软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2119−2135 [doi: 10.13328/j.cnki.jos.004640]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

大数据环境下多决策表的区间值全局近似约简*

徐菲菲[1], 雷景生[1], 毕忠勤[1], 苗夺谦[2], 杜海舟[1]

[1](上海电力学院 计算机科学与技术学院,上海 200090)

[2](同济大学 电子与信息工程学院,上海 200092)

通讯作者: 徐菲菲, E-mail: xufeifei@shiep.edu.cn

摘 要: 在电力大数据中,很多具体的应用如负荷预测、故障诊断都需要依据一段时间内的数据变化来判断所属类别,对某一条数据进行类别判定是毫无意义的.基于此,将区间值粗糙集引入到大数据分类问题中,分别从代数观和信息观提出了基于属性依赖度和基于互信息的区间值启发式约简相关定义和性质证明,并给出相应算法,丰富和发展了区间值粗糙集理论,同时为大数据的分析研究提供了思路.针对大数据的分布式存储架构,又提出了多决策表的区间值全局约简概念和性质证明,进一步给出多决策表的区间值全局约简算法.为了使得算法在实际应用中取得更好的效果,将近似约简概念引入所提的 3 种算法中,通过对 2012 上半年某电厂一台 600MW 的机组运行数据进行稳态判定,验证所提算法的有效性.实验结果表明,所提的 3 种算法均能在保持较高分类准确率的条件下从对象和属性个数两方面对数据集进行大幅度缩减,从而为大数据的进一步分析处理提供支撑.

关键词: 大数据;区间值;近似约简;多决策表;全局约简

中图法分类号: TP181

中文引用格式: 徐菲菲,雷景生,毕忠勤,苗夺谦,杜海舟.大数据环境下多决策表的区间值全局近似约简.软件学报,2014,25(9): 2119−2135. http://www.jos.org.cn/1000-9825/4640.htm

英文引用格式: Xu FF, Lei JS, Bi ZQ, Miao DQ, Du HZ. Approaches to approximate reduction with interval-valued multi-decision tables in big data. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2119−2135 (in Chinese). http://www.jos.org.cn/1000-9825/4640.htm

Approaches to Approximate Reduction with Interval-Valued Multi-Decision Tables in Big Data

XU Fei-Fei[1], LEI Jing-Sheng[1], BI Zhong-Qin[1], MIAO Duo-Qian[2], DU Hai-Zhou[1]

[1](College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)

[2](College of Electronic and Information Engineering, Tongji University, Shanghai 200092, China)

Corresponding author: XU Fei-Fei, E-mail: xufeifei@shiep.edu.cn

Abstract: For the big data on electric power, many specific applications, such as load forecasting and fault diagnosis, need to consider data changes during a period of time to determine their decision classes, as deriving a class label of only one data record is meaningless. Based on the above discussion, interval-valued rough set is introduced into big data classification. Employing algebra and information theory, this paper defines the related concepts and proves the properties for interval-valued reductions based on dependency and mutual information, and presents the corresponding heuristic reduction algorithms. The proposed methods can not only enrich and develop the interval-valued rough set theory, but also provide a new way for the analysis of big data. Pertaining to the distributed data storage architecture of big data, this paper further proposes the interval-valued global reduction in multi-decision tables with proofs of its properties. The corresponding algorithm is also given. In order for the algorithms to achieve better results in practical applications, approximate reduction is introduced. To evaluate three proposed algorithms, it uses six months' operating data of one 600MW unit in some power plant. Experimental results show that the three algorithms

* 基金项目: 国家自然科学基金(61272437, 60305094); 上海市教育委员会科研创新项目(12YZ140, 14YZ131); 上海市自然科学基金(13ZR1417500)

收稿时间: 2014-03-31; 定稿时间: 2014-05-14

proposed in this article can maintain high classification accuracy with the proper parameters, and the numbers of objects and attributes can both be greatly reduced.

Key words: big data; interval-value; approximate reduction; multi-decision tables; global reduction

随着云计算、物联网、移动互联网等新兴信息技术的发展,将人类带进了大数据时代,无处不在的大数据成为了各界关注的焦点[1−9].有调查指出,如今大规模的企业系统包括由分布在不同位置的上千台服务器所构成的完整数据中心[10].如何从分布式存储的大数据中快速、准确地挖掘其潜在的价值,将大数据转化为经济价值的来源,日益成为企业超越竞争对手的有力武器.

分布式存储的大数据呈现出许多鲜明的特征:数据体量巨大,数据种类繁多,流动速度快,价值密度低,这些对大数据的处理能力和效率提出了更高的需求.与以往的数据分析不同,对大数据的分析处理不再一味热衷于追求精确度和寻找因果关系[11].面对海量的即时数据,适当忽略微观层面上的精确度可以在宏观层面拥有更好的洞察力.同样,在大数据时代,寻求事物之间的相关关系而无须紧盯事物之间的因果关系,可以提供非常新颖且有价值的观点.

在很多实际大数据环境中,均存在着大量的不确定性因素,采集到的数据往往包含着噪声、不精确甚至不完整.粗糙集理论[12]是继概率论、模糊集、证据理论之后又一个处理不确定性的强有力的数学工具.作为一种软计算方法,其有效性已在各应用领域中得到证实,是人工智能理论及其应用领域中的研究热点之一[13−27].粗糙集与概率论、模糊集、证据理论有很多相同的特征,但相比于后三者,粗糙集无需任何的先验知识,只通过数据本身就可以获得知识,而概率论、模糊集和证据理论分别需要概率、隶属度和概率赋值等信息.

粗糙集研究中的核心问题之一是属性约简,通过属性约简,可以求得决策表的最小表达,即保持知识表达系统中分类能力不变的情况下,删除其中不相关或不重要的属性,这也是知识获取的关键.但已有证明,求解所有约简和求解最小约简都是 NP-hard 问题.目前提出的属性约简算法大都基于启发式的,且都是针对集中式单决策表(即一张完整决策表)的情况,并不适用于分布式存储的大数据分析与挖掘.目前,已有学者对粗糙集的属性约简算法在分布式平台下进行研究并实现[28,29].然而,这些算法仅仅是将约简算法本身在分布式平台的实现,仍然处理的是集中式单决策表,并未考虑数据集的分布式存储.对分布式存储的大数据环境下的约简算法研究还不多见.对大数据的条件属性进行约简,可以选取保持决策分类不变的最小条件属性子集,极大地减少大数据分析的工作量.分布式存储的带标签的大数据,每个站点都可看成是一张决策表,整体的大数据可认为是由多张决策表构成的,并且这些决策表的条件属性互不相同,但决策属性为同一个.因此,对分布式存储的大数据进行约简算法研究,可转化为求多决策表的约简方法研究.文献[30]针对分布式多决策表的近似约简进行了相关研究.文献[31]在前文基础上考虑到在某些应用场景中,各站点希望自己持有的本地决策表原始数据和敏感信息不被其他站点获取,加入隐私保护策略,设计了多决策表的隐私保护属性约简算法.由此可见,对多决策表(分布式存储的大数据)的研究离不开具体的应用.

随着智能电网建设的推进,电力大数据格局逐步形成.目前,获得电力运行大数据的主要形式来源于分散在各地不同的系统数据库,所获得的数据类型也以连续值属性为主.与传统的分类方法不同,对大数据的分类研究不再单独考虑某一条数据,而是以数据块的形式作为一个研究对象.这是因为仅仅依靠某一条数据来判断它的类别信息已意义不大,而是应该考虑某个时间段内的数据特征,从而判断该数据段所属的类别.例如,基于电力大数据对负荷进行预测,单条数据不具备负荷预测的特质,而是应该将待预测的数据段与某时间段的数据进行相似性比较,从而确定负荷预测值.因此,对大数据的分类研究应从数据块开始.为了快速有效地对电力大数据建立分类模型,将数值型条件属性的数据块近似表示成区间值形式,即通过该数据块的最大最小值对数据块进行近似描述(对非数值型的条件属性可转化为数值型处理),从而研究区间值的属性约简策略,建立分类模型.已有学者对区间值条件属性约简方法进行了研究[32−35],但这些方法均是针对一个集中数据集,并未考虑多决策表的情况,因此不适用于分布式存储的大数据环境.

本文将分布式存储的大数据看成是由多张决策属性相同、条件属性不同的决策表组成,在此基础上,将大数据进行分块使其区间化,研究多决策表的区间值全局近似约简方法.本文所做工作的意义在于:

1) 针对大数据的数据体量巨大、噪声多的特点,将粗糙集方法引入至大数据分析中,通过属性约简方法减少大数据分析所涉及的数据量;
2) 针对电力大数据以连续值属性为主,并且对大数据的分类研究实际应以数据块作为对象单位,提出将数据块近似描述为区间值形式,从而讨论了区间值决策表的启发式约简方法;给出基于依赖度的区间值属性约简相关概念和性质证明,并提出相应算法;给出基于互信息的区间值属性约简相关概念和性质证明,提出相应算法;为了增强算法实用性,提出区间值决策表的近似约简概念和方法;
3) 针对大数据的分布式存储,给出条件属性不同、决策属性相同的多决策表下的全局近似约简相关概念和性质证明,并提出相应的约简算法,从而对分布式存储的大数据求得满足分类结果近似不变的全局约简;
4) 将所设计的 3 种算法在电力大数据真实数据集中进行测试,并对结果进行分析和讨论;实验结果表明:3 种算法在合适的区间长度时,选取的属性子集均能保持较高的分类准确率;随着属性个数的增加,基于依赖度的区间值约简方法比基于互信息的区间值约简方法运行时间略长,多决策表下的全局约简运行时间最短.

本文第 1 节对多决策表以及区间值决策表的相关概念和性质进行介绍.第 2 节分别给出基于依赖度和基于互信息的区间值属性约简的相关定义和性质证明,并提出相应的算法;同时,将近似约简引入到上述方法中,增强算法的实用性.第 3 节给出多决策表下的区间值全局近似约简概念和性质证明,提出相应的算法.第 4 节将以上算法在电力大数据中进行实验、比较和分析,实验结果验证了算法的有效性.第 5 节对全文进行总结,并对未来的工作进行展望.

