基于变精度和浓缩布尔矩阵的属性约简

李 艳[1] 郭娜娜[1] 赵 浩[2]
(河北大学数学与信息科学学院机器学习与计算智能重点实验室 保定 071002)[1]
(西安交通大学电子与信息工程学院 西安 710049)[2]

摘 要 属性约简是粗糙集理论研究的重要内容。传统的基于差别矩阵的属性约简方法只能处理一致决策表,改进的差别矩阵针对决策表中一致和不一致的对象做不同的处理,从而解决了这一问题。浓缩布尔矩阵进一步节省了矩阵的存储空间并提高了矩阵的生成效率,从而可以快速计算得到约简。在此基础上,结合变精度的思想把部分不一致对象合理地加入到一致对象的集合中,从而增加了一致数据的信息量,并通过使用浓缩布尔矩阵有效降低了约简的计算消耗。实验表明,所提方法在运行速度和分类精度方面均表现出了优势。

关键词 粗糙集,差别矩阵,属性约简,浓缩布尔矩阵,变精度
中图法分类号 TP181 **文献标识码** A

Attribute Reduction Based on Variable Precision Rough Sets and Concentration Boolean Matrix

LI Yan[1] GUO Na-na[1] ZHAO Hao[2]
(Key Lab of Machine Learning and Computational Intelligence, School of Mathematics and Information Science, Hebei University, Baoding 071002, China)[1]
(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)[2]

Abstract Attribute reduction is the most important research topic in rough set theory. The traditional attribute reduction based on discer-nibility matrix can only handle consistent decision tables. Then the concept of improved discernibility matrix was proposed to effectively deal with both consistent and inconsistent decision tables. Further, the condensed Boolean matrix was defined to represent the discernibility matrix in order to save the storage space and improve the efficiency of matrix generation. Based on the previous work, the idea of variable precision was used to select some inconsistent objects in the developing of the discernibility matrix, thus more information can be considered in generating attribute reduction. The experimental results show that the proposed method performs advantages in both running speed and classification accuracy.

Keywords Rough set, Discernibility matrices, Attribute reduction, Concentration boolean matrix, Variable precision

1 引言

Rough 理论[1]是 Z. Pawlak 教授于 1982 年提出的一种用于处理不精确、不完全与不相容知识的数学理论,近年来被广泛应用于机器学习、数据挖掘及模式识别等多个领域。属性约简[2,4]是该理论的一个重要研究内容,其目的是在保持数据的分类能力的前提下只保留最少的信息。求属性约简最常用且高效的方法是基于差别矩阵的方法[5-8],其一般过程是:首先依据差别矩阵求出约简核,即对决策最关键的属性集合;在此基础上按照属性的重要度逐步增加其余属性,直到求出一个属性约简为止[9-12]。这种属性约简算法主要针对一致决策表[13],而对于不一致决策表,使用传统的差别矩阵求得的约简与定义是不等价的[14]。因此,文献[14]提出了基于改进的差别矩阵求解属性约简的方法,统一考虑了决策表一致与不一致两种情况下的属性约简,有效地弥补了传统的基于差别矩阵求解属性约简的不足。不难发现,差别矩阵中并不是所有的元素都对求解约简有用,如果每次扫描整个差别矩阵就会增加时间的开销,降低求属性约简的效率,尤其不适用于较大的数据集。针对这个问题,杨明等人对差别矩阵进行了浓缩[14],以此降低存储开销;但该方法仍存在生成效率低的问题,且存储空间有待进一步降低。于是文献[15]定义了浓缩布尔矩阵的概念,并建立了相应的属性约简算法。该算法通过布尔代数的形式有效降低了矩阵的存储空间,明显提高了约简效率。在处理不一致决策表时,此类方法把相容对象和不相容对象分开处理,即分割成完全一致与不一致的两个子表,再用两个子表分别计算差别矩阵。这会导致在进行浓缩布尔矩阵的计算时不一致子表中含有大量冗余的对象参与计算;而另一方面,如果不相容的对象较多,那么仅采用一致的子表求解浓缩布尔矩阵时会损失大量信息。考虑到这种情况,我们提出将变精度[14-19]的思想和浓缩布尔矩阵的约简方法相结合,把部分不一致对象加入到相容对象集中,用于计算浓缩布尔矩阵和进一步计算约简,由于增加了相容对象集的

