软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(7):1645−1654 [doi: 10.13328/j.cnki.jos.005036]

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

统计粗糙集*

陈 俞[1,2], 赵素云[1], 陈 红[1,2], 李翠平[1,2], 孙 辉[1,2]

[1](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)

[2](中国人民大学 信息学院 计算机系,北京 100872)

通信作者: 赵素云, E-mail: zhaosuyun@ruc.edu.cn

摘 要: 现有的模糊粗糙集方法,由于其基础理论复杂度的桎梏,无法应用到大规模数据集上.考虑到随机抽样是一种可以极大地减少运算量的统计学方法,将随机抽样引入到经典的模糊粗糙集理论中,建立了一种统计粗糙集模型.首先,提出了统计上、下近似的概念,它相比经典模糊粗糙集模型的优势在于,以随机抽样得到的小容量样本代替了大规模全集,从而显著降低了计算量.而且,随着全集数量的增大,抽样样本数量并不会显著增大.此外,还讨论了统计上、下近似的性质,揭示统计上、下近似和经典上、下近似之间的关系.并且,提出了一个定理,该定理保证了统计下近似与经典下近似的取值统计误差在允许的范围内.最后,通过数值实验验证了统计下近似在计算时间上的显著优势.

关键词: 随机抽样;近似算子;统计粗糙集;模糊粗糙集

中图法分类号: TP311

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/5036.htm

Statistical Rough Sets

CHEN Yu[1,2], ZHAO Su-Yun[1], CHEN Hong[1,2], LI Cui-Ping[1,2], SUN Hui[1,2]

[1](Key Laboratory of Data Engineering and Knowledge Engineering, MOE (Renmin University of China), Beijing 100872, China)

[2](Department of Computer Science, School of Information, Renmin University of China, Beijing 100872, China)

Abstract: This paper introduces random sampling into traditional fuzzy rough methods and proposes a random sampling based statistical rough set model. The work focuses on how to bring random sampling into traditional rough set. First, random sampling is used to propose a concept of *k*-limit, which can dramatically reduce the amount of computation during the computing of lower approximation value. Then, statistical upper and lower approximation is formulated. By mathematical reasoning, sufficient theorem and proof are used to valid the reliability of new model. Finally, numerical experiments illustrate the efficiency of the proposed statistical rough sets.

Key words: random sampling; approximate operator; statistical rough set; fuzzy rough set

当今,社会信息化和网络化的发展导致数据呈爆炸式增长.大数据的涌现,使人们处理计算问题时获得了前

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316205); 国家高技术研究发展计划(863)(2014AA015204); 国家自然科学基金(61532021, 61202114, 61272137); 中国人民大学科学研究基金(15XNLQ06)

Foundation item: National Basic Research Program of China (973) (2012CB316205); National High Technology Research and Development Program of China (863) (2014AA015204); National Natural Science Foundation of China (61532021, 61202114, 61272137); Research Funds of Renmin University of China (15XNLQ06)

收稿时间: 2015-09-26; 修改时间: 2016-01-12; 采用时间: 2016-02-22; jos 在线出版时间: 2016-03-22

CNKI 网络优先出版: 2016-03-22 13:23:31, http://www.cnki.net/kcms/detail/11.2560.TP.20160322.1323.005.html

所未有的大规模样本,但同时也不得不面对更加复杂的大数据处理:数据复杂且规模巨大[1−3].在现实的诸多应用中,数据的表现形式存在着诸多不确定性,比如说,数据的缺失、收集不准确、数据集粒度过于粗糙等都会导致不确定性数据的产生.在此背景下,模糊集、粗糙集等处理不确定性数据的理论和方法应运而生,并成为处理不确定性数据的有效工具[4,5].

粗糙集(rough set)首先由 Pawlak 提出,它主要是研究由不可分辨引起的不确定信息的智能处理的理论.目前,粗糙集理论已成为一种重要的不确定信息处理技术[6,7],该理论已经在机器学习和知识发现、数据挖掘、决策支持与分析等方面得到广泛应用[8].

模糊粗糙集理论中,作为模糊集与粗糙集的融合理论,兼顾数据的模糊性与不可分辨性[9−14].模糊粗糙集利用模糊粗糙近似算子将被隐藏在不确定数据中的信息,按照确定的规则呈现出来.模糊粗糙集可以更好地处理数据中的不确定性,但是该理论不能处理大规模数据.这是由于模糊粗糙集所依赖的理论核心——模糊粗糙近似算子的复杂度,是元素个数的平方量级[9,15,16].无法解决这个问题,粗糙集的研究就无法应用到大数据的背景当中去.因此,在本文中,我们将就这个问题进行研究,希望能够通过某种方式,降低其复杂度,使不确定性研究能够恰当地应用到大数据中去.

