概率差别矩阵与不完备信息系统属性约简*)

闫德勤
(辽宁师范大学计算机系 大连 116029)

摘 要 差别矩阵的概念是基于粗糙集理论对信息系统进行属性约简的一个重要内容。针对不完备信息系统的属性约简本文提出了一种概率差别矩阵的概念与构造方法,给出了相关的定理。在此基础上提出了一种利用概率差别矩阵对不完备信息系统属性约简的方法,并给出了应用举例。

关键词 粗糙集,概率差别矩阵,属性约简

Probability Discernibility Matrix and Attribute Reduction for Incomplete Information Systems

YAN De-Qin
(Department of Computer Science, Liaoning Normal University, Dalian 116029)

Abstract In this paper, the attribute reduction for incomplete information systems is approached. A new concept of probability discernibility matrix is proposed, and the construction method of probability discernibility matrix is given. By use of probability discernibility matrix, a new method of attribute reduction for incomplete information systems is proposed, and, as an application of this method an example is given.

Keywords Rough sets, Probability discernibility matrix, Attribute reduction

1 引言

由于不完备信息系统中一些属性的丢失或不确定,使得以等价类为基础的针对完备信息系统属性约简的粗糙集理论方法受到限制[5,6]。为对不完备信息系统进行属性约简,一些学者对信息系统中丢失属性或不确定数据从不同的角度提出了一些处理方法[5],也有一些对传统粗糙集扩展模型的提出。所有这些方法与模型都是在不同意义下对不完备信息系统进行属性约简的一种处理方式。差别矩阵的概念是基于粗糙集理论对信息系统进行属性约简的一个重要内容,本文提出了一种概率差别矩阵的概念与构造方法,给出了相关的定理。在此基础上提出了一种利用概率差别矩阵对不完备信息系统属性约简的方法,并给出了应用举例。

2 基本概念

设 $S=(U,Q,V,F)$为一信息系统,其中 $U=\{x_1,x_2,\cdots,x_n\}$是论域,Q是属性集合,V 是属性取值集合,F 是 $U\times Q\to V$ 的映射。设属性集合包含 m 个条件属性 $C=\{C_1,C_2,\cdots,C_m\}$和一个决策属性 D。若 D 的取值有 s 个,则由 D 导出的等价类构成 U 的一个划分:$\{Y_1,Y_2,\cdots,Y_s\}$。其中,$Y_i=\{x\in U|F(x,D)=i\}$,$i=1,2,\cdots,s$。

信息系统中每一 x_i 及其所对应的属性值称为一个(信息表示的)规则。本文中简称 x_i 规则。

在一个信息系统中,当 $i\neq j$ 时若存在 $F(x_i,C)=F(x_j,C)$但 $F(x_i,D)\neq F(x_j,D)$则称该系统为不相容的,此时 x_i 与 x_j 所对应的规则为不相容规则。

定义 1 设 $X\subseteq U$ 为论域的一个子集,$P\subseteq C$,X 的关于 P 的下近似为

$$P_-X=\{x\in U|[x]_P\subseteq X\}$$

其中$[x]_P$ 表示 U 中在等价关系 P 下的等价类元素构成的集合。

定义 2 设 U 为一个论域,P,Q 为 U 上的两个等价关系簇,Q 的 P 正域记为 $POS_P(Q)$,定义为

$$POS_P(Q)=\bigcup_{X\in U/Q}P_-(X)$$

定义 3 设 $P\subseteq C$,对于划分$\{Y_1,Y_2,\cdots,Y_k\}$的 P 的近似精度为

$$\gamma_P=\sum_{i=1}^{k}card(P_-Y_i)/card(U)$$

其中,$card()$表示集合的基数。

定义 4 设 $P\subseteq C$,若 $\gamma_P=\gamma_C$,且不存在 $R\subset P$,使得 $\gamma_R=\gamma_P$,则称为 P 为 C 的一个属性约简。所有 C 的属性约简的交称为 C 的核,记为 $Core(C)$。

定义 5[4] 给定信息系统 S,差别矩阵 $M=(m_{ij})$的元素定义为:当 $F(x_i,D)\neq F(x_j,D)$时,$m_{ij}=\{a\in C|F(x_i,a)\neq F(x_j,a)\}$,在其它情况下 m_{ij} 为空集。

