属性约简在高校就业决策分析中的应用

杨 飞[1],代广珍[2]

(1.安徽医科大学 计算中心,安徽 合肥 230032;
2.安徽大学 电子科学与技术学院,安徽 合肥 230039)

摘 要:粗糙集理论是一种采用新方式来研究不精确、不确定性知识的数学工具。属性约简的计算是粗糙集理论中的一个重要问题。描述基于粗糙集的属性约简的相关概念,包括核、约简、分类精度;通过分析多种属性约简算法,结合可辨识矩阵和逻辑运算,提出了一种属性约简算法;围绕高校中的管理信息系统,利用该算法抽取与学生就业相关的数据信息,给出了影响学生就业的各条件因素与工作方向之间的依赖关系和约简后的数据表;获取相关规则得出结论,取得了良好的效果。

关键词:粗糙集;属性约简;决策分析;就业

中图分类号:TP311.13;G434 **文献标识码**:A **文章编号**:1673-629X(2007)07-0223-03

Application of Attribute Reduction Algorithm in Collegiate Employment Decision Analysis

YANG Fei[1], DAI Guang-zhen[2]

(1. Computer Center of Anhui Medical University, Hefei 230032, China;
2. School of Electronic Science and Technique, Anhui University, Hefei 230039, China)

Abstract: Rough set theory is a new mathematical tool to research imprecise and uncertain knowledge. Attribute reduction's computation takes an important role in this theory. describes the concept of attribute reduction based on the rough set theory, including core, reduction, and classification precision; analyzes many kinds of attribute reduction algorithm; and then comes up with a new attribute reduction algorithm with the help of discernibility matrix and logic operation. It has been proved that in the area of MIS in university, with the employment-related data collected, this algorithm can be applied to achieve the implied relationship between each element and working direction as well as contracted data sheets. Finally it can help gain the correlation rule, draw the conclusion, obtain a good effect.

Key words: rough sets; attribute reduction; decision analysis; employment

0 引 言

随着我国经济建设的快速发展、科教兴国战略和人才强国战略的实施,我国高等教育规模在不断扩大。但是,由于工作岗位的限制,学生就业问题日益突出,各所高校为了自身更好的发展,一方面加强办学的软硬件设施,另一方面管理决策者需要通过一个平台来把握如何培养学生能力、提高学生素质,进而为提高学生的就业率做出更好的决策,在全方位的提高学生工作就业率,从而吸引更多、更好的生源。

在高校中一般都建立了完备的学生、职工档案管理系统,包括职工档案信息库、学生档案信息库、学生成绩信息库、学生就业信息库等。如何从这些历史数据中发掘出有利于学生教育和分配的决策信息,提供给管理决策者们,是我们所关心的问题。

粗糙集理论(Rough Set)是由波兰科学家 Z. Pawlak 教授在 1991 年提出研究不完整数据、不精确知识的表达、学习、归纳方法[1]。这一理论从新的视角出发对知识进行了定义,它把知识看作是关于论域的划分,并引入代数学中的等价关系来讨论知识,它为智能信息处理提供了有效的处理技术,目前已经在人工智能、知识获取、模式识别、分类等方面得到了显著成功的应用。知识约简是在保持知识库分类能力不变的前提下,删除不相关或者不重要的知识,使得在大量的数据信息中能够挖掘出高效的、有价值的模式信息来辅助决策。

收稿日期:2006-09-27

基金项目:安徽省教育厅自然基金资助项目(2006KJ078B);安徽医科大学校科研基金资助项目(2005KJ23)

作者简介:杨 飞(1976-),男,安徽蚌埠人,硕士,讲师,主要研究领域为数据挖掘、粗集理论。

文中探讨粗糙集的属性约简在高校就业决策分析中的应用。对于以“工作方向”为决策属性的学生就业信息表,首先进行数据预处理,然后利用属性约简算法给出了各条件因素与决策属性之间的依赖关系和约简后的数据表,并提取了有助于决策分析的规则,取得了良好的效果。

