软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2825−2835 [doi: 10.13328/j.cnki.jos.005337]

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

基于随机抽样的模糊粗糙约简*

陈 俞[1,2], 赵素云[1,2], 李雪峰[3], 陈 红[1,2], 李翠平[1,2]

[1](中国人民大学 信息学院,北京 100872)

[2](数据工程与知识工程教育部重点实验室(中国人民大学),北京 100872)

[3](中国人民大学 环境学院,北京 100872)

通讯作者: 赵素云, E-mail: zhaosuyun@ruc.edu.cn

摘 要: 传统的属性约简由于其时间复杂度和空间复杂度过高,几乎无法应用到大规模的数据集中.将随机抽样引入传统的模糊粗糙集中,使得属性约简的效率大幅度提升.首先,在统计下近似的基础上提出一种统计属性约简的定义.这里的约简不是原有意义上的约简,而是保持基于统计下近似定义的统计辨识度不变的属性子集.然后,采用抽样的方法计算统计辨识度的样本估计值,基于此估计值可以对统计属性重要性进行排序,从而可以设计一种快速的适用于大规模数据的序约简算法.由于随机抽样集以及统计近似概念的引入,该算法从时间和空间上均降低了约简的计算复杂度,同时又保持了数据集中信息含量几乎不变.最后,数值实验将基于随机抽样的序约简算法和两种传统的属性约简算法从以下 3 个方面进行了对比:计算属性约简时间消耗、计算属性约简空间消耗、约简效果.对比实验验证了基于随机抽样的序约简算法在时间与空间上的优势.

关键词: 模糊粗糙集;随机抽样;属性约简;统计粗糙集

中图法分类号: TP311

中文引用格式: 陈俞,赵素云,李雪峰,陈红,李翠平.基于随机抽样的模糊粗糙约简.软件学报,2017,28(11):2825−2835. http://www.jos.org.cn/1000-9825/5337.htm

英文引用格式: Chen Y, Zhao SY, Li XF, Chen H, Li CP. Fuzzy rough reduction based on random sampling. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2825−2835 (in Chinese). http://www.jos.org.cn/1000-9825/5337.htm

Fuzzy Rough Reduction Based on Random Sampling

CHEN Yu[1,2], ZHAO Su-Yun[1,2], LI Xue-Feng[3], CHEN Hong[1,2], LI Cui-Ping[1,2]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](Key Laboratory of Data Engineering and Knowledge Engineering (Renmin Universityof China), Ministry of Educaion, Beijing 100872, China)

[3](School of Environment, Renmin University of China, Beijing 100872, China)

Abstract: Traditional attribute reduction is less effective when applying to large-scale datasets because of its high time and space complexity. In this paper, random sampling is introduced into traditional rough reduction. First, statistical discernibility degree and

* 基金项目: 国家重点研发计划(2016YFB1000702); 国家重点基础研究发展计划(973)(2014CB340402); 国家高技术研究发展计划(863)(2014AA015204); 国家自然科学基金(61772536, 61772537, 61702522, 61532021); 国家社会科学基金(12&ZD220); 中国人民大学科学研究基金(中央高校基本科研业务费专项资金)(15XNLQ06); 国家高等学校学科创新引智计划(111)

Foundation item: National Key Research and Development Program of China (2016YFB1000702); National Program on Key Basic Research Project of China (973) (2014CB340402); National High-Tech R&D Program of China (863) (2014AA015204); National Natural Science Foundation of China (61772536, 61772537, 61702522, 61532021); National Social Science Foundation (12&ZD220); Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (15XNLQ06); Chinese National 111 Project Attracting

本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授推荐.

收稿时间: 2017-03-13; 修改时间: 2017-06-16; 采用时间: 2017-08-23

statistical rough reduction are proposed based on statistical rough approximation. Here the statistical rough reduction is not the traditional reduction any more, it is a subset which keeps the statistical discernibility degree almost invariant. By using random sampling to find the estimated value of statistical discernibility degree, all the condition attributes can be sorted. And then the reduction can be done on the sorted attributes by keeping the statistical discernibility degree almost invariant. Finally, numerical experimental comparison demonstrates that the random sampling based rough reduction is effective on both time and space consumption.

Key words: fuzzy rough set; random sampling; attribute reduction; statistical rough set

随着大数据时代的来临,数据挖掘技术蓬勃发展.近年来,由于不确定性数据比重的不断增大,不确定性数据挖掘越来越受到人们的重视.在不确定数据上进行降维,如基于模糊粗糙集的属性约简,近几年得到广泛关注.但是,现有的模糊粗糙集约简方法由于其基础理论复杂度的桎梏,无法直接应用到大规模数据集上.

模糊粗糙集是粗糙集在模糊集框架下的一种概述[1−4].属性约简作为模糊粗糙集的一种重要应用,已经得到了广泛和深入的研究[5−12].一般来说,有两种属性约简的方法:一种是基于依赖度函数的方法,另一种是基于辨识矩阵的方法.基于依赖度函数的属性约简方法,其核心思想是在保持依赖度不变的条件下去掉冗余属性.文献[7]首次提出模糊粗糙集上基于依赖度的约简算法,该方法在一些实际应用上表现相当不错,但是在实际应用中还存在着一些明显的不足.一个明显的不足就是算法时间复杂度极高.文献[9]中指出:文献[7]中的算法,由于其终止条件的设计不佳,在很多真实数据集上根本无法收敛.与文献[7]不同,文献[8]提出了一种属性约简方法,他们通过信息熵来衡量属性的重要性.但无论是依赖度函数,还是属性重要性,在大型数据集上都有着过高的时间消耗.另一个常用的属性约简方法是基于辨识矩阵[11,12],文献[11]提出了一种关于属性约简的统一框架:文中提出了基于模糊近似算子属性约简的正式的概念;同时,利用严格的数学证明,分析了属性约简的数学构造.通过这个框架,基于辨识矩阵的模糊属性约简方法得以提出.考虑到模糊粗糙集的灵敏度,文献[12]提出了近似算子的一个阈值.然而,文献[11,12]都面临着一个相同的问题,那就是基于辨识矩阵的方法空间消耗过大.