本文将随机抽样引入了传统的模糊粗糙集中,对近似逼近算子的效率进行了大幅提升.本文的主要内容为:巧妙地引入随机抽样,构建统计粗糙集模型.首先,我们通过引入随机抽样提出了一个限定区域 k-limit 的概念,限定区域可以极大地缩小下近似的计算规模.然后,基于 k-limit 我们又提出统计上、下近似概念.并且,在引入随机抽样的同时,用充分完备的定理和证明,验证了新概念和理论的可靠性,从而构建出统计粗糙集模型.

1 模糊粗糙集及约简

1.1 模糊粗糙集模型

假设全集 U 是一个有着有限个数元素的非空集合,记做 $U=\{x_1,x_2,...,x_n\}$.每一个元素都有着一系列条件属性,记做 $R=\{r_1,r_2,...,r_m\}$; 以及决策属性 D.于是$(U,(R\cup D))$被称作是一个决策系统 DS.

对于每一个 $P\subseteq R$,我们将二维关系 $P(x,y)$称为 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:$P(x,y)=P(y,x)$,以及 T-传递性:$P(x,y)\geqslant T(P(x,z),P(z,y))$.简单来说,$P$ 是用来代表其相似关系的,$F(\bullet)$用来代表模糊幂集.

模糊粗糙集是用来将模糊集和粗糙集结合起来的工具.它在文献[1]中首次被 Dubois 和 Prade 提出.然后其细节在文献[10,15,16]中被深入地加以研究.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.

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

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

在文献[9,10]中介绍了经典模糊粗糙集的下近似,实际上是该样本到不同类别样本之间的最小距离,而上近似则是到相同类别样本之间的最大相似度.模糊粗糙集的上、下近似的几何意义有助于我们设计基于随机抽样的统计粗糙集.

1.2 常见约简算法的实验比较分析

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.详细的算法比较请参考文献[15].

本实验中所使用的测试环境如下:

(1) 硬件:Intel(R) Xeon(R) CPU ES-2670,2.6GHz;

(2) 内存:252GB;

(3) 编程语言:C++;

(4) 操作系统:Linux.

实验中使用的数据库均从 UCI 选出.这些数据集的详细信息在表 1 中给出.在实验中使用了 10-交叉验证.

Table 1 Information of selected datasets

表 1 所选取数据库的详细信息

数据库	记录条数	属性数	类别
Image segmentation	2 310	18	7
Letter recognition	20 000	16	26
Wine quality-red	1 599	11	10
Waveform database generator	5 000	21	3

本节我们将从时间复杂度和空间复杂度上对比辨识矩阵算法和依赖度算法的可扩展性.在给出计算复杂度的比较之后,再给出具体的实验比较.

基于依赖度的约简算法的时间复杂度为 $O(|U|^2\times|R|^3)$,空间复杂度为 $O(|U|^2\times|R|)$.基于辨识矩阵的约简算法的时间复杂度为 $O(|U|^2\times|R|^2)$,空间复杂度为 $O(|U|^2\times|R|)$.由此可以看出,两种算法的空间复杂度相当,但是基于辨识矩阵的算法时间复杂度要明显小于基于依赖度的约简算法的时间复杂度.下面的实验比较说明了这一点.

(1) 运行时间的对比

表 2 显示了两种约简算法的运行时间.表 2 说明,依赖度函数的运行时间消耗要远大于辨识矩阵,特别是当数据集记录数增大,或者是属性数增多的时候.

Table 2 The comparison of running time

表 2 运行时间对比

数据库	依赖度方法(s)	辨识矩阵方法(s)
Image segmentation	8.74	2.62
Wine quality-red	24.73	2.73
Waveform database generator	1081.7	49.46
Letter recognition	9 107.22	750.32

为了更加清晰地比较两种算法的运行时间,我们用递增的数据集来比较两种算法.具体情况如图 1 所示,其中,横轴从 1~10,是将数据集均匀分成了 1~10 份,取其中的一份进行实验,逐渐增大,直到取 10 份进行实验,此时就是全部数据集.