本文受国家自然科学基金(61170040,61473111),河北省自然科学基金(F2014201100,A2014201003)资助。
李 艳(1976-),女,博士,教授,CCF 会员,主要研究方向为机器学习;郭娜娜(1992-),女,硕士生,主要研究方向为粒计算与知识发现;赵 浩(1994-),男,主要研究方向为自动控制。

信息量，因此能够提高约简后决策表的分类能力。

2　基本概念

粗糙集理论是基于等价关系定义的等价类来对论域进行划分，从而体现数据本身的分类能力，通过两个精确集即上、下近似对不确定的目标概念进行刻画，并在此基础上建立特征和决策之间的联系，发现隐藏的规则性决策知识。

下面简单介绍与本文工作相关的几个重要概念。

定义 1（决策表）　一个决策表表示为 $S=(U,C\cup D)$，其中 $C\cap D=\varnothing$，U 是一个非空有限的样例集合，每个样例用条件属性 C 和决策属性 $D=\{d\}$ 描述。

一致的决策表是指如果决策表中任意两个对象具有相同的条件属性值，那么它们也一定具有相同的决策属性值，否则就是不一致决策表。

决策表中的每个属性都有相应的属性值，通过对相同属性值进行划分，得到等价关系的概念，属性子集和等价关系是一一对应的。

定义 2（不可分辨关系和等价类）　设决策表为 $S=(U,C\cup D)$，对于任意一个非空属性子集 $B\subseteq C$，定义如下等价关系为不可分辨关系：

$$IND(B)=\{(x,y)\in U\times U:a(x)=a(y),\forall a\in B\}$$

这个等价关系把 U 划分成等价类的集合，表示为：

$$U/IND(B)=\{[x]_B:x\in U\}$$

其中，$[x]_B=\{y\in U:(x,y)\in IND(B)\}$ 称作基于等价关系 $IND(B)$ 的 x 的等价类。

粗糙集是通过两个精确集即上、下近似集来描述不精确集合的。

定义 3（上下近似和正域）　假设 U 基于决策类 D 形成的划分为 $U/IND(D)=\{D_1,\cdots,D_r\}$，其中 D_i 是具有相同决策值的对象的集合。则 $D_i\in U/IND(D)$ 上的关于等价关系 $IND(B)$ 的下近似和上近似分别定义为：

$$\underline{B}D_i=\bigcup\{[x]_B:[x]_B\subseteq D_i\}$$

$$\overline{B}D_i=\bigcup\{[x]_B:[x]_B\cap D_i\neq\varnothing\}$$

进一步，定义 D 的 B 正域为 $POS_U(B,D)=\bigcup_{k=1}^{r}\underline{B}D_k$。正域是关于 B 的 U 上所有一致对象的集合。

下近似或正域是由根据知识 C 判断肯定属于 D 的论域 U 中的元素组成的集合；上近似是由根据知识 C 判断肯定属于或可能属于 D 的论域 U 中的元素组成的集合。属性约简的原理是在原决策表的基础上去掉某些属性后其决策属性相对于条件属性的正域不变，即分类能力不变。

定义 4（属性约简）　知识表达系统中的属性并不是同等重要的，属性约简是指可以找到一个较小的属性集 $B\subseteq C$，使得可用 C 描述的对象集合必然可用 B 描述，从而消除冗余属性。其定义如下：给定一个决策表 $S=(U,C\cup D)$，子集 $B\subseteq C$ 是 C 的一个约简，如果它满足以下两个条件：

1) $POS_U(B,D)=POS_U(C,D)$

2) $\forall a\in B,POS_U(B-\{a\},D)\neq POS_U(B,D)$

即 B 为能够保持原有决策系统正域不变的最小的属性子集。

求属性约简最常用且高效的方法是基于差别矩阵的方法。差别矩阵的定义如下。

定义 5（差别矩阵）　给定决策表 $S=(U,C\cup D)$，$U=\{x_1,x_2,x_3,\cdots,x_n\}$，$D=\{d\}$，其差别矩阵是一个 $n\times n$ 的矩阵，记为 M_{imdis}，矩阵元素 m_{ij} 定义为 $m_{ij}=\{a\in C\mid F(x_i,a)\neq F(x_j,a)\wedge F(x_i,d)\neq F(x_j,d)\}$，$F(x_i,a)$ 是对象 x_i 在属性 a 上的属性值。

