基于系统熵的粗糙集属性约简新方法

李伟涛,刘琼荪

(重庆大学 数学与统计学院,重庆 401331)

摘 要:在系统熵的基础上,定义了一种新的属性重要度并提出了一种基于改进系统熵的粗糙集属性约简算法,实验分析表明,该属性重要度为启发式信息进行的属性约简,取得了理想效果。

关键词:粗糙集;属性约简;系统熵

中图分类号:TP18 **文献标识码**:A **文章编号**:1674-7720(2011)09-0084-03

A new method of rough set attribute reduction based on system entropy

Li Weitao, Liu Qiongsun

(College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China)

Abstract: This paper presents a new attribute significance and gives a kind of attribute reduction based on improved system entropy. Texts show that the perfect effect is acquired that the attribute reduction based on the attribute significance as heuristic information.

Key words: rough set; attribute reduction; system entropy

粗糙集(Rough Set)理论[1]是一种处理不确定、不完整知识的数学工具,最早是由 Pawlak 于 1982 年提出的。现在广泛应用于数据挖掘、智能控制、模式识别等领域[2-3]。属性约简是粗糙集理论中的核心内容之一,有许多学者致力于粗糙集属性约简算法的研究。其中应用较多的是基于差别矩阵及在此基础上的一些改进算法[4],虽然该算法可以得到所有的约简,但是只适合较小的数据集;基于代数观点的相对约简算法不能精确地度量粗糙集中的信息粒度划分;苗夺谦[5]等人提出基于互信息的属性约简算法,是建立在条件属性对决策属性的信息量基础上的。然而以上这些属性约简算法所依据的都是条件属性的分类能力,它们的出发点都是一样的,只是采用的标准有所不同。最近,有些学者提出新的属性约简定义,认为只关心条件属性的分类能力是不够的,决策属性的分类能力也应该充分考虑,即基于系统熵的属性约简定义[6],这种属性约简定义同时考虑到了条件属性和决策属性的分类能力,是一种较周全的属性约简模型。

本文从系统熵的角度出发,改进了原先的属性重要度定义,给出了新的属性重要性的度量方法,并构造了相应的启发式算法,并通过实例验证了算法的有效性。

1 粗糙集的基本概念

定义 1[7]:一个信息系统 S,表示为 $S=(U,\boldsymbol{A},V,f)$,其中 $U=\{X_1,\cdots,X_n\}$ 是论域;$\boldsymbol{A}$ 是属性集合;$V=\mathrm{U}V_a$,$\forall \in aA$,V_a 表示属性的值域;f 是一个信息函数,对 $x\in U$,$a\in A$ 有 $f(x,a)\in V_a$。若 $\boldsymbol{A}$ 可以分为条件属性集 C 和决策属性 D,即 $\boldsymbol{A}=CUD$,$CID=\phi$,则称该信息系统为决策表。

定义 2[7]:在一个信息系统中,对于每个属性子集 $B\subset A$ 可以定义一个不可区分关系 $IND(B)$:$IND(B)=\{(x,y)\in U\times U:\forall b\in B,b(x)=b(y)\}$,称为由 B 构造的不可分辨关系。

定义 3[7]:在信息系统 S 中,对于属性集 $X\subset U$,R 为等价关系,定义 2 个子集为:

$$\underline{R}X=\{Y\in U/R:Y\subset X\},\overline{R}X=\{Y\in U/R:YIX\neq\phi\}.$$

分别称它们为 X 的 R 下近似和 R 上近似集。

定义 4[7]:在信息系统 S 中,若 $P,Q\subset A$,则 Q 的 P 正域 $POS_P(Q)$ 定义为 $POS_P(Q)=\underset{X\subset U/R}{U}\underline{P}(X)$,其中 $\underline{P}(X)$ 为 X 的 P 下近似集。Q 的 P 正域是 U 中所有根据分类 U/P 的信息划分到关系 Q 的等价类的对象集合。

定义 5[6]:设五元组 $S=(U,C,D,V,f)$ 是一个决策表。$\forall P\subset C$,$U/P=\{P_1,\cdots,P_m\}$,$U/D=\{D_1,\cdots,D_n\}$,决策表 S 的系统熵 $H(U:P)$ 定义为:

