基于粗糙集理论的属性权重确定最优化方法研究

钟嘉鸣[1,2],李订芳[2]
ZHONG Jia-ming[1,2],LI Ding-fang[2]

1.湘南学院 网络中心,湖南 郴州 423000
2.武汉大学 数学与统计学院,武汉 430072
1.Network Center,Xiangnan University,Chenzhou,Hunan 423000,China
2.School of Mathematics and Statistics,Wuhan University,Wuhan 430072,China
E-mail:jmzhongcn@163.com

ZHONG Jia-ming,LI Ding-fang.Research on optimization method of attribute weight determining based on rough set theory.Computer Engineering and Applications,2008,44(20):51-53.

Abstract: In view of the deficiency in existing methods for Objective Attribute Weight Determining (OAWD) based on rough set theory,by integrating information view and the algebra view of rough set,optimization method of OAWD is given,then the optimal solution of integrated weight is got.Finally,an example is given to explain the availability of this method.
Key words: rough set;significance of the attribute;weight;algebra view;information view

摘 要:针对现有基于粗糙集理论属性客观权重确定方法的不足,将基于代数观和信息观的权重进行有机集成,建立了两者结合的属性权重最优化数学模型,从而得到综合权重的最优解。最后通过实例说明了该方法的有效性。
关键词:粗糙集;属性重要性;权重;代数观;信息观
DOI:10.3778/j.issn.1002-8331.2008.20.015 **文章编号:**1002-8331(2008)20-0051-03 **文献标识码:**A **中图分类号:**TP18

1 引言

在综合评判和决策分析中,属性权重的确定是其中很关键的一个环节,权重反映了各指标在评估决策中所处的地位或者说所起的作用,它直接影响到评估和决策的最终结果。目前常见的权重确定方法有主成分分析法、模糊综合分析法、数据包络分析法和神经网络等方法,但这些方法存在主观性和模糊性等明显局限[1,8]。因此有些学者提出了依据粗糙集理论[2,3]中的属性重要性概念来确定客观权重[1]。然而这些方法依然存在明显的不足。粗糙集理论中属性重要性的几个定义并不具备一致性[4,9]。基于属性依赖度的重要性,其标准过于"粗糙",而基于信息量的重要性,其标准过于"细致",这两种形式在实际应用中有时甚至出现矛盾结论[5]。因此单独地利用某个属性重要性概念作为确定权重的基础并不科学。探讨科学有效的属性权重确定方法显得非常必要。

在粗糙集理论中属性重要性概念可以归结为代数观定义和信息观定义两类,这两类定义具有互补的特性:属性重要性的代数定义考虑的是该属性对论域中确定分类子集的影响,而信息观定义考虑的是该属性对论域中不确定分类子集的影响。因此,可以在由代数观和信息观下的属性重要性确定权重的基础上,将两者进行有机的集成,从而最终确定属性的客观权重。本文利用最优化理论,建立基于代数观和信息观相结合的属性权重最优化数学模型,得到综合权重确定的最优化方法。

2 基于代数观的属性权重

定义 1[6] 称四元组 $T=(U,A,V,f)$ 为一个决策表系统。其中:$U\neq\varnothing$ 为论域;$A=C\cup D$,C 和 D 分别为条件和决策属性集,且 $C\cap D=\varnothing$;V 为属性的值域集,$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;$f:U\times A\to V$ 是一个信息函数,对 $\forall x\in U,a\in A$,存在 $f(x,a)\in V_a$。

每一个属性子集 $P\subseteq A$ 决定了一个二元不可区分关系 $IND(P)$:

$$IND(P)=\{(x,y)\in U\times U|\forall a\in A,f(x,a)=f(y,a)\}$$

性质 1[6] $IND(P)$是论域 U 上的等价关系,且

$$IND(P)=\bigcap_{a\in P}IND(\{a\})$$

性质 2[6] 令 $P,Q\subseteq A$,若 $P\subseteq Q$,则 $IND(Q)\subseteq IND(P)$。

关系 $IND(P)$,$P\subseteq A$,构成了 U 的一个划分,用 $U/IND(P)$ 表示,简记为 U/P,$U/IND(P)$中的任何元素$[x]_P$称为等价类。