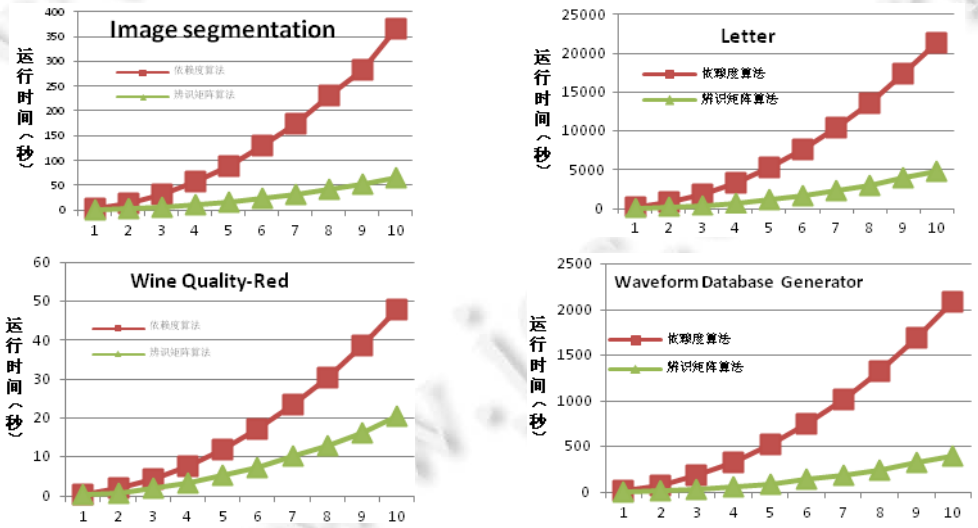


Fig.1 The chart of running time comparison

图 1 运行时间对比图

在图 1 中,我们可以清晰地看到,随着数据集的增大,基于依赖度函数的算法运行时间急剧增大,而辨识矩阵

算法增加得则相对较慢.从这些图中我们可以看出,在大规模数据集上,辨识矩阵算法比依赖度算法更有优势.

(2) 运行空间的对比

表 3 比较了两种约简算法运行空间消耗的情况.两种算法在较大规模数据集上都有着很高的空间消耗.

Table 3 The comparison of running space

表 3 运行空间对比

数据库	依赖度方法(MB)	辨识矩阵方法(MB)
Wine quality-red	411	404
Image segmentation	819.2	512
Waveform database generator	5 120	5 120
Letter recognition	104 990	82 389

由于依赖度的属性约简算法通常有着高空间和高时间复杂度,而虽然基于辨识矩阵的约简算法提升了时间方面的效率,但是由于空间消耗过大,仍然无法应用到大规模数据集上.因此,为了能够将属性约简应用到大规模数据集上,新方法的提出便势在必行.

2 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.假设 y 恰好就是 x 取到异类点(与 x 所属的类别不同的点)最小值的记录.很明显,x 和 y 必然有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,将整个数据库在该维度(属性)上排序,那么 x 和 y 一定会离得比较接近.于是我们提出两个概念,k-neighbor 和 k-limit,用来定义某元组的邻居.

定义 2.1(*k*-neighbor). 给出一个随机变量 X 和它的 n 个样本升序排列: $\{x_1,x_2,...,x_n\}$, 那么 x_i 的 k-neighbor 可以表达成:

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

定义 2.2(*k*-limit). 给出一个决策表$DT=(U,R,D)$,其中, $U=\{x_1,x_2,...,x_n\}, R=\{r_1,r_2,...,r_m\}$. 对于所有属性,$1\leqslant k\leqslant \arg(n/2)$将 x_i 在每一个属性上的 k-neighbor 集中到一个集合中.这个集合就叫作 x_i 的 k-limit(限定区域).这里的 k,就是限定长度.

基于 k-limit,我们可以定义一对新的上、下近似算子.

定义 2.3. 给出一个决策表$DT=(U,R,D)$,其中, $U=\{x_1,x_2,...,x_n\}, R=\{r_1,r_2,...,r_m\}$, $F(U)$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:

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

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

性质 2.1.

(1) K-统计下近似算子不小于经典的下近似算子;

(2) K-统计上近似算子不大于经典的下近似算子.