1 Rough 集的基本概念

关于粗糙集的基本概念,如不可分辨关系、知识粒度、粗糙集合的上下近似集、正域、负域等详细信息可见文献[2]。

定义 1 一个决策表信息系统 $S=<U,R,V,f>$,其中 U 是对象的集合,也称为论域,$R=C\cup D(C\cap D=\varnothing, D\neq\varnothing)$,子集 C 和 D 分别是条件属性集和决策属性集,V 是属性值的集合,$f:U\times R\rightarrow V$ 是一个信息函数,它表示每个实例对象与其相关属性的映射值。

定义 2 设 U 为一个论域,P,Q 是定义在 U 上的两个等价关系簇,若 $\mathrm{POS}_p(Q)=\mathrm{POS}_{P-\{r\}}(Q)$,则称 r 为 P 中相对于 Q 不必要的,否则称 r 为 P 中相对于 Q 绝对必要的。P 中所有 Q 绝对必要的关系组成的集合称为 P 的 Q 核,记为 $\mathrm{CORE}_Q(P)$。

定义 3 设 U 为一个论域,P,Q 是定义在 U 上的两个等价关系簇,若 P 的 Q 独立子集 $S\subset P$ 有 $\mathrm{POS}_S(Q)=\mathrm{POS}_P(Q)$,则称 S 为 P 的 Q 约简。

定义 4 设 $K=(U,R)$ 为知识库,且 $P,Q\subseteq R$,当 $k=r_p(Q)=\mathrm{card}(\mathrm{POS}_P(Q))/\mathrm{card}(U)$ 时,称知识 Q 是 k 度可导的($0\leqslant k\leqslant 1$),记为 $P_K\Rightarrow Q$。其中 $\mathrm{card}(\mathrm{POS}_P(Q))$ 表示根据 P,U 中所有一定能归入 Q 的元素的数目。所求出的 k 值用于表达 P 和 Q 之间的依赖度。

定义 5 假定集合 X 是论域 U 上的一个关于知识 B 的 Rough 集,定义其 B 精度(在不发生混淆的情况下,也简称精度)为

$$d_B(X)=|B^-(X)|/|B_-(X)| \tag{1}$$

其中,$X\neq\varnothing$;如果,$X=\varnothing$,可定义 $d_B(X)=1$。

定义 6 设集合 $F=\{X_1,X_2,\cdots,X_n\}(U=\bigcup_{i=1}^{n}X_i)$ 是 U 上定义的知识,B 是一个属性子集,定义 B 对 F 近似分类的精度 $d_B(F)$ 为

$$d_B(F)=\sum_{i=1}^{n}|B_-(X_i)|/\sum_{i=1}^{n}|B^-(X_i)| \tag{2}$$

在粗集理论中,知识约简和规则提取算法已被证明是 NP 问题。属性约简的计算是粗糙集理论中的一个重要问题。近年来,人们已提出了许多的属性约简算法,如 A. Skowron 提出了一种用分明矩阵表示知识的方法,并用它来解释、计算数据核和约简[3];叶东毅教授在文献[4]中通过对可辨识矩阵的改进提出了一种计算决策表核属性的方法;王国胤教授提出了基于条件熵的约简算法[5]、基于可辨识矩阵和逻辑运算的约简[2]及决策表在信息熵定义下的核属性计算方法[6]。

定义 7 令决策表系统为 $S=<U,R,V,f>$,$R=P\cup D$ 是属性集合,子集 $P=\{a_i\mid i=1,\cdots,m\}$ 和 $D=\{d\}$ 分别称为条件属性集和决策属性集,$U=\{x_1,x_2,\cdots,x_n\}$ 是论域,$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。$\boldsymbol{C}_D(i,j)$ 表示可辨识矩阵中第 i 行 j 列的元素,则可辨识矩阵 $\boldsymbol{C}_D$ 定义为:

$$\boldsymbol{C}_D(i,j)=\begin{cases}\{a_k\mid a_k\in p\wedge a_k(x_i)\neq a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ 0, & d(x_i)=d(x_j)\end{cases} \tag{3}$$