定义 2 设 $X\subseteq U$,R 是 U 上的等价关系,U/R 表示 R 对 U 的划分,子集 $\underline{R}X=\cup\{Y\in U/R|Y\subseteq X\}$,$\overline{R}X=\cup\{Y\in U/R|Y\cap X\neq\varnothing\}$

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70771708)。
作者简介:钟嘉鸣(1966-),男,副教授,武汉大学数学与统计学院访问学者,主要研究领域为智能计算、粗糙集理论及应用;李订芳(1965-),男,教授,博士生导师,主要研究领域为科学与工程计算、智能计算。
收稿日期:2008-01-21 **修回日期:**2008-04-16

分别称为 X 关于 R 的下近似和上近似。

定义 3 令 P 和 Q 为 U 中的等价关系，Q 的 P 正域记为 $pos_P(Q)$，即

$$pos_P(Q)=\bigcup_{X\in U/Q}\underline{P}X$$

Q 和 P 间的依赖度定义为：

$\gamma_P(Q)=|pos_P(Q)|/|U|$，其中 $0\leqslant\gamma_P(Q)\leqslant 1$

在决策表中，不同的属性可能具有不同的重要性，为了找出某些属性（或属性集）的重要性，可以从表中去掉一些属性，再来考察没有该属性后分类会怎样变化。若去掉该属性相应分类变化较大，则说明该属性的强度大，即重要性高；反之，说明该属性的强度小，即重要性低。

定义 4（属性重要性的代数观定义） 设决策表系统 $T=(U,A,V,f)$，$A=C\cup D$，且 $C\cap D=\varnothing$，C 和 D 分别为条件属性集和决策属性集，属性子集 $C'\subseteq C$ 关于 D 的重要性定义为：

$$\sigma_{CD}(C')=\gamma_C(D)-\gamma_{C-C'}(D)$$

特别当 $C'=\{a\}$ 时，属性 $a\in C$ 关于 D 的重要性为：

$$\sigma_{CD}(a)=\gamma_C(D)-\gamma_{C-\{a\}}(D)$$

定义 5（基于代数观的属性权重定义） 设决策表系统 $T=(U,A,V,f)$，$A=C\cup D$，且 $C\cap D=\varnothing$，C 和 D 分别为条件属性集和决策属性集，$C=\{a_1,a_2,\cdots,a_m\}$，则属性 a_i 的权重为：

$$\omega_i=\frac{\sigma_{CD}(a_i)}{\sum_{i=1}^{m}\sigma_{CD}(a_i)}$$

3 基于信息观的属性权重

定义 6 设决策表系统 $T=(U,A,V,f)$，$A=C\cup D$，且 $C\cap D=\varnothing$，C 和 D 分别为条件属性集和决策属性集，设 $X=U/IND(C)=\{X_1,X_2,\cdots,X_n\}$ 和 $Y=U/IND(D)=\{Y_1,Y_2,\cdots,Y_m\}$ 分别表示由等价关系 $IND(C)$ 和 $IND(D)$ 导出的 U 上的划分。

$$p(X_i)=\frac{|X_i|}{\sum_{j=1}^{n}|X_j|}=\frac{|X_i|}{|U|},i=1,2,\cdots,n$$

$$p(Y_j)=\frac{|Y_j|}{\sum_{i=1}^{m}|Y_i|}=\frac{|Y_i|}{|U|},j=1,2,\cdots,m$$

则 $(p(X_1),p(X_2),\cdots,p(X_n))$ 和 $(p(Y_1),p(Y_2),\cdots,p(Y_m))$ 分别为 C 和 D 在 X 和 Y 上的有限概率分布。

