

$SGF(a, C, D) = (\gamma_C(D) - \gamma_{C-(a)}(D)) / \gamma_C(C, D)$ 。表明从  $C$  中去除  $a$  属性后对分类决策的影响程度。

## 3 基于粗糙集的推理问题的量化分析方法

### 3.1 假设及说明

在进行推理问题的量化分析前, 首先给出假设:

假设 1 数据存储在由表构成的数据库中, 其敏感数据具有元素级安全标示。

假设 2 数据库看成只由一个表构成, 它包含了此数据库中所有的属性。

假设 3 系统仅有两个安全标示: 公开级(U级)与秘密级(C级)。

说明: 在多级安全数据库系统中, 安全标示一般分为元素级和元组级, 元素级的安全标示是最一般的情况。而将数据库看成只由一个表构成, 可以避免考虑数据库设计导致的推理通道。

为了将粗糙集理论运用在关系数据库系统中, 首先将一个关系表看成是一个知识表示系统 KRS, 其中表的列看成是 KRS 中的属性, 表的行看成是对象, 表中  $p$  列  $x$  行具有值  $p(x)$ , 表中的每一行表示了论域  $U$  中某对象的信息。

关系模型并不关心存储在关系表中的信息的含义, 哪些信息存储在关系表中可以不必在关系表中表示。而在 KRS 中, 即 KRS 中的每一项都与对象的特性、性质等明显的含义相关联。可以这样协调两种模型: 将关系表  $R$  中的每一个元组看成是论域  $U$  中的一个对象, 将关系表  $R$  中的主关键字  $K$  作为 KRS 的对象的标示符而其余的属性 ( $A-K$ ) 作为 KRS 中的属性。

### 3.2 推理特征函数的定义与分析

对一个  $C$  级的数据集合  $(x, P)$ , 它具有列集合  $p$  和行  $x$ , 其值集是  $P_x$ , 设属性集合  $B$  代表其值为  $C$  级的对象  $x$  的属性集合,  $D=P$ , 条件属性集合  $C=A-B$ , 表征其推理风险的特征值  $IRI$  定义如下<sup>[7]</sup>:

$$IRI(x, p) = \frac{\#([x]_{IND(C)} \cap [x]_{IND(D)})}{\#([x]_{IND(C)})} \quad (1)$$

这样  $IRI$  就提供了一种定量表征推理风险的方法。

从上面的定义可得出: 对于任何  $C$  级的数据元素集  $(x, P)$ ,  $0 \leq IRI(x, p) \leq 1$  成立。如果  $IRI(x, P) = 1$ , 则有  $[x]_{IND(C)} \subseteq [x]_{IND(D)}$ , 即对于所有  $[x]_{IND(C)}$  中的对象有一条确定性规则能够从  $C$  推出  $D$ 。由于  $x \in ([x]_{IND(C)} \cap [x]_{IND(D)})$  始终成立, 因此, 对于在  $[x]_{IND(C)}$  中的对象, 从  $C$  到  $D$  总是可以推导出一个可能的规则, 故  $IRI(x, p) \geq 0$  始终成立。

如果  $IND(C) \subseteq IND(D)$  (即属性集  $D$  函数依赖于属性集  $C$ ), 则对于任何  $P_x$ ,  $IRI(x, p) = 1$ , 即攻击者在已知属性集合  $C$  的情况下就可以推理出  $D$ 。当然也存在这样一种情况: 在  $C$  和  $D$  之间没有函数依赖而  $IRI$  值仍然为 1, 这是由于对

某些  $P_x$  的特定值存在(确定性规则)。为了排除数据干扰, 我们可以在条件属性集  $C$  中选择最重要的属性集  $C_s$  取代属性集  $C$  来计算  $IRI$ 。

由于在实际中攻击者只能看见一部分所需要的数据, 不可能看见全部所需要的数据。而  $IRI$  的计算是基于数据库中的所有相关数据, 故  $IRI$  对推理风险的估计是相当充分的。