$H(U:P)=-\sum_{i=1}^{n}\frac{|D_i|}{|U|}\sum_{j=1}^{m}\frac{|P_j\ I\ D_i|}{|D_i|}\log\frac{|P_j\ I\ D_i|}{|D_i|}$，其中$|D_i|$表示$D_i$的个数。

引理 1[6]：设五元组$S=(U,C,D,V,f)$是一个决策表，对$\forall B\subset C$，有$H(U:C)\geqslant H(U:B)$。

定义 6[6]：设五元组$S=(U,C,D,V,f)$是一个决策表，$a\in C$，则a的熵定义为：$HA^*(\{a\})=-\sum_{X_i\in U/IND(\{a\})}P(X_i)\log P(X_i)$其中$P(X_i)=\frac{|X_i|}{|U|}$。

定义 7[6]：设五元组$S=(U,C,D,V,f)$是一个决策表，对$\forall B\subset C$，若$H(U:C)=H(U:B)$，并且对$\forall b\in B$，均有$H(U:C)>H(U:B-\{b\})$，则称B为条件属性集C关于决策表属性集D的一个基于系统熵的属性约简，简称B为决策表的基于系统熵的属性约简。

若所用的属性约简算法得到的约简结果B同时满足$H(U:C)\geqslant H(U:B)$和$\forall b\in B$，均有$H(U:C)>H(U:B-\{b\})$，则称这个属性约简算法是一个完备的属性约简算法[8]。

定义 8[6]：设五元组$S=(U,C,D,V,f)$是一个决策表，$a\in C$，属性a的重要性定义为$sig(a)=\frac{H(U:C)-H(U:C-\{a\})}{HA^*(a)}$。

2 基于改进系统熵的启发式算法

2.1 改进的属性重要性度量方法

设五元组$S=(U,C,D,V,f)$是一个决策表，并且$R\subset C$，则对于任意属性$a\in C-R$的属性重要度如定义 8 所示。

算法 1：参考文献[8]提出的基于系统熵的属性约简算法，以系统熵增益率的变化量作为属性对决策的相对重要度为启发式信息，基于系统熵增益率的属性重要性度量方法为：

$$sig(a)=\frac{H(U:C)-H(U:C-\{a\})}{HA^*(a)}\tag{1}$$

由式(1)定义的属性重要度度量方法虽然考虑了属性值域的大小，及属性取值分布情况，但是由于数值的缘故，可能因为$HA^*(a)$较小，进而导致$sig(a)$取值过大的情况出现。如果出现这种情况，依此属性重要性选取的属性对于决策属性来说不一定是最重要的，这样可能会导致约简的个数增多，达不到所要求的约简效果。

基于以上属性重要性度量方法存在的一些问题，本文提出了一种新的属性重要性度量方法：

$$sig(a)=\frac{H(U:C)-H(U:C-\{a\})}{HA^*(D/a)}\tag{2}$$

这种新的度量方法同时兼顾了系统熵作为一种同时考虑了条件属性和决策属性的分类能力和数值大小对约简结果的影响，并充分考虑到了在属性子集R中添加属性$a\in C-R$后系统熵的增量（R自身的熵也被考虑在内）。这种新的属性重要性的定义有如下特点：(1)当系统熵增量大小相等时，$HA^*(D/a)$越小，相应的属性重要度越大；(2)当$HA^*(D/a)$相等时，$H(U:C)$越大，则相应的属性重要性越大。这两个特性使得本文提出的属性约简算法在绝大多数的情况下都能够做到使约简的个数相对较少。

2.2 基于改进系统熵的启发式算法

粗糙集的属性约简是指在不损失信息的情况下删除信息系统中冗余的属性，约简结果的集合$\boldsymbol{R}=\{R:R\subset C,H(U:R)=H(U:C)\}$。所以说同样的相关系数可以作为最终约简的终止条件。本文根据式(2)定义的属性重要度，提出一种改进的基于系统熵的粗糙集属性约简算法（算法 2）。该算法采用 backward 约简方式，依次剔除冗余的属性，直到满足终止条件$H(U:R)=H(U:C)$，或者$H(U:C)-H(U:R)\leqslant\omega$($\omega$为指定的任意小的正整数)。