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

文中基于可辨识矩阵和逻辑运算的约简,得到如下的属性约简算法:

1)计算决策表的可辨识矩阵 $\boldsymbol{C}_D$;

2)矩阵中存在单属性元素,取出该元素作为属性核,并且所有包含该元素的值置 0;

3)在矩阵中的所有取值为非空集合的元素 C_{ij},建立析取逻辑表达式 $L_{ij}=\bigvee_{a_r\in c_{ij}}a_r$;

4)对析取逻辑表达式 L_{ij} 进行合取运算,得到合取范式 $L=\wedge L_{ij}$;

5)获取属性约简。

2 应用实例

文中根据 2001 级临床医学专业的某个毕业班级学生就业数据作为数据源。

2.1 数据预处理

通过原始数据表学生档案表、学生成绩表、学生就业表,抽取与工作方向相关的属性及属性值。通过分析选取了专业成绩、身体状况、活动能力、实验操作能力、计算机水平、外语水平作为条件属性,工作方向作为决策属性。

其中专业成绩是各门专业课的加权成绩,为了避免成绩的随意性,经过与相关专业教师的探讨,对每门课程加了一个参数 β,确保课程的难易相关系数,这样加权成绩比较客观、公正。由于成绩是连续值,需要对成绩进行离散化处理,根据数值区间把成绩列为 A(85~100)、B(75~85)、C(60~75)、D(<60)四类;身体状况泛化为 1:健康,2:一般;活动能力这里泛化为 1:班

干部,2:群众;实验操作能力根据实验课实际成绩泛化为 1:强,2:一般,3:差;计算机水平根据国家等级考试由相应级别泛化为 1:一级,2:二级,3:三级,4:四级;外语水平根据四、六级考试设定为 F: 四级及以上水平,S:六级及以上水平;工作方向,根据我校学生的工作分布一般在省级、市级、县级医院,还有部分同学分到医药公司或自己建立诊所,把这部分同学规划到其它类,可以泛化为 A:省级,B:市级,C:县级,D:乡镇社区,E:考研,F:其它。为了简单起见,随机选择 15 个学生做为实例对象,如学生就业数据表(见表 1)。

表 1　某班级学生就业数据表

U	条件属怀(c)						决策属性(d)
	专业成绩(a_1)	身体状况(a_2)	活动能力(a_3)	实验操作能力(a_4)	计算机水平(a_5)	外语水平(a_6)	工作方向
X_1	A	1	2	2	2	S	E
X_2	A	1	1	1	2	S	A
X_3	B	1	1	1	3	S	A
X_4	B	1	2	2	2	F	B
X_5	C	1	2	1	3	F	F
X_6	C	1	2	2	2	F	C
X_7	B	1	1	2	2	S	E
X_8	B	1	2	1	2	S	B
X_9	A	1	1	2	2	F	B
X_{10}	B	1	2	2	2	F	B
X_{11}	B	1	2	2	2	S	E
X_{12}	C	1	2	2	2	F	C
X_{13}	B	1	2	1	2	S	B
X_{14}	B	1	2	1	2	F	B
X_{15}	B	1	2	1	2	F	B

2.2　属性约简

首先,看到属性"身体状况"的属性值全为 1,由:

$U = \{X_i\}(i = 1,2,3,\cdots,15)$　$C = \{a_i\}(i = 1,2,3,4,5,6)$　$D = \{d\}$

$U/\mathrm{IND}(D) = \{\{X_1, X_7, X_{11}\}, \{X_2, X_3\}, \{X_4, X_8, X_9, X_{10}, X_{13}, X_{14}, X_{15}\}, \{X_5\}, \{X_6, X_{12}\}\}$