## 4 一个 IRI 的计算例子

下面用一个例子说明  $IRI$  的计算。设有一个如表 1 所示的多级安全数据库表。此关系可以被看成以主关键字 EMP-ID 作为对象标示符的 KRS, KRS 中的属性是 *Grade*, *Location*, *Salary*, 其中  $C$  代表该数据项的安全级为秘密级别, 其余为公开级别。

首先令:  $C = \{Grade, Location\}$ ,  $D = \{Salary\}$

计算对象 5 和属性 *Salary* 的值 7000 ( $C$  级) 的  $IRI$  值:  $U/C = \{\{1, 5\}, \{3, 7\}, \{2\}, \{4\}, \{6\}, \{8\}, \{9\}, \{10\}\}$ ,  $U/D = \{\{1\}, \{2\}, \{3, 7\}, \{4, 6\}, \{5\}, \{8\}, \{9, 10\}\}$ 。因此,  $IRI(5, \{Salary\}) = 0.5$ , 同理可以得到:  $IRI(7, \{Salary\}) = 1$ 。

进一步分析条件属性 *Grade* 和 *Location* 的属性重要度:

$\#U = 10$ ,  $POS_C(D) = \{2, 3, 7, 4, 6, 8, 9, 10\}$ ,  $\gamma_C(D) = 8/10 = 0.8$ ,  $\gamma_{C-(Grade)}(D) = 0.1$ ,  $\gamma_{C-(Location)}(D) = 0.5$ 。由此可得:  $SGF(Grade, C, D) = (0.8 - 0.1)/0.8 = 0.875 \approx 0.88$ ,  $SGF(Location, C, D) = (0.8 - 0.5)/0.8 = 0.375 \approx 0.38$ 。

表 1 示例图

| EMP-ID | Grade | Location | Salary  |
|--------|-------|----------|---------|
| 1      | 2     | CA       | 8000    |
| 2      | 3     | NY       | 6000    |
| 3      | 5     | TX       | 5000    |
| 4      | 8     | TX       | 2000    |
| 5      | 2     | CA       | 7000, C |
| 6      | 8     | CA       | 2000    |
| 8      | 5     | CA       | 4000    |
| 9      | 7     | NY       | 3000    |
| 10     | 7     | MA       | 3000    |

进一步使用  $C_s = \{Grade\}$  来计算  $IRI$ :  $U/C_s = \{\{1, 5\}, \{2\}, \{3, 7, 8\}, \{4, 6\}, \{9, 10\}\}$ , 得到:  $IRI(5, \{Salary\}) = 0.5$ ,  $IRI(7, \{Salary\}) = 0.67$ 。

这给出了属性 *Grade* 和 *Salary* 之间一个更加精确的评价。

小结 文章简要介绍了多级安全数据库系统的推理问题, 介绍了粗糙集理论的基本概念, 基于粗糙集理论, 对多级安全数据库系统的推理规则给出了一种量化分析方法, 并以一个实际例子进行了分析。

## 参考文献

- Hinke T H, Harry S. Protecting databases from inference attacks. *Computer & Security*, 16(8): 687~708
- Marks D G. Inference in M.L.S Database Systems. *IEEE Trans. on Knowledge and Data Engineering*, 1996, 8(1)
- Nyanchama M, Osborn S L. Access Rights Administration in Role-Based Security System. In: J. Biskup, M. Morgenstern, et al. eds. *Database Security VIII: Status & Prospects*, Proc. of the Eighth Annual IFIP TC11 Workgroup Conf. on Database Security, IFIP, North-Holland, Aug. 1994. 35~76
- Thuraisingham B, Schlipper L. Security issues in data warehousing and data mining: panel discussion, *Database Security XI*, Chapman & Hall Press, 1998
- Li Lixin, Liao Changrong, Chen Weiming, et al. Inference Attack Analysis of M.L.S DBMS Using Rough Set Theory. In: Proc. of 7th joint Intl. Computer Conf. Press of Shantou University, P1339~1452
- 程岩, 黄梯云. 基于 Rough Set 理论发现最小依赖关系的方法研究. *计算机工程*, 2000, 20(3)
- Duntsch I, Gediga G. *Rough Set Data Analysis*. *Encyclopedia of Computer Science and Technology*, 2000, 43: 281~301