文章编号:1001 - 2486(2005)02 - 0052 - 04

基于 Rough 集理论的数据库推理通道动态消除*

张增军,戴江山,肖军模
(解放军理工大学 通信工程学院,江苏 南京 210007)

摘 要:提出了一种基于 Rough 集理论的数据库推理泄漏通道消除方法。在由数据库中所有数据生成的不完备决策表上,该方法应用 Rough 集理论,分析提取出敏感和非敏感数据之间的确定性推理关系,以此产生推理控制规则。利用这些规则对数据库系统返回给普通用户的数据动态地做最小修改,防止推理通道的产生。实验结果表明,该方法可扩展性强,在保证较高的数据库安全性的同时提高了数据可用性。

关键词:Rough 集理论;推理控制;推理通道;数据的可用性
中图分类号:TP309 文献标识码:A

Dynamic Elimination of Inference Channels in the Database Based on Rough Set Theory

ZHANG Zeng-jun, DAI Jiang-shan, XIAO Jun-mo
(Institute of Communication Engineering, PLA Univ. of Sci. & Tech., Nanjing 210007, China)

Abstract: This paper describes an approach to elimination of inference channels in the database based on rough set theory. The approach builds an incomplete decision table on all data in the database, then analyzes and discovers all the relations between non-sensitive and sensitive data with rough set theory. According to these relations, rules of inference control are generated and used to modify the data queried by generic users dynamically and most parsimoniously so as to eliminate inference channels. Experimental result shows that the approach is scalable and preserves security of inference control while improving availability of the data in the database.

Key words: rough set theory; inference control; inference channels; availability of data

推理泄漏的检测和控制是数据库安全研究的重点和难点之一。在数据库中,非敏感数据和敏感数据之间普遍存在着联系,而且数据库模式层上的约束关系,也最终表现为数据之间的联系[1],因此若能防止普通权限用户获取和利用这些联系产生推理通道,即能达到推理泄漏控制的目的。目前的一些方法,都是用统计概率来描述和度量属性、数据之间的依赖关系[2,3],需要很多先验知识,因此难以准确描述和有效地实施推理泄漏的控制,而且降低了数据的可用性。

1 Rough 集在不完备决策表中的决策规则获取

在现实世界的数据中,通常会有一些属性的值是缺省的(missing)。Rough 集理论中对缺省值可以作为丢失处理,也可以认为它们是无关紧要的。但前者获得的分类规则错误率相对较小[4]。因此,本文将缺省值作丢失处理,用‘?’表示。

定义 1 四元组 $T=(U,A,f,V)$ 表示一个决策表,其中,U 是对象的非空有限集合;A 是属性的非空集合,$A=C\cup D$,$C\cap D=\Phi$,C 称为条件属性,D 称为决策属性;集合 $V=Y_{a\in A}V_a$,V 被称为属性集 A 的值域;映射 $f:U\times A\to V_a$ 为每个对象的每个属性赋一个值。当存在一个或多个对象的属性值缺省时,T 称为不完备(incomplete)决策表,$f:U\times A\to V_a\cup\{?\}$ 称为不完备映射。

定义 2[4] 在 Rough 集理论中,一个属性—值对 (a,v) 的组(block)记为:$[(a,v)]=\{x\mid x\in U,$

* 收稿日期:2004 - 11 - 02
基金项目:江苏省基金资助项目(BK2004015)
作者简介:张增军(1977—),男,博士生。

$f(x,a)=v$, $f(x,a)\neq ?$; 对于 $B\subseteq A$ 和 $x\in U$,特征集合 $K_B(x)=\mathop{\mathrm{I}}_{a\in B}\{y\mid y\in U,\ \forall a\in B, f(x,a)=v, f(x,a)\neq ?, y\in[(a,v)]\}$。

称 $X\subseteq U$ 为一个概念,其对象通常具有相同的决策属性值。在不完备决策表 T 中,对于给定的 $B\subseteq A$,Rough 集理论使用 X 的 B – 下近似集和 B – 上近似集描述概念 X 的范畴。定义如下:

定义 3 概念 X 的 B – 下近似集 $\underline{B}X=\bigcup\{K_B(x)\mid x\in X, K_B(x)\subseteq X\}$;$X$ 的 B – 上近似集 $\overline{B}X=\bigcup\{K_B(x)\mid K_B(x)\cap X\neq\Phi\}$。