1 相关基本概念

本节主要介绍分布式环境中多决策表以及区间值决策表的相关概念和性质.

1.1 多决策表的相关概念和性质

设有 m 个站点 $S_1,S_2,\dots,S_m$,相应的局部决策表 DT_i(或成员决策表)的属性集分别为 $C_1\cup D,C_2\cup D,\dots,C_m\cup D$, $\bigcap_{i=1}^{m}C_i=\varnothing$, 各局部决策表具有相同的对象集 U 且均隐含一个对象标识属性.通过该属性,可将各局部决策表连接成一个单决策表 $DT=\langle U,C\cup D,V,f\rangle$, $C=\bigcup_{i=1}^{m}C_i$, 并假设唯一的决策属性 D 的取值范围是 $1,2,\dots,l$.由 D 导出的决策类构成 U 的一个划分 $\{\psi_1,\psi_2,\dots,\psi_l\}$.其中:$\psi_i=\{u\in U:f(u,D)=i\}$,$i=1,2,\dots,l$;$U$ 中的对象个数为 n.

定义 1.1[31]. 全局决策表 DT 是四元组$\langle U,C\cup D,V,f\rangle$.其中:$U$ 是一组对象的非空有限集合,称为论域;设有 n 个对象,则 U 可表示为 $U=\{u_1,u_2,\dots,u_n\}$;C 为条件属性集,D 为决策属性集;$V=\bigcup_{a\in(C\cup D)}V_a$, V_a 为属性 a 的值域集;f 是 $U\times(C\cup D)\to V$ 的映射.

定义 1.2[31]. 在站点 $S_i(i=1,2,\dots,m)$,局部决策表 DT_i 是四元组 $DT=\langle U,C_i\cup D,V,f\rangle$.其中:$C_i$ 为条件属性集,D 为决策属性集,$V=\bigcup_{a\in(C_i\cup D)}V_a$, V_a 为属性 a 的值域集,f 是 $U\times(C_i\cup D)\to V$ 的映射.

由于在大数据的复杂环境中,要求得全局决策表的精确约简所花费的代价较高,对大数据的分析应更多地考虑时间因素,因此定义ε-近似约简如下(由于基于信息熵的定义方法比代数观下的更加直观,本文所涉及的研究主要基于信息论观点):

定义 1.3. 对于给定的全局决策表 DT 和$\varepsilon(\varepsilon\geqslant 0)$,若$|H(D|C)-H(D|A)|\leqslant\varepsilon(A\subseteq C)$,且$|H(D|C)-H(D|B)|>\varepsilon(\forall B\subset A)$,则 A 为决策表的一个ε-近似约简.其中,$H(P|Q)$表示为条件信息熵,且 $P,Q\subseteq C\cup D$.

上述定义中,如果条件属性集合 C 的值域为有限离散集合,则 $H(P|Q)$可依据等价类的分布情况来计算.而在大数据环境中,条件属性集合 C 往往都是连续的,可选用 Pazon 窗方法或文献[25]采用的模糊粗糙集方法计算连续值的条件熵.对大数据构建粗糙集分类模型的首要任务就是求得全局的ε-近似约简.

定义 1.4. 设 X 为论域 U 的一个子集,即 $X\subseteq U,P\subseteq C,X$ 关于 P 的全局下近似为 $\underline{P}X(C)=\{u\in U:[u]_p\subseteq X\}$,其中:

$$[u]_p=\{x\in U|\forall a\in P,f(u,a)=f(x,a)\}.$$

性质 1.1. 若 $A\subseteq C,B\subseteq C$,且 $A\subseteq B$,则 $H(D|A)\geqslant H(D|B)$.

1.2 区间值决策表的相关概念和性质

目前对区间值信息系统的研究大多都基于无分类标签的信息系统[34−36],也有学者对决策属性为区间值的决策系统进行了探讨.本文基于电力大数据的特点,讨论条件属性为区间值,而决策属性为类别标签的情况.

定义 1.5. 设区间值决策表 $DT=\langle U,C\cup D,V,f\rangle$,非空有限属性集 $C\cup D$ 包括条件属性集 $C=\{a_1,a_2,\ldots,a_h\}$和决策属性集 $D=\{d\}$两部分;$V=V_C\cup V_D$,其中,V_C 为条件属性值集合,V_D 为决策属性值集合;$f:U\times C\to V_C$ 为区间值映射,$f:U\times D\to V_D$ 为单值映射.

表 1 为一个区间值决策表[33],其中:论域 $U=\{u_1,u_2,\ldots,u_{10}\}$,条件属性集 $C=\{a_1,a_2,a_3,a_4,a_5\}$,决策属性集 $D=\{d\}$;条件属性值 $f(a_k,u_i)=[l_i^k,u_i^k]$ 是区间值,如 $f(a_2,u_3)$=[7.03,8.94];决策属性值 $d(u_i)$是单值,如 $d(u_3)$=2.

Table 1 An interval-valued decision table

表 1 区间值决策表

U	a_1	a_2	a_3	a_4	a_5	d
u_1	[2.17,2.96]	[5.32,7.23]	[3.35,5.59]	[3.21,4.37]	[2.46,3.59]	1
u_2	[3.38,4.50]	[3.38,5.29]	[1.48,3.58]	[2.36,3.52]	[1.29,2.42]	2
u_3	[2.09,2.89]	[7.03,8.94]	[3.47,5.69]	[3.31,4.46]	[3.48,4.61]	2
u_4	[3.39,4.51]	[3.21,5.12]	[0.68,1.77]	[1.10,2.26]	[0.51,1.67]	3
u_5	[3.70,4.82]	[2.98,4.89]	[1.12,3.21]	[2.07,3.23]	[0.97,2.10]	2
u_6	[4.53,5.63]	[5.51,7.42]	[3.50,5.74]	[3.27,4.43]	[2.49,3.62]	2
u_7	[2.03,2.84]	[5.72,7.65]	[3.68,5.91]	[3.47,4.61]	[2.53,3.71]	1
u_8	[3.06,4.18]	[3.11,5.02]	[1.26,3.36]	[2.25,3.41]	[1.13,2.25]	3
u_9	[3.38,4.50]	[3.27,5.18]	[1.30,3.40]	[4.21,5.36]	[1.11,2.23]	1
u_{10}	[1.11,2.26]	[2.51,3.61]	[0.76,1.85]	[1.30,2.46]	[0.42,1.57]	4

经典粗糙集采用等价关系对论域进行划分,然而区间值决策表中,相同区间值形成的等价类很难对论域形成合理的划分.因此,引入相似率来表示 2 个区间值的相似程度,为论域的分类提供度量标准.

定义 1.6. 设区间值决策表 $DT=\langle U,C\cup D,V,f\rangle,a_k\in C$, $f(a_k,u_i)=[l_i^k,u_i^k]$,其中,$l_i^k\leqslant u_i^k$.当 $l_i^k=u_i^k$ 时,表示对象 u_i 在属性 a_k 上的取值为常数.若对任意的 u_i 和任意的条件属性 a_k, $l_i^k=u_i^k$,则该决策表为传统的决策表.定义对象 u_i 与 u_j 关于属性 a_k 的相似度[35]为

$$r_{ij}^k=\begin{cases}0, & [l_i^k,u_i^k]\cap[l_j^k,u_j^k]=\varnothing\\ \dfrac{card([l_i^k,u_i^k]\cap[l_j^k,u_j^k])}{card(\max\{u_i^k,u_j^k\}-\min\{l_i^k,l_j^k\})}, & [l_i^k,u_i^k]\cap[l_j^k,u_j^k]\neq\varnothing\end{cases},$$

其中,$card(\cdot)$表示区间值的长度.显然,$0\leqslant r_{ij}^k\leqslant 1$.如果 $r_{ij}^k=0$,则条件属性值 $f(a_k,u_i)$与 $f(a_k,u_j)$相离;若 $0<r_{ij}^k<1$,则条件属性值 $f(a_k,u_i)$与 $f(a_k,u_j)$部分相离或真包含;若 $r_{ij}^k=1$,则条件属性值 $f(a_k,u_i)$与 $f(a_k,u_j)$是完全不可分辨的.

条件属性值相似度描述了区间值环境下不同对象之间的等价程度.

定义 1.7[35]. 设 $DT=\langle U,C\cup D,V,f\rangle$是一区间值决策表,给定阈值水平$\lambda\in[0,1]$和任意属性子集 $A\subseteq C$,定义 U 上的二元关系 $R_A^\lambda:R_A^\lambda=\{(x_i,x_j)\in U\times U:r_{ij}^k>\lambda,\forall a_k\in A\}$ 称为关于 A 的λ-容差关系.

性质 1.2. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,给定阈值水平$\lambda\in[0,1]$和任意属性子集 $A\subseteq C$,显然, R_A^λ 是自反的和对称的,但不一定是传递的.

性质 1.3. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,任意属性子集 $A\subseteq C$,有 $R_A^\lambda=\bigcap_{a_k\in A}R_{\{a_k\}}^\lambda$.

记 $R_A^\lambda(u_i)$ 表示区间值对象 u_i 在属性集 A 下的λ-相容类,以表 1 为例,当λ=0.7,$A=a_1$ 时,根据定义 1.6 和定义 1.7 计算可得:

$$R^{0.7}_{\{a_1\}}(u_1)=\{u_1,u_3,u_7\}$$
$$R^{0.7}_{\{a_1\}}(u_2)=\{u_2,u_4,u_9\}$$
$$R^{0.7}_{\{a_1\}}(u_3)=\{u_1,u_3,u_7\}$$
$$R^{0.7}_{\{a_1\}}(u_4)=\{u_2,u_4,u_9\}$$
$$R^{0.7}_{\{a_1\}}(u_5)=\{u_5\}$$
$$R^{0.7}_{\{a_1\}}(u_6)=\{u_6\}$$
$$R^{0.7}_{\{a_1\}}(u_7)=\{u_1,u_3,u_7\}$$
$$R^{0.7}_{\{a_1\}}(u_8)=\{u_8\}$$
$$R^{0.7}_{\{a_1\}}(u_9)=\{u_2,u_4,u_9\}$$
$$R^{0.7}_{\{a_1\}}(u_{10})=\{u_{10}\}$$

由于λ-容差关系满足自反和对称但不满足传递性,在计算λ-相容类时只需考虑当前对象之后的记录,对之前的对象可通过对称关系获取,在大数据环境下可极大地节省计算λ-相容类的时间.如果 A 由多个属性组成,可根据性质 1.3,先分别计算区间值对象在每个属性下的λ-相容类(满足λ-容差关系的对象集合),再通过交运算得到多属性的λ-相容类.

