序信息系统中基于粗糙集的证据获取与合成

樊兵娇[1] 徐伟华[1,2]
(重庆理工大学数学与统计学院 重庆 400054)[1]
(南京理工大学高维信息智能感知与系统教育部重点实验室 南京 210094)[2]

摘 要 通过在决策序信息系统中引入证据理论,提出一种基于粗糙集的证据获取与合成方法。利用证据信任度计算近似条件概率分配,根据属性重要度和证据支持度计算权重,然后用合成公式对近似条件概率分配进行合成,得到决策。

关键词 粗糙集,近似条件概率分配,序信息系统,证据理论,证据支持度
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.012

Evidence Acquirement and Combination Method Based on Rough Set in Ordered Information System

FAN Bing-jiao[1] XU Wei-hua[1,2]
(School of Mathematics and Statistics, Chongqing University of Technology, Chongqing 400054, China)[1]
(Key Laboratory of Intelligent Perception and Systems for High-dimensional Information
(Nanjing University of Science and Technology), Ministry of Education, Nanjing 210094, China)[2]

Abstract A novel method of evidence acquirement and combination based on rough set was proposed by introducing evidence theory into the ordered decision information system. Confidence degrees of evidence are used to calculate approximate conditional probability assignments. Evidence weights are calculated according to the attribute significances and support degrees of evidence. Decisions are gained by using combinational rule to integrate approximate conditional probability assignments.

Keywords Rough set, Approximate conditional probability assignment, Ordered information system, Evidence theory, Support degrees of evidence

1 引言

粗糙集理论[9]是近年来发展起来的一种软计算工具,主要用来处理不精确、不确定和模糊知识,它已被成功应用于数据挖掘、模式识别、人工智能与智能信息处理等领域,并越来越受到国际学术界的关注。经典粗糙集是以完备信息系统为研究对象,以等价关系(满足自反性、对称性、传递性)为基础,通过等价关系将论域分成互不相交的等价类,划分越细,知识越丰富,信息越充分。

然而,在实际问题中有许多信息系统由于各种原因(如噪声、信息缺损等)并不是基于等价关系的,即 Pawlak 粗糙集模型中的等价关系极大地限制了粗糙集理论的研究与应用,于是人们将等价关系放宽为相容关系、相似关系等。特别地,Greco、Matarazzo 和 Slowinski 于 1998 年提出了基于优势关系的粗糙集研究方法(DRSA),其主要是利用优势关系代替经典粗糙集中的等价关系建立序信息系统来考虑现实中存在的对属性值排序的问题[3-5]。而且,近年来对推广经典粗糙集的研究也取得了可喜的成果[1,2,6,7,10]。

证据理论又称为 Dempster-Shafer 理论,最初由 Dempster 于 1967 年提出,并于 1976 年由 Shafer 扩充和推广。证据理论也是信息融合技术中一种有效处理不确定性问题的工具,它用集合赋值来代替概率论中的单点赋值,可以看作一种广义概率论。但与传统的概率论相比,其能更好地把握所研究问题的不确定性。证据理论对不同的证据进行合成,从而要比仅通过单一证据进行判定的准确性更高。但是证据理论本身也存在很多问题,比如证据理论合成规则将合成的证据同等对待,这与实际不相符,从而限制了证据理论的应用范围。

在序信息系统中,粗糙集理论主要是通过优势关系来划分论域,进而利用被近似集合的上、下近似来研究。而证据理论主要是通过建立满足两条公理的 mass 函数得到相关证据,进而分析所研究的序信息系统。因此,可以将粗糙集与证据理论的优点融合起来,来解决序信息系统中的证据获取与合成问题。为此,本文提出了一种序信息系统中基于粗糙集获取近似条件概率分配的方法,对于不同的证据,通过对其近似条件概率分配合成求得基本可信度分配,使之更合理。在证据合成方面,由于实际中证据具有不同的重要性,因此需要计算证据权重。本文根据获得的近似条件概率分配,利用优势关系下粗糙集中属性的重要性以及证据支持度来合成证据,

