结合模糊集理论的粗糙集属性约简算法*

钱 锋, 陈海山, 姜青山
(厦门大学 软件学院, 福建 厦门 361005)

摘 要: 结合模糊关系的理论, 对粗糙集理论的属性约简算法进行研究, 提出了一个新的属性约简算法, 并给出了一个应用实例。

关键词: 粗糙集; 模糊集; 属性约简算法
中图分类号: TP311 文献标志码: A 文章编号: 1001-3695(2007)11-0093-03

Attribute reduction algorithm of rough set combined fuzzy set theory

QIAN Feng, CHEN Hai-shan, JIANG Qing-shan
(School of Software, Xiamen University, Xiamen 361005, China)

Abstract: This paper discussed the attribute reduction in rough set combined fuzzy relation theory, and then proposed a new attribute reduction algorithm and gave an illustrative example.

Key words: rough set; fuzzy set; attribute reduction algorithm

波兰数学家 Z. Pawlak[1] 于 1982 年提出的粗糙集理论是一种新的处理不精确、不完全与不相容的数学方法, 能有效地处理各种不完备信息, 并从中发现隐含知识, 揭示潜在的规律。粗糙集以不可分辨关系为基础, 研究不同类中对象组成的集合之间的关系。属性约简是粗糙集理论的核心问题和重要课题之一。

随着数据挖掘(data mining, DM)和知识发现(knowledge discovery in database, KDD)的概念在 1989 年被提出, 随之出现了新一代的技术和工具用于 DM 和 KDD 领域。在 DM 和 KDD 的诸多方法中, 粗糙集理论与方法是复杂系统中一种较为有效的方法。因为它与概率方法、模糊集方法和证据理论方法等其他处理不确定性问题理论最显著的区别是它无须提供问题所需处理的数据集合之外的任何先验信息, 所以它对数据的不确定性描述和处理一般来说是比较客观的。

信息系统约简主要是使信息量减少, 将一些无关或多余的信息丢弃, 而不影响其原有的功能。目前粗糙集应用的有效算法的研究主要集中在信息系统属性约简和用以规则提取的值约简方面。属性约简是指在保持信息系统分类或决策能力不变的条件下, 删除冗余属性, 用以得出正确的、简洁的规则。求解最小属性约简是 NP-hard 问题[2]。不过在实际应用中, 得出相对属性约简就可以了。

研究人员已经提出很多属性约简算法[2~7]。其中, 不论是基于约简后属性数最少还是约简后规则最简, 都没有考虑到数据领域知识的特殊性和用户要求的灵活性。正如前面所说粗糙集不依赖任何先验信息比较客观一样, 本文结合模糊关系让它具有一定的领域知识, 让本文属性约简算法具有更实际的决策需要和用户要求。实验证明, 用户可以根据专家领域知识调整阈值, 得到用户满意的属性约简结果。

1 粗糙集理论

1.1 信息系统

知识表达在数据处理中有着十分重要的地位, 知识表达系统也被称为信息系统。形式上四元组 $S=(U, A, V, f)$ 是一个信息系统。其中: U 是对象的非空有限集合, 称为论域 $U=\{x_1, x_2, \cdots, x_n\}$; A 是属性的非空有限集合, $A=\{a_1, a_2, \cdots, a_m\}$; $V=\bigcup_{a\in A} V_a$ 是属性的值域集, V_a 是属性 $a\in A$ 的值域; f 是信息函数, $f: U\times A \rightarrow V$, 它为每个对象的每个属性赋予一个信息值, 即 $\forall a\in A, x\in U, f(x, a)\in V_a$。

如果 $A=C\cup D$, $C\cap D=\varnothing$ (C 表示条件属性集, D 表示决策属性集), 则该类信息系统又称为决策系统。决策系统是一类最为常见的信息系统。