基于定义 5 中的差别矩阵的属性约简，矩阵中单元素集取并即为核，约简是在核的基础上依次加入其它属性直到与辨识矩阵中的每一个位置的属性集的交不为空为止。

定义 5 给出的差别矩阵的定义只能处理一致的决策表，即假设有相同条件属性的对象也有一致的决策。然而实际问题中不一致决策表更加常见。这种情况下，使用传统的差别矩阵得到的约简与其定义不符。因此，一种改进的差别矩阵被提出[14]（见定义 6）。通过将一致和不一致的对象集分开处理，使得改进后的差别矩阵可以处理不一致的决策表，得到相应的约简。容易证明，在决策表一致的情况下，其与传统的差别矩阵相同。

定义 6（改进的可辨识矩阵）[14]　决策表 $S=(U,C\cup D)$，$U=\{x_1,x_2,x_3,\cdots,x_n\}$，$D=\{d\}$，其可辨识矩阵是一个 $n\times n$ 的矩阵，记为 M_{imdis}，矩阵元素 m_{ij} 定义为：

$$m_{ij}=\begin{cases}\{a\in C:F(x_i,a)\neq F(x_j,a)\}, & \text{当 } F(x_i,D)\neq F(x_j,D)\\ & \text{其中，}x_i\in U_1,x_j\in U_1\\ \{a\in C:F(x_i,a)\neq F(x_j,a)\}, & \text{其中，}x_i\in U_1,x_j\in U_2'\\ \varnothing, & \text{其他情况}\end{cases}$$

其中，$U_1=\bigcup_{i=1}^{k}\underline{C_{Di}}(U)$ 为一致的对象集；$U_2=U-U_1$ 为不一致的对象集。目前处理不一致决策表计算约简有两种方法。一种是在计算差别矩阵时不考虑不相容对象集 U_2，而是直接将 U_1 和 U_2 分开，根据定义计算差别矩阵[19]，称为原始的处理不一致决策表方法，简称为原始方法。而上述改进的差别矩阵将 U_2 中每个对象按顺序加入到 U_2'（初始为空）中在加入的同时判断此对象是否与 U_2' 中的每个对象相容，如果相容，则加入 U_2'，否则不加入，直到遍历完整个 U_2。因为这个概念中对于相容对象的加入是随机的，所以称这种方法为随机加入法。

性质 1　若 $U_2\neq\varnothing$，则 $|U_2'|<|U_2|$。

由上述定义和性质可知，改进的差别矩阵虽然能处理不一致的决策表，但不是所有的矩阵元素对求解属性约简起作用，有些元素是冗余的，致使矩阵的存储和计算时间耗费较大。

浓缩差别矩阵是在差别矩阵的基础上对差别矩阵的优化，只保留少部分矩阵元素对求解属性约简起作用，从而降低存储开销与计算量。

3　基于变精度和浓缩布尔矩阵的约简方法

如前所述，对于不一致的决策表，传统方法无法处理，而随机加入不一致对象的方法是基于改进的差别矩阵，随机加入某一个等价类中的对象参与约简的计算相对于原始方法并没有增加额外的有用信息，只是减少了矩阵存储空间和提高了生成效率。这种方法（见定义 6）首先将论域 U 划分为相容对象集 U_1 和不相容对象集 U_2；然后随机选择不一致对象放入 U_2'，即按次序从 U_2 中取出对象 x 放入 U_2' 中，x 需要满足与 U_2' 中已有的每个对象是相容的，直到遍历完 U_2 中的对象为止。最终的 U_2' 即是对不一致对象处理的结果。最后，将 U_1 和 U_2' 中的对象按定义 6 计算改进的差别矩阵并相应求得约简。可以看出，这样选择的不一致对象即是不一致对象集 U_2 中第一个等价类的对象。所以在选择加入时并没有判断

此类对象对于分类能力的贡献。此外，矩阵的存储耗费仍然很大。考虑到这种情况，本节提出结合变精度和浓缩布尔矩阵的方法，以进一步节省空间和计算时间；而且按变精度加入不一致对象的信息，从而增加了整体的信息量，使得分类精度更高。下面首先介绍浓缩差别矩阵的概念。