LEM2 算法使用了属性—值对的表示方式,输入一个概念 X 的下近似或上近似,算法将搜索整个属性—值对空间,得到概念 X 的一个局部覆盖(local covering)[4],并转化为一个决策规则集。决策规则的形式为:$\wedge(a,v)\rightarrow(d,w)$,其中 $a\in C$, $v\in V_a$, $d\in D$, $w\in V_d$。

定义 4 对于任一决策规则 $r:\wedge(a,v)\rightarrow(d,w)$,$LC(r)$ 是该规则条件部分所有属性—值对的集合,则 r 的最大属性—值对 $maxattr(r)=\{(a,v)\mid(a,v)\in LC(r),\forall(a_i,v_i)\in LC(r), card([(a,v)]\cap[(d,w)])\geqslant card([(a_i,v_i)]\cap[(d,w)])\}$,是用该规则分类时最重要的条件;$r$ 的支持度 $supp(r)=card(\mathop{\mathrm{I}}_{m=1}^{card(LC(r))}[(a_m,v_m)]\cap[(d,w)])/card(U)$,其中,$(a_m,v_m)\in LC(r)$,$card$ 表示集合的元素数。

使用 LEM2 算法,由概念的下近似集获得确定性的规则,由上近似得到的决策规则是不确定的。概念 X 中的对象能由确定性决策规则集惟一地判定。算法得到的规则都是最优的,即对于任意规则 r,不存在规则 $r':\wedge(a',v')\rightarrow(d,w)$,使得 $LC(r')\subset LC(r)$,且在 U 中为真。因此,规则 r 缺少任一条件属性值,则将失去决策意义,不能表达任何推理知识。

2 基于 Rough 集理论的数据库推理通道消除

基于 Rough 集理论的数据库推理通道动态消除方法分为三个部分:数据预处理、推理规则获取和动态推理控制,结构如图 1 所示。

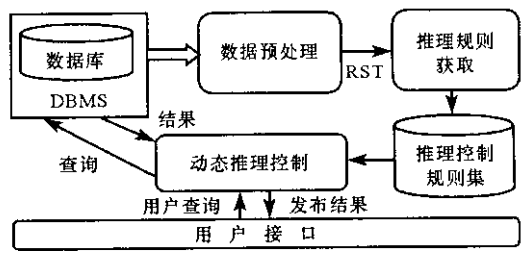

图 1 基于 Rough 集理论的数据库推理通道动态消除方法结构

Fig.1 Architecture of the approach to dynamic elimination of inference channels in the database based on RST

数据预处理过程是将数据库中的所有相关数据合并成一个不完备决策表。将敏感信息作为决策属性,由推理规则获取模块使用 Rough 集理论技术,从中提取所有决策规则,并转化为推理控制规则集。动态推理控制模块作为 DBMS 的代理,根据推理控制规则集中的规则,对 DBMS 返回给用户的查询结果做一些隐藏处理,然后再发布给用户,以消除数据中蕴含的推理泄漏通道。

2.1 数据预处理

假定数据库中包含一系列在同一个域上的相关关系表,其中的敏感和非敏感数据是以属性为单位标记的。数据预处理功能主要包括:(1)通过数据字典中定义的各关系表的主码、外码,将数据库中的所有数据形成一个全体关系样本表(Universal Relation Paradigm),包含所有属性。这一过程的具体实现可参见文献[5]。(2)对全体关系样本表中值域为实数的属性进行离散化。离散化的方法包括等频率划分算法、Naive scaler 算法等[6]。(3)对缺省值的处理。因为在数据库中,有些对象的部分属性值可能不存在或未被记录,因此在全体关系样本表中将可能存在缺省值,而且对任何用户都是未知的,应该认为是丢失的,并赋值为“?”。

2.2 运用 Rough 集理论获取推理控制规则

将数据库的全体关系样本表看成是一个不完备决策表,$Td=(U,A,f,V)$,其中所有属性组成属性集 A,每一行代表一个对象,所有的对象组成论域空间 U。$v=f(x,a)$ 是对象 $x\in U$ 在属性 $a\in A$ 上的值,所有的 $f(x,a)$ 组成 V。Td 的决策属性集 D 是数据库中不允许普通权限用户访问的敏感属性集,条件属性集 $C=A-D$。

