基于覆盖算法的条件信息熵表示及属性约简

单雪红[1,2],吴 涛[2,3],李国成[2]
SHAN Xue-hong[1,2],WU Tao[2,3],LI Guo-cheng[2]

1.宿州学院 数学系,安徽 宿州 234000
2.安徽大学 数学科学学院,合肥 230039
3 安徽大学 智能计算与信号处理教育部重点实验室,合肥 230039
1.Deparement of Mathematics,Suzhou College,Suzhou,Anhui 234000,China
2.School of Mathematics Science,Anhui University,Hefei 230039,China
3.Key Lab IC&SP,Anhui University,Hefei 230039,China
E-mail:shanxuehong99@yahoo.com.cn

SHAN Xue-hong,WU Tao,LI Guo-cheng.Conditional information entropy based on covering algorithm and attributes reduction.Computer Engineering and Applications,2009,45(31):115-117.

Abstract: Processing data can be partitioned using covering algorithm.In this paper,a new concept of conditional information entropy is put forward,and then the new significance of an attributes is defined based on this entropy.In a consistent decision table,the equivalence between algebraic significance and conditional information entropy significance of attributes is proved.But it is incorrect in the inconsistent decision table.A heuristic algorithm for knowledge reduction is designed.The experimental results show that this algorithm can find the minimal or optimal reduction.

Key words: covering algorithm;rough set theory;reduction;conditional information entropy

摘 要:利用覆盖算法对数据进行处理,得到论域 U 的一个划分,定义一种基于覆盖的条件信息熵,由新的条件信息熵定义新的属性重要性,并证明了对于一致决策表,它与代数定义下的重要性是等价的。以新的属性重要性为启发信息设计约简算法,并给出计算新的条件信息熵的算法。实验结果表明该约简算法能快速搜索到最优或次优约简。

关键词:覆盖算法;Rough 集理论;知识约简;条件信息熵
DOI:10.3778/j.issn.1002-8331.2009.31.034 **文章编号**:1002-8331(2009)31-0115-03 **文献标识码**:A **中图分类号**:TP18

1 引言

Rough 集理论建立在分类机制的基础上,将知识理解为对数据的划分,是在特定空间上由等价关系构成的划分[1-2]。知识约简是 Rough 集理论中的核心问题之一,搜索所有约简或最小约简被证明是一个 NP 完全问题[1]。因此一般采用启发式算法搜索最优或次优约简[3-4]。覆盖算法是一种基于 M-P 神经元的构造性神经网络算法,是把样本集 U 的分类问题转化成在样本空间构造覆盖簇$\{C_i\}$,使每个覆盖只盖住同一类点 C_i 且覆盖整个$\{X_1,X_2,\cdots,X_n\}$,覆盖网络中有多少覆盖领域,就会形成多少个等价的样本集,它构成了 U 的一个划分。根据这个划分定义新的条件信息熵和新的属性重要性,并与代数观点作了一些比较,同时以此重要性为启发信息设计约简算法。

2 Rough 集的基本概念

下面简要介绍 Rough 集理论的相关概念,其详细定义可参阅文献[3,5]。

在决策表 $S=(U,C\cup D,V,f)$中,C 为条件属性,D 为决策属性。对与任意的 $B\subseteq C\cup D$,由 B 确定的不可区分关系为 $IND(B)=\{(x,y)\in U\times U|\ \forall\ b\in B,b(x)=b(y)\}$对象 X 在 $IND(B)$ 中的等价类为$[X]_B$。$IND(B)$在 U 上导出的划分记为 $U/IND(B)$,简记为 U/B。

定义 1 决策表 $S=(U,C\cup D,V,f)$属性集合 $A\subseteq C$ 对决策属性 D 的相对正域是 $POS_A(D)=\bigcup_{x\in U/D}A\_X$ 并记 $r_A(D)=|POS_A(D)|/|U|$。

定义 2 决策表 $S=(U,C\cup D,V,f)$中决策规则 d_x 和决策表中任意其它决策规则 d_y,由 $d_x|C=d_y|C$,可得 $d_x|D=d_y|D$,则称决策

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318108,No.2007BC311003);国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675031);中国博士后基金面上项目(No.20070411028);安徽省高等学校省级自然科学研究项目(No.KJ2008B093,No.KJ200845ZC);安徽大学学术创新团队和安徽大学人才队伍建设经费资助。