证明:对于$\forall x\in U, k\leqslant \text{int}(n/2)\Rightarrow k\text{-limit}(x)\subseteq U\Rightarrow$

$$\underline{R_{\vartheta}^{k}}A(x)=\inf_{u\in k\text{-limit}(x)}\vartheta(R(u,x),A(u))\geqslant\inf_{u\in U}\vartheta(R(u,x),A(u))=\underline{R_{\vartheta}}A(x);$$

$$\overline{R_{T}^{k}}A(x)=\sup_{u\in k\text{-limit}(x)}T(R(u,x),A(u))\leqslant\sup_{u\in U}T(R(u,x),A(u))=\overline{R_{T}}A(x);$$

$$\underline{R_{S}^{k}}A(x)=\inf_{u\in k\text{-limit}(x)}S(N(R(u,x)),A(u))\geqslant\inf_{u\in U}S(N(R(u,x)),A(u))=\underline{R_{S}}A(x);$$

$$\overline{R_{\sigma}^{k}}A(x)=\sup_{u\in k\text{-limit}(x)}\sigma(N(R(u,x)),A(u))\leqslant\sup_{u\in U}\sigma(N(R(u,x)),A(u))=\overline{R_{\sigma}}A(x).$$ □

性质 2.2. 当 $k\to\text{int}(n/2)$ 时,k-统计近似算子的极限也趋向于经典的近似算子.

定义 2.3 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 2.1 和性质 2.2 展示出,在 k 足够大时,k-统计近似算子可以逼近经典的近似算子.但是如何最小化限定距离 k 是一个问题.考虑到随机抽样可以做到在统计上无偏估计经典上、下近似值,并可极大地缩小计算量.基于随机抽样,可确定限定距离 k 的大小.

定义 2.4. 对于 $A=\{\lambda_1,\lambda_2,...,\lambda_n\}, A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$, 其中, $\lambda_i,\lambda_i'\in[0,1],\forall i=1,2,...,n$. 如果

$$|\lambda_i-\lambda_i'|\leqslant\delta(0\leqslant\delta\leqslant 1),$$

我们就说 (λ_i,λ_i') 是一个δ-近似对.

定义 2.5. 对于 $A=\{\lambda_1,\lambda_2,...,\lambda_n\}, A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$, 其中, $\lambda_i,\lambda_i'\in[0,1],\forall i=1,2,...,n$. 如果集合 A 和 A'中δ-近似对的数量是 p 对,那么我们称 p/n 是 A 和 A'之间的相似度,记做 $p=\gamma_{A\approx A'}$.

用样本比例 p 作为总体的比例估计 P,则 p 是 P 的无偏估计.下面的性质说明了这一点.同时,我们还给出了 p 和 P 的方差的统计性质.

性质 2.3. 对于简单随机抽样,$E(P)=E(p)=P$.

证明略,详细论证请见文献[17]第 30 页定理 2.1 及其推论 2.2.

性质 2.4. 对于简单随机抽样,p 的方差为 $V(p)=\frac{PQ}{a}\frac{(n-a)}{(n-1)}, Q=1-P$, 并且 $V(p)=\frac{PQ}{a}\frac{(n-a)}{(n-1)}, Q=1-P$ 是 $V(P)$的无偏估计.

证明略,详细论证请见文献[17]第 31 页定理 2.2 及其推论 2.5、推论 2.8.

在 p 和 P 具有以上统计性质时,以及在用样本比例 p 估计总体比例 P 时,样本容量 a 的确定方法可以如下设计.

定理 2.1. 给出一个决策表 $DT=(U,R,D)$,其中,$U=\{x_1,x_2,...,x_n\}.S=\{s_1,s_2,...,s_a\}$ 是随机地从全集 U 中抽样得出的样本集.当 $a>\frac{t^2p(1-p)}{d^2}$ 时,满足以下条件,则至少以 e 的置信度,使得 $P=p\pm d\%$.

(1) $P=\gamma_{A\approx A'}$, 其中, $A=\{\lambda_1,\lambda_2,...,\lambda_n\}$ 是 $\{x_1,x_2,...,x_n\}$ 的下近似值. $A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$ 是 $\{x_1,x_2,...,x_n\}$ 的 k-统计下近似.

(2) $p=\gamma_{B\approx B'}$, 其中, $B=\{\beta_1,\beta_2,...,\beta_n\}$ 是 $\{s_1,s_2,...,s_a\}$ 的下近似值. $B'=\{\beta_1',\beta_2',...,\beta_n'\}$ 是 $\{s_1,s_2,...,s_a\}$ 的 k-统计下近似.

(3) t 是标准正态分布的 1-e 双侧分位数.