$U/\mathrm{IND}(C) = \{\{X_1\}, \{X_2\}, \{X_3\}, \{X_4, X_{10}\}, \{X_5\}, \{X_6, X_{12}\}, \{X_7\}, \{X_8, X_{13}\}, \{X_9\}, \{X_{11}\}, \{X_{14}, X_{15}\}\}$

$\mathrm{POS}_C(D) = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}\} = U$,可以知道 U 是 C 上相对于 D 是一致的。

同样可以求得 $\mathrm{POS}_{C-\{a_2\}}(D) = \mathrm{POS}_C(D)$,知道 a_2 为 C 中相对于 D 是不必要的,可以删除该属性及其属性值,得到一个新表,限于篇幅不再画出。

在新表中求得可辨识矩阵,存在单属性元素 a_1, a_4, a_6,即 $\mathrm{CORE}_D(C) = \{a_1, a_4, a_6\}$,调整后的新矩阵如图 1 所示(为简化起见,省略的元素皆为 0)。

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_1															
x_2								a_3a_5					a_3a_5		

图 1　调整后的可辨识矩阵

在新的矩阵中包含两个非空元素,经过逻辑析取化简得:

$L = a_3 \vee a_5$

把刚才求出的核属性分别加入到 L 中合取后,有

$(a_1 \wedge a_3 \wedge a_4 \wedge a_6) \vee (a_1 \wedge a_4 \wedge a_5 \wedge a_6)$

删除冗余的行,得出两个约简后的表,限于篇幅只画出一个表,见表 2。

表 2　约简后的表

U	条件属性(c)				决策属性(d)
	专业成绩(a_1)	活动能力(a_3)	实验操作能力(a_4)	外语水平(a_6)	工作方法
X_1	A	2	2	S	E
X_2	A	1	1	S	A
X_3	B	1	1	S	A
X_4	B	2	2	F	B
X_5	C	2	1	F	F
X_6	C	2	2	F	C
X_7	B	1	2	S	E
X_8	B	2	1	S	B
X_9	A	1	2	F	B
X_{10}	B	2	2	S	E
X_{11}	B	2	1	F	B

2.3　规则提取

从约简后的两个表中,可以得出与"工作方向"有关联的潜在规则。

规则 1:IF 专业成绩 = 'A'或'B' and 外语是六级或以上水平 and 实验操作能力一般 Then 工作方向选择考研。

规则 2:IF 专业成绩 = 'A'或'B' and 外语是六级或以上水平 and 实验操作能力强 Then 在省级医院工作。

规则 3:IF 专业成绩 = 'B' and 外语是四级或以上水平 and 实验操作能力强 Then 在市级医院工作。

规则 4:IF 专业成绩 = 'B'或'C' and 计算机是 3 级或以上水平 and 实验操作能力强 Then 在其它医药公司或自己办公司单独创业。

3　结　论

从以上实例分析可以看到,利用 Rough Set 中的属性约简,可以快速地发掘与学生就业的相关潜在因素,通过提取的规则,有助于决策者们规划相关专业的课程设置,调整教学思路,把握教学方向,为培养多方

(下转第 229 页)

方正书版注解中有一些特殊符号,如换行符、换段符等在 FoxPro 环境下无法输入,解决的办法是在方正书版环境下输入这些符号,然后用复制粘贴的方法拷贝到 FoxPro 环境中,尽管这些符号在 FoxPro 环境中无法正常显示,但不会影响程序的功能。

表结构中的注解字段可以事先定义好,也可以在录入完成后通过修改数据库结构的方法插入这些字段,注解字段要留有足够的宽度。有时需要对数据库中记录的顺序进行调整,而现有的字段又不能满足排序要求,这时可以增加一个专门用来排序的临时字段,排序完成后将该字段删除,以免影响排版结果。