作者简介:单雪红(1980-),女,硕士研究生,研究方向为智能计算与信息处理;吴涛(1970-),博士,副教授,主要从事机器学习、智能计算的应用研究工作;李国成(1976-),硕士研究生。

收稿日期:2008-06-24 **修回日期**:2009-09-20

规则 d_x 是一致的,否则,称决策规则 d_x 是不一致的,如果 S 中每条决策规则都是一致的,则称决策表 S 是一致的,否则称决策表 S 是不一致的。

定义 3(属性重要性的代数观点) 在决策表 $S=(U,C\cup D,V,f)$ 中,$A\subseteq C$,则任意 $a\in C-A$ 的属性重要性为 $Sig_1(a,A,D)=r_{A\cup a}(D)-r_A(D)$。

定义 4 决策表 $S=(U,C\cup D,V,f)$ 属性集合 $A\subseteq C$,$a\in A$,如果 $POS_A(D)=POS_{A-\{a\}}(D)$,则称 a 为 A 中 D 不必要的,否则称 P 中 D 必要的。如果 P 中的每一个 P 都为必要的,则称 P 为 D 独立的。

定义 5 决策表 $S=(U,C\cup D,V,f)$ 属性集合 $A\subseteq C$,A 为 C 的 D 约简当且仅当 A 是 C 的独立子集且 $POS_A(D)=POS_C(D)$。

C 的所有 D 约简的交称为 C 的 D 核,简称为相对核。记为 $core_D(C)$。

3 覆盖算法

张铃教授于 1997 年给出了 M-P 神经元的几何意义[6],指出用三层神经网络构造分类器等价于求出一组领域,这组领域能将不同类的点分开,并进一步给出覆盖算法。该算法的主要思路是:先求一个领域 $C1$,它只覆盖一类中的点,而不覆盖其他类的点,对余下的点求二类覆盖领域 $C2$,它只覆盖二类中的点而不覆盖其他类的点,如此交叉进行覆盖,直到样本集中的点均被领域覆盖了为止。

覆盖算法在使用中要求样本空间中的向量的长度都相等,即样本空间中的样本点都位于 $n+1$ 维空间中某个中心在原点的球面 Sn 上(若不然,可通过变换 $T:D\to Sn$,$T(x)=(x,r^2-|x|^2)$,其中 $d\geqslant\max\{|x||x\in D\}$,将样本点投影到球面 Sn 上)则 $\boldsymbol{w}^{\mathrm{T}}x-\theta>0$ 表示球面上由超平面 $\boldsymbol{w}^{\mathrm{T}}x-\theta=0$ 所分割的正半空间的部分,称为球面上的“球形领域”,若 W 与 x 等长,则 W 就是这个球形领域的中心。每个球形领域作为一个神经元。取其激励函数为 $\sigma(x)=\begin{cases}1 & x>0\\ 0 & \text{其他}\end{cases}$,则激励函数就是“球形领域”的特征函数。

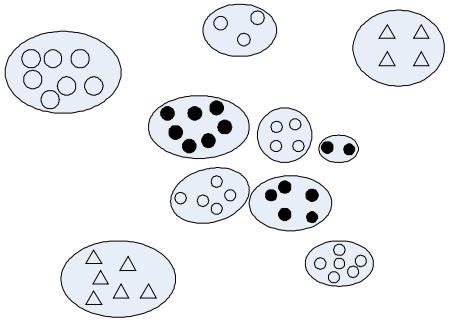

图 1 原样本覆盖图

通过这种变换就将神经网络的最优设计问题转化成某种最优覆盖问题。

为了形象说明覆盖算法问题[7],图 1 从二维覆盖图进行说明。

样本共分为黑实心点、空心圆点、空心三角点 3 类,分别用 $C1,C2,C3$ 表示,分别有 3、5、2 个覆盖领域,每个领域的覆盖样本数不同。覆盖领域内的样本有以下特点:一是同一覆盖领域内的样本具有相同的类标记,样本间位置相近,因此样本具有相似性;二是同一类标记的样本,如果差异较大,不会在同一覆盖领域内,会形成多个覆盖领域;三是投影位置相近的样本,如果类别不同,也不可能聚集在同一 覆盖领域。这些独特的特征使得覆盖网络中的每个覆盖领域就是一个很好的模式。