证明:在简单随机抽样中,可以用样本比例 p 作为总体比例估计 P.由于 $E(P)=E(p)=P$,故而 p 是 P 的无偏估计.由于 p 的方差为 $V(p)=\frac{PQ}{a}\frac{(n-a)}{(n-1)}, Q=1-P$; 绝对误差限度为 $d=t\sqrt{V(\tilde{\theta})}$; 上两式合并可得:

$$d=t\sqrt{V(\tilde{\theta})}=t\sqrt{\frac{PQ}{n}\frac{N-n}{N-1}}.$$

经过化简可得:

$$a=\frac{t^2\frac{PQ}{d^2}}{1+\frac{1}{n}\left(t^2\frac{PQ}{d^2}-1\right)}.$$

由于 $1\leqslant n, \frac{1}{n}\to 0$, 因此, $a>\frac{t^2PQ}{d^2}$. □

由上述定理可以看出,样本容量的确定与绝对误差限度 d 成反比,与分布双侧分位数 t 和置信度$(1-\alpha)$成正比.即要求的精度越高,所需的样本容量越大.也就是说,该定理保证了用抽样集合计算出的限制距离 k 以很高的置信度和很小的误差近似等于真实的、由全集 U 计算出的限制距离.

同时,在保证置信度α确定的前提下,该定理给出了样本容量的下限.也就是说,通过随机抽样,可以采用相对

小的样本集,快速确定限制距离 k.

3 基于统计粗糙集模型的下近似算法设计

在统计粗糙集模型中,为了计算统计下近似,我们必须先得到限定距离 k,然后确定 k-limit 的范围.下面的一种算法即用来计算限定距离 k.

3.1 计算限定距离k的算法设计

一个很朴素的解决方案是,首先对 k 赋一个较小的初值,然后尝试使用此 k 值下的 k-limit 计算统计下近似值,然后与真实下近似进行对比,如果有一定比例(达到阈值)的统计下近似值和真实下近似值相同(或者在可控误差范围内),那么,就认为这个限定距离是合适的.我们引入随机抽样,从而不需要计算全部的真实下近似,而只需要计算样本集中的真实下近似即可.

算法 3.1. 计算限定距离 k.

输入:$DT=(U,R,D), U=\{x_1,x_2,...,x_n\}, R=\{r_1,r_2,...,r_m\}, d,e$;

输出:限定距离 k.

第 1 步.计算 $a=\max_p\left(\frac{t^2p(1-p)}{d^2}\right)$,然后随机从全集 U 中抽取样本集 $S=\{s_1,s_2,...,s_a\}$;

第 2 步.计算 S 中所有样本的真实下近似,$\lambda_i=\underline{R_g}[s_i]_D(s_i)$, $\forall i\in\{1,2,...,a\}$; 令 $A=\{\lambda_1,\lambda_2,...,\lambda_n\}$;

第 3 步.$k\leftarrow 1$,对于 $\forall i\in\{1,2,...,a\}$, 计算 s_i 的 k-limit;

第 4 步.对于 $\forall i\in\{1,2,...,a\}$, 计算 $\lambda_i'=\underline{R_g^1}[s_i]_D(s_i)$, 令 $A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$;

第 5 步.计算 $p=\gamma_{A\approx A'}$;

第 6 步.while $p<1$, do:

(6.1) $k\leftarrow k\times 2$;

(6.2) 对于 $\forall i\in\{1,2,...,a\}$, 计算 s_i 的 k-limit;

(6.3) 对于 $\forall i\in\{1,2,...,a\}$, 计算 $\lambda_i'=\underline{R_g^1}[s_i]_D(s_i)$, 令 $A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$;

(6.4) 计算 $p=\gamma_{A\approx A'}$;

第 7 步.输出限定距离 k.

在算法 3.1 中,我们使用随机抽样去验证此时的限定距离 k 是否已经足够得到准确的统计下近似值.定理 2.1 则保证了,用样本集得到的限定距离 k,与全集中得到的 k 统计误差极小.

得到限定长度之后,下面我们将利用它来计算统计下近似.

3.2 计算统计下近似算法设计

根据上面得到的限定距离 k,我们将设计一种算法来计算统计下近似值.

算法 3.2. 计算统计下近似值.

输入:$DT=(U,R,D), U=\{x_1,x_2,...,x_n\}, R=\{r_1,r_2,...,r_m\}$, 限定距离 k;

输出:所有元素的 k-统计下近似值.

第 1 步.$i\leftarrow 1$;

第 2 步.while $i\leqslant n$, do

(2.1) 计算 x_i 的限定区域 k-limit

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

(2.3) $i\leftarrow i+1$

第 3 步.输出 $A'=\{\lambda_1',\lambda_2',...,\lambda_n'\}$.

4 实 验

本文实验均是在 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,内存为 252GB.

