第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·安全技术· 文章编号：1000—3428(2009)23—0127—03 文献标识码：A 中图分类号：TP311.12

基于β重要度的数据隐含化

熊树洁，邱桃荣，龚科华，白小明
(南昌大学计算机系，南昌 330031)

摘 要：为权衡敏感或重要数据的公开及关键内容的隐含间的关系，提出一种基于变精度粗糙集模型中的β重要度和粒度原理的数据隐含方法，通过获取信息表中的重要属性，对次要属性的属性值进行扩展，从而使用户信息的粒度变粗，达到对信息表的数据隐含化效果。仿真实验结果表明，该方法是可行的。

关键词：β重要度；数据隐含化；变精度粗糙集；粒度

Data Anonymisation Based on β-Importance

XIONG Shu-jie, QIU Tao-rong, GONG Ke-hua, BAI Xiao-ming
(Department of Computer, Nanchang University, Nanchang 330031)

【Abstract】 In order to balance the relation between publicity of sensitive or important information and the anonymisation of privacy information, This paper proposes the β-importance based on Variable Precision Rough Set(VPRS) model and granularity theory to anonymise data. The important attributes are obtained from the information table, and the values of unimportant attributes are handled to extend. In term of users, the granularity of information becomes rough, and the anonymisation is achieved. Simulation experimental results show this method is feasible.

【Key words】 β-importance; data anonymisation; Variable Precision Rough Set(VPRS); granularity

1 概述

近年来，在互联网上收集和使用个人信息变得越来越容易。个人信息的公开对未授权的用户来说，即使不是故意的，也可能导致个人隐私权的问题。因此，如何很好地权衡信息的完整性和私密的保护性是数据隐含化这个过程中的核心问题。文献[1]采用 K-Anonymisation 技术处理这个问题，文献[2]则采用资料隐私保护 Cellsecu 系统处理数据隐含化问题，并应用到包含患者详细基本资料以及病历记录的医疗数据库系统中。然而，尝试采用粗糙集粒计算理论去处理数据隐含化问题却并不多见。

本文通过分析研究变精度粗糙集模型和一些有关知识粒度的问题，了解到数据隐含化其实就是由细粒度到粗粒度转变的同时，在一定程度上保证信息表的分类能力不变，这样就能很好地权衡私密保护和信息完整 2 个方面，即对用户和信息管理员采取不同的策略，因此，提出一种“β 重要度”概念和粒度原理保护重要或关键信息的方法。

2 知识粒度

为方便起见，这里讨论文献[3]提出的经典 Rough 集理论所主要针对的完备信息系统。

定义 1 一个完备的决策表，这里记为 $DT=(U,At,L,V,I)$，其中，U 是对象的非空有限集合称为论域；$At=C\cup D$ 是属性的非空有限集合，即 $C=\{c_1,c_2,\cdots,c_n\}$ 和 $D=\{d\}$ 分别称为条件属性集和决策属性集；对于每个 $a\in At$ 都有 $a:U\to V_a$，其中，V_a 称为 a 的值域；L 表示用属性定义的决策逻辑语言；I 是信息函数的集合，它描述的是个对象在某个属性下的取值和一个属性值一一映射的情况，即 $\forall a\in At,I_a:U\to V_a$。

定义 2[4] 在决策逻辑语言 L 中，原子式表示为 $a=v$，其中，$a\in At,v\in V_a$，则：

(1)如果 φ，ψ 是公式，则 $\neg\varphi$，$\varphi\wedge\psi$，$\varphi\vee\psi$ 也是公式。

(2)决策表中的对象 x 满足 $a=v$，当且仅当 $I_a(x)=v$。

如果 φ 是个公式，所有满足在 DT 上的公式 φ 的对象集合记为 $m(\varphi)$，即 φ 看作是对象集合 $m(\varphi)$ 的描述。由此，L 的公式与 U 的子集间的联系就建立起来了。根据决策逻辑语言 L，在决策表 DT 中，一个可定义的知识就是 $(\varphi,m(\varphi))$，其中，$\varphi\in L$。特别地，作为对象集合 $m(\varphi)$ 的一个描述，公式 φ 就是知识 $(\varphi,m(\varphi))$ 的内涵；而知识的外延就是所有满足公式 φ 的对象的集合 $m(\varphi)$。

定义 3 DT 是决策表，设 $B\subseteq C$ 是个条件属性子集，可以定义一个等价关系 E_B，它满足：$xE_By\Leftrightarrow I_a(x)=I_a(y)$，对于所有 $a\in B\Leftrightarrow I_B(x)=I_B(y)$。