信息系统的数据以关系表的形式来表示。关系表的行对应所要研究的对象, 列对应对象的属性。对象的信息是通过指定对象的各属性值来表达。

1.2 不可分辨关系

设 R 是 U 上的一个等价关系, U/R 表示 R 的所有等价类构成的集合, $[x]_R$ 表示包含元素 $x\in U$ 的 R 等价类。对于属性子集 $R\subseteq A$, 则 R 在 U 上的不可分辨关系定义为 $IND(R)=\{(x, x')\in U\times U \mid \forall a\in R, a(x)=a(x')\}$。

收稿日期: 2006-09-02; 修返日期: 2006-12-13 基金项目: 国家自然科学基金资助项目(60275023); 厦门大学科学研究基金资助项目(Y07002)

作者简介: 钱锋(1981-), 男, 江苏张家港人, 硕士研究生, 主要研究方向为数据挖掘、粗糙集约简算法(leo-qianfeng@hotmail.com); 陈海山(1957-), 男, 副教授, 硕导, 主要研究方向为算法分析、数据库理论及其应用技术; 姜青山(1962-), 男, 教授, 博导, 主要研究方向为数据挖掘、图像处理、数学模型、数据库系统、聚类分析、地球信息系统、数字通信、模糊集理论与应用、统计计算.

不可分辨关系是一种等价关系,它把 U 划分为有限个集合,称为等价类。在每个集合中,对象间是不可分辨的。用 $U/\mathrm{IND}(R)$ 表示 U 的所有等价类。对于元素 $x\in U$,它的等价类定义为 $[x]_R=\{y\mid(x,y)\in \mathrm{IND}(R)\}$。容易看出,在信息系统中,一个属性对应一个等价关系,一个表可以看做是定义的一族等价关系,即知识库。

1.3 集合近似和粗糙集

给定一个信息系统 $S=(U,A)$,对于任意一个对象集合 $X\subseteq U$ 以及属性集合 $R\subseteq A$。X 的 R 下近似定义为 $R\_(X)=\{x\in U\mid[x]_R\subseteq X\}$;$X$ 的 R 上近似定义为 $R^-(X)=\{x\in U\mid[x]_R\cap X\neq\varnothing\}$。$X$ 的 R 下近似表示所有一定属于 X 的对象集合;X 的 R 上近似表示所有可能属于 X 的对象集合。X 的边界区定义为 $BN_R(X)=R^-(X)-R\_(X)$,这表示既不能确定又不能划入 X 对象的集合。如果 $BN_R(X)$ 非空,则称 X 是粗糙集。

粗糙集的精度定义为 $\alpha_R=|R\_(X)|/|R^-(X)|$。其中,$|\cdot|$ 表示集合的基数。对有限集合来说,集合基数就是集合中元素的个数,即可将 X 对等价关系集 R 的粗糙程度用 R 下近似集合成员个数与 R 上近似集合成员个数的比值来量度。显然 $0\leqslant\alpha_R\leqslant 1$。如果 $\alpha_R=1$,则称集合 X 相对于 R 是清晰的;如果 $\alpha_R<1$,则称集合 X 相对于 R 是粗糙的。

1.4 信息约简

信息约简有两个基本概念,即约简(reduction)和核(core)。对于信息系统 $S=(U,A)$,任何最小集 $R\in A$ 且 $\mathrm{IND}(R)=\mathrm{IND}(A)$ 定义为信息系统的一个约简。记 $\mathrm{RED}(A)$ 表示所有的约简集,所有约简的交称为信息系统的核。

对于决策系统 $S=(U,C\cup D)$,C 为条件属性集合,D 为决策属性集合。$B\subset C$,定义 B 相对于 D 的正域为 $\mathrm{POS}_B(D)=\{B\_(X)\mid X\in U/\mathrm{IND}(D)\}$。其中,$U/\mathrm{IND}(D)$ 为 D 对 U 划分所得到的等价类集合。

2 模糊集关系

前面所述的粗糙集模型有一个很大的缺点,对于其中的等价关系 R 来说,应用在很多现实例子中是有困难的。一个重要的办法是先用其他的关系来处理,然后再转换成等价关系。本文中用模糊集中的相似关系来解决这个问题。所谓相似关系是指满足自反性和对称性的二元模糊关系,不满足传递性。为此需要将模糊相似关系 R 改造成模糊等价关系 $t(R)$;然后在适当的阈值上截取,便可得到所需 U 的一个分类。