从上述情况可以看出:在模糊粗糙集中,基于依赖度或者辨识矩阵的算法通常有着高时间或者高空间复杂度.这是由于作为其理论支撑的基本概念复杂度过高,比如说,在模糊粗糙集中十分重要的下近似算子与辨识信息的度量,其复杂度是全集数量的平方量级.因此,基于下近似算子与辨识信息度量的约简和分类器在大规模数据集上无法得到有效的应用.随机抽样是一种可以极大地减小运算量的统计学方法,且随机抽样已被引入模糊粗糙集理论,提出了一种统计近似的知识表示方法[13].然而基于随机抽样的属性约简方法并未进一步讨论.因此,本文将随机抽样引入到经典的模糊粗糙约简理论中,建立一种新的统计粗糙辨识度量方法,并设计出基于随机抽样的统计约简的算法.

本文的主要贡献在于:将随机抽样和传统的属性约简方法结合,借助于统计下近似算子,提出了统计辨识度和统计属性约简的概念;然后,利用随机抽样快速计算出统计辨识度的估计值,从而设计用以计算统计属性约简的算法,该算法极大地降低了属性约简的复杂度和计算量.

本文第 1 节回顾模糊粗糙约简.第 2 节提出统计不可辨识的几个基本概念,如统计不可辨识度与统计粗糙约简等;并给出一个定理,该定理指出,随机抽样的方法可用于发现属性不可辨识度的无偏估计值.第 3 节设计基于随机抽样的序约简算法.通过第 4 节的数值实验发现,基于随机抽样的序约简算法相对于经典的属性约简算法所需的时间、空间较小,该约简算法更适用于大规模数据集.第 5 节总结全文的工作.

1 相关工作

1.1 模糊粗糙集模型

首先,我们回顾一下模糊粗糙集的基本概念[14].

全集 U 是一个有限个数示例的非空集合,记作 $U=\{x_1,x_2,\ldots,x_n\}$;每一个示例都有一系列条件属性,记作 $R=\{r_1,r_2,\ldots,r_p\}$;决策属性 D.于是,$(U,R\cup D)$被称为一个决策表 DS.

对于每一个 $P\subseteq R$,我们将二维关系 $P(x,y)$表示成为 P 的模糊相似关系,对于任意 $x,y,z\in U$,一个模糊相似关系

满足自反性:$P(x,x)=1$;对称性:$P(x,y)=P(y,x)$;T-传递性:$P(x,y)\geq T(P(x,z),P(z,y))$.简单来说,$P$ 是用来代表其相似关系的.

模糊粗糙集是用于将模糊集和粗糙合起来的工具,它在文献[2]中被 Dubois 等人提出来.然后,其细节在文献[15,16]中被深入地研究.总体上来说,如今的模糊粗糙集可以被下列 4 个近似算子概括.

$$\overline{R_T}A(x)=\sup_{u\in U}T(R(x,u),A(u));\ \underline{R_\vartheta}A(x)=\inf_{u\in U}\vartheta(R(x,u),A(u));$$

$$\overline{R_\sigma}A(x)=\sup_{u\in U}\sigma(N(R(x,u)),A(u));\ \underline{R_S}A(x)=\inf_{u\in U}S(N(R(x,u)),A(u)),$$

其中,T,ϑ,σ,S 和 N 均为模糊逻辑算子,详见文献[11,12].依据对偶性,$(\overline{R_T}A,\underline{R_\vartheta}A)$是一对近似算子,$(\overline{R_\sigma}A,\underline{R_S}A)$是另一对近似算子.为简便起见,本文接下来只讨论$(\overline{R_T}A,\underline{R_\vartheta}A)$的性质与应用.$(\overline{R_\sigma}A,\underline{R_S}A)$采用类似的方法,也可以得到类似的结果.

1.2 常见约简算法研究

常见的属性约简方法主要有两种:一种是基于依赖度函数的方法,另一种是基于辨识矩阵的方法.下面我们简单地回顾上述两种方法.

1.2.1 基于依赖度的约简

本节我们给出一些属性约简的相关定义,比如正域、依赖度和约简.

在决策表 $DS=(U,R\cup D)$中,对于所有的 $x\in U$, $POS_R(D)(x)=\bigcup_{z\in U}\underline{R_\vartheta}([z]_D)(x)$ 被称为 D 关于 R 的正域.这里,

$$[z]_D=\{x\in U \text{ s.t. } D(z,x)=1\}.$$

$Dep(POS_R(D))=\sum_{x\in u}POS_R(D)(x)$ 被称为决策属性集 D 在 R 上的依赖度.

属性约简的核心思想就是:在保持正域不变的条件下,最大程度地减少冗余的属性.

在一个决策表 $DS=(U,R\cup D)$,$P\subseteq R$,P 被称为 R 对于 D 的一个约简,如果 P 满足下列式子:

(1) $Dep(POS_P(D))=Dep(POS_R(D))$;

(2) 对于任何 $b\in P$,$Dep(POS_{P-\{b\}}(D))\neq Dep(POS_R(D))$.

保持依赖度不变,就等价于正域不变.因此,也可以用下面的方式描述约简.

在决策表 $DS=(U,R\cup D)$中,$P\subseteq R$ 称为 R 对于 D 的一个约简,如果 P 满足下列式子:

(1) $POS_P(D)=POS_R(D)$;

(2) 对于任何 $b\in P$,$POS_{P-\{b\}}(D)\neq POS_R(D)$.

基于依赖度设计的约简方法被称为依赖度算法.依赖度算法通过一个值来度量数据集的辨识信息含量,在计算依赖度的过程中,需要计算每一个示例的下近似.计算所有示例的下近似所需要的时间是数据规模的平方级,因而该算法所需的时间代价较大.