定义 4 设任意一个对象子集 $X\subseteq U$，如果在 DT 上至少存在一个公式 φ 满足 $m(\varphi)=X$，则称 X 为可定义基本粒。X 表示满足等价关系 E_B 的对象的集合，即为等价类。

定义 5 在论域 U 上，划分 $\pi=\{X_i\mid 1\leqslant i\leqslant n\}$，其中，$X_i$ 满足以下条件：

(1) X_i 为非空集合；

基金项目：国家自然科学基金资助项目(50863003)；江西省科技攻关计划基金资助重点项目(20061B01002)；江西省教育厅科技基金资助项目([2007]28)

作者简介：熊树洁(1985－)，女，硕士研究生，主研方向：数据库技术，人工智能；邱桃荣，教授；龚科华，硕士研究生；白小明，副教授

收稿日期：2009-07-03 **E-mail**：PYL618@163.com

(2)对任意的 $i \neq j$， $X_i \cap X_j = \phi$；

(3) $\cup\{X_i \mid 1 \leqslant i \leqslant n\} = U$。

根据定义 5 可知，如果每个等价类 X_i 都是个可定义基本粒，则这个划分 π 称为在 DT 上的一个可定义划分。任意一个条件属性子集 B 在 U 上的一个划分，记为 $\pi(B)$，并称其为 U 上的一个信息粒。因此，从知识粒度方面看，可对决策表进行粒度划分，以便进行由细粒度到粗粒度的转化。

定义 6 假如 π_1 中每个等价类都被包含在 π_2 的某个等价类里，则划分 π_1 是另一个划分 π_2 的细化，或者说 π_2 是 π_1 的一个粗化，记为 $\pi_1 \leqslant \pi_2$。

3 β重要度概念

Pawlak粗糙集模型的一个局限性是它所处理的分类必须是完全正确的或肯定的，因为它是严格按照等价类来分类的，所以它的分类是精确的，即“包含”或“不包含”，而没有某种程度上的“包含”或“属于”。为克服这些局限性，Ziarko 提出 VPRS 模型。本文根据在规则挖掘这个领域内的经验值来选取 β 阈值的，经过后续实验的检验也表明 β 值选取合适。

本文着重讨论引入了 β 值的 β 重要度概念的涵义及作用。有了 β 值这个衡量标准，一方面有利于用粗糙集理论从认为不相关的数据中发现相关数据，这样就可以有效地防止由一些看似不相关的数据，通过数据的重组而重新获得有关个人隐私数据的危险。另一方面由于引入阈值 β 的关系，在获取知识的时候允许某种程度上的“包含”，这样就能有效地减少有用信息的丢失。

一般地，在决策表中，不同的条件属性相对于决策属性有不同的重要性。因此，本文结合基于 Pawlak 属性重要度和变精度粗糙集模型中的 β 阈值，提出 β 重要度的相关概念。

定义 7 对于决策表 DT，设 $X,Y \subseteq U$ 是 U 上的 2 个子集，X 至少以 β 程度包含于 Y，形式上定义如下：

$$V_\beta(X,Y) = \begin{cases} \max\left(\dfrac{|X_i \cap Y_j|}{|X_i|}\right) & |X_i| \neq 0 \\ 1 & |X_i| = 0 \end{cases}$$

本文中将阈值 β 定义为划分正确率，取值为[0.5,1)。

定义 8 对于决策表 DT 下的 β 正区域：

$$POS_c(D,\beta) = \bigcup_{V_\beta(\pi(C),\pi(D)) \geqslant \beta} \{X_i \mid X_i \subseteq \pi(D), X_i \in \pi(C)\}$$

其中，$\pi(C)$ 为在条件属性集 C 下产生的划分；$\pi(D)$ 是在决策属性 $\{d\}$ 上产生的划分；$POS_c(D,\beta)$ 表示条件属性 c 在条件属性集 C 上相对于决策属性 D 的 β 正区域。

定义 9 对于决策表 DT，设 $c \in C$ 是单个条件属性，则 β 划分质量如下：

$$\gamma(c,\beta) = \frac{|POS_c(D,\beta)|}{|U|}$$

其中，$|U|$ 表示对象的个数；$|POS_c(D,\beta)|$ 为条件属性 c 在条件属性集 C 上相对于决策属性 D 的 β 正区域中对象的个数。

通常，采用一个知识相对于另一知识的正区域概念来刻画属性(或属性集)相对于决策属性的重要度。由于本文引入了 β 阈值，因此 β 重要度概念定义如下：

定义 10 对于决策表 DT，对任意单个条件属性 c(其中，$c \in C$)，从条件属性集合 C 中删除属性 c 后对划分的影响：

$$SIG(C,\{c\},\beta) = \gamma(C,\beta) - \gamma(C-\{c\},\beta)$$

根据决策表中的 β 重要度的定义可知：如果值越大，说明删除属性 c 对条件属性集 C 相对于决策属性 D 的划分能力影响越大，即属性 c 对条件属性集合 C 相对于决策属性 D 越重要。