可用传递闭包的方法将 R 改造成 $t(R)$。此时 $t(R)$ 满足了传递性,于是模糊相似矩阵 R 就被改造成了一个模糊等价关系矩阵 $t(R)$[7]。

定义模糊相似矩阵为 $R=(r_{ij})_{n\times n}$。其中:$r_{ij}\in[0,1]$,$r_{ij}=r_{ji}$,$r_{ii}=1(i,j=1,2,\cdots,n)$。用数 r_{ij} 来刻画对象 x_i、x_j 的相似程度。在实际中关键是如何确定 x_{ij} 的值。本文主要采用文献[8]中的绝对减数法来构造 $r_{ij}=1-c\times\sum_{k=1}^{m}|x_{ik}-x_{jk}|$。其中,$c>0$ 为常数,可根据实际情况选定,使 $r_{ij}\in[0,1]$。

3 结合模糊关系理论的粗糙集属性约简算法

属性约简是粗糙集用于数据分析的重要方法。它主要消减属性个数,即将一些无关或多余的属性去掉,而不影响数据原有的功能,再将约简后的信息重新组合而产生新的决策规则。

约简对于在模型中分类对象最终构建一系列规则是重要的,找出所有约简的问题是 NP-hard问题。目前已研究了许多搜寻约简的方法,如基于差别矩阵、属性依赖度、条件信息熵、遗传算法、特征选择的属性约简算法等。本文结合模糊集的相似关系,克服了原模型等价关系过于严格的约束,可以让用户根据决策的需要和专家领域的知识来更改阈值,得到更满意的属性约简结果。

理论上的粗糙集只能对数据库中的离散属性进行处理,而绝大多数现实的数据库既包含了离散属性,又包含了连续属性。实际应用中数据必须以类别的形式出现。因此,连续数据必须首先进行离散化处理。离散的结果可能会降低原始数据的精度,但将会提高它的一般性。

具体的算法如下:

a)对所需分析的原始数据进行预处理,如缺失数据的加补、重复对象的合并等;然后进行离散化生成包括条件属性和结论属性集合的满足粗糙集数据处理要求的二维关系规则表[4]。

b)计算每个属性 a_i 的一组数字取值集合,设属性 a_i 有 n 个不同的属性取值,根据这 n 个不同的属性取值,确定对论域 U 的划分 $U/\mathrm{IND}(a_i)$,$U/\mathrm{IND}(a_i)=\{E_{1i},E_{2i},\cdots,E_{ni}\}$。属性 a_i 的数字取值集合为 $\{|E_{1i}|,\cdots,|E_{ni}|\}$。其中运算符 $|\cdot|$ 是求集合中元素的个数。

c)用上面求得的属性 a_i 的数字取值集合,采用绝对值减数法,求得它们之间的相似矩阵 $R=(r_{ij})_{n\times n}$。其中 r_{ij} 用来刻画对象 x_i、x_j 之间的相似程度。

d)使用模糊理论中矩阵复合运算用平方法求出相似矩阵 $[R]$ 的传递闭包 $[t(R)]$。$[R]^2=[R]\circ[R]$,$[t(R)]=[R]^n$。其中 $n=2^k(k=1,2,3\cdots)$。

e)对于矩阵依据属性间的关联强度,适当选取阈值 α,得到主条件属性集。$[t(R)]_\alpha=[r_{ij}(\alpha)]_{n\times n}$;$r_{ij}(\alpha)=\begin{cases}1 & r_{ij}\geqslant\alpha\\ 0 & r_{ij}<\alpha\end{cases}$。

f)根据计算出来的新的分类属性进行属性约简,导出相应的决策规则。

4 应用实例

市场上汽车数据库积累了大量有关家庭用车交易的数据。由于原始数据比较详尽,为了说明问题且尽量简化,经过预处理和离散化后如表 1所示。有 20个汽车样本,5个条件属性(分别为汽车排气量、汽车的车体、汽车的加速性能、汽车的最高时速、汽车的耗油量)和 1个决策属性为汽车价格。