文[4]指出当差别矩阵中的某个元素为单元素时,该属性属于核。

若信息系统中的每个属性值都是已知的,则称为完备的信息系统。在有些情况下,由于种种原因信息系统中的某些属性值不能得到或确定,则称信息系统为不完备信息系统。

3 概率差别矩阵及相关定理

对于完备信息系统,可以利用差别矩阵进行属性约简[4,6],而对于不完备信息系统由于未知属性值的存在不能采用已有的方法构造差别矩阵[5]。为利用差别矩阵对不完备信息系统进行属性约简,下面提出一种关于不完备信息系统的差别矩阵构造方法,以该方法构造的差别矩阵称为概率差别矩阵。

设 $S=(U,Q,V,F)$为一个不完备信息系统,V 是属性取值集合,信息系统中的不确定属性值用 $*$ 表示,系统中属性取值的种类用 v 表示。构造概率差别矩阵基于不确定属性值取到信息系统中任意一个确定属性值的可能性为 v 分之一的概率假设。

定义 6 给定不完备信息系统 S,概率差别矩阵 $M=$

*)国家自然科学基金(60372071)资助;辽宁师范大学校基金资助。闫德勤 博士,教授,主要研究领域为模式识别、数据挖掘等。

(m_{ij})的元素定义为：

在 $F(x_i,D)\neq F(x_j,D)$ 条件下，当 $F(x_i,a)$ 与 $F(x_j,a)$ 都是确定值时：

$m_1=\{a\in C|F(x_i,a)\neq F(x_j,a)\}$；

当 $F(x_i,a)$ 与 $F(x_j,a)$ 有一个是 * 时：

$m_2=\{(1-\frac{1}{v})a\in C\}$；

当 $F(x_i,a)$ 与 $F(x_j,a)$ 两个都是 * 时：

$m_3=\{(1-\frac{1}{v^2})a\in C\}$；

$m_{ij}=m_1\cup m_2\cup m_3$。

在其它情况下 m_{ij} 为空集。

在概率差别矩阵中，m_{ij} 称为元素项(集合)，其内容是具有系数(或权值)的属性符号。

设 $P\subseteq C$ 为属性集合，对于规则 x_i 与 x_j，记 $a\subseteq P$ 为 $F(x_i,a)$ 与 $F(x_j,a)$ 都是确定值所对应的条件属性集合，记 $b\subseteq P$ 为 $F(x_i,a)$ 与 $F(x_j,a)$ 中只有一个为 * 所对应的条件属性集合，记 $c\subseteq P$ 为 $F(x_i,a)$ 与 $F(x_j,a)$ 都是 * 所对应的条件属性集合。

定义 7 给定不完备信息系统 S，规则 x_i 与 x_j 在属性集合 P 下的相似度 s_{ij} 定义如下：

对于所有属性 $e\in a$ 如果同时满足 $F(x_i,d)=F(x_j,d)$ 或 a 为空集，则 $s_0=1$，否则；$s_0=0$；$s_{ij}=\frac{s_0}{v^{k+2m}}$。

其中，$k=card(b)$，$m=card(c)$。

设由 D 导出的等价类构成 U 的一个划分$\{Y_1,Y_2,\cdots,Y_s\}$ 中 Y_i 的下标集合为 δ_i，对于 $t\in\delta_i$ 令 $\bar{s}_t=\sum_{j\in\delta_i}s_{tj}$，若 $\bar{s}_t>1$ 则令 $\bar{s}_t=1$。

定义 8 设 $P\subseteq C$，对于划分$\{Y_1,Y_2,\cdots,Y_k\}$的 P 的近似精度为

$$\lambda_P=\sum_{i=1}^{k}card(P\_Y_i)/card(U)$$

其中，$card(P\_Y_i)=card(\delta_i)-\sum_{t\in\delta_i}\bar{s}_t$。

定义 9 设 $P\subseteq C$，若 $\lambda_p=\lambda_C$，且不存在 $R\subset P$，使得 $\lambda_R=\lambda_P$，则称 P 为 C 的一个属性约简。所有 C 的属性约简的交称为 C 的核，记为 $Core(C)$。

概率差别矩阵中，元素项只包含一个属性时，称为单属性项，所包含的属性称为单属性，包含两个以上的属性时为多属性项。

定理 1 概率差别矩阵中的单属性为相应不完备信息系统的核属性。

证明：概率差别矩阵中，单属性为对应不同决策属性值的两个规则在该属性下的属性值不确定相同，但在其它条件属性下的属性值一定相同的属性。从概率差别矩阵的定义 6 以及定义 8、9 可知，若删去单属性必定使 λ_c 减小。因此，单属性必是核属性。证毕。