定义 1.8. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,任意属性子集 $A\subseteq C,X\subseteq U$,定义 X 关于 A 的粗糙上、下近似为

$$\overline{R}^{\lambda}_{A}(X)=\{u_i\in U,R^{\lambda}_{A}(u_i)\cap X\neq\varnothing\},$$
$$\underline{R}^{\lambda}_{A}(X)=\{u_i\in U,R^{\lambda}_{A}(u_i)\subseteq X\}.$$

以上定义和性质实际并未涉及到决策属性,仅仅是将无标签的区间值信息系统的概念简单地移植到区间值决策表中.

2 区间值决策表的启发式约简

文献[33]提出了一种基于区分函数的区间值决策表约简算法,然而该算法的计算复杂度较高,很难用于处理大数据.本节针对大数据分析中无须过度追求精确度的特点,分别从代数观和信息观给出了区间值决策表的启发式约简概念和性质证明,并提出相应算法.同时,为了增强算法的实用性,将近似约简概念引入,并提出相应方法.

2.1 代数观下区间值决策表约简的相关概念和性质

根据定义 1.8,我们可以定义决策属性关于区间值条件属性子集的上、下近似为:

定义 2.1. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,由 D 导出的决策类构成 U 的一个划分$\{\psi_1,\psi_2,\ldots,\psi_l\}$.任意条件属性子集 $A\subseteq C$,定义决策属性 D 关于 A 的上、下近似为

$$\overline{R}^{\lambda}_{A}(D)=\bigcup_{i=1}^{l}\overline{R}^{\lambda}_{A}(\psi_i),$$
$$\underline{R}^{\lambda}_{A}(D)=\bigcup_{i=1}^{l}\underline{R}^{\lambda}_{A}(\psi_i),$$

其中,$\overline{R}^{\lambda}_{A}(X)=\{u_i\in U,R^{\lambda}_{A}(u_i)\cap X\neq\varnothing\}$,$\underline{R}^{\lambda}_{A}(X)=\{u_i\in U,R^{\lambda}_{A}(u_i)\subseteq X\}$,$R^{\lambda}_{A}(u_i)$ 表示区间值对象 u_i 在属性集 A 下的λ-相容类.

决策属性 D 的下近似也称为正域,记为 $POS^{\lambda}_{A}(D)$.正域的大小反映的是分类问题在给定属性空间中的可分离程度.正域越大,表明各相容类的重叠区域越少.为了度量属性的重要度,定义决策属性 D 相对于区间值条件属性 A 的λ-依赖度为

$$\gamma_A^\lambda(D)=\frac{|\underline{R}_A^\lambda(D)|}{|U|},$$

其中,|·|表示集合的基.$0\leqslant\gamma_A^\lambda(D)\leqslant 1$表示了区间值对象集合中根据条件属性 A 的描述,那些能够被某一类决策完全包含的对象所占全体对象的比率.显然,正域越大,决策属性 D 对条件属性 A 的依赖性越强.

性质 2.1. 给定区间值决策表 $DT=\langle U,C\cup D,V,f\rangle$和$\lambda$,如果 $B\subseteq A\subseteq C$ 且 $u_i\in POS_B^\lambda(D)$,则 $u_i\in POS_A^\lambda(D)$ 成立.

证明:假设 $u_i\in\underline{R}_B^\lambda(D_j)$,其中,$D_j$ 表示决策类别为 j 的对象集合,即 $R_B^\lambda(u_i)\subseteq D_j$.由于 $B\subseteq A\subseteq C$, $R_A^\lambda(u_i)\subseteq R_B^\lambda(u_i)$, 因此, $R_A^\lambda(u_i)\subseteq R_B^\lambda(u_i)\subseteq D_j$.从而有 $u_i\in POS_A^\lambda(D)$. □

性质 2.2. $\gamma_A^\lambda(D)$ 是单调的.如果 $A_1\subseteq A_2\subseteq\ldots\subseteq C$,则 $\gamma_{A_1}^\lambda(D)\leqslant\gamma_{A_2}^\lambda(D)\leqslant\ldots\leqslant\gamma_C^\lambda(D)$.

证明:根据性质 2.1 可知: $\forall u_i\in POS_{A_1}^\lambda(D)$,我们有 $u_i\in POS_{A_2}^\lambda(D),\ldots,u_i\in POS_C^\lambda(D)$. 可能存在 $u_j\notin POS_{A_1}^\lambda(D)$,但 $u_j\in POS_{A_2}^\lambda(D),\ldots,u_j\in POS_C^\lambda(D)$, 因此有 $|POS_{A_1}^\lambda(D)|\leqslant|POS_{A_2}^\lambda(D)|\leqslant\ldots\leqslant|POS_C^\lambda(D)|$. 由于 $\gamma_A^\lambda(D)=\frac{|POS_A^\lambda(D)|}{|U|}$, 所以有 $\gamma_{A_1}^\lambda(D)\leqslant\gamma_{A_2}^\lambda(D)\leqslant\ldots\leqslant\gamma_C^\lambda(D)$. □

定义 2.2. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$A\subseteq C$,$\forall a_k\in A$,如果 $\gamma_{A-\{a_k\}}^\lambda(D)<\gamma_A^\lambda(D)$, 称属性 a_k 相对于属性集 A 是必要的;否则,如果 $\gamma_{A-\{a_k\}}^\lambda(D)=\gamma_A^\lambda(D)$, 称属性 a_k 相对于属性集 A 是多余的.如果$\forall a_k\in A$ 都是必要的,称属性集 A 是独立的.

如果 $\gamma_{A-\{a_k\}}^\lambda(D)=\gamma_A^\lambda(D)$,表明从决策表中去掉属性 a_k,决策表的正域不会发生改变,即各类的可区分性不变.也就是说,属性 a_k 没有给分类带来任何的贡献.因此,a_k 是多余的.相反地,如果删除 a_k,决策表的决策正域变小了,则表明各类的可区分性变差了.此时,a_k 不能被删除.

定义 2.3. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$A\subseteq C$,称属性子集 A 是条件属性集 C 的一个λ-约简,如果 A 满足:

(1) $\gamma_A^\lambda(D)=\gamma_C^\lambda(D)$;

(2) $\forall a_k\in A,\gamma_{A-\{a_k\}}^\lambda(D)<\gamma_A^\lambda(D)$.

该定义的条件(1)要求λ-约简不能降低决策表的区分能力,λ-约简应该与决策表中全部条件属性具有相同的分辨能力;条件(2)要求在一个λ-约简中不存在多余的属性,所有的属性都应该是必要的.这一定义与经典粗糙集模型中的定义在形式上是完全一致的.然而,该模型定义了区间值空间中的λ-约简,而经典粗糙集是定义在离散空间中的.

定义 2.4. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$A_1,A_2,\ldots,A_s$ 是该决策表的所有λ-约简,则定义 $Core=\bigcap_{i=1}^{s}A_i$ 为决策表的核.

2.2 基于依赖度的区间值决策表λ-约简算法

如果要找出区间值决策表的全部λ-约简,需要计算 2^h-1 个属性子集,判断它们是否满足λ-约简的条件.其中,h 是条件属性的个数.这对于拥有上百个,甚至上千个属性的大数据而言,计算量是不可容忍的.本文将基于依赖度的概念构造启发式约简算法,极大地降低算法复杂度.由于依赖度描述了条件属性对分类的贡献,因此可以作为属性重要度的评价标准.

定义 2.5. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$A\subseteq C$,$a_k\in C-A$,定义 a_k 相对于 C 的重要度为

$$SIG(a_k,A,D)=\gamma_{A\cup\{a_k\}}^\lambda(D)-\gamma_A^\lambda(D).$$

有了属性重要度的定义,我们可以构造区间值λ-约简的贪心算法.该算法以空集为起点,每次计算全部剩余属性的属性重要度,从中选取属性重要度值最大的属性加入到λ-约简集合中,直到所有剩余属性的重要度为 0,即加入任何新的属性,依赖度不再发生变化为止.前向搜索算法能够保证重要的属性先被加入到λ-约简中,从而不损失重要的特征.后向搜索算法难以保证这个结果,因为对于有大量冗余属性的区间值决策表而言,即使那些

重要的属性被删除也不一定会降低整个决策表的区分能力.因此,最终可能保留了大量区分能力很弱、但作为一个整体依然能够保持原始数据的分辨能力的特征,而不是少量区分能力很强的特征.基于依赖度的区间值决策表的λ-约简算法描述见算法 1.

算法 1. 基于依赖度的区间值决策表λ-约简(λ-reduction in interval-valued decision table based on dependence,简称 RIvD).

输入:$DT=\langle U,C\cup D,V,f\rangle,\lambda$;

输出:λ-约简 red.

Step 1. 令 $red=\varnothing$;

Step 2. 对所有属性 $a\in C$,计算属性 a 下的λ-相容类 $R^{\lambda}_{\{a\}}$;

Step 3. 对任意的 $a_k\in C-red$,计算 $SIG(a_k,red,D)=\gamma^{\lambda}_{red\cup\{a_k\}}(D)-\gamma^{\lambda}_{red}(D)$; //定义 $\gamma^{\lambda}_{\varnothing}(D)=0$

Step 4. 选择 a_i,满足: $SIG(a_i,red,D)=\max\limits_k(SIG(a_k,red,D))$;

Step 5. 如果 $SIG(a_i,red,D)>0$,$red=red\cup\{a_i\}$,转至 Step 3;

否则,返回 red,结束.

设条件属性 C 的个数为 h,区间值对象个数为 n,则该算法的时间复杂度为 $O(n^2+hn)$.