浓缩差别矩阵是对差别矩阵的优化，它只保留少部分对求解属性约简起作用的矩阵元素，从而降低了存储开销。浓缩布尔矩阵是在浓缩差别矩阵的基础上提出来的。

定义 7(浓缩差别矩阵)[10] 对于给定差别矩阵 M，其浓缩差别矩阵定义为 $IME(M)=\{m \mid m(m\neq\emptyset)\in M\}$，且不存在 $m'\in M(m'\neq\emptyset)$ 使得 $m'\subset m$。

由定义 7 可知，浓缩差别矩阵中任意 m 之间不存在相互包含关系，通过筛选只存储了直接影响到约简的矩阵元素。但是元素 m 是字符串的集合，使得约简的运算效率和存储空间依旧不够理想。为了进一步解决这一问题，在浓缩差别矩阵的基础上提出了浓缩布尔矩阵算法[11]，它使用布尔代数 0 和 1 代替传统的字符串，大大减少了计算量，提高了算法的效率，节省了存储空间。布尔矩阵定义如下。

定义 8(布尔矩阵) 布尔矩阵 BM 中每个元素定义为

$$m((x_i,x_j),C_k)=\begin{cases}1, & \text{当 } F(x_i,D)\neq F(x_j,D) \text{ 且 } F(x_i,C_k)\neq F(x_j,C_k),\\ & x_i,x_j\in U_1\\ 1, & \text{当 } F(x_i,C_k)\neq F(x_j,C_k),x_i\in U_1,x_j\in U_2'\\ 0, & \text{其他}\end{cases}$$

其中，$i=1,2,3,\cdots,n$，$j=1,2,3,\cdots,n$，$k=1,2,3,\cdots,m$。

该布尔矩阵的含义为：$m((x_i,x_j),C_k)=1$ 表示属性 C_k 能区分对象 x_i 与 x_j，而 $m((x_i,x_j),C_k)=0$ 表示属性 C_k 不能区分对象 x_i 与 x_j。

设 m'，m 均为 C 的子集，判断 m' 是 m 的子集的方法是让 m' 和 m 的各位做或运算，其运算结果可能有 3 种情况：

1)结果中各位与 m' 的各位相同，则说明 $m\subset m'$；

2)结果中各位与 m 的各位相同，则说明 $m'\subset m$；

3)结果与 m 和 m' 都不相同，则说明 m 和 m' 不存在包含与被包含的关系。

在此基础上，浓缩布尔矩阵可定义为如下形式。

定义 9(浓缩布尔矩阵) $IME(BM)=\{m \mid m(m\neq\emptyset)\in BM\}$ 且不存在 $m'\in BM(m'\neq\emptyset)$ 使得 m 与 m' 各位进行或运算的结果与 m 各位相同。

可见，浓缩布尔矩阵只需要利用 0，1 之间的或运算就可以快速判断出矩阵元素之间的包含关系，选择出影响计算约简的矩阵元素并存储起来，大大提高了差别矩阵的生成效率。

在此基础上，我们采用变精度的思想来对不一致的对象集 U_2 进行处理。区别于随机加入的方法，我们首先判断 U_2 中的等价类对于分类的重要程度，引入一个阈值来判断是否选择某类不一致对象参与到约简的计算过程中，从而既增大了计算约简的信息量，又通过浓缩布尔矩阵降低了存储量，提高了矩阵的生成效率。方法的详细步骤如下。

1)对给定的决策表，计算相容对象集 U_1 和可加入的不相容对象集 U_2'，并置核属性集 $core=\emptyset$，浓缩布尔矩阵 $IME(BM)=\emptyset$。其中 $U_1=POS_U(C,D)$，$U_2=U-U_1$，$U_2'=deal(U_2)$，其中，函数 $deal(U_2)$ 决定了如何选择不一致对象加入到 U_1，详细描述如下。