1.2.2 基于辨识矩阵的约简

在模糊粗糙集中,辨识矩阵是一个 $n\times n$ 的矩阵 c_{ij},用 $M(U,R,D)$来表示.

(M1): $c_{ij}=\{a:T(a(x_i,x_j),\lambda)=0\},\lambda=\underline{R_\vartheta}[x_i]_D(x)$,对于$D(x_i,x_j)=0$;

(M2):$c_{ij}=\varnothing$,对于 $D(x_i,x_j)=1$.

在基于辨识矩阵的属性约简中,其核心思想是保持辨识矩阵中的辨识能力不变.基于辨识矩阵,可以这样定义约简:在决策表 $DS=(U,R\cup D)$中,$P\subseteq R$,P 被称为 R 对于 D 的一个约简,如果 P 满足下列条件:

$$P\cap c_{ij}=\varnothing,\text{对于所有 } c_{ij}\neq\varnothing.$$

基于辨识矩阵设计的约简方法被称为辨识矩阵算法.该算法计算了每一对示例的辨识信息,并将它们存为矩阵形式,这是辨识矩阵算法的特点.然而,矩阵的形式的信息度量方法也成为辨识矩阵方法在大规模数据上应用的瓶颈.这是由于辨识矩阵算法需要数据规模平方级大小的空间来存储矩阵.

1.3 统计粗糙集模型

对于$\forall x\in U$,下近似值就是到不同类别点的最小距离[17,18].假设 y 恰好就是 x 取到异类点最小值的示例,那么很明显地,x 和 y 必然在某些维度上非常接近.因为如果 x 和 y 在所有维度上都距离很远,那么 x 就不会在 y 上取到异类点的最小距离.基于这个逻辑,如果我们将整个数据库在该维度,即属性)上排序,那么 x 和 y 会离得比较接近.下面我们会提出两个概念 k-neighbor 和 k-limit,用来定义某示例的邻居.

定义 **1**. 给出一个随机变量 X 和它的 n 个样本升序排列$\{x_1,x_2,\ldots,x_n\}$,那么 x_i 的 k-neighbor 可以表达成

$$\begin{cases} x_1,\ldots,x_i,\ldots,x_{i+k}, & \text{if } i-k<1 \\ x_{i-k},\ldots,x_i,\ldots,x_n, & \text{if } i+k>n,\ 1\le k\le \arg(n/2). \\ x_{i-k},\ldots,x_i,\ldots,x_{i+k}, & \text{others} \end{cases}$$

定义 **2**. 给定决策表 $DT=(U,R,D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$,$R=\{r_1,r_2,\ldots,r_m\}$.对于所有属性 $1\le k\le \arg(n/2)$,将 x_i 的 k-neighbor 集中到一个集合中,这个集合就称为 x_i 的 k-limit(限定区域).这里,k 就是限定长度.

定义 **3**. 给定决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$,$R=\{r_1,r_2,\ldots,r_m\}$,$F(U)$是 U 的模糊幂函数集合.对于$\forall x\in U,A\in F(U)$,x 的 k-统计下近似算子和 k-统计上近似算子可以如下表示:

$$\underline{R_T^k}A(x)=\inf_{u\in k\text{-}limit(x)}\vartheta(R(u,x),A(u)),\overline{R_T^k}A(x)=\sup_{u\in k\text{-}limit(x)}T(R(u,x),A(u)),$$

$$\underline{R_S^k}A(x)=\inf_{u\in k\text{-}limit(x)}S(N(R(u,x)),A(u)),\overline{R_\sigma^k}A(x)=\sup_{u\in k\text{-}limit(x)}\sigma(N(R(u,x)),A(u)).$$

定义 3 提供了一种新的方法,用更少的计算量来拟合经典的近似算子.性质 1 和性质 2 说明:在 k 足够大时,k-统计近似算子可以逼近经典的近似算子.文献[13]还详细地给出了基于随机抽样的统计下近似计算的方法以及理论支持.然而,如何基于统计近似算子,借助随机抽样的手段构建快速的统计粗糙约简算法,文献[13]并未提及.因此,本文接下来的工作将围绕如何基于统计粗糙集与随机抽样的方法构建基于随机抽样的粗糙约简算法.

2 基于随机抽样的属性约简

为了提高属性约简算法的效率,我们首先定义一些新信息度量的概念,用以定义统计不可辨识性和基于随机抽样的统计约简.

定义 **4**. 对于决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$.对于$(x_i,x_j)\in U$ 并且 x_j 是 x_i 的 k-neighbor,如果 $D_r(x_i,x_j)<\lambda_i$,那么我们说 x_i 对于 x_j 在 r 上是 k-统计不可辨识,或者说(x_i,x_j)称为 r 上的 k-统计不可辨识对.如果 $D_r(x_i,x_j)\ge\lambda_i$,那么我们说 x_i 对于 x_j 在 r 上是 k-统计可分辨,或者说(x_i,x_j)称为 r 上的 k-统计可分辨对.其中,$D_r(x_i,x_j)$是 x_i 和 x_j 在属性 r 上的距离,可以是欧氏距离;λ_i 是 x_i 的 k-统计下近似.

定义 **5**. 对于决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$,$\forall x_i\in U$.

(1) $\forall r\in R$, $IND_r^k(x_i)=\{(x_i,x_j)\mid D_r(x_i,x_j)<\lambda_i,1\le j\le n\}$,$IND_r^k(x_i)$ 称为$|INP_r(x_i)|$上的不可辨识集合,$IND_r^k(x_i)$ 包含所有 x_i 在$|INP_r(x_i)|$上的 k-统计不可辨识对.

(2) $\forall r\in R$,$|IND_r^k(x_i)|$代表着 $IND_r^k(x_i)$ 中元素的个数,我们称其为 x_i 在$|INP_r(x_i)|$上的 k-统计不可辨识度.