应用 Rough 集理论技术,获取不完备决策表 Td 中最优决策规则的集合 R,并按照设定的支持度阈

值STH，对 R 中规则进行过滤。决策规则集生成算法如下：

step1 若 $D=\Phi$，则算法结束；否则，任选 $d\in D$，令 $D=D-\{d\}$，$U_X=U$；

step2 若 $U_X=\Phi$，则转step1；否则，任选 $x\in U_X$，计算 $X=[(x,d)]$，然后令 $U_X=U_X-X$；

step3 计算 $\underline{C}X$，将 $\underline{C}X$ 输入LEM2算法，得到决策规则集 R_X；

step4 对于 R_X 中每条规则 r，计算它的支持度 $supp(r)$，若 $supp(r)<$STH，则从 R_X 中删除该规则 r，令 $R=R\cup R_X$，转step2。

R 中的决策规则大部分都对应着现实中的一些知识或经验，有些虽然没有明显的实际意义，但它们是在特定的数据集中所表现出来的，可能蕴含未发掘的必然性。因此，普通权限用户可以利用这些规则和非敏感数据，产生一条从非敏感数据到敏感数据的推理通道，通过推理得到非授权的敏感信息。为了控制此类问题的发生，需限制普通权限用户利用这些规则进行推理。本文提出的方法根据这些规则生成推理控制规则，即若普通用户查询结果的数据记录满足某一规则，则隐藏该规则某一个条件对应的属性值。因为 R 中的规则都是最优的，所以经隐藏处理后，用户将无法获取并利用数据中蕴含的规则产生推理通道。但在有些情况下，用户仍能根据规则中的其他属性值，以较大可信度推理出敏感数据。为减小这种不确定推理泄漏的可能性，采用了隐藏规则中一个最大属性—值对的方法。

从 R 中产生推理控制规则集的过程如下：

对于 R 的每一条最优决策规则 r：$(a_1,v_1)\wedge(a_2,v_2)\wedge\cdots\wedge(a_n,v_n)\rightarrow(d,w)$，其中，$1\leqslant n\leqslant card(C)$：

(1)根据定义4计算 $maxattr(r)$，任选 $(a_m,v_m)\in maxattr(r)$；

(2)将 r 改写为if-then形式的控制规则：if $a_1=v_1$ and $\cdots$ and $a_n=v_n$ then HIDE a_m，加入控制规则集中。

在生成 R 的过程中，阈值STH的设置与安全要求有关，对安全性要求高，可降低STH的值，考虑更多的规则；提高STH的值，将忽略发生概率较小的偶然性规则，在保证所需安全性的条件下，可以提高数据库数据的可用性。

2.3 利用推理规则的动态推理泄漏控制

动态推理控制模块是根据推理控制规则集中的规则，对普通权限用户查返回的数据并进行动态修改，以消除其中蕴含的推理通道。为了便于描述，设数据库中各个关系表中描述同一对象的元组 t 具有相同的主码 $Key(t)$。动态推理控制的具体过程如下：

当用户发出查询“select Y from R where W”时（Y 是属性集，W 是选择条件），推理控制模块检查推理控制规则集中是否存在控制规则 r_s：if $a_1=v_1$ and $\cdots$ and $a_n=v_n$ then HIDE a_m，使得 $a_m\in Y$：

(1)若存在，则向DBMS发出查询语句“select * from R where W”。设DBMS返回元组集 $\{t_1,\cdots,t_s\}$，则推理控制模块再向DBMS发出查询语句“select $a_1,\cdots,a_n$ from ALL where (Key = Key(t_1),$\cdots$,Key = Key(t_s))”，即从数据库的所有表中，查询元组 $t_1,\cdots,t_s$ 在属性 $a_1,\cdots,a_n$ 上的值，其中，“Key = Key(t_i)”表示查询条件是主码值等于元组 t_i 的主码。从DBMS返回的数据中，对所有 $t\in\{t_1,\cdots,t_s\}$ 检查 t 是否和 r_s 匹配，若匹配，则根据 r_s 将 $f(t,a_m)$ 隐藏，用“—”表示。最后将 $\{t_1,\cdots,t_s\}$ 在属性集 Y 的投影发布给用户。

(2)若不存在，向DBMS转发用户查询，并将DBMS返回的结果直接发布给用户。

3 实验及分析

试验数据库数据采用文献[3]中的实例数据（表1、4、5、6）。Rough集的分析工具采用波兰Poznan工业大学智能决策支持系统实验室开发的软件系统ROSE2（http://idss.cs.put.poznan.pl/rose/），该软件是一个图形界面下的Rough集分析软件，可采用属性—值对的表示方法，对决策属性自动分类，并利用LEM2算法对不完备决策表计算出最优决策规则集。