考虑不相容对象集 U_2，对于其中的一个对象 x，必存在另一对象 y 与 x 的条件属性相同但决策不同，即 y 与 x 不相容。首先根据条件属性 C 对 U_2 进行划分，假设分为 n 个不同的等价类，记为 $C_1,C_2,\cdots,C_n$，称为条件等价类。因为 U_2 中对象的不相容性，因此每个 C_i 中的对象根据决策属性值又可以形成多个决策类。图 1 所示为第 i 个条件等价类 C_i，根据决策属性值其又被划分为 m 个决策类，记为 D_{i1}，D_{i2}，$\cdots$，D_{im}。这样，在每个决策类 D_{ij} 中的对象两两是相容的，即同时具有相同的条件属性值和决策属性值。

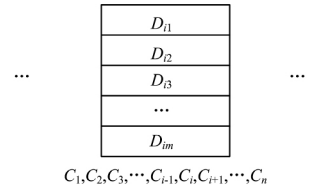

图 1

$deal(U_2)$ 函数根据条件属性等价类中每个决策类所占比重来判别某个决策类中对象的重要程度，从而确定如何从每个条件属性等价类 $C_i(i=1,\cdots,n)$ 中选择对象加入到 U_1。如果在 C_i 中存在某个决策等价类 D_{ij}，其中的对象占据 C_i 较大的比例，则认为 D_{ij} 对决策的影响也较大。因此，设置一个阈值 t，对于某个条件等价类 C_i，D_{ij} 为其中最大的决策类，若 $|D_{ij}|/|C_i|\geqslant t$，则把 D_{ij} 中的任一个对象加入到 U_1 中；否则把 C_i 中的任一个对象加入到 U_2'(初始值为空)中，以此处理 U_2 中的每个条件等价类。在传统差别矩阵的方法中只有相容对象参与约简的计算，即在传统方法中，每个条件等价类 C_i 也必然只对应一个决策类 D_i，即有 $|D_i|/|C_i|=1$。所以采用上述重要度阈值的方法就相当于将传统方法中的 U_1 扩大，增加了信息量丰富的对象。将原来重要度为 1 的要求放宽为 t，相当于扩大了正域，在这个意义上借鉴了变精度的思想。其中阈值是依据经验和具体数据给出。

2)生成初始的浓缩布尔矩阵。遍历集合 U_1 中的对象，对任意 $x_i\in U_1$，$x_j\in U_1(i<j)$，若 $F(x_i,D)\neq F(x_j,D)$，则将 x_i，x_j 的各个条件属性对应进行“异或”操作，属性值相同则为 0，不同则为 1，将所得的结果存入数组 arr 中，然后再添加到 $IME(BM)$ 中。用相同的方法遍历 U_1 和 U_2' 中的对象。

3)计算核属性集。遍历矩阵 $IME(BM)$ 中每一行的元素值，如果只有一个元素值为 1，那么将此元素值对应的属性加入到 Core 中，遍历完 $IME(BM)$ 之后得到初始的 Core，然后去掉 Core 中重复的元素得到最终的核属性集。

4)对上面得到的布尔矩阵 $IME(BM)$ 进行浓缩。把 Core 属性在 $IME(BM)$ 中对应值为 1 的所有行去掉。然后再用定义 9 中的判断方法遍历矩阵 $IME(BM)$，使 $IME(BM)$ 中任意两行不存在包含关系。最终得到 $IME(BM)$。

性质 2 当 $t=0$ 时，$U_2'=\emptyset$；当 $t=1$ 时，U_1 不变，U_2' 为最少数量相容对象集合。

基于变精度和浓缩布尔矩阵的约简方法相对于最初方法和随机加入方法增加了 U_1 中一致对象的个数，从而增加了属性约简的信息量；同时也去掉了 U_2 中冗余的对象，使最少的有用的对象参与布尔矩阵的运算，减少了运算时间。

4 实验结果与分析

为测试本文算法的约简效果，选取了 UCI 数据库中的

spect. train，tae，backuplarge，spect. test 以及 kohkiloyeh 5 个数据集。其中在 backuplarge 中去掉了缺失数据。分别采用最初方法、随机加入的方法和按阈值加入法测试并对比约简效果。其中，阈值 t 从 0 增到 1(每次增加 0.1)，依次判断，当属性约简的精度最高时，相应的 t 值即作为本数据的阈值。对于不同的数据，t 不尽相同。