(3) $\forall Q\subseteq R$, $IND_Q^k(x_i)=\bigcap_{r\in Q}IND_r^k(x_i)$,$IND_Q^k(x_i)$ 包含所有 x_i 在 Q 上的 k-统计不可辨识对,称为 x_i 在 Q 上的不可辨识集.

(4) $\forall Q\subseteq R$,$|IND_r^k(x_i)|$称为 x_i 在 Q 上的 k-统计不可辨识度.

由定义 4 可得:某属性的不可辨识度越大,该属性可以分辨的示例就越多.

性质 **1**.

(1) 当 $P\subseteq Q\subseteq R$, $IND_P^k(x_i)\supseteq IND_Q^k(x_i)\supseteq IND_R^k(x_i)$;

(2) 当 $P\subseteq Q\subseteq R$,$|IND_P^k(x_i)|\ge|IND_Q^k(x_i)|\ge|IND_R^k(x_i)|$.

定义 **6**. 对于决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$.

(1) $\forall r_j\in R, IND_r^k=\bigcup_{i=1}^{n} IND_r^k(x_i), IND_r^k$ 包含着所有 x_i 在$|INP_r(x_i)|$上的 k-统计不可辨识对,我们称 IND_r^k 为属性 r 的不可辨识集.

(2) $\forall r_j\in R, |IND_r^k|$ 是 IND_r^k 中元素的个数,称为属性 r 的 k-统计不可辨识.

(3) $\forall Q\subseteq R, IND_Q^k=\bigcap_{r\in Q} IND_r^k, IND_Q^k$ 包含着属性集 Q 中每一个属性的 k-统计不可辨识对,称为属性集 Q 上不可辨识集.

(4) $\forall Q\subseteq R, |IND_Q^k|$ 是 IND_Q^k 中元素的个数,称为属性集 Q 的 k-统计不可辨识度.

由定义 6 可以看出,不可辨识度越大,该属性集就有越多无法分辨的示例对.

性质 **2**.

(1) 如果 $P\subseteq Q\subseteq R$,那么 $IND_P^k\supseteq IND_Q^k\supseteq IND_R^k$;

(2) 如果 $P\subseteq Q\subseteq R$,那么 $|IND_P^k|\geqslant|IND_Q^k|\geqslant|IND_R^k|$.

定义 **7**(统计属性约简). 对于决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$,$R=\{r_1,r_2,\ldots,r_m\}$.如果子集 $Q\subseteq R$ 满足下列条件,那么我们就说子集 Q 是 R 的一个统计属性约简.

(1) $IND_Q^k=IND_R^k$ 或者 $|IND_Q^k|=|IND_R^k|$;

(2) $\forall r\in R, |IND_{Q-\{r\}}^k|>|IND_Q^k|$.

定义 7 说明了本文中统计属性约简是保持统计不可辨识的信息含量不变的最小属性子集.

定理 **1**. 给定决策表,其中,$U=\{x_1,x_2,\ldots,x_n\}$,$S=\{s_1,s_2,\ldots,s_a\}$表示随机从全集 U 中抽样得到的一个示例集合.当 $a>\dfrac{t^2\sigma^2}{d^2\overline{Y}^2}$ 时,满足以下条件,则至少以 e 的置信度,使得:

$$\left|\frac{\dfrac{ind(r_i)}{a}-\dfrac{IND(r_i)}{n}}{\dfrac{IND(r_i)}{n}}\right|<d.$$

(1) $IND(r_i)$是 r_i 在 U 上的不可辨识度;

(2) $ind(r_i)$是 r_i 在 S 上的不可辨识度;

(3) $\sigma^2,\overline{Y}$ 分别是 r_i 在 U 上的不可辨识度的方差与均值,由预调研得到;

(4) t 是标准正态分布的 $1-e$ 双侧分位数;

(5) d 是绝对误差.

证明:对于简单随机抽样,$\bar{y}$ 的方差为 $V(\bar{y})=\dfrac{N-n}{Nn}\sigma^2$,其中,$\sigma^2=\dfrac{1}{N-1}\sum_{i=1}^{N}(Y_i-\overline{Y})^2$,$N$ 为总体个数,n 为样本个数;$\bar{y}$ 的相对误差记为 $d=\dfrac{\sqrt{V(\bar{y})}}{\overline{Y}}$.将 $V(\bar{y})$ 代入相对误差,可得 $d=\dfrac{t}{\overline{Y}}\sqrt{\dfrac{N-n}{Nn}\sigma^2}$,从而有 $n=\dfrac{Nt^2\sigma^2}{Nd^2\overline{Y}^2+t^2\sigma^2}$ [19].当 N 足够大时,$n=\dfrac{t^2\sigma^2}{d^2\overline{Y}^2}$. □

定理 1 保证了在样本集 S 中计算的 k-统计不可辨识度可以有很高的置信度,以很小的误差拟合全集 U 中的 k-统计不可辨识度.

前面我们回顾了经典的下近似、属性约简算法的不足——无法应用到大规模数据集中.因此,为了满足现实应用中的大规模数据约简的需要,我们需要基于统计粗糙集模型,对经典的属性约简方法进行改进.而缩小属性约简的计算范围,可以提升运算的效率——在海量数据下,任何计算,其范围是全集的时候,必然会导致运算量过大.

为了进一步对约简的效率进行提升,在约简时用抽样集代替全集,计算每个属性的重要性排序.并且,应用定理 1 保证在样本集中计算出的重要性排序,有很高的置信度,以很小的误差拟合全集中的重要性排序.从而以

很小的代价计算出统计属性约简.

本节中提出的统计辨识信息相对于经典模糊粗糙信息度量的优势在于:原本涉及到大规模计算(全集范围)的部分,都巧妙地以随机抽样得到的小容量样本代替.由于随机抽样得到的样本数量和全集相比数量极小,更值得一提的是,随着全集数量的增大,样本数量并不会显著增大,因此,全集数量越大,统计粗糙约简的优势就越大.