到稿日期:2014-04-20 返修日期:2014-06-20 本文受国家自然科学基金(61105041,61472463,61402064),重庆市自然科学基金资助项目(cstc2013jcyjA40051),南京理工大学高维信息智能感知与系统教育部重点实验室基金(30920140122006)资助。
樊兵娇(1989—),女,硕士生,主要研究领域为人工智能的数学基础,E-mail:fanbingjiao890617@126.com;徐伟华(1979—),男,博士,硕士生导师,主要研究领域为模糊数学、人工智能、粗糙集、应用数学等,E-mail:chxuwh@gmail.com。

有效考虑序信息系统中的权重因子，解决了无权重时引起的重要信息丢失问题，使得合成结果更加符合实际情况。

2 序信息系统粗糙集与证据理论

粗糙集理论[9]是波兰学者 Z. Pawlak 于 1982 年提出的，它能有效地分析和处理不精确、不一致、不完整等各种不完备信息，并从中发现隐含的知识，揭出潜在的规律。在 Pawlak 近似空间意义下的信息系统中，每个属性集决定了一个二元不可区分关系，即等价关系。然而，在实际生活中有许多系统并不是基于等价关系的，而是基于优势关系的，即对每个属性值域有一个递增或者递减的偏序关系，如一个班级的各科成绩情况等问题。这时就需要建立基于优势关系的信息系统，即序信息系统。

定义 1[8] 称一个三元组 $I=(U,AT,f)$ 为一个信息系统，其中：

$U=\{x_1,x_2,\cdots,x_n\}$ 是有限对象集，称为论域；

$AT=\{a_1,a_2,\cdots,a_p\}$ 是有限属性集；

$f=\{f_l\mid f_l:U\to V_{a_l},\forall a_l\in A\}$ 是 U 与 A 的关系集，其中 V_{a_l} 是 a_l 的有限值域。

在一个信息系统中，如果在某个属性值域上建立了偏序关系，则称这个属性为一个准则。当所有的属性都为准则时，该信息系统成为序信息系统。

一般，序信息系统用 $I^{\geqslant}=(U,AT,f)$ 来表示。

在序信息系统 $I^{\geqslant}=(U,AT,f)$ 中，对于 $B\subseteq AT$，令 $R_B^{\geqslant}=\{(x,y)\in U\times U\mid f_l(x)\geqslant f_l(y),\forall a_l\in B\}$，则 $R_B^{\geqslant}$ 称为序信息系统 $I^{\geqslant}$ 的优势关系。

若记 $[x_i]_B^{\geqslant}=\{x_j\in U\mid (x_i,x_j)\in R_B^{\geqslant}\}=\{x_j\in U\mid f_l(x_j)\geqslant f_l(x_i),\forall a_l\in B\}$，$U/R_B^{\geqslant}=\{[x_i]_B^{\geqslant}\mid x_i\in U\}$，则称 $[x_i]_B^{\geqslant}$ 为对象 x_i 关于属性子集 B 的优势类，$U/R_B^{\geqslant}$ 为该序信息系统对象集 U 关于属性子集 B 的一个分类。

对于任意集合 $X\subseteq U$，在优势关系下也可以用一对上、下近似来近似描述概念 X。

$\underline{R_B^{\geqslant}}(X)=\{x\in U\mid [x]_B^{\geqslant}\subseteq X\}$

$\overline{R_B^{\geqslant}}(X)=\{x\in U\mid [x]_B^{\geqslant}\cap X\neq\varnothing\}$

定义 2[8] 决策序信息系统 $I^{\geqslant}=(U,AT\cup D,f,g)$，$D=\{d\}$，$g:U\to V_d(\forall d\in D)$ 是 U 与 D 的关系函数，其中 V_d 是 d 的有限值域，U/R_D 为决策序信息系统中对象集 U 关于决策集合 D 的一个划分。决策属性 D 对条件属性 AT 的依赖度定义为 $r(AT,D)=|POS_{AT}(D)|/|U|$，其中 $POS_{AT}(D)=\bigcup\{\underline{R_{AT}^{\geqslant}}(Y_i)\mid Y_i\in U/R_D\}$，$Y_i=\{x\in U\mid g(x)=d_i\}$，$d_i\in V_D$。