以上为代数观点下的区间值λ-约简算法.在传统粗糙集中,对于一致决策表的启发式算法,已经证明代数观点与信息论观点等同.然而对于不一致决策表而言,信息论观点下对象的划分依然可以改变知识的条件信息熵,即基于条件信息熵的属性约简与影响不一致对象划分的粒度有一定的关系.主要体现在基于条件信息熵的属性约简可以增加一些属性,而这些属性影响了不一致对象划分的粒度.因此,粗糙集的信息论观点包含了其代数观点,为决策表的知识获取和规则提取提供了更加有效的途径.因此,非常有必要对基于条件信息熵的区间值属性约简作进一步研究.

2.3 信息观下区间值决策表约简的相关概念和性质

由于在区间值决策表中,λ-容差关系取代了等价关系,不再构成论域的划分而是覆盖,因此,我们先定义区间值决策表的λ-知识粗糙熵,进而定义λ-信息熵及λ-条件信息熵等概念.知识粗糙熵表征了知识整体的统计特征,是总体的平均不确定性的量度;信息熵也是度量信息的平均不确定性的量度,与知识粗糙熵的和为 $\log_2|U|$;条件信息熵表示如果已经完全知道某变量(集)的前提下,另一变量(集)的信息熵还有多少.为了计算条件信息熵,需要用到联合信息熵的概念.

定义 2.6. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\dots,u_n\}$.任意属性子集 $A\subseteq C$,则区间值决策表的λ-知识粗糙熵定义为

$$H_{Rough}(R^{\lambda}_A)=\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2 f^{\lambda}_A(u_i),$$

其中, $f^{\lambda}_A(u_i)$ 表示 u_i 在所有 $u_j(1\leqslant j\leqslant|U|)$的$\lambda$-相容类中出现的次数.

性质 2.3. 若 R 是基于知识 A 的等价关系,则有 $H_{Rough}(R^{\lambda}_A)=H_{Rough}(A)$.

证明:如果 R 是基于知识 A 的等价关系,则对象 u_i 所在的λ-相容类就是等价类.设属性集 A 将论域划分为 k 个不同的等价类$\{X_1,X_2,\dots,X_k\}$,则有:

$$H_{Rough}(R^{\lambda}_A)=\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2 f^{\lambda}_A(u_i)=\frac{1}{|U|}\sum_{j=1}^{k}|R(u_j)|\times\log_2|R(u_j)|=\sum_{j=1}^{k}\frac{|R(u_j)|}{|U|}\times\log_2|R(u_j)|=H_{Rough}(A).$$

知识粗糙熵与信息熵的和为论域的信息量 $\log_2|U|$,所以等价关系下知识粗糙熵为

$$\log_2|U|+\sum_{i=1}^{k}\frac{|R(u_i)|}{|U|}\log_2\frac{|R(u_i)|}{|U|}=\sum_{i=1}^{k}\frac{|R(u_i)|}{|U|}\times\log_2|R(u_i)|.$$ □

性质 2.4. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\dots,u_n\}$.$B\subseteq A\subseteq C$,则有:

$$H_{Rough}(R^{\lambda}_A)\leqslant H_{Rough}(R^{\lambda}_B).$$

性质 2.4 可由定义 2.6 直接推理得到.性质 2.4 说明,区间值决策表的λ-知识粗糙熵随着知识分辨能力的增强而单调下降.

有了上述对区间值决策表λ-知识粗糙熵的定义,根据知识粗糙熵与信息熵之和为 $\log_2|U|$,我们可以定义区间值决策表的λ-信息熵为:

定义 2.7. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\ldots,u_n\}$.任意属性子集 $A\subseteq C$,则区间值决策表的λ-信息熵定义为

$$H(R_A^\lambda)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{f_A^\lambda(u_i)}{|U|}.$$

性质 2.5. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\ldots,u_n\}$.$B\subseteq A\subseteq C$,则有 $H(R_A^\lambda)\geqslant H(R_B^\lambda)$.

证明:如果 $B\subseteq A\subseteq C$,则有 $R_A^\lambda\subseteq R_B^\lambda$,则存在 $u_i\in U$,使得 $f_B^\lambda(u_i)\leqslant f_A^\lambda(u_i)$.根据定义 2.7,则有 $H(R_A^\lambda)\geqslant H(R_B^\lambda)$.

证毕. □

性质 2.5 说明:λ-相容类形成对论域的覆盖块越小,知识所包含的信息量就越大.

定义 2.8. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\ldots,u_n\}$,$P,Q\subseteq C\cup D$,则 P,Q 的λ-联合信息熵可表示为

$$H(R_P^\lambda\cup R_Q^\lambda)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{f_{P\cup Q}^\lambda(u_i)}{|U|},$$

其中, $f_{P\cup Q}^\lambda(u_i)$ 表示区间值对象 u_i 在属性集 $P\cup Q$ 下的 $u_j(1\leqslant j\leqslant|U|)$$\lambda$-相容类中出现的次数.

定义 2.9. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\ldots,u_n\}$,$P,Q\subseteq C\cup D$,且 $P\neq Q$,则知识(属性集合)Q 相对于知识(属性集合)P 的λ-条件信息熵的定义为

$$H(R_Q^\lambda\mid R_P^\lambda)=\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{f_P^\lambda(u_i)}{f_{P\cup Q}^\lambda(u_i)}.$$

定理 2.1. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$U=\{u_1,u_2,\ldots,u_n\}$,$A\subseteq C$,$a_k\in A$,属性 a_k 是不必要的,其充分必要条件是 $H(D\mid R_A^\lambda)=H(D\mid R_{A-\{a_k\}}^\lambda)$.

证明:

- 必要条件

假设存在 $a_k\in A$ 是不必要的,对于任意 $u_i\in U$,则有 $R_A^\lambda(u_i)=R_{A-\{a_k\}}^\lambda(u_i)$,易得 $H(D\mid R_A^\lambda)=H(D\mid R_{A-\{a_k\}}^\lambda)$.

- 充分条件

假设存在 $a_k\in A$ 满足 $H(D\mid R_A^\lambda)=H(D\mid R_{A-\{a_k\}}^\lambda)$. 如果对于任意的 $a_k\in A$ 都是必要的,即存在 $u_i\in U$,使得不等式 $R_A^\lambda(u_i)\neq R_{A-\{a_k\}}^\lambda(u_i)$ 成立.又由于 $A-\{a_k\}\subset A$,有 $H(D\mid R_A^\lambda)<H(D\mid R_{A-\{a_k\}}^\lambda)$, 这与假设 $H(D\mid R_A^\lambda)=H(D\mid R_{A-\{a_k\}}^\lambda)$ 相矛盾.由此可知:对于任意的 $a_k\in A$,当 $H(D\mid R_A^\lambda)=H(D\mid R_{A-\{a_k\}}^\lambda)$ 时,a_k 是不必要的. □

定义 2.10. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$A\subseteq C$,称属性子集 A 是条件属性集 C 的一个λ-约简,如果 A 满足:

(1) $H(D\mid R_A^\lambda)=H(D\mid R_C^\lambda)$;

(2) $\forall a_k\in A, H(D\mid R_A^\lambda)\neq H(D\mid R_{A-\{a_k\}}^\lambda)$.

区间值的λ-条件信息熵描述的是一个属性集对另一属性集的依赖程度.由定理 2.1 可知,λ-条件信息熵可以应用到区间值决策表的λ-约简中.

2.4 基于互信息的区间值λ-约简算法

为了能够进行有效的知识约简,必须要建立一个衡量属性重要性的标准.在传统粗糙集理论的信息观点下,提出在决策表中添加某个属性所引起的互信息的变化大小可以作为该属性重要性的度量.

设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$B\subseteq C$.那么,在 B 中添加一个区间值条件属性 $a\in C-B$ 之后,互信

息的增量为

$$I(B\cup\{a\};D)-I(B;D)=H(D\mid R_B^\lambda)-H(D\mid R_{B\cup\{a\}}^\lambda).$$

这里,$I(x;y)$表示 x 与 y 的互信息.

定义 2.11. 设 $DT=\langle U,C\cup D,V,f\rangle$是区间值决策表,$\lambda\in[0,1]$,$B\subseteq C$.则对于任意区间值条件属性 $a\in C-B$ 的重要性 $SGF(a,B,D)$定义为

$$SGF(a,B,D)=I(B\cup\{a\};D)-I(B;D)=H(D\mid R_B^\lambda)-H(D\mid R_{B\cup\{a\}}^\lambda).$$

若 $B=\varnothing$,则 $SGF(a,B,D)$变为 $SGF(a,D)=H(D)-H(D\mid R_{\{a\}}^\lambda)=I(a;D)$,即为区间值条件属性 a 与决策 D 的互信息.$SGF(a,B,D)$的值越大,说明在已知区间值条件属性子集 B 的条件下,区间值条件属性 a 对决策 D 就越重要.

有了上述理论准备,我们就可以完整地提出基于互信息的区间值λ-约简算法.同样的,我们采用前向贪心算法设计,以空集为起点,依据上述定义的属性重要性,逐次选择最重要的属性添加到约简子集中,直到终止条件满足.

算法 2. 基于互信息的区间值决策表λ-约简(λ-reduction in interval-valued decision table based on mutual information,简称 RIvMI).

输入:$DT=\langle U,C\cup D,V,f\rangle$,$\lambda$;

输出:λ-约简 red.

Step 1. 令 $red=\varnothing$;

Step 2. 对所有属性 $a\in C$,计算属性 a 下的λ-相容类 $R_{\{a\}}^\lambda$;

Step 3. 对任意的 $a_k\in C-red$,计算 $SIG(a_k,red,D)=H(D\mid R_{red}^\lambda)-H(D\mid R_{red\cup\{a_k\}}^\lambda)$;

//当 $red=\varnothing$时,计算 $SGF(a_k,D)=H(D)-H(D\mid R_{\{a_k\}}^\lambda)$

Step 4. 选择 a_i,满足: $SIG(a_i,red,D)=\max_k(SIG(a_k,red,D))$;

Step 5. 如果 $SIG(a_i,red,D)>0$,$red=red\cup\{a_i\}$,转至 Step 3;

否则,返回 red,结束.