这种寻优方法的计算复杂度是线性的，因此不会从本质上降低算法的效率。但是这样得到的最优参数不是最精确的，由于篇幅有限，在未来的工作中可以考虑更好的方法。

原始方法的思想是指最初针对不一致决策表做的改进(见定义 6)。随机加入法的思想是指在文献[14]中对不相容对象的处理。按阈值加入法是本文主要研究的方法，它主要是结合变精度的思想对不相容对象集进行的改进。为了节省程序的总体运行时间和使 3 种算法具有可比较性，实验中使原始方法和随机加入法的属性约减基于浓缩布尔矩阵，因此这 3 种方法都是在浓缩布尔矩阵的基础上进行的，唯一不同的是对 U_2 的处理。约简的效果用测试精度和运行时间来衡量，精度越高并且运行时间越短表明算法越好。

对于每个数据集，随机抽取其 70%的数据作为训练集对算法进行属性约简，得出模型，然后用剩余的 30%的数据进行测试。测试精度＝正确识别样本数/测试样本总数。为了保证结果的准确性，随机进行 10 次训练和测试，将 10 次结果的测试精度的平均值作为对算法精度的估计。实验结果如表 1 所列。

表 1　3 种方法的运行时间和分类精度

数据集	原始方法		随机加入法		按阈值加入法	
	时间/s	精度/%	时间/s	精度/%	时间/s	精度/%
spect. train 80＊24	2.205	0.35978	0.796	0.3598	0.769	0.4559
tae 151＊7	0.416	0.4186	0.398	0.4186	0.366	0.4419
backuplarge 212＊29	17.155	0.5809	15.617	0.5809	15.218	0.7032
spect. test 187＊24	2.503	0.36953	2.353	0.36953	2.292	0.4915
kohkiloyeh 100＊7	0.126	0.7222	0.119	0.7222	0.118	0.8947
平均值	4.481	0.4902	3.8566	0.4902	3.7526	0.5974

对于各个数据集，相比原始方法和随机加入方法，我们把按阈值加入法的属性约减精度的提高率记为 α，其中 α＝(按阈值加入法精度－原始方法或随机加入法精度)/按阈值加入法精度。按阈值加入法相比原始方法在时间上的缩减率记为 β_1＝(原始方法时间－按阈值加入法时间)/原始方法时间，相比随机加入法在时间上的缩减率记为 β_2＝(随机加入法时间－按阈值加入法时间)/随机加入法时间。于是在每个数据集上，α，β_1 和 β_2 的值如表 2 所列。

表 2　按阈值加入法的精度提高率和时间缩减率/%

数据集	α	β_1	β_2
spect. train 80＊24	26.7	65.12	3.39
tae 151＊7	5.6	12.02	8.04
backuplarge 212＊29	21	11.29	2.55
spect. test 187＊24	24.8	8.43	2.6
kohkiloyeh 100＊7	23.9	6.35	0.84
平均值	21.8	20.642	3.484

从表 1 可以看出，原始方法利用改进差别矩阵进行属性约简，运行时间最长且精度最低；随机加入的方法运行时间有所改善，但精度没有提高，这是因为随机加入的方法相对于原始方法做的改进只是去掉了 U_2 中的不相容对象，并没有增加额外有用的信息，所以只会减少运行时间而不会提高精度。相比之下，用本文提出的按阈值加入法进行属性约简时运行时间短并且模型的精度高。从表 2 中可以看出按阈值加入法的精度和运行时间分别得到了提高和改善，但是相对于随机加入法在时间上改进的不明显。本文提出的算法相对于随机加入法来说，重点在于不增加时间的前提下能够进一步提高精度。没有明显减少运行时间的主要原因如下：

1)这两种方法都是在浓缩布尔矩阵的基础上求约简，没有产生浓缩布尔矩阵和差别矩阵两种方法的时间差。

2)与数据的结构有关系。如果数据集的不一致对象的个数相比总的样例个数少许多，那么对不一致对象集处理后的结果很可能不会有明显的变化，即 U_2'的个数不会产生明显的变动，那么会使按阈值加入法的运行时间相比随机加入法没有明显的减少。还有一种情况是条件等价类 C_i 中的决策类的个数偏多，几乎接近 C_i 中对象的个数，那么用按阈值加入法对 U_2 进行处理得出的 U_2'接近 U_2，所以在这种情况下两种方法的运行时间也不会有明显的差异。