定义 7[7] 属性集 C 的信息熵 $H(C)$ 定义为：

$$H(C)=-\sum_{i=1}^{n}p(X_i)\mathrm{lb}(p(X_i))$$

当某个 p_i 为 0 时，规定 $0\cdot\mathrm{lb}0=0$。

定义 8[7] 属性集 $D(Y=U/IND(D)=\{Y_1,Y_2,\cdots,Y_m\})$ 相对于属性集 $C(X=U/IND(C)=\{X_1,X_2,\cdots,X_n\})$ 的条件熵 $H(D|C)$ 定义为：

$$H(D|C)=\sum_{i=1}^{n}p(X_i)H(Y|X_i)$$

其中 $H(Y|X_i)=-\sum_{j=1}^{m}p(Y_j|X_i)\mathrm{lb}p(Y_j|X_i)$，$p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

定义 9（基于信息观的属性重要性定义） 设 $T=(U,C\cup D,V,f)$ 是一个决策表系统，其中 C 是条件属性集合，$D=\{d\}$ 是决策属性集合，且 $A\subset C$，则对任意属性 $a\in C-A$ 的重要性 $SGF(a,A,D)$ 定义为：

$$SGF(a,A,D)=H(D|A)-H(D|A\cup\{a\})$$

其中 $H(D|A)$ 表示属性集 D 相对于属性集 A 的条件熵。若 $A=\phi$，则 $SGF(a,A,D)=H(D|A)-H(D|\{a\})$，称为条件属性 a 和决策 D 的互信息，记为 $I(a,D)$。$I(a,D)$ 的值越大，说明属性 a 对于决策 D 就越重要。

定义 10（信息观下属性权值定义） 一个决策表系统 $T=(U,C\cup D,V,f)$，其中 $C=\{a_1,a_2,\cdots,a_m\}$ 是条件属性集合，$D=\{d\}$ 是决策属性集合，设 $I(a_i,D)$ 表示条件属性 a_i 与决策属性 D 的互信息，则属性 a_i 的权值为：

$$\omega_i=\frac{I(a_i,D)}{\sum_{i=1}^{m}I(a_i,D)}$$

至此根据 Rough Set 理论中属性重要性概念的代数定义和信息熵定义，得到了关于属性权重的代数观和信息观定义。这两种定义具有互补的特性：属性重要性的代数定义考虑的是该属性对论域中确定分类子集的影响，而信息定义考虑的是该属性对于论域中不确定分类子集的影响。为了综合考虑属性对论域中确定分类子集和不确定分类子集的影响，下面建立基于代数观和信息观的属性权重最优化数学模型，得到了属性权重的最优解。

4 基于代数观和信息观的属性权重最优解

设一个决策表系统 $T=(U,C\cup D,V,f)$，其中 $C=\{a_1,a_2,\cdots,a_m\}$ 是条件属性集合，$D=\{d\}$ 是决策属性集合，α_i、β_i $(i=1,2,\cdots,m)$ 分别是属性 a_i 基于代数观和信息观的权重，ω_i 为两者的综合权重，$\sum_{i=1}^{m}\alpha_i=1$，$\sum_{i=1}^{m}\beta_i=1$，$\sum_{i=1}^{m}\omega_i=1$，$0\leqslant\alpha_i\leqslant 1$，$0\leqslant\beta_i\leqslant 1$，$0\leqslant\omega_i\leqslant 1$，$(i=1,2,\cdots,m)$。建立最优化模型：

$$\min\{\sum_{i=1}^{m}\left[\mu\left(\frac{1}{2}(\omega_i-\alpha_i)^2\right)+(1-\mu)\left(\frac{1}{2}(\omega_i-\beta_i)^2\right)\right]\}\qquad(1)$$

其中 $\omega_i\in\Omega=\{\omega_i|\sum_{i=1}^{m}\omega_i=1,0\leqslant\omega_i\leqslant 1,(i=1,2,\cdots,m)\}$，$0\leqslant\mu\leqslant 1$。