算法 2 和算法 1 具有相同的时间复杂度,设条件属性 C 的个数为 h,区间值对象个数为 n,则该算法的时间复杂度为 $O(n^2+hn)$.

为了更好地解决现实生活中的问题,就不能使用过于苛刻的约简条件,所以算法 1 和算法 2 中的约简条件 $SIG(a_i,red,D)>0$ 可改为 $0<SIG(a_i,red,D)<\varepsilon$,$\varepsilon$需要根据具体的数据提前设定.这种改进将在一定程度上使约简的结果更加接近现实生活,更加实用.ε值的大小会直接影响分类的结果,进而影响算法结果的应用.ε值过小,会导致选取的条件属性过多,影响算法的实用性;ε值过大的话会导致选取的条件属性过少而影响算法的精度.

3 多决策表下的区间值λ-全局近似约简

第 2 节中的算法均只能对一个整体数据集进行处理,而大数据均是分布式存储在不同的位置.因此,我们进一步讨论多决策表下的区间值λ-全局约简方法.本节讨论信息论观点下的多决策表区间值λ-全局约简方法,代数观点下的约简方法类似.

3.1 多决策表下的区间值λ-全局约简相关概念和性质

在分布式环境中,网络通信代价是影响多决策表属性约简效率的关键.因而,有效减小网络通信量,是求解分布式环境下多决策表全局约简的关键任务.虽然将各站点的局部决策表传送到一中心站点可简单实现属性约简的求解,但该做法的网络通信量大,尤其在面对大数据环境(规模巨大且含有较高维数)的局部决策表(单个站点)时,需要传送大量的数据.

由定义 2.9 可知,区间值的λ-条件信息熵 $H(D\mid R_A^\lambda)(A\subseteq C)$ 仅与λ-相容类 R_A^λ 及 $R_{A\cup D}^\lambda$ 有关,因而采用有效的λ-相容类存储机制并只传送相应的λ-相容类的策略,可有效地避免传送所有的局部决策表.为此,对于λ-相容类

R_A^λ 中的 $R_A^\lambda(u_i)(1\leqslant i\leqslant n)$,采用如下的三元组存放:

(站点标识,$|R_A^\lambda(u_i)|$,$R_A^\lambda(u_i)$ 中的区间值对象标号 ID 递增序列).

其中,|·|表示区间值对象的个数.

对于不同站点 S_g 和 S_e 上的λ-相容类 $R_A^\lambda(A\subseteq C_g),R_B^\lambda(B\subseteq C_e)$,利用上述存储方式可得如下引理:

引理 3.1. 对于不同站点 S_g 和 S_e 上的λ-相容类 $R_A^\lambda(A\subseteq C_g),R_B^\lambda(B\subseteq C_e)$,有:

$$R_{A\cup B}^\lambda=\{R_A^\lambda(u_i)\cap R_B^\lambda(u_j):R_A^\lambda(u_i)\cap R_B^\lambda(u_j)\neq\varnothing,1\leqslant i\leqslant n,1\leqslant j\leqslant n\}.$$

可见:采用λ-相容类传送方式,求 $R_{A\cup B}^\lambda$ 的网络通信量至多为 $n+\max(|R_A^\lambda|,|R_B^\lambda|)$ (n 为局部决策表的对象数);而采用传送子局部决策表的方法,相应的网络通信量至少为 $\min(|A|,|B|)\times n$,在大数据环境下,所选出的属性子集个数远大于 1,即 $\min(|A|,|B|)>>1$.进一步地,利用如下引理 3.2 和定理 3.1,仅需传送 $R_A^\lambda(A\subseteq C_g)$ 或 $R_B^\lambda(B\subseteq C_e)$ 中的部分λ-相容类,即可求解 $R_{A\cup B}^\lambda$.

引理 3.2. 对于不同站点 S_g 和 S_e 上的λ-相容类 $R_A^\lambda=\{X_1,X_2,...,X_s\}(A\subseteq C_g),R_B^\lambda=\{Y_1,Y_2,...,Y_t\}(B\subseteq C_e)$, $U/D=\{\psi_1,\psi_2,\ldots,\psi_l\}$,若 $X_w\subseteq\psi_k(1\leqslant w\leqslant s,1\leqslant k\leqslant l)$,则 $X_w\cap Y_j\subseteq\psi_k$.

定理 3.1. 对于不同站点 S_g 和 S_e 上的λ-相容类 $R_A^\lambda=\{X_1,X_2,...,X_s\}(A\subseteq C_g),R_B^\lambda=\{Y_1,Y_2,...,Y_t\}(B\subseteq C_e)$, $U/D=\{\psi_1,\psi_2,\ldots,\psi_l\}$,若 $X_w\subseteq\psi_k(1\leqslant w\leqslant s,1\leqslant k\leqslant l)$,则 $\forall Y_j\in R_B^\lambda,X_w\cap Y_j\neq\varnothing$, 有:

$$p(X_w\cap Y_j)\sum_{j=1}^{d}p(X_w\cap Y_j\cap\psi_k)\log_2 p(X_w\cap Y_j\cap\psi_k)=0,$$

其中,d 为与 X_w 相交不为空的λ-相容类 Y_j 的个数.

证明:由于 $X_w\cap Y_j\neq\varnothing$,由引理 3.2 可知 $X_w\cap Y_j\subseteq\psi_k$,则 $p(X_w\cap Y_j\cap\psi_k)=1$ 成立,所以定理 3.1 成立. □

定理 3.2. 对于不同站点 S_g 和 S_e 上的λ-相容类 $R_A^\lambda=\{X_1,X_2,...,X_s\}(A\subseteq C_g),R_B^\lambda=\{Y_1,Y_2,...,Y_t\}(B\subseteq C_e)$, $U/D=\{\psi_1,\psi_2,\ldots,\psi_l\}$,若设 $Y_v(1\leqslant v\leqslant d)\subseteq\psi_{y(v)}(1\leqslant y(v)\leqslant l),X_w(1\leqslant w\leqslant q)\subseteq\psi_{x(w)}(1\leqslant x(w)\leqslant l)$($d$ 的含义同上),则:

$$H(D|R_{A\cup B}^\lambda)=\sum_{i=q+1}^{s}\sum_{j=d+1}^{t}p(X_i\cap Y_j)\sum_{k=1}^{l}p(X_i\cap Y_j\cap\psi_k)\log_2 p(X_i\cap Y_j\cap\psi_k).$$

定理 3.2 可由定理 3.1 和定义 2.9 得证.

由定理 3.2 可知:对于 $R_A^\lambda=\{X_1,X_2,...,X_s\}(A\subseteq C_g),R_B^\lambda=\{Y_1,Y_2,...,Y_t\}(B\subseteq C_e)$, $U/D=\{\psi_1,\psi_2,\ldots,\psi_l\}$,$X_w(1\leqslant w\leqslant q)\subseteq\psi_{x(w)}(1\leqslant x(w)\leqslant l)$.为求 $H(D|R_{A\cup B}^\lambda)$, 仅需要将λ-相容类 $X_{q+1},X_{q+2},\ldots,X_s$ 从站点 g 传送到站点 e,因而需要的网络通信量至多为 $\sum_{i=q+1}^{s}|X_i|+s-q$. 而 $s-q\lhd|R_A^\lambda|$, $\sum_{i=q+1}^{s}|X_i|<\sum_{i=1}^{s}|X_i|$,且通常大数据环境下 s–q 相对于 $\sum_{i=q+1}^{s}|X_i|$ 可忽略不计,为方便起见,记 $Z(R_A^\lambda)=\{X_i\in R_A^\lambda:X_i\cap\psi_k\neq\varnothing\wedge X_i\cap\psi_r\neq\varnothing,1\leqslant k\neq r\leqslant l\}$, 表示不包含在某决策类中的相容类;记从某站点传送λ-相容类 R_A^λ 到另一站点需要的网络通信量为 $NZ(R_A^\lambda)$.

由基于互信息的约简算法可知:在算法运行过程中,随着重要属性的不断扩展,网络传输代价将快速降低.

3.2 多决策表下的区间值λ-全局近似约简算法

依据第 2.4 节的基于互信息的区间值属性λ-近似约简方法和上述的相容类传送策略,可设计多决策表下的区间值λ-全局近似约简算法,见算法 3.

算法 3. 多决策表下的区间值λ-全局近似约简(λ-global approximate reduction in interval-valued multi-decision tables,简称 GARIv).

输入:$DT=\langle U,C\cup D,V,f\rangle,\lambda$;

输出:λ-全局近似约简 red.

Step 1. 令 $red=\varnothing$;

Step 2. 各站点上并行计算λ-相容类 $R_{\{a_k\}}^\lambda(a_k\in C_i)$;各站点并行计算 $Z(R_{\{a_k\}}^\lambda)(a_k\in C_i)$,找到各站点 S_i 的使得

$H(D|R^{\lambda}_{\{a_k\}})$最小的属性 a_i;并在站点 S_j 得到使得 $H(D|R^{\lambda}_{\{a_j\}})\leqslant H(D|R^{\lambda}_{\{a_i\}})(a_i\in C_i)$ 的属性 a_j(即,a_j是各站点选出的 a_i 中条件熵最小的一个), $red=red\cup\{a_j\}$, $H(D|R^{\lambda}_{red})=H(D|R^{\lambda}_{\{a_j\}})$;

Step 3. 若$(C_i-red)\neq\varnothing$,$i\neq j$,将站点 S_j 得到的 $Z(R^{\lambda}_{red})$ 传送到各站点 S_i,在各站点并行计算 $Z(R^{\lambda}_{red\cup\{a_k\}})$ $(a_k\in(C_i-red))$,找到各站点 S_i 使得 $H(D|R^{\lambda}_{red\cup\{a_k\}})$ 最小的属性 a_i;并在站点 S_j 得到使得 $H(D|R^{\lambda}_{red\cup\{a_j\}})\leqslant H(D|R^{\lambda}_{red\cup\{a_i\}})$ $(a_i\in(C_i-red),i\neq j)$的属性 a_j(即,a_j是各站点选出的 a_i 中条件熵最小的一个),记:

$$SIG(a_j,red,D)=H(D|R^{\lambda}_{red})-H(D|R^{\lambda}_{red\cup\{a_j\}});$$

Step 4. 如果 $SIG(a_j,red,D)>\varepsilon$且 $red\neq\bigcup_{i=1}^{m}C_i$ (m 为站点数),$red=red\cup\{a_j\}$,转 Step 3;

否则,输出λ-全局近似约简 red.