从上面的分析可知同一覆盖领域的样本具有很强的相似性,因此可以定义覆盖领域等价关系的概念,即将同一覆盖领域样本看成是不可分辨的等价关系。

设 x,y 是覆盖网络的两个样本,若 (x,y) 属于同一覆盖领域 C_j^t,则称 x 与 y 在整个覆盖网络中是不可分辨的。U/R 是 U 上由 R 生成的等价类全体,它构成了 U 的一个划分。即 $U/R=\{X_1,X_2,\cdots,X_n\}$,其中 $X_i\subseteq U$ 是等价类,对应的是第 i 个覆盖领域中的样本,$X_i\neq\Phi$,$X_i\cap X_j=\Phi$,且 $i\neq j,i,j=1,2,\cdots,n$,$\cup X_i=U$。这样有多少覆盖领域,就能形成相同数量的等价类。

4 条件信息熵

定义 6 给定样本集 S 分为 k 类,表示为集合 $S=\{S_1,S_2,\cdots,S_k\}$,$S_i(i=1,2,\cdots,k)$ 为属于第 i 类的样本集。如果覆盖集 $C=\{C_1,C_2,\cdots,C_m\}$,$C_j(j=1,2,\cdots,m)$ 均为覆盖,满足 $C_i\cap C_j$ 为空集($i\neq j$),并且每个 C_j 只和一个 S_i 相交以及 C_j 的并覆盖整个 S,则称 C 为 S 的划分覆盖集,由此划分得到的等价关系记为 R_C。

R_C 是根据 S 中的数据计算出来的等价关系, 但 Rough 理论一般根据属性集来获得对应的等价关系。为方便起见,R_C 可以看作一个属性,由这个属性导出的等价关系便是 R_C。

定义 7(新的条件信息熵) R_C 在 U 上的子集组成的 σ 代数上的概率分布定义为 $[R_c:p]=\begin{bmatrix}C_1 & C_2 & \cdots & C_m\\ p(C_1) & p(C_2) & \cdots & p(C_m)\end{bmatrix}$。其中 $P(C_i)=|C_i|/|U|$,$i=1,2,\cdots m$。对任意的 $P\subseteq C$(条件属性),设 $U/P=\{X_1,X_2,\cdots,X_n\}$,可得新的条件信息熵 $H(R_C|P)=-\sum_{i=1}^{n}p(X_i)\times\sum_{j=1}^{m}p(C_j|X_i)\log(p(C_j|X_i))$。其中 $p(C_j|X_i)=|C_j\cap X_i|/|X_i|$。

定义 8(新的属性重要性) 设在决策表 $S=(U,C\cup D,V,f)$ 中,$A\subseteq C$,则任意 $a\in C-A$ 的属性重要性为 $Sig_2(a,A,R_C)=H(R_C|A)-H(R_C|A\cup a)$。

根据文献[4]中的引理,可得如下定理:

定理 1 设在决策表 $S=(U,C\cup D,V,f)$ 中,$A\subseteq C$,任意 $a\in C-A$,如果 $Sig_2(a,A,R_C)=0$,则 $Sig_1(a,A,D)=0$。

由定义 3 可知,$Sig_1(a,A,D)$ 只对正域的基数进行定量描述。定理 1 说明,如果增加属性后正域基数变大,即 $Sig_1(a,A,D)\neq0$,这必然会在 $Sig_2(a,A,R_C)$ 上有所反应,此时 $Sig_2(a,A,RC)\neq0$,这就说明 $Sig_2(a,A,R_C)$ 包含了比 $Sig_1(a,A,D)$ 更多的信息。

由覆盖算法领域内样本的特点可知,划分 R_C 能将属于不同决策类的对象分开, 也能把同一决策类的差异较大的对象分开。

可见 $Sig_2(a,A,R_C)$ 有以下优点:

(1)$Sig_2(a,A,R_C)$ 定义在 R_C 上,R_C 能区分开不同决策类的对象和同一决策类的差异较大的对象,因此增加属性 a 后能够分开不同决策类的对象和同一决策类的差异较大的对象。

(2)由于 $Sig_2(a,A,R_C)$ 能更准确全面地描述信息,以它为启发信息的约简算法更有可能得到最小或次优约简。如下面的例 2 以 $Sig_2(a,A,R_C)$ 为启发信息的约简可得最小约简 $\{a,e,f\}$,而以 $Sig_1(a,A,D)$ 为启发信息的约简为 $\{a,b,c,f\}$。