本文实验将应用 6 个 UCI 数据集[18],用来验证统计下近似算法的效率.具体数据集参数见表 4.在具体实施随机抽样时,考虑到数据分布的特征,我们采用的是分层随机抽样法.根据不同类别记录的比值,按比例分配每一类随机抽样的个数.此时,不同类别随机抽样个数的总数仍是根据定理 2.1 计算出的样本量的数值.

Table 4 The information of datasets

表 4 数据集描述

数据集	记录条数	特征数
Image segmentation	2 310	19
Letter recognition	20 000	16
Wine quality white	4 898	11
Wine quality red	1 599	11
MAGIC gamma telescope	19 020	10
Waveform database generator	5 000	21

4.1 统计下近似的运行时间

表 5 代表了统计下近似和真实下近似的算法的时间消耗.由表 5 我们可以看出,真实算法和统计算法对于 6 个数据集进行下近似的时间消耗,这体现了随机抽样算法在机选(统计)下近似时的高效率.

Table 5 The running time to compute the statical lower approximation

表 5 计算(统计)下近似的时间消耗

数据集	真实下近似(s)	统计下近似(s)
Image segmentation	2.84	0.72
Letter recognition	282.93	52.4
Wine quality white	11.41	2.13
MAGIc gamma telescope	138	26.66
Wine quality red	1.19	0.33
Waveform database generator	20.48	3.8

为了更好地展示计算的时间和空间,我们把这 6 个数据集分割成 10 个等分.第 1 部分被看成是 1st data set,第 2 部分被看成是 2nd data set,…,第 10 部分被看成是 10th data set.我们将用这些被分割过的数据集来观察所提出的统计下近似的算法和真实下近似的算法在数据集不断增大时的表现.图 2 中,清晰地展示了随着数据集的增大两种算法的下近似时间消耗情况.

从表 5 和图 2 我们可以清晰地看出,随着数据集的增大,下近似的计算时间也在逐步增大.我们可以从图中看出,基于随机抽样的算法,相对于真实下近似算法,时间消耗明显较少.这是由于,通过随机抽样以及限定区域 k 来计算统计下近似.此外,基于随机抽样的算法随着数据集的增大,它的时间消耗增大并不明显.而真实下近似的计算时间随数据集的增大,其时间消耗急速增加.这体现出了基于随机抽样统计下近似的优越性——数据集越大,其时间优势越明显.

在统计粗糙集模型中,我们针对下近似的计算,通过限定区域 k-limit 和随机抽样进行了优化,使之成为统计下近似,在保证其精度的情况下,大幅度缩减了计算(统计)下近似的时间消耗.验证了在计算下近似时间效率方面,随机抽样算法的优越性.并且,可以看到,数据集越大,随机抽样算法在计算下近似的时间消耗方面其优势越显著.

综上,我们可以得出结论,随机抽样算法在下近似时间效率方面,相比传统的下近似算法都有着大幅提升,并且这种提升随着数据集的增大会更加显著.因此,随机抽样算法确实弥补了传统下近似方法无法应用到大规模数据集上的不足.而且,随机抽样算法的可扩展性较强,可适用于大规模数据集.

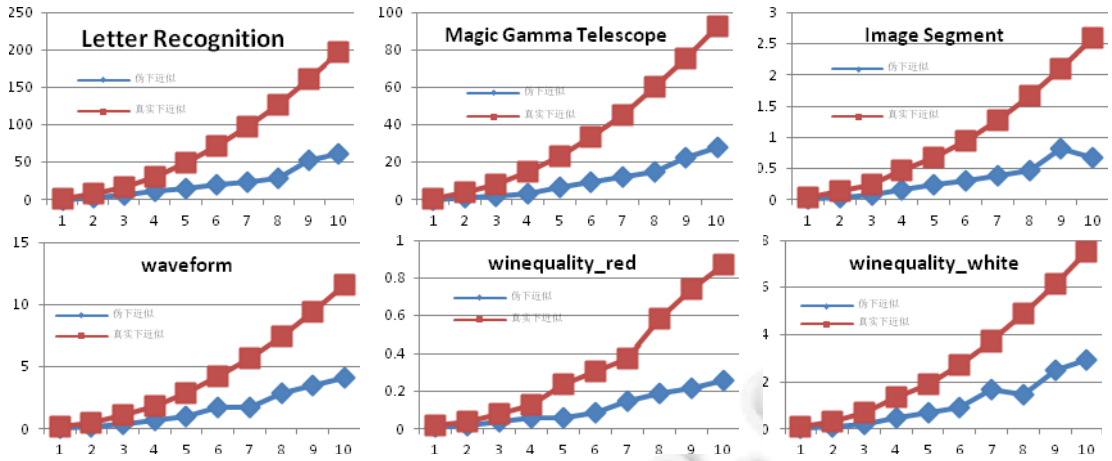


Fig.2 The trend chart of lower approximation running time changing with the size of datasets

图 2 计算下近似所消耗时间随数据集大小变化趋势

4.2 统计下近似的效果比较

在本小节,我们的目的是展示统计下近似的知识发现的效果.我们从以下两个方面展示统计下近似的知识发现的效果:一方面是基于统计下近似的约简的运行时间,另一方面是基于统计下近似的约简的精度.