在实际应用中,可采用将数量较小的λ-相容类传送到数量相对大的λ-相容类所在的站点,来进一步优化该算法.

4 实验与分析

火电站是一个多子系统串连的复杂大系统,主要设备是锅炉、汽轮机和发电机,完成从热能到机械能、最后到电能的转换过程.现代电站的机组均采用了分散控制系统 DCS,许多老电站也大多进行了 DCS 的改造.随着 DCS 系统在电力行业的普遍推广,电站大都分布式存储了大量运营生产数据.DCS 产生海量的生产数据,逐步形成电力大数据格局.生产数据在时间上具有很强的规律性,通过发电站的历史数据分析找到电站运行规律,可为发电站的运行、检修和事故处理提供决策依据.现有的数据挖掘技术在电站生产数据上进行了较多的尝试,也取得了一定的成果.然而,现有的数据挖掘方法大都没有注意到电站运行数据的特点,在没有对数据进行稳态判定的情况下直接进行数据挖掘.由于工况划分不够明确,导致挖掘结果跟实际运行数据的可比性不是很好.本实验根据电站生产特点,对生产数据进行稳态判定,建立分类模型,通过对分类结果的准确率及建立分类模型的时间来评价所提算法的有效性.

4.1 实验数据

本实验选用某电厂的一台 600MW 机组进行实验.所有数据均存放在 2 个不同的工业实时数据库中,监测电厂长期运行状态,包括汽轮机部分、锅炉部分和管道部分等各方面的数值数据.为了测试本文所提算法 1(RIvD)和算法 2(RIvMI),先将各实时数据库中的历史数据作了集成.该电厂数据采集频率为 1 分钟,即每分钟产生一条数据.选用 2012 年上半年数据作为实验对象,除去机组检修停机时间,共产生 107 184 条记录.集成后的数据共有 427 个属性,除去系统自动生成关键字 ID 号和数据保存时间,共有 425 个条件属性(均为数值型).对运行数据根据稳态工况参数判定公式进行稳态和非稳态标注,形成决策属性,从而得到一张大型的决策表.为了测试算法 3(GARIv)的有效性,原各系统数据不作处理,为每个数据库数据添加同一决策属性.

为了评价算法的性能,我们对数据的区间设计多种划分方法.如每隔 10 分钟,20 分钟,…,90 分钟为一个区间.若在划分过程中,遇到某一个区间对应不同的决策类,则将同一个决策类的数据划分为一个小的区间,下一个区间从不同决策类开始.

4.2 实验环境

所有实验都运行在 Intel Xeon(R) Processor(Four Core,2.5GHz,16GRAM)工作站上,利用 JAVA 进行编写.为了测试 GARIv 算法,在该工作站搭建两台虚拟机作为两个站点.为了保证实验比较的公平性,采用十折交叉确认估计分类准确率.

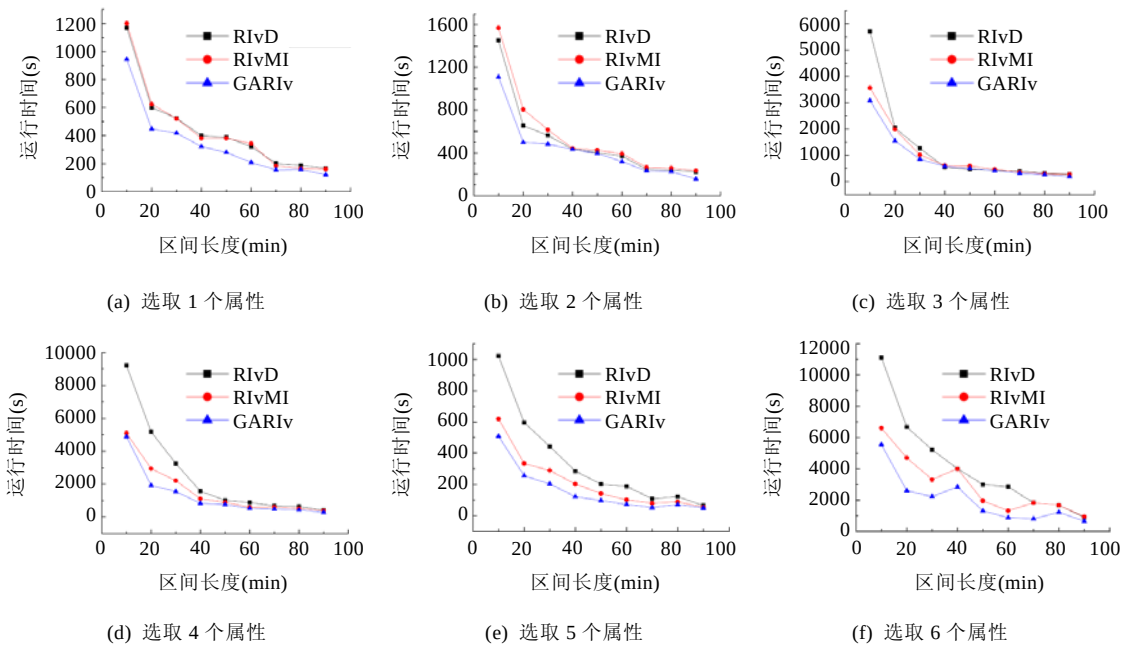
4.3 评价指标

由于ε-近似约简只影响所选属性子集的长度(即所选子集个数),并不影响按重要性选取属性的先后顺序,所

以本文并未对ε的取值进行讨论.对电力大数据构建分类模型,除了考虑算法的运行时间,还应考虑算法的平均分类准确率.本文主要针对区间值启发式约简进行研究,因此在评价算法的准确率时,测试数据和训练数据均按照同样时间进行分块,对每个属性求最大最小值,对数据块采用区间值记录保存.选取与训练数据块相似度最高的决策类作为测试数据块的决策类,与测试数据块真实的决策类进行比较,计算正确分类的比例.对整个过程,仍采用十折交叉确认计算分类的平均准确率.

4.4 参数的选择和设置

首先,将算法 1~算法 3 在所选数据集上进行实验,记录各算法在不同区间长度选取不同属性个数所需的时间.图 1(a)~图 1(f)分别表示当λ=0.7 时,区间长度为 10 分钟,20 分钟,…,90 分钟选取不同属性个数的运行时间图.图 2(a)~图 2(c)表示λ=0.5,0.7,0.9 时,不同区间长度下选取 3 个属性时所需的运行时间;图 2(d)~图 2(f)表示不同λ取值时,随着区间长度变化,选取 6 个属性时所需的运行时间.数据区间化时间不计算在内,仅考虑区间值约简算法选取属性的运行时间.



(a) 选取 1 个属性 (b) 选取 2 个属性 (c) 选取 3 个属性

(d) 选取 4 个属性 (e) 选取 5 个属性 (f) 选取 6 个属性

Fig.1 The running time of different reduction algorithms when λ=0.7

图 1 λ=0.7 时不同约简算法运行时间

从图 1 可以看出:

- 随着区间长度的增加,数据对象成倍数减少,3 个算法运行时间也极大地降低;
- 但随着区间长度的增加,区间之间的重合度将增加,容易造成λ-相容类的元素个数增加,添加一个属性时,对相容类的交运算量增加,因此,算法的运行时间没有呈线性变化.甚至在某些时候,特别是区间长度越长,随着区间长度的增加,运行时间没有因为对象的减少而减少,反而增加.这也可能是因为区间的长度虽然增加,但由于λ-相容类的元素个数增加,从而导致运行时间也有所增加;
- RIvD 算法与 RIvMI 算法在属性个数较少时,运行时间较为相似;但随着属性个数的增加,RIvD 算法的运行时间较多于 RIvMI.这可能是因为随着属性个数的增加,通过交运算后相容类的个数增加,而计算正域时需要重新判断每个相容类是否属于正域,导致 RIvMI 算法的计算时间增加.对 RIvMI 算法来说,虽然增加一个属性相容类个数会随着增加,但在计算新的条件熵时,可在原条件熵的基础上通过交运算来计算新的条件熵,因此随着属性个数增加,RIvMI 算法比 RIvD 算法所需时间普遍较少.