算法 2：

输入：五元组$S=(U,\boldsymbol{C},D,V,f)$，(其中$\boldsymbol{C}$为条件属性集合，$D$为决策属性集合)，初始属性集合$\boldsymbol{R}$为空；

输出：属性约简结果$\boldsymbol{R}$。

对$\forall a\in C$计算$HA^*(D/a)$；

```
Result:=A
While(|Result|>1)Do
{
  b  End:=Ture;
    ∀  a∈Result:
{
  If(a是冗余的)
  {计算 sig(a);  b  End:=false;}
}
  If(b  End:=Ture)跳出循环

  Whlie  else
{
    Attr:=a,
    得到 sig(a)就是最小的值;
     Result:=  Result-{Attr};
    }
  }
```

此算法的时间复杂性为$O(|A|^3|U|\log|U|)$。

3 仿真实例和相关比较

为了验证上述算法的有效性，从 UIC 数据库中选取了三个具有离散属性的数据库实例进行验证。分别采用文中所提到的两种不同属性重要性定义的约简算法对其进行属性约简。约简结果如表 1 所示。其中C为该属性集合所包含的条件属性的个数，算法 1 和算法 2 分别是以系统熵增益率和本文改进的系统熵增益率为属性重要性度量方法的启发式属性约简算法。从表中可以看到本文所提出的算法在大多数情况下获得的相对约简

属性个数较少。

表 1 属性约简结果

数据集	C	实例	算法 1	算法 2
Zoo	16	101	9	4
Winsongolf	4	14	3	3
Chess	36	3 175	27	16

为了进一步验证文中所改进算法的特点，使用 Zoo 数据集如表 2 所示。其中论域 $U=\{1,\cdots,101\}$，条件属性 $C=\{$hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, catsize$\}$，$D=\{type\}$ 为决策属性。

表 2 Zoo 数据

对象	hair	feathers	…	type
1	1	0		1
2	1	0		1
3	0	0		4
4	0	0		7
⋮	⋮	⋮	…	⋮
101	0	1		2

如果按照式(1)所提出的属性重要性来度量各个属性的重要性，经计算得出属性重要性最大的是{milk}。而依据本文所提出的属性重要性得到的结果是 {eggs}，算法 1 所得到的属性约简结果是：$R_a=\{$feathers, milk, airborne, aquatic, backbone, breathes, fins, legs$\}$。

依照本文算法 2 所得到的属性约简结果是：$R_b=\{$milk, eggs, aquatic, legs$\}$。这是因为利用式(1)计算属性重要性的时候只考虑了属性本身的值的分布而没有考虑属性的相对信息熵，如果某一属性的相对信息熵较小会导致该属性的属性重要度较大，从而会使所选属性并不是最重要的，或者造成错选。

本文从系统熵的角度出发，定义了一种新的度量属性重要性的方法，构造了相应的启发式算法。相对于原算法，本文算法优势明显，通过实例证明，在大多数情况下本文的算法所得到的属性约简个数较少。

参考文献

[1] PAWLAK Z. Rough sets [J]. Int computer & science, 1982; 11(5): 341–356.

[2] 常犁云，王国胤，吴渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报，1999，10(11)：1206–1211.

[3] Hu Xiaohua, CERCONE N. Learning in relational databases a rough set approach [J]. International Journal of Computational Intelligence, 1995, 11(2): 320–340.

[4] RAUSZER S. The discernibility matrices and functions in information systems [M]. Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht Kluwer, 1992, 31–362.

[5] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999，36(6)：681–684.

[6] Zhao Jun, Wu Zhongfu, Li Hua. System entropy and its application in feature selection [J]. The Journal of China Universities of Posts and Telecommunications, 2004, 11(1): 100–105.

[7] 苗夺谦，李道国. 粗糙集理论算法与应用[M]. 北京：清华大学出版社，2008.

[8] 王雄彬，郑雪峰，等. 基于系统熵的属性约简的简化差别矩阵方法 [J]. 计算机应用研究，2009，26 (7)：2461–2464.

(收稿日期：2010–12–14)

作者简介：

李伟涛，男，1984 年生，硕士，主要研究方向：智能计算仿真及其应用，应用数学。

刘琼荪，女，1956 年生，教授，主要研究方向：智能计算、数据挖掘及应用统计。