定理 1 最优化模型（1）在可行域 Ω 上有唯一解，且其解为：

$$\omega_i=\mu\alpha_i+(1-\mu)\beta_i,i=1,2,\cdots,m\qquad(2)$$

证明 作 Lagrange 函数：

$$L(\omega_i,\lambda)=\sum_{i=1}^{m}\left[\mu\left(\frac{1}{2}(\omega_i-\alpha_i)^2\right)+(1-\mu)\left(\frac{1}{2}(\omega_i-\beta_i)^2\right)\right]+\lambda\left(\sum_{i=1}^{m}\omega_i-1\right)$$

令 $\frac{\partial L}{\partial\omega_i}=0$，$\sum_{i=1}^{m}\omega_i-1=0$，$(i=1,2,\cdots,m)$，得方程组

$$\begin{cases}\mu(\omega_i-\alpha_i)+(1-\mu)(\omega_i-\beta_i)=0\\ \sum_{i=1}^{m}\omega_i-1=0\end{cases}\qquad(i=1,2,\cdots,m)$$

解此方程组得

$$\omega_i=\mu\alpha_i+(1-\mu)\beta_i,i=1,2,\cdots,m$$

从而定理得证。

5 实例

表 1 给出了一决策表系统 $T=(U,C\cup D,V,f)$，其中有限论

域 $U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8\}$,条件属性集 $C=\{a,b,c\}$,决策属性集 $D=\{d\}$;$V=\{0,1,2\}$。

表 1 决策表系统

U	a	b	c	d
u_1	1	1	0	0
u_2	1	0	1	1
u_3	1	0	2	1
u_4	0	1	0	0
u_5	0	0	1	0
u_6	0	1	2	1
u_7	0	0	1	1
u_8	0	1	2	0

解

(1)求代数观下的客观权重

$U/C=\{\{u_1\},\{u_2\},\{u_3\},\{u_4\},\{u_5,u_7\},\{u_6,u_8\}\}$

$U/D=\{\{u_1,u_4,u_5,u_8\},\{u_2,u_3,u_6,u_7\}\}$

$U/(C-\{a\})=\{\{u_1,u_4\},\{u_2,u_5,u_7\},\{u_3\},\{u_6,u_8\}\}$

$U/(C-\{b\})=\{\{u_1\},\{u_2\},\{u_3\},\{u_4\},\{u_5,u_7\},\{u_6,u_8\}\}$

$U/(C-\{c\})=\{\{u_1\},\{u_2,u_3\},\{u_4,u_6,u_8\},\{u_5,u_7\}\}$

$POS_C(D)=\{u_1,u_2,u_3,u_4\}$

$POS_{(C-\{a\})}(D)=\{u_1,u_4\}$

$POS_{(C-\{b\})}(D)=\{u_1,u_2,u_3,u_4\}$

$POS_{(C-\{c\})}(D)=\{u_1,u_4\}$

$\gamma_C(D)=\frac{4}{8};\gamma_{C-\{a\}}(D)=\frac{2}{8};\gamma_{C-\{b\}}(D)=\frac{4}{8};\gamma_{C-\{c\}}(D)=\frac{2}{8}$

从而得代数观下各权重分别为:

$\alpha_a=0.5,\alpha_b=0,\alpha_c=0.5$

(2)求信息观下的客观权重

由 $U/D=\{\{u_1,u_4,u_5,u_8\},\{u_2,u_3,u_6,u_7\}\}$

$U/\{a\}=\{\{u_1,u_2,u_3\},\{u_4,u_5,u_6,u_7,u_8\}\}$

$U/\{b\}=\{\{u_1,u_4,u_6,u_8\},\{u_2,u_3,u_5,u_7\}\}$

$U/\{c\}=\{\{u_1,u_4\},\{u_2,u_5,u_7\},\{u_3,u_6,u_8\}\}$

可得

$$H(D)=-\left(\frac{4}{8}\ \mathrm{lb}\ \frac{4}{8}+\frac{4}{8}\ \mathrm{lb}\ \frac{4}{8}\right)=1$$