条件属性 a 相对于决策属性 D 的重要度定义为 $SGF(a)=r(AT,D)-r(AT-\{a\},D)$。

定义 3[8] 设 $I^{\geqslant}=(U,AT,f)$ 为序信息系统，$B\subseteq AT$，对任意 $X\in U/R_B^{\geqslant}$，若记 $h(X)=\{x\in U\mid [x]_B^{\geqslant}=X\}$，则 $I^{\geqslant}$ 中的 mass 函数可定义为 $m:U/R_B^{\geqslant}\to[0,1]$，其中 $m(X)=|h(X)|/|U|$。

由以上定义可知，序信息系统中的 mass 函数仍满足两条基本公理。也就是说，对任意 $X\in U/R_B^{\geqslant}$，下列公理成立：

(M1) $m(\varnothing)=0$

(M2) $\sum\limits_{X\in U/R_B^{\geqslant}}m(X)=1$

3 序信息系统中基于粗糙集的证据获取与合成

在传统的序信息系统证据理论中，存在一些问题：首先，mass 函数不仅不容易获取，而且存在一定的主观性，因而在一定程度上限制了证据理论的应用；其次，将证据理论合成规则合成的证据同等对待，与实际不符，同样限制了证据理论的应用范围。

本文是在序信息系统的背景下，提出了一种基于粗糙集的近似条件概率分配获取方法，可以有效克服证据获取的主观性问题；提出了一种新的计算合成证据重要性的方法，用来解决合成证据被同等对待的问题，体现每个证据不同的重要性，使证据的合成结果更准确。

定义 4 设 $I^{\geqslant}=(U,AT\cup D,f,g)$ 为决策序信息系统，$\forall a\in AT$ 为条件属性，$D=\{d\}$ 为决策属性，$U/R_D=\{Y_1,Y_2,\cdots,Y_m\}$，对于 $\forall x\in U$，其证据信任度为 $\mu_j=|[x]_a^{\geqslant}\cap Y_j|/|[x]_a^{\geqslant}|$，$j=1,2,\cdots,m$。

证据信任度表示的是在满足实例 x 的一个条件属性值的前提下，满足决策属性 D 某一个决策值的比例。证据信任度越大，则在实例 x 下该决策的可信度就应该越大，利用证据信任度来获得近似条件概率分配解决了证据获取的主观性问题。

定义 5 设 $I^{\geqslant}=(U,AT\cup D,f,g)$ 为决策序信息系统，$U/R_D=\{Y_1,Y_2,\cdots,Y_m\}$，$\forall a\in AT$，则对于证据 x，在条件属性 a 下得到的近似条件概率分配为 $m(d_i/x)=|[x]_a^{\geqslant}\cap Y_i|/|[x]_a^{\geqslant}|$，$i=1,2,\cdots,m$。其中 $Y_i=\{x\in U\mid g(x)=d_i\}$，$d_i\in V_D$。

由定义 5 明显可得 $m(\varnothing/x)=0$，$\sum\limits_{i=1}^{n}m(d_i/x)=\sum\limits_{i=1}^{n}|[x]_a^{\geqslant}\cap Y_i|/|[x]_a^{\geqslant}|=1$。因此该式满足 mass 函数的定义，可作为序信息系统下的 mass 函数。

在决策序信息系统 $I^{\geqslant}=(U,AT\cup D,f,g)$ 中，每个属性的重要性是不相同的，基于决策表中属性的重要度，将其引入到证据合成公式中，把属性的重要性作为权重因子。用证据合成公式合成近似条件概率分配时，属性的重要性越大，则该属性下的近似条件概率分配在合成时就应具有较大的权重。