3 基于随机抽样的约简算法

本节针对统计属性约简设计了基于随机抽样的约简算法.我们将计算统计属性约简分成两步:第 1 步是根据依赖度函数对属性重要性进行排序,第 2 步是根据所得到的顺序进行约简.

3.1 对属性进行排序

对于决策表 $DS=(U,R\cup D)$,其中,$U=\{x_1,x_2,\ldots,x_n\}$.$|IND_r^k|$是 r 上的 k-统计不可辨识度.从第 2 节对 k-统计不可辨识度的描述中可知:$|IND_r^k|$越小,属性 r 可以辨识的元素越多,它的重要性可能就越高.

算法 **1**. 对属性排序.

输入:$DS=(U,R\cup D)$,$U=\{x_1,x_2,\ldots,x_n\}$,$R=\{r_1,r_2,\ldots,r_m\}$,排序属性队列 $SR\leftarrow\varnothing,\eta,e$.

输出:SR.

第 1 步:计算出样本数量 $a=\max_P\left(\frac{t^2\sigma^2}{\eta^2\bar{Y}^2}\right)$,随机的从全集 U 中抽取 a 个样本,得到 $S=\{s_1,s_2,\ldots,s_a\}$.

第 2 步:计算 $IND_{r_j}^k=\bigcup_{i=1}^{n}IND_{r_j}^k(s_i)$,$\forall j\in[1,2,\ldots,m]$:

(2.1) 计算 x_i 的限定区域 k-limit;

(2.2) 计算 $\lambda_i'=\underline{R_{\vartheta}^k}[x_i]_D(x_i)$;

(2.3) $i\leftarrow i+1$.

第 3 步:$q\leftarrow 0$.

第 4 步:当 $q<m$,执行:

(4.1) $q\leftarrow q+1$;

(4.2) $R'=R-SR$;

(4.3) $\forall r_j\in R'$,计算 $IND_{SR\cup r_j}^k$;

(4.4) $SR\leftarrow SR\cup sr_q$,其中,$|IND_{SR\cup sr_q}^k|=\min_{r_{j\in R'}}(|IND_{SR\cup r_j}^k|)$.

第 5 步:输出 $SR=\{sr_1,sr_2,\ldots,sr_m\}$.

在本节算法中,我们使用随机抽样所得到的样本,而非全集来决定属性约简的顺序.同时,我们使用统计不可分辨来衡量属性的重要性.定理 1 保证在很高的置信度下,以很小且可控的误差用样本集中的数据来拟合全集中的数据.

3.2 计算属性约简

算法 **2**. 得到 k-统计属性约简.

输入:$DS=(U,R\cup D)$,$U=\{x_1,x_2,\ldots,x_n\}$,$R=\{r_1,r_2,\ldots,r_m\}$,$SR=\{sr_1,sr_2,\ldots,sr_m\}$,$k$.

输出:RED.

第 1 步:$Red\leftarrow sr_1$;$q\leftarrow 1$.

第 2 步:对于$\forall x_i\in U$,计算 $IND_{Red}^k(x_i)$:

(2.1) 计算 x_i 的限定区域 k-limit;

(2.2) 计算 $\lambda_i'=\underline{R_{\vartheta}^k}[x_i]_D(x_i)$;

(2.3) $i\leftarrow i+1$.

第 3 步:当 $|IND^k_{Red\cup sr_{q+1}}|-|IND^k_{Red}|>0$ 和 $q<m$ 时,执行:

$q\leftarrow q+1, Red\leftarrow Red\cup sr_q$.

第 4 步:输出 RED.

这里,我们需要强调以下两点.

第一,我们使用不可辨识对来存储不可辨识关系.

与传统的辨识矩阵关系不同的是,我们只存储第 1 个属性,也就是最重要的属性的不可辨识对.在不断约简这些不可辨识对的时,只有尚存的不可辨识对,用新的属性去测试是否可以被辨识.这种做法有两个好处:其一是只需要存储一个属性下的不可辨识对,减少了空间消耗;其二是没有无用的操作,所有的比较操作都是有效的.

第二,我们使用位图来存储这些不可辨识对.

位图可以用 1bit 存储一个不可辨识对,这样可以节省 32 倍的空间.在比较时,我们也是用位操作来进行操作,由于位操作比正常的运算快很多,因此这样也可以有更好的时间空间性能.

需要指出的是,本文提出的算法的计算复杂度为 $O(\max(2k|R|^2|U|,|R|^2|S|^2))$,而经典算法的计算复杂度为 $O(|R|^2|U|^2)$.因此当 $2k$ 远小于$|U|$时,我们提出的基于随机抽样的算法比经典算法节约了大量的时间和空间.

4 数值实验

在本节中,随机抽样算法(简记为 Sampling)与两种经典启发式属性约简算法,即依赖度函数算法(简记为 Dependency)和辨识矩阵算法(简记为 Matrix)进行对比.我们从 3 个方面来评估 3 种算法的性能:算法的时间效率、空间效率和算法得到约简的性能.

4.1 实验环境及数据集

所有的实验均是在 Linux 下,由 C++编码完成.实验所使用的硬件参数是,CPU 为 Intel® Xeon® CPU ES-2670 2.6GHz.

在本实验中,几个 UCI 数据集用来验证每一个算法的效率[20],具体数据集参数见表 1.

Table 1 Description of the datasets

表 **1** 数据集描述

	数据集名称	名称缩写	示例个数	属性个数
1	Wine quality-red	WQ-Red	1 599	11
2	Image segmentation	Seg	2 310	19
3	Waveform	Waveform	5 000	21
4	Letter recognition	LR	20 000	16
5	Wine quality-white	WQ-white	4 898	11
6	MAGIC gamma telescope	Magic	19 020	10
7	Shuttle	Shuttle	58 000	9
8	Mirflickr*	Mir	25 000	64

为了更好地展示计算的时间和空间,在数据集的示例数逐渐增加的情况下,观察 3 种算法的运算时间与所占用的空间变化趋势.