相反，对于不一致程度较大的数据集，改进效果会较为明显。但不管哪种情况，采用本方法的精度均会有所提高。

总之，按阈值加入法无论在约减精度上还是在时间复杂度上相对于原始方法和随机加入法都有所提高和改进，从而证明了本文所提出的按阈值加入法的可行性。

结束语　本文以浓缩布尔矩阵属性约简算法为基础，主要针对不一致决策表，采用类似于变精度的思想，通过引入重要性度量将不相容的对象有选择地加入到差别矩阵的计算中，从而增加约简的信息量，提高分类精度，同时采用浓缩布尔矩阵提高了计算效率。今后将进一步考虑优势关系下的决策表，提出相应的差别矩阵的优化算法。

参 考 文 献

[1] PAWLAK Z. Rough set[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[2] WEI L, LI H R, ZHANG W X. Knowledge reduction based on the equivalence relations defined on attribute set and its power set[J]. Information Sciences, 2007, 177(15): 3178-3185.

[3] QIAN Y H, LIANG J Y, PEDRYCZ W. An efficient accelerator for attribute reduction from incomplete data in rough set framework[J]. Pattern Recognition, 2011, 44: 1658-1670.

[4] CHEN X, ZIARKO W. Experiments with rough set approach to face recognition[J]. International Journal of Intelligent Systems, 2011, 26(6): 499-517.

[5] ABBAS A, LIU J. Designing an intelligent recommender system using partial credit model and Bayesian rough set[J]. International Arab Journal of Information Technology, 2012, 9(2): 179-187.

[6] CHANG B, HUNG H. A study of using RST to create the supplier selection model and decision making rules[J]. Expert Systems with Applications, 2010, 37(12): 8284-8295.

[7] LIU D, YAO Y Y, LI T R. Three-way investment decisions with decision-theoretic rough sets[J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 66-74.

[8] 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.
[9] JELONEK J, KRAWIEC K, SLOWINSKI R. Rough set reduction of attributes and their domains for neural networks[J]. Computational Intelligence, 1995, 11(2): 339-347.
[10] WANG J, WANG J. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.
[11] 王治和,崔晓慧. 改进的差别矩阵启发式属性约减算法[J]. 计算机工程与设计,2016,37(4):1032-1036.
[12] SKOWRONA, RAUSZER C. The discernibility matrices and functions in information systems [M]. Dordrecht: Kluwer Academic Publishers, 1992: 331-362.
[13] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001:30-62.
[14] 杨明,杨萍. 差别矩阵浓缩及其属性约简求解方法[J]. 计算机科学,2006,33(9):181-183.
[15] 殷志伟,张健沛. 基于浓缩布尔矩阵的属性约简算法[J]. 哈尔滨工程大学学报,2009,30(3):307-311.
[16] CHEN D G, YANG Y Y, DONG Z. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied Soft Computing, 2016, 45: 129-149.
[17] JAMES N, LIU K, HU Y X, et al. A set covering based approach to find the reduct of variable precision rough set[J]. Information Sciences, 2014, 275(3): 83-100.
[18] 黄卫华,晏林,冯云再. 广义变精度粗糙集模型与其它粗糙集模型的比较[J]. 兰州文理学院学报,2015,29(5):6-8.
[19] 杨勇,朱影. 一种基于 MapReduce 的粗糙集并行属性约简算法[J]. 重庆邮电大学学报,2015,27(1):90-96.

(上接第 49 页)

调整了计划生育政策。

近日,接连发生的电信诈骗事件受到广泛关注,引起了政府和公众对社会安全问题的担忧。徐玉玉被骗 9900 元学费死亡、清华教师遭诈骗犯以“公检法”名义诈骗 1760 万元等事件无不揭露了电信诈骗的危害所在。这类事件的发生除了有关部门监管不严之外,大数据应用的滞后是主要原因之一。

电信诈骗源于可疑的短信、电话和邮件等,据统计,在日常生活中 50%的电话为无用电话,其中 40%是推销电话,10%是诈骗电话。从诈骗电话的号码源类型来看,固定电话呼出的诈骗电话数量超过半数,其次是 400 和 800 电话占 27.1%;手机呼出的诈骗电话仅占 15.4%;1.2%的诈骗电话来自境外呼入。运营商知道这些规律以后应该标记来电,提醒用户谨防受骗。此外,诈骗电话有一定规律,如往往针对同一对象以不同电话号码、不同用户身份连番实施,如果电信运营商以保护公众通信安全为己任,就可以通过大数据分析针对可疑通话提醒用户,减少用户损失。