定义 6 设 $I^{\geqslant}=(U,AT\cup D,f,g)$ 为决策序信息系统，$\forall a\in AT$ 为条件属性，D 为决策属性，$U/R_D=\{Y_1,Y_2,\cdots,Y_m\}$，证据 x 在条件属性 a 下的证据支持度为 $\mu=\max(|[x]_a^{\geqslant}\cap Y_j|/|U|)$，$j=1,2,\cdots,m$。

证据支持度表示的是满足证据 x 的一个条件属性值，且满足决策属性 D 某一个决策值的实例在论域 U 中所占的比例。证据支持度越大，则该决策发生的可能性越大，可靠性越大。基于证据支持度的这种性质，将其作为证据合成中的另一权重因子。

由以上分析研究可知，一方面，属性的重要度越大，该属性下的近似条件概率分配在合成时就应具有较大的权重；另一方面，证据支持度越大，该决策发生的可能性越大，可靠性越大。因此，给定证据 x，在条件属性 $a_j\in AT$ 下得到的近似条件概率分配 $m(d_i/x)$ 权重计算公式为：

$$w_j=\frac{\mu_j+\lambda_j}{\sum\limits_{j=1}^{m}(\mu_j+\lambda_j)},\ j=1,2,\cdots,n;i=1,2,\cdots,m$$

其中，λ_j 为条件属性 $a_j\in AT$ 的重要度，μ_j 为给定证据 x 时条件属性 a_j 的证据支持度。

当决策序信息系统 $I^{\geqslant}=(U,AT\cup D,f,g)$ 和证据 x 给定时，对于条件属性 $a_j\in AT$，近似条件概率分配为 $m_j(d_i/x)$，$j=1,2,\cdots,n;i=1,2,\cdots,m$，其合成公式为：

$$m(A)=\begin{cases}0, & A=\varnothing \\ \dfrac{\sum\limits_{\cap A_i=A}\prod\limits_{j=1}^{n}w_j m_j(A_i)}{1-\sum\limits_{1\leqslant j\leqslant n\cap A_i=\varnothing}\prod\limits_{j=1}^{n}w_j m_j(A_i)}, & A\neq\varnothing\end{cases}$$

其中，w_j 为给定证据 x 时，条件属性 $a_j\in AT$ 对应的近似条件概率分配权重。

4 实例分析

下面通过一个实例分析验证提出的序信息系统中基于粗糙集的证据获取与合成方法的有效性和优越性。

例 如表 1 所列，$I^{\geqslant}$ 为决策序信息系统，给定证据 $x=\{3,2,2,2\}$，$y=\{1,2,1,2\}$，通过决策表来推断其决策。

表 1 决策序信息系统

U	A				
	a_1	a_2	a_3	a_4	d
x_1	1	2	2	2	1
x_2	3	2	2	2	1
x_3	1	2	1	2	2
x_4	2	2	1	2	2
x_5	1	1	1	1	3
x_6	1	2	1	2	3
x_7	2	1	1	1	3
x_8	2	2	2	1	3
x_9	1	2	1	2	3
x_{10}	3	1	1	2	3
x_{11}	3	2	2	2	3
x_{12}	3	2	2	2	3

利用第 3 节知识分别计算条件属性 $a_j(j=1,2,3,4)$ 的重要度，给定证据 x 和 y 时条件属性的证据支持度，进而通过上一节的公式计算得到近似条件概率分配 $m_j(d_i/x)$ 和 $m_j(d_i/y)$，$j=1,2,3,4$，$i=1,2,3$，进一步得到证据 x 和 y 的决策结果，如表 2 所列。

表 2 证据决策结果

		{1}	{2}	{3}	{1,3}	{2,3}
方法 1	证据 x	0	0	0	0	0
	证据 y	0	0	0	0	0
方法 2	证据 x	0.0853	0	0.9147	0	0
	证据 y	0	0.0502	0.9498	0	0
方法 3	证据 x	0.0453	0	0.9547	0	0
	证据 y	0	0.0416	0.9584	0	0