在信息系统中，一个属性的重要性是由其对信息系统近似精度的影响程度决定的。对近似精度的影响大的属性比影响小的属性重要。可约简的属性称为不重要属性。

定理 2 概率差别矩阵中：

(1)若两个属性都是单属性，则在所有单属性元素项中系数和大的属性更重要。

(2)若一个属性所在的每个元素项中总存在至少一个系数为 1 的属性，则这个属性一定不重要。

证明：(1)如果两个属性都是单属性，由定义 7 和 8 可知它们对 λ_C 的影响大小由其单属性项中系数和决定。和大的对 λ_C 的影响就大，因此就更重要。(2)若一个属性所在的每个元素项中总存在至少一个系数为 1 的属性，说明该属性在不同规则(比如规则 x_i 与 x_j)下确定地或有可能地有不同取值的时候总确定的存在在别的属性下有两个确定不同的属性值。由定义 7 可知，$s_0=0$。因此 $s_{ij}=0$。由定义 7 和 8 可知该属性对 λ_C 没有任何影响。所以该属性不重要。证毕。

4 约简例子

对于给定的不完备信息系统进行属性约简的第一步是根据定义 6 构造出相应的差别矩阵。然后根据定理 1 和 2 分别找出核属性和不重要属性。保留核属性，删除不重要属性。对于其它的属性分别比较其对近似精度影响的大小不断把最重要的加入到约简集合中。

表 1 是一个不完备信息系统。由定义 6 可得到相应的概率差别矩阵的元素项。由于差别矩阵是对称的，下面给出的是该差别矩阵的下三角部分。不为空集的元素项是 4～8 行，1～6 列。下三角阵中未写出部分的元素项为空集。其中，$\tau=\frac{1}{v}$。在表 1 中，属性取值只有 0 和 1 两种，$v=2$。对应表 1 的下三角概率差别矩阵如下：

$$M=\begin{pmatrix}
 & & & & & \\
 & & & & & \\
 & & & & & \\
 & & & & & \\
b & b,\tau d & b,\tau c & & & \\
c,d & \tau c,\tau d & \tau c & & & \\
\tau a,b,\tau c,d & \tau a,b,\tau c,\tau d & \tau a,b,\tau^2 c,d & & & \\
\tau c,b,d & \tau c,b,\tau d & \tau^2 c,b,d & \tau c,d & b,\tau c & \tau a,\tau^2 c \\
a,c,d & a,c,\tau d & a,\tau c,d & a,b,c,d & a & \tau a,\tau c
\end{pmatrix}$$

从概率差别矩阵 M 中可以看到表 1 的不完备信息系统中属性 a,b,c 都是单属性，因此是核属性。经计算，表 1 的不完备信息系统条件属性的近似精度为 $\lambda_C=\frac{1}{8}(5+\tau^3)$。删除属性 d 后近似精度 λ_C 的值不改变。所以不完备信息系统表 1 的属性约简为 a,b,c。

表 1 不完备信息系统

U	a	b	c	d	e
x_1	1	1	1	0	1
x_2	1	1	1	*	1
x_3	1	1	*	0	1
x_4	1	0	1	0	0
x_5	1	1	0	1	0
x_6	*	0	*	1	0
x_7	1	0	*	1	2
x_8	0	1	0	1	2

结论 为利用差别矩阵的概念方法对不完备属性信息系统进行属性约简，本文提出了概率差别矩阵的概念和构造方法。同时给出了利用概率差别矩阵对不完备信息系统属性约简的具体方法。在本文方法的基础上我们又进一步提出了模糊差别矩阵的概念和构造方法，应用于对模糊(连续)属性以

一种新型的神经网络构造方法 RCBNN*)

谢振华 李 宁 商 琳 陈兆乾 陈世福

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 一个好的神经网络结构可以大大提高它的处理能力和收敛速度,所以神经网络的构造方法一直是人们研究的热点问题。本文利用粗集理论的数据分析能力和决策树对数值属性的分割能力,提出一种基于粗集与决策树的新型神经网络构造方法 RCBNN。经试验表明,使用该方法构造的神经网络,具有易于构造、可理解性好、收敛速度快且构造的网络规模较小的特点。

关键词 神经网络,粗集,决策树

RCBNN: A New Constructing Method for Neural Network

XIE Zhen-Hua LI Ning SHANG Ling CHEN Zhao-Qian CHEN Shi-Fu

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

Abstract A good structure can greatly improve the power and convergent speed of neural network. According this, the constructing method is always a hotspot in neural network research. This paper proposes a constructing method for neural network based on Rough Set and Decision Tree. Rough Set has good capability of data analysis; Decision Tree is good at segmentation of continues-valued attributes. This method makes good use of these advantages. Experimental results show that the neural network designed by this method is easy constructible, good understandable, fast convergent, and small dimensional.