4.2 实验结果与分析

4.2.1 属性约简时间消耗对比

在本节中,约简时间消耗随着数据集大小的变化趋势绘制在图 1 中.

在图 1 中,基于依赖度函数的部分时间消耗值由于过于庞大,以至于超过图表的顶部.这是因为基于依赖度函数的时间复杂度较高,在大规模数据集上其时间消耗过于庞大,甚至达到了某限定时间内无法收敛的地步.

在图 1 中,随着数据集的增大,基于依赖度的算法时间消耗急剧上升,辨识矩阵算法时间消耗上升较快,随机抽样算法则上升速度十分平缓,而且随机抽样算法的时间消耗总是远小于另外两种算法的时间消耗.这体现了

随机抽样算法在进行属性约简时的时间优越性,特别是在大规模数据集下的计算时间可控.

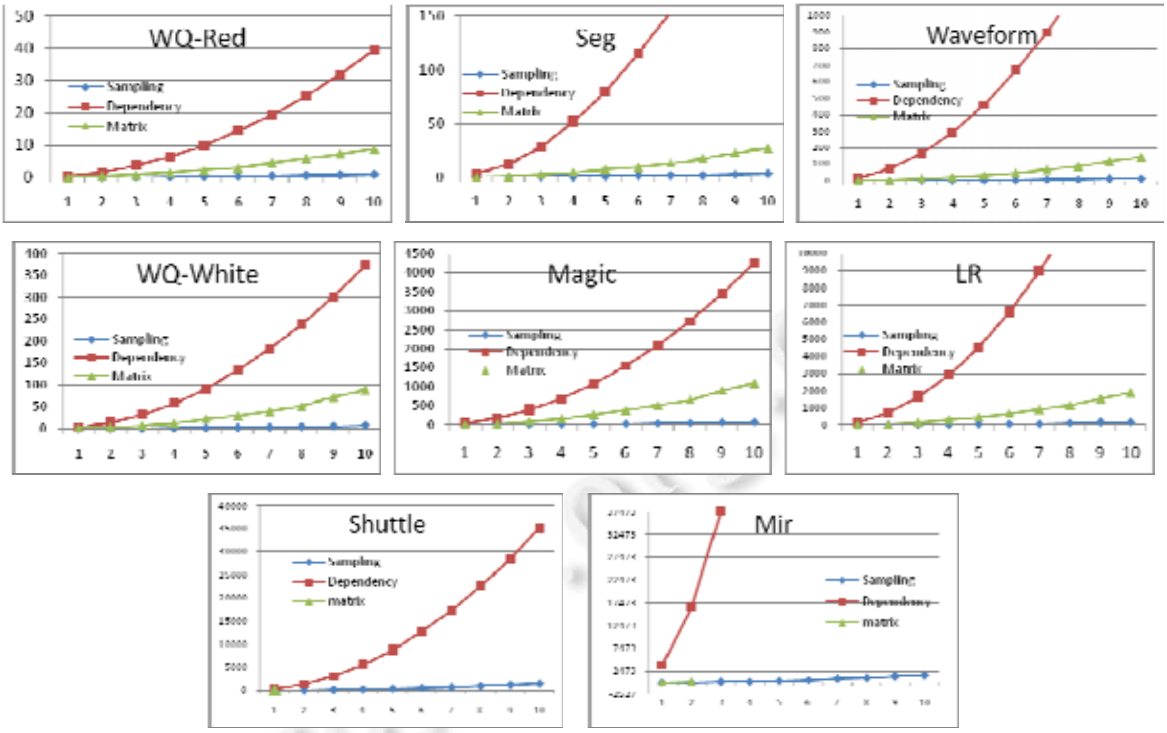

Fig.1 Trend of computational time of Sampling, Dependency and Matrix

图 1 Sampling 算法、Dependency 算法和 Matrix 算法的计算时间消耗趋势

4.2.2 空间消耗对比

为什么传统的辨识矩阵约简算法无法应用到大规模数据集呢?一个主要的因素就是太多的空间消耗.因此,本节对比了随机抽样算法和基于辨识矩阵算法的空间消耗势.

表 2 给出两种算法在 8 个数据集上进行约简的空间消耗对比.

Table 2 Space consuming comparison of reduction between Matrix and Sampling

表 2 Matrix 约简算法和 Sampling 约简算法的空间消耗比较

	Sampling (KB)	Matrix (KB)	Ratio (Matrix/Sampling)
LR	107 520	80 732 628	750.861 5
Seg	3 072	1 255 931	408.831 71
Waveform	9 216	5 868 630	636.787 11
WQ-red	1 128	82 944	73.531 915
WQ-white	8 192	4 086 282	498.813 72
Magic	96 256	50 328 484	522.860 75
Mir	46 080	–	–
Shuttle	16 384	–	–

从表 2 中我们可以清晰地看到:随机抽样算法的空间消耗非常小,所降低的空间消耗量级约在 2 个数量级到 3 个数量级之间;并且,可以隐约看出,数据量越大,随机抽样算法所降低的量级也越多.图 2 给出了随机抽样算法和辨识矩阵算法的空间消耗随着数据集大小变化的趋势.

从图 2 中可以看出,基于辨识矩阵的算法与随机抽样相比,有着很高的空间消耗.事实上,很多时候,当数据集的大小不断增长时,辨识矩阵算法的空间消耗急剧增加,而随机抽样算法的消耗较为平稳.这是因为随机抽样的样本集大小并不会随着全集的增加急剧增大,而是基本保持稳定,所以使得整体空间消耗较为稳定.