对文字进行校对工作时,可以用数据库直接打印出校对稿,也可以先运行转换程序把数据库文件转换成小样文件,然后利用方正书版软件编译生成的编排结果打印出校对稿,但对原始数据的修改工作都应该在数据库中进行。

2　结　论

对于信息量大、规律性很强的出版物的排版制作,应该尽量交给计算机来自动完成,这样才能够压缩制作的周期,同时保证排版效果的完全一致,而且人为出错的几率也会大大降低。数据库编程排版方法改变了传统将内容和排版注解录入小样文件的排版方法,而是将内容录入到数据库中,通过程序自动添加排版注解,将数据库文件转换为符合方正书版要求的小样文件,最后通过批处理排版软件编译处理,获得用来输出的排版结果文件。与传统的排版方法相比,这种方法具有以下特点:

1)将大量重复性容易出错的工作交给程序完成,不仅减少了工作量,提高了工作效率,而且排除的版面一致性好。

2)采用关系型数据库管理系统(RDBMS)管理编排数据,可以有效地管理数据,充分利用数据库的功能对数据进行分类、排序、查重等整理工作,非常适合边收集资料边进行编排的工作。

3)在实现的过程中,不仅完成了编排任务,同时还建成了数据库,这个数据库还可以用作其他用途。

4)将 FoxPro 命令与 DOS 命令相结合的方法,有效地解决了对 memo 备注字段内容的连续转换,满足了对文摘、内容简介等大段文字内容的排版要求,扩大了适用范围。

5)编排的信息虽然统一地存放到同一个数据库中,但在编排的制作时可以通过数据库的选择功能对数据进行选择输出,以满足不同的需求。

总之,数据库编程排版方法的实现过程灵活、方便,对系统的环境要求不高,可以在没有网络环境下的单台微机上实现,也可以在网络环境中应用。数据库编程排版方法也可以推广应用到其它批处理或叙述标记系统(descriptive markup system)[8]的排版软件中,如 LaTex 排版系统等等。

参考文献:

[1] 徐容宽. Visual FoxPro 6.0 简明教程[M]. 南京:东南大学出版社,2005.

[2] 梁锐城. Visual FoxPro 数据库应用教程[M]. 北京:科学出版社,2005.

[3] 北京大学计算机科学技术研究所,北京大学新技术公司培训部. BD 排版语言[M]. 北京:[出版者不祥],1992.

[4] 杨　勇,周庆才. 电子排版技术:方正飞腾 4.0[M]. 北京:电子工业出版社,2004.

[5] 易桂生. 计算机文字排版技术[M]. 北京:科学出版社,2003.

[6] 丁革建. BD 排版语言与数据库输出表格设计[M]. 微机发展,1996,6(6):44－45.

[7] 仲秋雁,李　弘. FoxPro 3.0 实用程序设计[M]. 大连:大连理工大学出版社,1994.

[8] 罗振东,葛向阳. 排版软件 LATEX 简明手册[M]. 第 2 版. 北京:电子工业出版社,2004.

(上接第 225 页)

位人才,为营造多渠道就业环境搭建了一个好的平台。通过该方法,对高校招生、教职工科研方面进行了分析,并取得了许多有价值的信息。

通过实例也可以看到,利用可辨识矩阵和逻辑运算的约简算法,比较繁琐,下一步工作是对该属性约简算法做进一步改进,提高挖掘效率。

参考文献:

[1] Pawlak Z. Rough Sets——Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] 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.

[3] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[C]//In: Slowinski R. Intelligent Decision Support－Handbook of Applications and Advances of the Rough Sets Theory. [s.l.]: Kluwer Academic Publishers, 1992.

[4] 叶东毅,陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7):1086－1088.

[5] 王国胤,于　洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7):759－766.

[6] 王国胤. 决策表核属性的计算方法[J]. 计算机学报,2003, 26(5):611－615.