同文献[3]一样，设属性AIDS为敏感的。首先根据记录的主码，合并四个关系表，并生成决策表，用“?”代替缺省值。设阈值参数STH为0，即要求最高安全性，考虑系统生成的所有决策规则。用ROSE2直接生成的最优决策规则集合，生成if-then控制规则集。然后根据这些规则对数据库表进行隐藏修改，计算修改率（数据库数据被隐藏的数量/整个数据库数据的数量），并与文献[3]的方法结果作了比较。

此外,在保持文献[3]给出的属性之间的依赖概率条件下,随机产生了记录,计算出了需要隐藏的数据数量和修改率。每种情况实验 10 次,求平均值。实验结果如表 1 所示。

表 1 两种方法对数据库数据修改量的比较

Tab.1 Comparison between modified quantities of two approaches

	在文献[3]数据库条件下		在不同记录数的情况下			
	文献[3]的方法	本文的方法	40	100	300	500
修改的数据量	26	9	17.6	41.8	130.1	219.2
修改率	0.144	0.05	0.049	0.046	0.048	0.049

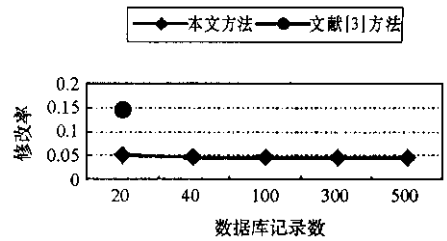

图 2 修改率的变化曲线图

Fig.2 Graph of rate of modified data

由表 1 可知,与文献[3]方法的实验结果相比,在相同数据集中,本文的方法在最高安全性要求的条件下对数据库数据的修改量减少了 65%,从而提高了数据库数据的可用性。图 2 表示在数据库记录数不断增加的情况下,本方法的数据修改率变化趋势。可以看出,即使数据库中实例数据大幅度增长,该方法对数据库数据的修改率都将保持相对固定,具有较强的可扩展性。

4 结 论

提出的基于 Rough 集理论的数据库的推理泄漏通道动态消除方法,具有以下特点:

(1) 提高了数据库数据的可用性。由于利用 Rough 集理论获得的决策规则集是最小冗余的,而且根据这些规则对用户获得的数据库记录做最小的动态修改,因此需要隐藏的非敏感信息量明显减少。

(2)较强的可扩展性,可以适用于大规模数据库中推理泄露的控制。因为通常数据库所描述的现实对象间的关系是确定的,反映在数据中的具体蕴含关系也是确定的,因此该方法获得的推理规则集也将相对固定。但在更为复杂的实例数据和属性同时增加情况下,该方法的可扩展性有待进一步验证。

(3) 提高了数据库数据的安全性。Rough 集理论作为一种分析数据内在规律和特征的技术,在该方法中被用于自动分析、获取所有非敏感数据和敏感数据之间的确定性决策规则。而这些规则大多对应着现实世界的一些规律,普通权限用户可能通过分析已知数据获得并利用这些规律产生推理泄漏通道。该方法利用最优决策规则生成的推理控制规则,隐藏少量数据,使得普通用户无法获得这些规律,从而防止了推理通道的产生,可以提高数据库控制推理泄漏的安全性。

(4) 易于实现。本方法同文献[3]方法相比,不需要先验知识,更为灵活、可行。

参 考 文 献:

[1] Marks D. Inference in MLS Database System[J]. IEEE Trans. Knowledge and Data Eng., 1996,8(1):46 – 55.

[2] Shafer G. Detecting Inference Attacks Using Association Rules[EB]. http://www.glennshafer.com/courses/downloads/raman.pdf,April, 2004.

[3] Chang L, Moskowitz L S. A Study of Inference Problem in Distributed Database Systems [A]. In:Proc. of IFIP Data Security and Applications [C],Cambridge, UK, 2002. 229 – 243.

[4] Grzymala-Busse J W, Siddhaye S. Rough Set Approaches to Rule Induction from Incomplete Data[A]. Proceedings of the IPMU '2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System[C],Perugia,Italy, July 4, 2004 2: 923 – 930.

[5] Ullman J D. Principle of Database and Knowledge-base System [M]. Vols. Ⅰ and Ⅱ, Rockville, MD:Computer Science Press, 1988, 1989.

[6] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001.