在表 2 中，方法 1 是通过将经典信息系统中等价关系直接换为优势关系，经计算得到的证据 x 和 y 的近似条件概率分配值均为零；方法 2 是在证据合成时将证据同等对待，经计算虽然结果与表 1 接近，但与实际意义不相符；方法 3 即为本文第 3 节提出的方法，其结果说明证据 x 和 y 的决策结果更偏重于决策{3}，这与通过实际调查得到的表 1 相符合，并且从另外一个角度提供了一种已知对象的条件属性值来做决策的方法，进一步丰富了理论基础，强化了本文的意义，使该方法的有效性和优越性更明了。

结束语 粗糙集理论能有效地分析和处理不精确、不一致、不完整等各种不完备信息，证据理论是处理不确定决策问题的重要方法。根据两者之间的关系，本文提出了一种序信息系统中基于粗糙集的近似条件概率分配获取方法，解决了对于不同的证据从决策表得到相同的基本可信度分配的问题；并且给出了证据(属性)的近似条件概率分配权重计算方法，解决了组合证据无权重的问题。实例分析表明，与其他方法相比，本文方法更具有效性。

参 考 文 献

[1] Yee L, Wu Wei-zhi, Zhang Wen-xiu. Knowledge acquisition in incomplete information systems: a rough set approach[J]. European Journal of Operational Research, 2006, 168(1): 164-180

[2] Liang Ji-ye. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004, 12(1): 37-46

[3] Greco S, Matarazzo B, Slowinski R. Rough approximation of a preference relation by dominance relation[J]. European Journal of Operation Research, 1999, 117: 63-83

[4] Greco S, Matarazzo B, Slowinski R. A new rough set approach to multicriteria and multiattribute classificatioin[C]// Polkowsik L, Skowron A, eds. Rough Sets and Current Trends in Computing (RSCTC'98). Lecture Notes in Artificial Intelligence, Vol 1424, Berlin: Springer-Verlag, 1998: 60-67

[5] Greco S, Matarazzo B, Slowinski R. A new rough sets approach to evaluation of bankruptcy risk[C]// Zopounidis X, ed. Operational Tools in the Management of Financial Risks. Dordrecht: Kluwer, 1999: 121-136

[6] Shao Ming-wen, Zhang Wen-xiu. Dominance relation and rules in an incomplete ordered information system [J]. International Journal of Intelligent Systems, 2005, 20: 13-27

[7] Xu Wei-hua, Zhang Wen-xiu. Measuring roughness of generalized rough sets induced by a covering[J]. Fuzzy Sets and Systems, 2007, 158: 2443-2455

[8] 徐伟华. 序信息系统与粗糙集[M]. 北京：科学出版社，2013
Xu Wei-hua. Ordered Informaton System and Rough Set[M]. Beijing: Science Press, 2013

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

[10] 张文修，梁怡，吴伟志. 信息系统与知识发现[M]. 北京：科学出版社，2003
Zhang Wen-xiu, Liang Yi, Wu Wei-zhi. Information System and Knowledge Discovery[M]. Beijing: Science Press, 2003

(上接第 49 页)

[8] Ganter B, Wille R. Application of Combinatorics and Graph Theory to the Biological and Social Sciences[M]. Roberts F ed. New York: Springer, 1983: 139-167

[9] Ganter B, Wille R. Formal Concept Analysis [M]. Mathematical Foundations[M]// New York: Springer-Verlag, 1999

[10] 魏玲. 粗糙集与概念格约简理论与方法[D]. 西安：西安交通大学，2005
Wei Ling. Reduction Theory and Approach to Rough Set and Concept Lattice [D]. Xi'an: Xi'an Jiaotong University, 2005

[11] 王霞. 基于不可约元的概念格的概念约简[D]. 西安：西安交通大学，2008
Wang Xia. Attribute reduction in concept lattice based on irreducible elements [D]. Xi'an: Xi'an Jiaotong University, 2008

[12] 徐伟华. 序信息系统与粗糙集[M]. 北京：科学出版社，2013
Xu Wei-hua. Ordered Information Systems and Rough Set [M]. Beijing: Science Press, 2013