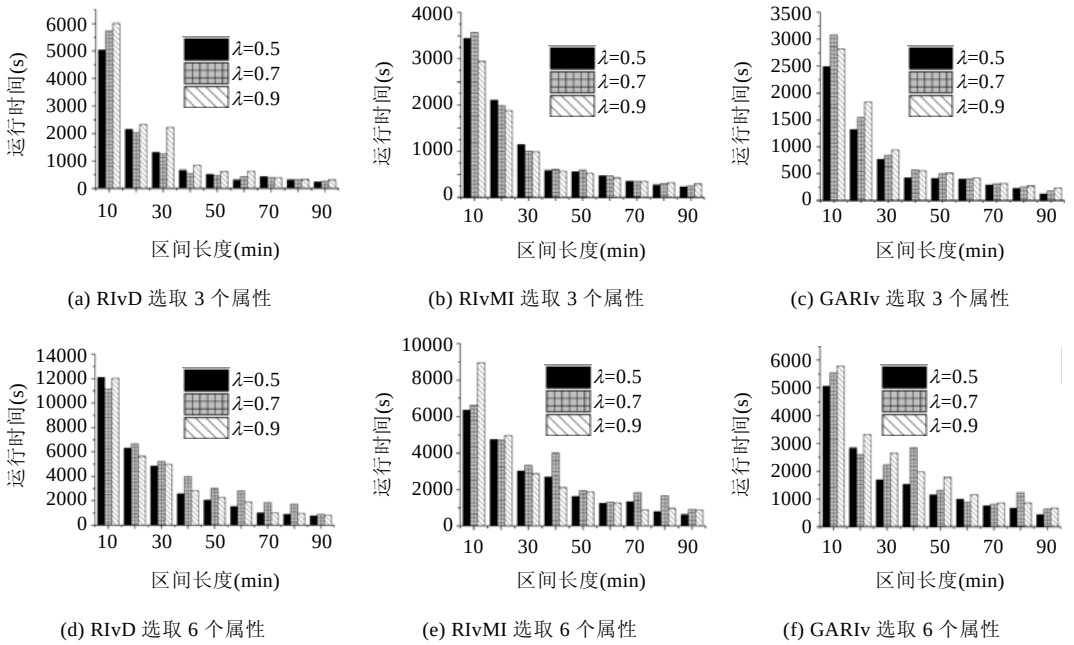


Fig.2 The running time of reduction algorithms with different λ

图 2 不同λ取值时各约简算法运行时间

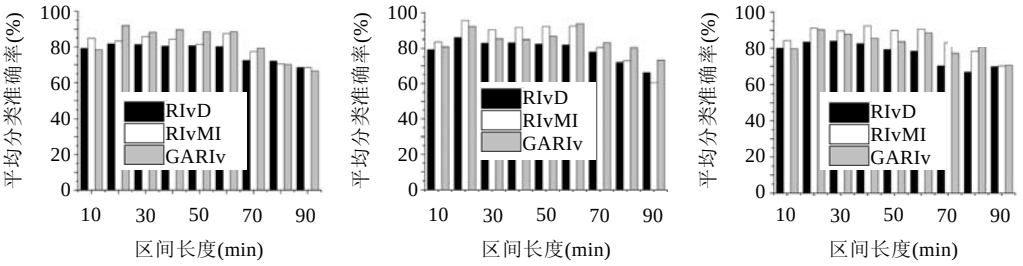
从图 1 还可看出,GARIv 算法的运行时间最少.理论上来说,GARIv 算法所需的时间应为 RIvMI 算法的 1/2(共 2 个站点并行计算),但由于在计算相容类时,涉及到部分相容类的传输与交运算,以及虚拟机的配置低于整个工作站,所以 GARIv 算法的实际运行时间高于理论值.但总体来说,GARIv 算法的运行时间低于 RIvMI.随着站点个数的增加,GARIv 算法的实际运行时间应明显低于 RIvMI.

从图 2 可以看出,3 种算法的运行时间受λ取值影响不具规律性.随着λ的增加,相容类越细,即相容类元素个数越少.对 GARIv 算法来说,所需要传输的相容类也就越多,所以同样条件下,运行时间略长.

将 3 种算法在ε=0.01 时选出的不同属性子集进行平均分类准确率比较,结果如图 3 所示.图 3(a)为λ=0.5 时的算法平均分类准确率;图 3(b)为λ=0.7 时各算法的平均分类准确率;图 3(c)为λ=0.9 时各算法的平均分类准确率.从图中可以看出,3 种算法的平均分类准确率在合适区间长度上基本均能达到 80%以上.当λ=0.7、区间长度为 20 分钟时,各算法总体平均分类准确率最高.由于 RIvD 算法和 RIvMI 算法的重要性度量方法不同,导致所选取属性子集也不同.从图 3 可以看出,RIvMI 算法所选取的属性子集比 RIvD 算法的分类准确率稍高.这可能是因为数据区间化后产生不一致现象,而基于依赖度(正域)的方法不适合处理不一致问题,因此分类准确率比 RIvMI 略低.GARIv 算法相当于将 RIvMI 算法在垂直划分的数据集上并行进行约简算法,但两种算法所选取的属性子集并不相同.这是因为 RIvMI 算法在选取属性时,当某些属性的重要度一样时,则先选择最左边的属性到约简集合中.而 GARIv 算法所处理的数据集可看成是 RIvMI 算法数据集垂直方向的划分,因此属性的排序不同,两个算法所选取的属性子集也不同.

同时,从图 3 还可以看出:当区间长度超过 1 小时后,虽然运行时间减少,但平均分类准确率大幅下降.这主要是因为数据区间长度如果过大,数据的区间值不能反映数据块的数据特征;同理,如果数据区间长度过小,不仅使得运行时间较长,而且数据块信息不够充分,同样会导致低分类准确率.因此,对区间值约简算法而言,区间长度的选取对算法整个结果影响较大,应根据不同的应用设置不同的区间选取粒度.

由此可知:我们所提的区间值约简算法适用于处理呈连续分布的数据,而无法处理跳跃式分布的数据.



(a) λ=0.5, ε=0.01 (b) λ=0.7, ε=0.01 (c) λ=0.9, ε=0.01

Fig.3 The average accuracies of classification for different reduction algorithms with different λ

图 3 不同约简算法在不同λ取值的平均分类准确率

通过对以上实验所选取的属性子集进行比较,发现它们中大都包含发电机功率、主汽压力(机侧)、主汽温度(机侧)、再热热段温度(机侧)、#1 高加进汽温度这 5 个属性,这与实际判断稳态所涉及的指标相吻合.

为了进一步考察本文所提出的 3 种算法的有效性,将以上实验的平均分类准确率与传统分类器 KNN(k 近邻)、RBF(径向基函数神经网络)、REPTree(缩减误差修剪树)的平均分类准确率进行比较,结果见表 2.

Table 2 The average accuracies of classification

表 2 平均分类准确率

算法		区间长度(单位:分钟)								
		10	20	30	40	50	60	70	80	90
RIvD	λ=0.5	79.2%	81.7%	81.4%	80.5%	80.7%	80.2%	72.4%	72.1%	68.5%
	λ=0.7	79.2%	**85.9%**	82.8%	83.1%	82.3%	81.7%	77.8%	71.9%	66.2%
	λ=0.9	80.1%	83.5%	84.0%	82.6%	79.4%	78.5%	70.3%	66.9%	69.9%
RIvMI	λ=0.5	84.8%	83.3%	85.8%	84.3%	81.5%	87.5%	77.4%	70.5%	68.5%
	λ=0.7	83.6%	**95.7%**	90.3%	91.7%	92.4%	92.5%	80.3%	72.8%	60.4%
	λ=0.9	84.3%	91.2%	89.7%	92.4%	89.9%	90.6%	83.1%	78.4%	70.3%
GARIv	λ=0.5	78.5%	92.0%	88.3%	89.7%	88.5%	88.6%	79.3%	70.2%	66.5%
	λ=0.7	80.7%	92.4%	85.4%	84.8%	86.7%	***93.6%***	83.1%	80.3%	73.1%
	λ=0.9	79.8%	90.3%	87.8%	85.5%	83.7%	88.6%	77.3%	81.2%	70.6%
1NN		89.5%								
2NN		89.2%								
3NN		86.8%								
5NN		81.9%								
10NN		73.5%								
RBF		54.7%								
REPtree		52.9%								

从表 2 中可以看出:在传统分类方法中,KNN 的分类效果最好;对 RBF 神经网络和 REPTree 决策树两种方法的分类准确率都较低,这是因为对电力大数据的判稳应该基于某一段数据,而对某一条数据而言,无法正确判断其是否在稳态或非稳态中.KNN 的分类效果较好,主要因为决策类的取值都是分段的,同一段数据的决策类相同,因此计算 k 近邻时,离测试数据距离最近的 k 个点所对应的决策类与相邻测试数据可能相同;并且随着 k 值的增加,易出现跨类的现象,导致平均分类准确率下降.而本文所提出的 3 个算法均是基于区间的,更加符合判稳的条件,因此平均分类准确率比传统方法高.由此也可看出,本文所提的 3 个算法对大数据的分类问题是有效的.根据不同的应用选取不同的区间长度,确定λ的值,根据所需的属性个数确定ε的值.对分布式存储的大数据,可直接采取 GARIv 算法对数据进行处理,求得全局近似约简.

5 总结与展望

本文针对电力大数据分类问题的特点,提出了基于依赖度和互信息的区间值约简算法,并针对大数据的分

布式存储,提出了信息论下的区间值全局近似约简概念和方法,在电力大数据的判稳中进行应用,取得了较好的结果.由于电力大数据的分析应用中大多都需考虑某个数据段的变化而不是某一条数据,将数据集进行区间化,不仅可以极大减少大数据的数据量、降低大数据分析的难度,同时也符合电力大数据的具体应用.而将数据集进行属性约简,在不影响整个数据集分类条件下,也从维度对大数据进行了缩减,降低了大数据的数据量,降低了大数据分析难度.从实验结果来看:3 种算法均是有效的,为区间值约简方法提供了新思路,同时也为大数据的分类问题提供了解决方案.

在后续的研究工作中,我们将围绕以下几个方面展开研究:

- 对 3 种算法的参数选择进行更加详细的讨论,通过更多的实验给出合理和有效的参数选择方法;
- 同时,将 3 种算法分别在 Map-Reduce 分布式平台下给出相应的算法,加大算法并行化处理的程度,从而真正实现大数据的分析处理;
- 在电力大数据的负荷预测、故障诊断等方面进一步将所提算法进行应用,进一步验证算法的有效性,并为电力大数据的分析提供新思路,继而更好地在其他大数据实际问题中展开应用.

致谢 在此,我们向对本文工作给予支持和建议的同行,尤其是对本文给予评审并提出宝贵意见的专家们表示感谢.

References:

[1] Lynch C. Big data: How do your data grow? Nature, 2008,455(7209):28−29. [doi: 10.1038/455028a]

[2] The role of stream computing in big data architectures. 2013. http://ibmdatamag.com/2013/01/the-role-of-stream-computing-in-bigdata-architectures/

[3] Li GJ, Cheng XQ. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012,27(6): 647−657 (in Chinese with English abstract).

[4] Wang YZ, Jin XL, Cheng XQ. Network big data: Present and future. Chinese Journal of Computers, 2013,36(6):1125−1138 (in Chinese with English abstract). [doi: 10.3724/SP.J.1.16.2013.01125]

[5] Wang S, Wang HJ, Tan XP, Zhou H. Architecting big data: Challenges, studies, forecasts. Chinese Journal of Computers, 2011, 34(10):141−1752 (in Chinese with English abstract). [doi: 10.3274/SP.J.1016.2011.0174]

[6] Li JZ, Liu XM. An important aspect of big data: Data usability. Journal of Computer Research and Development, 2013,50(6): 1147−1162 (in Chinese with English abstract).