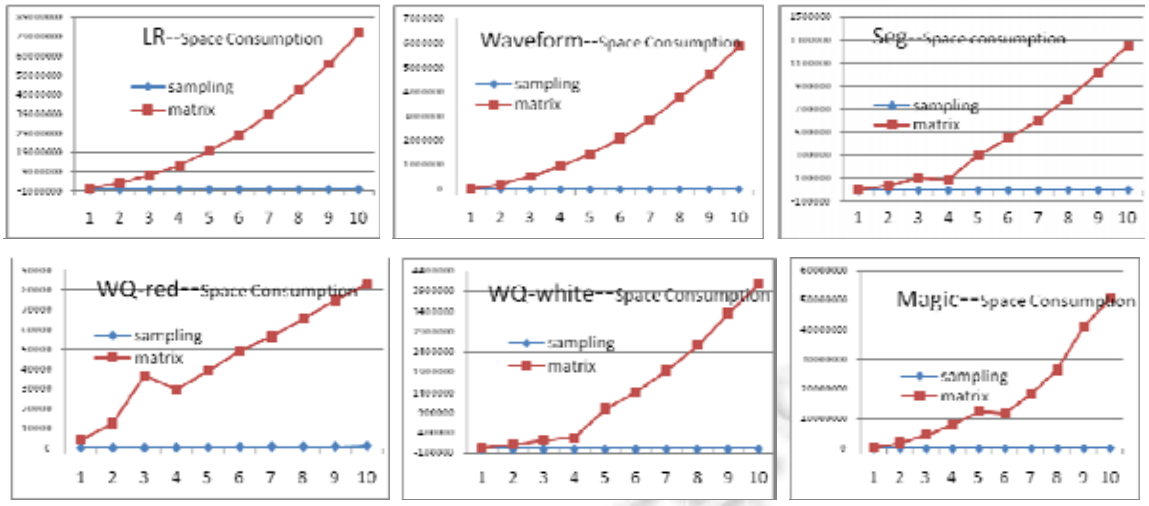

Fig.2 Space computing trend of reduction between Matrix and Sampling

图 2 Matrix 约简算法和 sampling 约简算法的空间消耗趋势

4.2.3 约简效果对比

我们将用 KNN 算法作为分类器,对原始数据集、依赖度算法约简后数据集、辨识矩阵算法约简后数据集、随机抽样算法约简后数据集进行分类,观察分类精度的变化.所有数据集在进行分类时,都是用了 10-交叉验证.

从表 3 中可以看出,3 种算法在约简后剩下的属性数量上大致相同;而在精度上,基于随机抽样的算法与其他两者相比,分类精度略有下降,但是下降不多.这是由于在随机抽样过程中,不可避免地有一些信息量的损失,但是定理 1 保证了其损失的精度在可控范围.

Table 3 Reduction performance comparison of Sampling, Dependency and Matrix

表 3 Sampling 算法、Dependency 算法和 Matrix 算法的约简性能比较

		All	Matrix	Dependency	Sampling
Waveform	Attr. No.	21	13	14	12
	Accuracy	0.857	0.822	0.825	0.824 04
Seg	Attr. No.	19	11	10	8
	Accuracy	0.904	0.904	0.901	0.900
WQ-red	Attr. No.	11	6	6	6
	Accuracy	0.653	0.634	0.616	0.621
WQ-white	Attr. No.	11	7	7	7
	Accuracy	0.681	0.653	0.667	0.653
Magic	Attr. No.	10	6	5	–
	Accuracy	0.829	0.823	0.814	0.775
LR	Attr. No.	16	9	10	11
	Accuracy	0.930	0.920	0.925	0.897
Average	Attr. No.	**14.667**	**8.667**	**8.667**	**7.333**
	Accuracy	**0.809**	**0.793**	**0.791**	**0.778**

综上可得,随机抽样算法虽然在某些时候对于约简精度会有小幅度但可控的下降,但其在属性约简时间效率、属性约简空间效率方面,相对于传统的属性约简算法都有着大幅度的提升.

4.2.4 抽样下近似的稳定性分析

基于抽样的约简的稳定性取决于基于抽样的不可辨识度的稳定性,本节我们实验验证抽样下近似的稳定性.我们实验比较了真实下近似与抽样下近似之间的差异性.实验结果汇总至表 4.

Table 4 Absolute error |*P*−*p*| of the approximate ratio on sample (*p*) and population (*P*)

表 **4** 不可辨识度在样本(*p*)与总体(*P*)上的差异(绝对误差(|*P*−*p*|))的统计稳定性

Upper quantile (*t*)	Absolute error (*d*)	|*P*−*p*|						|*P*−*p*|<*d*?
		Shuttle	LR	Seg	Waveform	WQ	Mirflickr	
1.65	0.05	**0.009**	**0.011**	**0.002**	**0.015**	**0.008**	**0.000**	Yes, all less than 0.05
1.65	0.025	**0.006**	**0.003**	**0.004**	**0.001**	**0.000**	**0.010**	Yes, all less than 0.025
1.96	0.05	**0.005**	**0.004**	**0.013**	**0.001**	**0.004**	**0.004**	Yes, all less than 0.05
1.96	0.025	**0.001**	**0.008**	**0.002**	**0.002**	**0.000**	**0.002**	Yes, all less than 0.025
2.326	0.05	**0.010**	**0.008**	**0.003**	**0.001**	**0.001**	**0.013**	Yes, all less than 0.05
2.326	0.025	**0.001**	**0.003**	**0.001**	**0.004**	–	**0.005**	Yes, all less than 0.025

基于定理 1,给定绝对误差 *d* 与置信度 *e* 的双侧分位数 *t*,即可得到样本容量的下界.基于此下界抽样获得样本上的不可辨识度在样本上与总体上的真实不可辨识度的差值,即|*P*−*p*|,统计上应小于 *d*.表 4 给出了在给定对误差 *d* 与置信度 *e* 的双侧分位数 *t* 后,样本上的不可辨识度在样本上与总体上的真实不可辨识度的差值,即|*P*−*p*|.表 4 的最后一列显示确实如此.这说明,不可辨识度在样本上与总体上是统计稳定的.因而,基于样本不可辨识度所获得的约简也是具有统计稳定性的.