4 算法描述

对使用信息表中的普通用户说，数据隐含化就是指，对使用信息表中的普通用户来说，信息表进行由细粒度到粗粒度的转变；而对信息管理员来说要能保证在一定程度上信息表的分类能力不变，以此来权衡私密保护性和信息完整性。本文采用基于 β 重要度的决策表属性约简算法来得到初始决策表中的重要属性，然后采用文献[5]提出的改进 K-means 算法对原始决策表的条件属性进行聚类，接着对每类中的非重要属性的取值进行泛化，得到新的决策表，最后比较原始决策表和新的决策表，根据定义 6 判别是否产生从细粒度到粗粒度的转化。

4.1 数据隐含化方法

由上述可知，数据隐含化方法的主要步骤可描述如下：

第 1 步 给定一完备的决策表 DT。

第 2 步 预处理决策表 DT。

第 3 步 运用基于 β 重要度的属性约简算法对决策表进行约简。

第 4 步 采用 K-means 聚类算法对原始决策表 DT 的条件属性进行聚类。

第 5 步 根据第 4 步聚类结果所得到的类别，将每类中除了第 3 步得到的重要属性外的所有非重要属性的属性值进行泛化为该类中该属性的值域。

第 6 步 得到泛化后的新的决策表，即信息隐含化后的决策表 DT’。

这是本文提出的数据隐含化方法的总体过程，以下是在数据隐含化过程中所使用到的算法的描述。

4.2 基于β重要度的决策表属性约简算法

算法步骤如下：

输入 一张完备的决策表 $DT = (U, At, L, V, I)$，阈值 $\beta = 0.8$(由经验值给定)。

输出 条件属性 C 相对于决策属性 D 的一个相对约简 $RED_C(D)$。

第 1 步 $RED_C(D) = \phi$。

第 2 步 按照定义 7 中的公式计算条件属性中单个属性 c_i 的 V_β 值，并判断其是否大于等于阈值 β，成立就转到第 3 步，否则重复第 2 步。

第 3 步 按照定义 8 中的公式计算正区域 $POS_{c_i}(D,\beta)$。

第 4 步 根据第 3 步求出的结果来计算单个属性 c_i 的 β 划分质量 $\gamma(c_i,\beta)$。

第 5 步 根据第 4 步求出的结果来求重要度 $SIG(C,\{c_i\},\beta)$，若 $SIG>0$，则令 $RED_C(D) \Leftarrow RED_C(D) \cup \{c_i\}$，转到第 2 步。

第 6 步 输出最小约简 $RED_C(D)$。

4.3 改进的K-means聚类算法

详细的算法步骤参见文献[5]，文中作者提出 2 种改进算法，这里引用改进算法 A。算法 A 的提出解决了 K-means 算法对初始值的依赖，同时在一定程度上减少了算法陷入局部最优的可能。

4.4 非重要属性值的泛化

本文选择对非重要属性的属性值来进行泛化的方法。因此，根据改进的 K-means 聚类算法所得出的聚类结果，将每类中的非重要属性的属性值替换为该类中该属性的值域。

5 实验分析

5.1 数据隐含化实验过程

本文引用文献[4]中的决策表(见表 1)作为原始信息表进行实验验证。

表 1 决策表 DT

论域 U	条件属性 C			决策属性 d
	height(c_1)	hair(c_2)	eyes(c_3)	
x_1	Short	blond	blue	+
x_2	Short	blond	brown	-
x_3	Tall	red	blue	+
x_4	Tall	dark	blue	-
x_5	Tall	dark	blue	-
x_6	Tall	blond	blue	+
x_7	Tall	dark	brown	-
x_8	Short	blond	brown	-

(1)预处理原始决策表 DT，过程如表 2 所示。

表 2 对表 1 的预处理

属性	值	离散化
height	Short, tall	1, 2
hair	Blond, dark, red	2, 4, 5
eyes	Blue, brown	1, 3
D	-, +	1, 2

(2)对原始决策表进行属性约简。

根据 4.2 节得出的结果如下:

符合 SIG>0 的属性为{hair, eyes}，且这 2 个属性的 β 重要度均为 1，即最小约简 $RED_C(D)=\{c_2,c_3\}$。

(3)对原始决策表进行聚类。

根据文献[5](取 h=0.7)，所得聚类结果如表 3 所示。

表 3 聚类结果

类别	论域 U
类 0	x_1, x_2, x_6, x_7, x_8
类 1	x_3, x_4, x_5

(4)泛化非重要属性的属性值。

对聚类所得结果的类别中的非重要属性的属性值泛化，所得新的决策表 DT'如表 4 所示。

表 4 新决策表 DT'

论域 U	条件属性 C			决策属性 d
	height(c_1)	hair(c_2)	eyes(c_3)	
x_1	tall，short	blond	blue	+
x_2	tall，short	blond	brown	-
x_3	tall	red	blue	+
x_4	tall	dark	blue	-
x_5	tall	dark	blue	-
x_6	tall，short	blond	blue	+
x_7	tall，short	dark	brown	-
x_8	tall，short	blond	brown	-

5.2 实验分析

得到决策表 DT 中各个属性子集的划分如下:

$$\pi(height)=\{\{x_1,x_2,x_8\},\{x_3,x_4,x_5,x_6,x_7\}\}=\{m(height=short),m(height=tall)\}$$

$$\pi(height,hair)=\{\{x_1,x_2,x_8\},\{x_3\},\{x_4,x_5,x_7\},\{x_6\}\}=\{m(height=short\wedge hair=blond),\cdots\}$$

$$\pi(height,hair,eyes)=\{\{x_1\},\{x_2,x_8\},\{x_3\},\{x_4,x_5\},\{x_6\},\{x_7\}\}=\{m(height=short\wedge hair=blond\wedge eyes=blue),\cdots\}$$

…

而新的决策表 DT'的各属性子集划分如下:

$$\pi'(height)=\{U\}=\{m(height=tall,short),m(height=tall)\}$$

$$\pi'(height,hair)=\{\{x_1,x_2,x_6,x_8\},\{x_3\},\{x_4,x_5,x_7\}\}=\{m(height=tall,short\wedge hair=blond),\cdots\}$$

$$\pi'(height,hair,eyes)=\{\{x_1,x_6\},\{x_2,x_8\},\{x_3\},\{x_4,x_5\},\{x_7\}\}=\{m(height=tall,short\wedge hair=blond\wedge eyes=blue),\cdots\}$$

…

可以看出，各属性子集的划分均满足定义 6，比如: $\pi(height)\leqslant\pi'(height)$，$\pi(height,hair)\leqslant\pi'(height,hair)$，$\pi(height,hair,eyes)\leqslant\pi'(height,hair,eyes)$ 等，即原始决策表 DT 是新的决策表 DT'的一个细化，而新的决策表 DT'即是原始决策表 DT 的一个粗化。

从以上对实验结果的分析来看，本文通过对原始决策表 DT 和新的决策表 DT'的属性子集进行划分，再判断所得结果是否满足定义 6 所示的偏序关系，满足则表示粒度由细变粗了，即决策表中的信息变粗糙了。所以，对普通用户来说，所能得到的信息变粗糙了，即私密信息隐藏了；而对于信息管理原来说，信息的泛化是在聚类出的同类中进行泛化的，所以，能保证信息分类能力在一定程度上保持不变，即信息完整性得到一定保护。

6 结束语

在现实生活中，数据通常都不是完备的，因此，今后对不完备信息系统的私密信息的隐藏技术还有待发展。另外，为更好地衡量隐含化的程度，提出一个通用有效的标准也是很必要的。在粒度的变化上，则希望能够根据用户的使用等级而查询到不同粒度的隐含化信息。这都是今后的研究重点。

参考文献

[1] Loukides G. An Efficient Clustering Algorithm for K-anonymisation[J]. Journal of Computer Science and Technology, 2008, 23(2):188-202.

[2] Wang Dawei, Liau Churn-Jung, J, Hsu Tsan-Sheng. Medical Privacy Protection Based on Granular Computing[J]. Artificial Intelligence in Medicine, 2004, 32(2): 137-149.

[3] Pawlak Z. Rough Set Approach to Knowledge-based Decision Support[J]. European Journal of Operational Research, 1997, 99(11): 48-57.

[4] Yao Jingtao. Granular Computing as a Basis for Consistent Classification Problems[J]. Communications of Institute of Information and Computing Machinery, 2002, 5(2): 101-106.

[5] 张建辉. K-means 聚类算法研究及应用[D]. 武汉: 武汉理工大学, 2007.

编辑 陈 文