Keywords Neural network, Rough set, Decision tree

1 引言

波兰科学家 Pawlak 提出的粗集(Rough set)[1]是描述不完整、不精确和含噪声数据的有力工具,其主要特点是不需要任何与数据有关的先验和附加知识。从本质上看,它反映了认知过程在非确定性、非模型化信息处理方面的机制和特点,从而成为一种有效的非单调推理工具。粗糙集模型是一种结构化的、非数值化的信息处理方法,处理的是数据的符号描述,以属性、语义决策规则等形式构造知识表达,其主要特点是不需要任何与数据有关的先验和附加知识,这一点明显区别于模糊集理论中的隶属函数等方法。目前,粗糙集主要应用于属性约简、规则生成及预测等几个方面。但粗集理论容错能力与推广能力相对较弱,适于处理离散的非确定数据,对于连续数据的处理能力有限[2],且因对数值型属性作为离散值看待,规则生成后的规模大;决策树学习是应用最广的归纳推理算法之一,对噪声数据有很好的健壮性,其学习速度快,规则学习能力强,本文所采用的 C4.5 决策树算法[3]能通过对连续数据的分割,有效地缩小规则的规模,但决策树容易产生过拟合现象,虽然通过裁减可一定程度避免过拟合,但精度的损失较大;人工神经网络最具吸引力的是其数值逼近能力,并能够处理定量的、数值化的信息。人工神经网络关注的是建模和学习过程中获得的数据的细节信息,具有较强的自组织能力、容错能力和推广能力,但不能优选条件属性,而粗集的属性约简能力恰好能弥补这一缺陷。已有研究者成功地将粗集理论与神经网络结合,构造出新型的神经网络结构[4,5]。本文利用粗集与决策树理论的优点,并结合人工神经网络的特点提出了一种基于粗集与决策树的神经网络构造方法——RCBNN,该方法易于构造、可理解性好、计算简单、收敛速度快且构造的神经网络规模较小,通过实验表明,该方法能够得到较满意的结果。

2 粗集理论与决策树算法

给定一个有穷对象集 U、有穷属性集 A、各属性值的集合 V 以及信息函数 f 构成一个信息系统,表示为 (U,A,V,f),

*)基金项目:国家自然科学基金(60273033);江苏省自然科学基金(BK2003067)。**谢振华** 硕士研究生,主要研究方向:神经计算。**李 宁** 硕士,讲师,主要研究方向:机器学习。**商 琳** 博士研究生,主要研究方向:人工智能、数据挖掘。**陈兆乾** 教授,博士生导师,主要研究领域:人工智能、机器学习。**陈世福** 教授,博士生导师,主要研究领域:人工智能、机器学习。

及 Vague 集属性信息系统的属性约简,其结果将另文发表。

参 考 文 献

1 Pawlak Z. Rough set approach to multi-attribute decision analysis. European Journal of Operational Research, 1994, 72: 443~459
2 Pawlak Z. Rough set theory and its applications to data analysis. Cybernetics and System, 1998, 29(27): 661~688
3 Chen M S, Han J, Yu P S. Data Mining: An overview from a database perspective. IEEE Transaction on Knowledge and Data Engineering, 1996, 8(6):866~883
4 Hu XiaoHua, Cercone N. Learning in relational database: a rough set approach. Computational Intelligence, 1995, 11(2): 323~337
5 Grzymala-Busse J W, Hu M. A comparison of several approaches to missing attribute values in data mining. In: Proc. of the 2nd Int'l. Conf. on Rough Sets and Current Trends in Computering. Berlin: Springer Verlag, 2000. 378~385
6 王国胤. Rough 集理论与知识获取. 西安:西安交通大学出版社, 2001
7 曾黄麟. 粗集理论及其应用. 重庆:重庆大学出版社,1996