银行业也是阻止电信诈骗的一道有力屏障。针对诈骗方式分析,犯罪嫌疑人通常是将一笔数额较大的款项以最快的速度分散到若干账户并在极短时间内通过不同网点的取款机取走。只要具备银行业的领域知识,不难从银行交易大数据中发现这种规律。以前银行业对可疑交易的分析主要是针对洗钱行为,一般银行的可疑交易系统实际上只是反洗钱系统,对于目前日益猖獗的诈骗犯罪,银行业应该考虑尽快采用大数据技术开发反诈骗系统,最大限度地保障用户的金融资产安全。

结束语 由于对大数据研究与应用的系统分析不够,以及领域专家的知识结构限制,目前的大数据研究基本上处于不同专家各抒己见、不同专家都在渲染自己所熟悉的研究阶段,所以大数据研究与应用还是声势大而实效小,投入大而成果小。解决上述问题的途径包括:1)信息技术专家、应用领域专家、模型与统计专家密切配合,并且各类专家都要增强对其他专家知识的理解;2)要有一批能将几个环节的知识融会贯通的“杂家”。各领域的专家要放下身段,学习一些相关领域的知识,身体力行、扎扎实实地进入一两个应用领域做好一两个项目,才能进一步推进大数据的研究与应用。

参 考 文 献

[1] 彭怡,寇纲. 基于领域知识的数据挖掘理论框架研究[C]// 第三届(2008)中国管理学年会--信息管理分会场论文集. 北京:中国管理现代化研究会,2008.
[2] Cloud information. 2020 global information will be more than 40ZB, reaching 12 years 12 times[EB/OL]. [2016-09-05] https://www.aliyun.com/zixun/content/1_1_15290.html.
[3] 冯登国,张敏,李昊. 大数据安全与隐私保护[J]. 计算机学报,2014,37(1):246-258.
[4] Nature. Big Data[EB/OL]. [2016-09-05]. http://www.nature.com/news/specials/bigdata/index.html.
[5] Science. Special online collection: Dealing with data[EB/OL]. [2016-09-05]. http://www.sciencemag.org/site/special/data.
[6] Gartner report: big data will be popular in Chinese[EB/OL]. [2016-09-05]. http://mobile.163.com/16/0824/06/BV7BBSM800118024.html.
[7] KUMAR R. Two computational paradigm for big data [EB/OL]. [2016-09-05]. http://kdd2012.sigkdd.org/sites/images/sum-merschool/Ravi-Kumar.pdf.
[8] Information Week Report. The big data managementchallenge [R/OL]. [2016-09-05]. http://reports.informationweek.com/abstract/81/8766Business-intelligence-and-information-management/research-the-big-Data-management-challenge.html.
[9] GROBELNIK M. Big-data computing: Creating revolutionary breakthroughs in commerce, science, and society [R/OL]. [2016-09-05]. http://videoLectures.net/eswc2012_grobelnik_big_data.
[10] BARWICK H. The “four Vs” of Big Data, Implementing InformationInfrastructure Symposium [EB/OL]. [2016-09-05]. http://www.computerWorld.com.au/article/396198/iiis_four_vs_big_data.
[11] 李广建,化柏林. 大数据分析与情报分析关系辨析[J]. 中国图书馆学报,2014,40(213):14-22.
[12] 陈杰. 大数据场景下的云存储技术与应用[J]. 中兴通讯技术,2012,18(6):47-51.
[13] ROUSSEAU R. A view on big data and its relation to Informetrics[J]. Chinese Journal of Library and Information Science, 2012(3): 12-26.
[14] 庄峻,褚燕. 基于领域知识的企业知识模型[J]. 华东电力,2013,41(5):1101-1104.
[15] 石杰. 基于知识的企业战略管理系统及其模型研究[D]. 西安:西北工业大学,2003.
[16] WANG Y, et al. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations[OL]. http://dx.doi.org/10.1016/j.techfore.