5 总 结

现有的模糊粗糙集约简方法由于其基础理论复杂度的桎梏,不可避免地使得约简效率极低,无法应用到大规模数据集上.随机抽样则是一种可以极大地减少运算量的统计学方法,因此,本文将随机抽样引入到经典的模糊粗糙约简理论中,设计了一种基于随机抽样的粗糙约简算法.通过数值实验发现:随机抽样算法虽然在某些时候对于约简精度会有小幅度但可控的下降,但其在属性约简时间效率、属性约简空间效率方面,相对于传统的属性约简算法都有着大幅度提升,并且这种提升随着数据集的增大更加显著.因此,随机抽样算法弥补了传统属性约简方法无法应用到大规模数据集上的缺陷,适合应用于大规模数据集.

References:

[1] Zadeh LA. Fuzzy sets. Information Control, 1965,8:338−353.

[2] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. Int'l Journal of General Systems, 1990,17:191−208. [doi: 10.1080/03081079008935107]

[3] Chen DG, Wang XZ, Yeung DS, Tsang ECC. Rough approximations on a complete completely distributive lattice with applications to generalized rough sets. Information Sciences, 2006,176:1829−1848. [doi: 10.1016/j.ins.2005.05.009]

[4] Morsi NN, Yakout MM. Axiomatics for fuzzy rough sets. Fuzzy Sets and Systems, 1998,100(1-3):327−342. [doi: 10.1016/S0165-0114(97)00104-8]

[5] Wu WZ, Zhang WX. Constructive and axiomatic approaches of fuzzy approximation operators. Information Sciences, 2004,159: 233−254. [doi: 10.1016/j.ins.2003.08.005]

[6] Wu WZ, Zhang M, Li HZ, Mi JS. Knowledge reduction in random information systems via Dempster-Shafer theory of evidence. Information Sciences, 2005,174:143−164. [doi: 10.1016/j.ins.2004.09.002]

[7] Jensen R, Shen Q. Fuzzy-Rough attributes reduction with application to web categorization. Fuzzy Sets and Systems, 2004,141(3): 469−485. [doi: 10.1016/S0165-0114(03)00021-6]

[8] Hu QH, Yu DR, Xie ZX. Information-Preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters, 2006,27:414−423. [doi: 10.1016/j.patrec.2005.09.004]

[9] Bhatt RB, Gopal M. On fuzzy rough sets approach to feature selection. Pattern Recognition Letters, 2005,26:1632−1640. [doi: 10.1016/j.patrec.2004.09.044]

[10] Slowinski R, Vanderpooten D. Similarity relation as a basis for rough approximations. In: Wang PP, ed. Proc. of the Advances in Machine Intelligence and Soft-Computing. Durham: Department of Electrical Engineering, Duke University, 1997. 17−33.

[11] Tsang ECC, Chen DG, Yeung DS, Wang XZ, Lee JWT. Attributes reduction using fuzzy rough sets. IEEE Trans. on Fuzzy Systems, 2008,16(5):1130−1141. [doi: 10.1109/TFUZZ.2006.889960]

[12] Zhao SY, Tsang ECC, Chen DG. The model of fuzzy variable precision rough sets. IEEE Trans. on Fuzzy System, 2009,17(2): 451−467. [doi: 10.1109/TFUZZ.2009.2013204]

[13] Chen Y, Zhao SY, Chen H, Li CP, Sun H. Statistical rough sets. Ruan Jian Xue Bao/Journal of Software, 2016,27(7):1645−1654 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5036.htm [doi: 10.13328/j.cnki.jos.005036]

[14] Yeung DS, Chen DG, Tsang ECC, Lee J WT, Wang XZ. On the generalization of fuzzy rough sets. IEEE Trans. on Fuzzy System, 2005,13(3):343−361. [doi: 10.1109/TFUZZ.2004.841734]

[15] Morsi NN, Yakout MM. Axiomatics for fuzzy rough sets. Fuzzy Sets System, 1998,100(1-3):327−342. [doi: 10.1016/S0165-0114(97)00104-8]

[16] Radzikowska M, Kerre EE. A comparative study of fuzzy rough sets. Fuzzy Sets System, 2002,126(2):137−155. [doi: 10.1016/S0165-0114(01)00032-X]

[17] An S, Hu QH, Yu DR, Liu JF. Soft minimum-enclosing-ball based robust fuzzy rough sets. Funmamenta Informaticae, 2012,115 (2-3):189−202. [doi: 10.3233/FI-2012-649]

[18] Hu QH, Zhang L, An S, Zhang D, Yu DR. On robust fuzzy rough set models. IEEE Trans. on Fuzzy System, 2012,20(4):636−651. [doi: 10.1109/TFUZZ.2011.2181180]

[19] Jin YJ, Du ZF, Jiang Y, eds. Sampling Technique. 4th ed., Beijing: China Renmin University Press, 2015 (in Chinese).

[20] http://www.ics.uci.edu/~mlearn/MLRepository.html

附中文参考文献:

[19] 金勇进,杜子芳,蒋妍,编著.抽样技术.第 4 版,北京:中国人民大学出版社,2015.

[13] 陈俞,赵素云,陈红,李翠平,孙辉.统计粗糙集.软件学报,2016,27(7):1645−1654. http://www.jos.org.cn/1000-9825/5036.htm [doi: 10.13328/j.cnki.jos.005036]

陈俞(1992－),男,安徽六安人,硕士,主要研究领域为数据挖掘,模糊粗糙集.

赵素云(1979－),女,博士,副教授,CCF 专业会员,主要研究领域为基于模糊集,粗糙集理论,概率统计论的不确定信息处理方法研究.

李雪峰(1994－),男,本科生,主要研究领域为数据库,数据挖掘.

陈红(1965－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据仓库与数据挖掘,传感器网络数据管理,流数据管理.

李翠平(1971－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据仓库和数据挖掘,社会网络分析.