表 1中包含了 20个目标对象及其属性,论域 $U=\{1,2,3,\cdots,20\}$ 表示目标集,条件属性 $C=\{a_1,a_2,a_3,a_4,a_5\}$,决策

属性$D = \{d\}$。其中:

a_1表示家庭汽车的排气量大小——1小排气量; 2中小排气量; 3中大排气量; 4为大排气量。

a_2表示家庭用车的体积——1为小体积; 2为中小体积; 3为中大体积; 4为大体积。

a_3表示汽车的加速性能——1为较差的加速性; 2为中等的加速性; 3为较好的加速性; 4为很好的加速性。

a_4表示汽车的最高时速——1为较慢速度; 2为中等速度; 3为较快速度; 4为很快速度。

a_5表示汽车的耗油量——1为低耗油量; 2为较低耗油量; 3为中等耗油量; 4为高耗油量。

d表示汽车的价位: 1为普通价位; 2为较高价位; 3为高价位。

表 1　家庭汽车数据表

记录号	排气量	车体	加速性能	最高时速	耗油量	价位
(U)	(a_1)	(a_2)	(a_3)	(a_4)	(a_5)	(d)
1	2	2	4	1	3	1
2	3	3	3	1	3	2
3	2	2	4	2	2	1
4	4	1	1	4	3	3
5	2	3	3	3	1	2
6	1	2	3	2	4	1
7	4	4	4	2	4	3
8	4	4	3	2	3	3
9	4	4	2	3	4	3
10	3	3	2	3	2	3
11	3	2	2	3	4	2
12	3	3	2	3	4	2
13	2	3	2	4	4	2
14	2	1	1	2	3	2
15	1	1	3	2	1	1
16	1	1	3	1	2	1
17	4	3	1	2	1	1
18	3	3	2	1	1	2
19	2	2	3	1	2	1
20	1	3	2	2	1	2

根据表 1中的数据可以得出属性集合C的各属性取值集合: $a_1 = (4, 6, 5, 5)$, $a_2 = (4, 5, 8, 3)$, $a_3 = (3, 7, 7, 3)$, $a_4 = (5, 8, 5, 2)$, $a_5 = (5, 4, 5, 6)$。然后就是建立相应的模糊相似矩阵。此时可取 $c = 0.1$, 用绝对值减数法。

$$[R] = \begin{bmatrix} 1 & 0.4 & 0.4 & 0.4 & 0.6 \\ 0.4 & 1 & 0.6 & 0.2 & 0.2 \\ 0.4 & 0.6 & 1 & 0.4 & 0 \\ 0.4 & 0.2 & 0.4 & 1 & 0.2 \\ 0.6 & 0.2 & 0 & 0.2 & 1 \end{bmatrix} \rightarrow [t(R)] = \begin{bmatrix} 1 & 0.4 & 0.4 & 0.4 & 0.6 \\ 0.4 & 1 & 0.6 & 0.4 & 0.4 \\ 0.4 & 0.6 & 1 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 1 & 0.4 \\ 0.6 & 0.4 & 0.4 & 0.4 & 1 \end{bmatrix}$$

如果取 $0.4 < \alpha \leqslant 0.6$ 有 $[t(R)]_\alpha = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$。此时$C$可以分为$\{a_1, a_5\}$, $\{a_2, a_3\}$, $\{a_4\}$3类。如果把 $C' = \{a_1, a_5\}$ 作为主要属性时可以把样本分成 14类 $U/C' = \{U_1, U_2, \cdots, U_{14}\}$。其中: $U_1 = \{u_1, u_{14}\}$, $U_2 = \{u_3, u_{19}\}$, $U_3 = \{u_4, u_8\}$, $U_4 = \{u_7, u_9\}$, $U_5 = \{u_{11}, u_{12}\}$, $U_6 = \{u_{15}, u_{20}\}$, $U_7 = \{u_2\}$, $U_8 = \{u_5\}$, $U_9 = \{u_6\}$, $U_{10} = \{u_{10}\}$, $U_{11} = \{u_{13}\}$, $U_{12} = \{u_{16}\}$, $U_{13} = \{u_{17}\}$, $U_{14} = \{u_{18}\}$。$U/D = \{Y_1, Y_2, Y_3\}$。其中: $Y_1 = \{u_1, u_3, u_6, u_{15}, u_{16}, u_{17}, u_{19}\}$, $Y_2 = \{u_2, u_5, u_{11}, u_{12}, u_{13}, u_{14}, u_{18}, u_{20}\}$, $Y_3 = \{u_4, u_7, u_8, u_9, u_{10}\}$。