下面,我们将观察约简时间随数据集大小变化的情况.

在表 6 中我们可以清晰地发现,依赖度算法和辨识矩阵算法空间消耗大致相同,在一个数量级上.而随机抽样算法的空间消耗相对于另外两种来说非常之小,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.这体现了随机抽样算法在进行属性约简时的优越性,特别是在大规模数据集下体现出的优越性.

Table 6 The comparison of space complexity of statistical lower approximation based reduction

表 6 基于统计下近似的属性约简的空间消耗比较

	依赖度约简(MB)	辨识矩阵约简(MB)	随机抽样约简(MB)
Image segmentation	720	512	2.1
Letter recognition	60 655	62 301	58
MAGIc gamma telescope	42 707	41 300	50
Waveform database generator	5 120	5 120	6

表 7 和表 8 列出了 3 种约简方法所得约简结果的分类精度,可以看到所有属性的分类精度最佳,而经典的两种方法——辨识矩阵约简方法和依赖度约简方法的约简结果的分类精度相近.基于统计抽样的约简方法在不同的抽样情况下所得到的约简的分类精度不同,但却相近.而且它们均略微低于经典的两种约简方法的精度.

Table 7 The comparison of classification performance of statistical lower approximation based reduction (1)

表 7 基于统计下近似的属性约简的分类精度比较(1)

Waveform 所有属性	21	0.856 626	Segment 所有属性	19	0.903 514	Winequality-Red 所有属性	11	0.653 125
辨识矩阵约简	13	0.821 8	辨识矩阵约简	11	0.903 517	辨识矩阵约简	6	0.633 75
依赖度约简	14	0.824 8	依赖度约简	10	0.900 912	依赖度约简	6	0.615 625
随机抽样的约简 1	12	0.824 04	随机抽样约简 1	8	0.900 056	随机抽样约简 1	6	0.620 625
随机抽样的约简 1	12	0.830 236	随机抽样约简 2	10	0.922 979	随机抽样约简 2	6	0.620 625
随机抽样约简 3	13	0.839 832	随机抽样约简 3	7	0.847 694	随机抽样约简 3	6	0.620 625
随机抽样约简 4	11	0.825 433	随机抽样约简 4	7	0.847 694	随机抽样约简 4	6	0.620 625
随机抽样约简 5	12	0.826 434	随机抽样约简 5	8	0.900 056	随机抽样约简 5	6	0.620 625
随机抽样约简 6	13	0.820 436	随机抽样约简 6	8	0.900 056	随机抽样约简 6	6	0.620 625

Table 7 The comparison of classification performance of statistical lower approximation based reduction (1) (Continued)

表 7 基于统计下近似的属性约简的分类精度比较(1)(续表)

Winequality-White 所有属性	11	0.680 717	Magic 所有属性	10	0.829 189	Letter 所有属性	16	0.930 255
辨识矩阵约简	7	0.652 962	辨识矩阵约简	6	0.823 172	辨识矩阵约简	9	0.920 067
依赖度约简	7	0.667 449	依赖度约简	5	0.813 934	依赖度约简	10	0.924 749
随机抽样约简 1	7	0.652 545	随机抽样约简 1	3	0.774 538	随机抽样约简 1	11	0.897 461
随机抽样约简 2	7	0.652 545	随机抽样约简 2	3	0.789 34	随机抽样约简 2	10	0.903 486
随机抽样约简 3	7	0.650 709	随机抽样约简 3	3	0.772 031	随机抽样约简 3	12	0.893 486
随机抽样约简 4	7	0.652 545	随机抽样约简 4	3	0.774 873	随机抽样约简 4	11	0.906 712
随机抽样约简 5	7	0.651 936	随机抽样约简 5	2	0.748 946	随机抽样约简 5	11	0.906 712
随机抽样约简 6	7	0.651 936	随机抽样约简 6	3	0.787 364	随机抽样约简 6	11	0.906 712

Table 8 The comparison of classification performance of statistical lower approximation based reduction (2)