$$H(D|\{a\})=-\frac{1}{8}(4-5\ \mathrm{lb}\ 5)=0.951\ 3$$

$$H(D|\{b\})=-\frac{1}{4}(3\ \mathrm{lb}\ 3-8)=0.811\ 3$$

$$H(D|\{c\})=-\frac{1}{4}(2-3\ \mathrm{lb}\ 3)=0.688\ 8$$

$I(a,D)=0.048\ 7,I(b,D)=0.188\ 7,I(c,D)=0.311\ 2$

从而得信息观下的客观权重分别为:

$\beta_a=0.088\ 8,\beta_b=0.344\ 0,\beta_c=0.567\ 2$

再由公式(2):$\omega_i=\mu\alpha_i+(1-\mu)\beta_i$,取 $\mu=0.7$,得属性 a、b、c 的综合客观权重分别为:

$\omega_a=0.376\ 6,\omega_b=0.103\ 2,\omega_c=0.520\ 2$

6 结束语

利用粗糙集理论中基于代数观和信息观的属性重要性概念,提出将在此基础上得到的客观权重进行有机的结合,从而得到权重确定的科学有效方法,建立了两者结合的最优化数学模型,得到了综合权重的最优解,彻底克服了以往方法的不足。该最优化模型同样可以应用于主、客观权重相结合的情形,从而使最终得到的权重结果更加合理、科学。

参考文献:

[1] 黄光明,张巍.基于 Rough Set 的综合评价方法研究[J].计算机工程与应用,2004,40(2):36-38.

[2] Pawlak Z.Rough set[J].Communication of ACM,1995,38(11):88-95.

[3] Pawlak Z.Rough set:theoretical aspects of reasoning about data[M].[S.l.]:Kluwer Academic Publishers,1992.

[4] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简[J].计算机学报,2002,25(7):759-766.

[5] 石峰,娄臻亮,张永清,等.一种改进的粗糙集属性约简启发式算法[J].上海交通大学学报,2002,36(4):478-481.

[6] 梁吉业,曲开社,徐宗本.信息系统的属性约简[J].系统工程理论与实践,2001(12):76-80.

[7] 王国胤.Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001.

[8] 袁亚湘,孙文瑜.最优化理论与方法[M].北京:科学出版社,2007.

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

(上接 44 页)

能够很快的跳出局部最优,提高算法收敛速度;

(2)结合演化算法,利用微粒种群信息找出更有效的逃逸方法,进一步增强算法的全局搜索能力。

参考文献:

[1] Kennedy J,Eberhart R C.Particle Swarm Optimization [C]//Proc IEEE Int Conf on Neural Networks IV,Piscataway,NJ:IEEE Service Center,1995:1942-1948.

[2] Suganthan P N.Particle swarm optimizer with neighbourhood operator[C]//Proceedings of the Congress on Evolutionary Computation,Washington DC,USA,1999:1958-1961.

[3] Kennedy J.Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance[C]//Proceedings of IEEE Congress on Evolutionary Computation,Piscataway,NJ:IEEE Service Center,1999:1931-1938.

[4] Riget J,Vesterstroem J S.A diversity-guided particle swarm optimizer-the ARPSO,No.2002-02[R].Department of Computer Science,University of Aarhus,EVALife,2002.

[5] 曾建潮,崔志华.一种保证全局收敛的 PSO 算法[J].计算机研究与发展,2004,41(8):1333-1338.

[6] 赫然,王永吉,王青,等.一种改进的自适应逃逸微粒群算法与实验分析[J],软件学报,2005,16(12):2036-2044.

[7] 曾建潮,介婧,崔志华.微粒群算法[M].北京:科学出版社,2004.

[8] Solis F,Wets R.Minimization by random search techniques[J].Mathematics of Operations Research,1981,6(1):19-30.