经归纳可以得到如下决策规则:

a)确定性的规则有 r_1: $(a_1 = 2 \cap a_5 = 2) \cup (a_1 = 1 \cap a_5 = 4) \cup (a_1 = 1 \cap a_5 = 2) \cup (a_1 = 4 \cap a_5 = 1) \rightarrow d = 1$; r_2: $(a_1 = 2 \cap a_5 = 1) \cup (a_1 = 2 \cap a_5 = 4) \cup (a_1 = 3 \cap a_5 = 1) \cup (a_1 = 3 \cap a_5 = 3) \cup (a_1 = 3 \cap a_5 = 4) \rightarrow d = 2$; r_3: $(a_1 = 3 \cap a_5 = 2) \cup (a_1 = 4 \cap a_5 = 3) \cup (a_1 = 4 \cap a_5 = 4) \rightarrow d = 3$。

b)不确定的规则有 r_4: $(a_1 = 1 \cap a_5 = 1) \cup (a_1 = 2 \cap a_5 = 3) \rightarrow d = 1$, 规则的确定性因子为 0.5; r_5: $(a_1 = 1 \cap a_5 = 1) \cup (a_1 = 2 \cap a_5 = 3) \rightarrow d = 2$, 规则的确定性因子为 0.5。

依上面规则可以看出, 排气量越大的汽车价格越贵, 相对来说油耗大的车也比较贵一点。这是很符合客观实际的。人们购车时有一个共同点, 就是比较看重车的排气量。一般来说, 车的排气量越大越好, 其中尤以个人购车最为显著。同样可以用属性$\{a_2, a_3\}$来划分目标对象集, 可以得出车体比较大、加速性能好的汽车价格相对较高。其实各因素的影响因购车人群的不同而有所不同。家庭购车比较倾向于车体较大的汽车, 普通家庭主要考虑车的价格, 而高薪家庭可能更看重车本身的品质。非家用个人购车, 多为经济实力较雄厚的个人, 他们多有特殊职业, 由于对工作效率的追求, 可能更以动力性和速度为考虑方面。

5　结束语

本文在粗糙集和模糊集理论的基础上提出一种处理粗糙数据的属性约简算法, 探讨了基于粗糙集属性约简算法在决策领域的应用, 对数据挖掘和决策分析都是一个有益的尝试。由于其中引入了模糊关系和模糊集中的复合矩阵运算, 用户可以根据决策的需要和领域知识更改阈值, 得到用户满意的属性结果。通过在家庭用车的应用实例验证了改进算法的可行性、有效性。在粗糙集中如何更好地离散化数据和与其他的软计算方法结合使用等方面, 有待进一步研究。

参考文献:

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

[2] WONG S K M, ZIARKO W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33(11/12): 693-696.

[3] DU Wei-feng, LI Hai-ming. Another kind of fuzzy rough sets[C]// Proc of IEEE International Conference on Granular Computing. 2005: 145-148.

[4] 苗夺谦. Rough Set理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302.

[5] 常犁云, 王国胤, 吴渝. 一种基于 Rough Set理论的属性约减及规则提取方法[J]. 软件学报, 1999, 10(11): 1206-1211.

[6] 代建华, 李元香. 粗集中属性约简的一种启发式遗传算法[J]. 西安交通大学学报, 2002, 36(12): 1286-1290.

[7] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.

[8] 胡宝清. 模糊理论基础[M]. 武汉: 武汉大学出版社, 2004.