表 8 基于统计下近似的属性约简的分类精度比较(2)

	所有属性	辨识矩阵约简	依赖度约简	随机抽样约简 1	随机抽样约简 2	随机抽样约简 3	随机抽样约简 4	随机抽样约简 5	随机抽样约简 6
平均精度	0.808 904	0.792 545	0.791 245	0.778 211	0.786 535	0.770 73	0.771 314	0.775 785	0.781 188

综上,我们可以得出结论,随机抽样算法虽然在某些时候会对约简精度有小幅度但可控的下降,但其在下近似时间效率、属性约简时间效率方面,相比传统的属性约简算法都有着大幅提升,并且这种提升随着数据集的增大会更加显著.因此,随机抽样算法确实改善了传统属性约简方法无法应用到大规模数据集上的缺陷.

5 结论与展望

本文将随机抽样引入到经典的模糊粗糙集理论中,建立了统计粗糙集模型,设计了计算统计下近似的算法.本文的创新点在于:将随机抽样和经典的模糊粗糙集相结合,将随机抽样代入到模糊粗糙集的基本理论中,提出了限制区域 k-limit(x)的概念,将计算范围限制到更小、更有效的范围,提出了统计上、下近似等概念,构建了统计粗糙集模型.

References:

[1] 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]

[2] 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]

[3] 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]

[4] Pawlak Z. Rough sets. Int'l Journal of Information and Computer Science, 1982,11(5):314−356. [doi: 10.1007/BF01001956]

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

[6] An AJ, Stefanowski, Ramanna S, Butz C, Pedrycz W. Rough sets, fuzzy sets, data mining and granular computing. In: Proc. of the RSFDGrC 2007 (the 11th Int'l Conf. on Rough Ssets, Fuzzy Sets, Data Mining and Granular Computing). Toronto: Springer-Verlag, 2007. [doi: 10.1007/978-3-540-72530-5]

[7] Wang GY, Li TR, Grzymala-Busse J, Miao DQ, Skowron A, Yao YY. Rough sets and knowledge technology. In: Proc. of the RSKT 2008 (the 3rd Int'l Conf. on Rough Sets and Knowledge Technology). Berlin: Springer-Verlag, 2008. [doi: 10.1007/978-3-540-79721-0]

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

[9] An S, Hu QH, Yu DR, Liu JF. Soft minimum-enclosing-ball based robust fuzzy rough sets. Fundamenta Informaticae, 2012,115(2-3):189−202.

[10] 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]

[11] Chen Y. Random sampling based fuzzy rough reduction and application [MS. Thesis]. Beijing: Renmin University of China, 2015 (in Chinese with English abstract).

[12] Zhang QH, Wang GY. Rough set theory and its application. Communication of Communication of China Artificial Intelligence Association, 2011 (in Chinese with English abstract).

[13] Chen XQ, Jin XL, Wang YZ, Guo JF, Zhang TY, Li GJ. Survey on big data system and analytic technology. Ruan Jian Xue Bao/ Journal of Software, 2014,25(9):1889−1908 (in Chinese with English abstract). http://www.jos.org.cn/4674.htm [doi: 10.13328/j.cnki.jos.004674]

[14] Li GJ. Scientific value of big data research. Communications of the CCF, 2012,8(9):8−15 (in Chinese with English abstract).

[15] 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]

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

[17] Jin YJ, Du ZF, Jiang YB. Sampling Techniques. 4th ed., Beijing: The Press of Renmin University of China, 2015 (in Chinese).

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

附中文参考文献:

[11] 陈俞,基于随机抽样的模糊粗糙集约简方法及其应用研究[硕士学位论文].北京:中国人民大学,2015.

[12] 张清华,王国胤.粗糙集理论及其应用概述.中国人工智能学会通讯,2011.

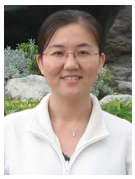
[13] 程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰.大数据系统和分析技术综述.软件学报,2014,25(9):1889−1908. http://www.jos.org.cn/html/4674.htm [doi: 10.13328/j.cnki.jos.004674]

[14] 李国杰.大数据研究的科学价值.中国计算机学会通讯,2012,8(9):8−15.

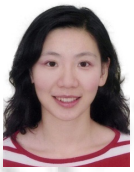
[17] 金勇进,杜子芳,蒋妍.抽样技术.第 4 版,北京:中国人民大学出版社,2015.

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.

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,数据挖掘,社会网络分析.



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



孙辉(1977－),女,博士,讲师,CCF 会员,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.