[7] Sun DW, Zhang GY, Zheng WM. Big data stream computing: Technologies and instances. Ruan Jian Xue Bao/Journal of Software, 2014 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4558.htm [doi: 10.13328/j.cnki.jos.004558]

[8] Meng XF, Ci X. Big data management: Concepts, techniques and challenges. Journal of Computer Research and Development, 2013,50(1):146−169 (in Chinese with English abstract).

[9] Shen DR, Yu G, Wang XT, Nie TZ, Kou Y. Survey on NoSQL for management of big data. Ruan Jian Xue Bao/Journal of Software, 2013,24(8):1786−1803 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4416.htm [doi: 10.3724/SP.J.1001. 2013.04416]

[10] Rabl T, Sadoghi M, Jacobsen HA. Solving big datachallenges for enterprise application performance management. Proc. of the VLDB Endowment, 2012,5(12):1724−1735. [doi: 10.14778/2367502.2367512]

[11] Mayer V, Cukier K. A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt, 2013.

[12] Pawlak Z. Rough sets. Int'l Journal of Compute and Information Science, 1982,11(4):341−356. [doi: 10.1007/BF01001956]

[13] Wang GY, Yao YY, Yu H. A survey on rough set theory and applications. Chinese Journal of Computers, 2009,32(7):1229−1246 (in Chinese with English abstract). [doi: 10.3274/SP.J.1016.2009.01229]

[14] Mac Parthalain N, Jensen R, Shen Q. Rough and fuzzy-rough methodsfor mammographic data analysis. Intelligent Data Analysis —An Int'l Journal, 2010,14(2):225−244.

[15] Zhu W. Generalized rough sets based on relations. Information Sciences, 2007,177(22):4997−5011. [doi: 10.1016/j.ins.2007.05.037]

[16] Zhang WX, Wu WZ, Liang JY, Li DY. Rough Set Theory and Method. Beijing: Science Press, 2001 (in Chinese).

[17] Mi JS, Wu WZ, Zhang WX. Constructive and axiomatic approaches of theory of rough sets. Pattern Recognition and Artificial Intelligence, 2002,15(3):280−284 (in Chinese with English abstract). [doi: 10.3969%2fj.issn.1003-6059.2002.03.005]

[18] Zhu W. Topological approaches to covering rough sets. Information Sciences, 2007,177(6):1499−1508. [doi: 10.1016/j.ins.2006.06.009]

[19] Zhang WX, Yao YY, Liang Y. Rough Setand Concept Lattice. Xi'an: Xi'an Jiaotong University Press, 2006 (in Chinese).

[20] Qian YH, Liang JY, Yao YY, Dang CY. MGRS: A multi-granulation rough set. Information Sciences, 2010,180(6):949−970. [doi: 10.1016/j.ins.2009.11.023]

[21] Suyun Z, Tsang E, Degang C. The model of fuzzy variable precision rough sets. IEEE Trans. on Fuzzy Systems, 2009,17(2):451−467. [doi: 10.1109/TFUZZ.2009.2013204]

[22] Huang B, Hu ZJ, Zhou XZ. Dominance relation-based fuzzy-rough model and its application to audit risk evaluation. Control and Decision, 2009,24(6):899−902 (in Chinese with English abstract).

[23] Zhang DB, Wang YN, Huang HX. Rough neural network modeling based on fuzzy rough model and its application to texture classification. Neurocomputing, 2009,72(10-12):2433−2443. [doi: 10.1016/j.neucom.2008.12.003]

[24] Xu FF, Miao DQ, Wei L. Fuzzy-Rough attribute reduction via mutualinformation with an application to cancer classification. Computers & Mathematics with Applications, 2009,57(6):1010−1017. [doi: 10.1016/j.camwa.2008.10.027]

[25] Hu QH, Zhao H, Yu DR. Efficient symbolic and numerical attribute reduction with rough sets. Pattern Recognition and Artificial Intelligence, 2008,6:732−738 (in Chinese with English abstract).

[26] Liang JY, Qian YH, Pedrycz W, Dang CY. An efficient accelerator for attribute reduction from incomplete data in rough set framework. Pattern Recognition, 2011,44(8):1658−1670. [doi: 10.1016/j.patcog.2011.02.020]

[27] Wang WH, Zhou DH. An algorithm for knowledge reduction in rough sets based on genetic algorithm. Journal of System Simulation, 2001,13:91−94 (in Chinese with English abstract).

[28] Qian J, Miao DQ, Zhang ZH, Zhang ZF. Parallel algorithm model for knowledge reduction using MapReduce. Journal of Frontiers of Computer Science and Technology, 2013,7(1):35−45 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.1206048]

[29] Zhang JB, Li TR, Pan Y. PLAR: Parallel Large-Scale Attribute Reduction on Cloud Systems. Institute of Electrical & Electronic Engineers, 2013.

[30] Yang M, Yang P. Approximate reduction based on conditional information entropyover vertically partitioned multi-decision table. Control and Decision, 2008,23(10):1103−1108 (in Chinese with English abstract).

[31] Ye MQ, Hu XG, Wu CR. Privacy preserving attribute reduction based on conditional information entropy over vertically partitioned multi-decision tables. Journal of Shandong University (Natural Science), 2010,45(9):14−26 (in Chinese with English abstract).

[32] Zhang N, Miao DQ, Yue XD. Approaches to knowledge reduction in interval-valued information systems. Journal of Computer Research and Development, 2010,47(8):1362−1371 (in Chinese with English abstract).

[33] Chen ZC, Qin KY. Attribute reduction of interval-valued information system based on the maximal tolerance class. Fuzzy Systems and Mathmatics, 2009,23(6):126−132 (in Chinese with English abstract).

[34] Guo Q, Liu WJ, Jiao XF, Wu L. A novel interval-valued attribution reduction algorithm based on fuzzy cluster. Fuzzy Systems and Mathmatics, 2013,27(1):149−153 (in Chinese with English abstract).

[35] Gong WL, Li DY, Wang SG, Cheng LT. Attribute reduction of interval-valued information system based on fuzzy discernibility matrix. Journal of Shanxi University (Natural Science), 2011,34(3):381−387 (in Chinese with English abstract).

附中文参考文献:

[4] 王元卓,靳小龙,程学旗.网络大数据:现状与展望.计算机学报,2013,36(6):1125−1138. [doi: 10.3724/SP.J.1.16.2013.01125]

[5] 王珊,王会举,覃雄派,周烜.架构大数据:挑战、现状与展望.计算机学报,2011,34(10):141−1752. [doi: 10.3274/SP.J.1016.2011.0174]

[6] 李建中,刘显敏.大数据的一个重要方面:数据可用性.计算机研究与发展,2013,50(6):1147−1162.

[7] 孙大为,张广艳,郑纬民.大数据流式计算:关键技术及系统实例.软件学报,2014. http://www.jos.org.cn/1000-9825/4558.htm [doi: 10.13328/j.cnki.jos.004558]

[8] 孟小峰,慈祥.大数据管理:概念、技术与挑战.计算机研究与发展,2013,50(1):146−169.

[9] 申德荣,于戈,王习特,聂铁铮,寇月.支持大数据管理的 NoSQL 系统研究综述.软件学报,2013,24(8):1786−1803. http://www.jos.org.cn/1000-9825/4416.htm [doi: 10.3724/SP.J.1001.2013.04416]

[13] 王国胤,姚一豫,于洪.粗糙集理论及应用研究综述.计算机学报,2009,32(7):1229−1246. [doi: 10.3274/SP.J.1016.2009.01229]

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

[17] 米据生,吴伟志,张文修.粗糙集的构造与公理化方法.模式识别与人工智能,2002,15(3):280−284. [doi: 10.3969%2fj.issn.1003-6059.2002.03.005]

[19] 张文修,姚一豫,梁怡.粗糙集与概念格.西安:西安交通大学出版社,2006.

[22] 黄兵,胡作进,周献中.优势模糊粗糙模型及其在审计风险评估中的应用.控制与决策,2009,24(6):899−902.

[25] 胡清华,赵辉,于达仁.基于粗糙集的符号与数值属性的快速约简算法.模式识别与人工智能,2008,6:732−738.

[27] 王文辉,周东华.基于遗传算法的一种粗糙集知识约简算法.系统仿真学报,2001,13:91−94.

[28] 钱进,苗夺谦,张泽华,张志飞.MapReduce 框架下并行知识约简算法模型研究.计算机科学与探索,2013,7(1):35−45. [doi: 10.3778/j.issn.1673-9418.1206048]

[30] 杨明,杨萍.垂直分布多决策表下基于条件信息熵的近似约简.控制与决策,2008,23(10):1103−1108.

[31] 叶明全,胡学钢,伍长荣.垂直划分多决策表下基于条件信息熵的隐私保护属性约简.山东大学学报(理学版),2010,45(9):14−26.

[32] 张楠,苗夺谦,岳晓冬.区间值信息系统的知识约简.计算机研究与发展,2010,47(8):1362−1371.

[33] 陈子春,秦克云.区间值信息系统基于极大相容类的属性约简.模糊系统与数学,2009,23(6):126−132.

[34] 郭庆,刘文军,焦贤发,吴磊.一种基于模糊聚类的区间值属性约简算法.模糊系统与数学,2013,27(1):149−153.

[35] 龚伟林,李德玉,王素格,程利涛.基于模糊区分矩阵的区间值信息系统属性约简.山西大学学报(自然科学版),2011,34(3):381−387.

徐菲菲(1983－),女,江西南昌人,博士,副教授,CCF 会员,主要研究领域为粗糙集,粒计算,云计算,数据挖掘,模式识别.
E-mail: xufeifei@shiep.edu.cn

雷景生(1966－),男,博士,教授,CCF 高级会员,主要研究领域为数据挖掘,数据库技术,云计算.
E-mail: jshlei@126.com

毕忠勤(1977－),男,博士,副教授,主要研究领域为云计算,大数据处理.
E-mail: bizhongqin@gmail.com

苗夺谦(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为粗糙集,大数据,云计算.
E-mail: miaoduoqian@163.com

杜海舟(1980－),男,副教授,主要研究领域为电厂数据挖掘,电网规划.
E-mail: du_hz@126.com