定理 2 决策表 $S=(U,C\cup D,V,f)$ 是一致的,其充分必要条

件是 $H(R_C|C)=0$。

证:设 $U/C=\{X_1,X_2,\cdots X_n\}$,$U/R_C=\{Y_1,Y_2,\cdots Y_m\}$。

(必要性)任取 $X_i\in U/C$,由于 S 是一致的,则对任意的 x,$y\in X_i$,有 $d_x|R_C=d_y|R_C$,也即存在 $Y_j\in U/R_C$,使得 $X_i\subseteq Y_j$,所以对任意 $Y_j\in U/R_C$,存在$\{1,2,\cdots,n\}$的子集 E_k,使得 $Y_j=\bigcup_{i\in E_k}X_i$,所以 $Y_j\cap X_i=X_i$,或者 $Y_j\cap X_i=\Phi$ 成立。($i=1,2,\cdots,n;j=1,2,\cdots m$),则 $P(Y_j|X_i)=1$ 或者 $P(Y_j|X_i)=0$,所以 $H(R_C|C)=0$。

(充分性)设 $H(R_C|C)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\mathrm{lb}(p(Y_j|X_i))=0$,由于 $p(X_i)\neq 0,i=1,2,\cdots,n$,对任意的 $i,j(i=1,2,\cdots,n;j=1,2,\cdots m)$有 $P(Y_j|X_i)=1$ 或 $P(Y_j|X_i)=0$,则 $Y_j\cap X_i=X_i$,或 $Y_j\cap X_i=\Phi$ 之一成立。所以,如果 $X_i\in U/C$,则一定存在 $Y_j\in U/R_C$,使得 $X_i\subseteq Y_j$ 也即对任意决策规则 d_x 和 d_y,如果 $d_x|C=d_y|C$,则 $d_x|R_C=d_y|R_C$,所以决策表 S 是一致的。

定理 2 说明,决策表是否一致可以由条件信息熵的值来判断。

推论 1 决策表 $S=(U,C\cup D,V,f)$是不一致的,其充分必要条件是 $H(R_C|C)>0$。

定理 3 设 $S=(U,C\cup D,V,f)$是一致决策表,$A\subseteq C$,则 A 是 C 的一个 D 约简的充分必要条件是:

(1)对任意的 $a\in A$ 有 $H(R_C|A-\{a\})>0$。

(2)$H(R_C|A)=0$。

证:由约简的定义 A 是 C 的一个 D 约简的充分必要条件:(1)A 关于 D 独立;(2)$POS_A(D)=POS_C(D)$,而条件(2)可改为决策表 $S1=\{U,A\cup D,V,f\}$是一致的。因此这两个条件等价于(1)对任意的 $a\in A$ 有 $H(R_C|A-\{a\})>0$;(2)$H(R_C|A)=0$。

以上定理说明对于一致决策表,其属性约简用新的条件信息熵表示和代数表示是等价的。对不一致决策表,考察下面的例子

例 1 在表 1 的决策表中,条件属性 $C=\{a,b,c\}$,决策属性 $D=\{d\}$。

表 1　决策表

U	a	b	c	d
1	0	0	0	1
2	0	0	1	2
3	2	1	2	3
4	1	0	2	4
5	0	0	0	5
6	2	2	2	3
7	2	0	1	2
8	0	0	1	5

用覆盖算法对表 1 进行处理的得到的划分:

$U/R_C=\{Y_1,Y_2,\cdots Y_5\}=\{\{1\},\{2\},\{3,6\},\{4\},\{5\},\{8\},\{7\}\}$

$U/\{a,b\}=\{Z_1,Z_2,\cdots Z_5\}=\{\{1,2,5,8\},\{3\},\{4\},\{6\},\{7\}\}$

$U/\{a,b,e\}=\{X_1,X_2,\cdots X_6\}=\{\{1,5\},\{2,8\},\{3\},\{4\},\{6\},\{7\}\}$

$Pos_C(D)=\{3,4,6,7\}=Pos_{C-\{e\}}(D)$由代数观点知属性 e 是 D 可省的,但是

$$H(R_C|C)=-\sum_{i=1}^{6}p(X_i)\sum_{j=1}^{5}p(Y_j|X_i)\log(p(Y_j|X_i))=$$

$$\frac{2}{8}\times(\frac{1}{2}\mathrm{lb}\frac{1}{2}+\frac{1}{2}\mathrm{lb}\frac{1}{2})+\frac{2}{8}\times(\frac{1}{2}\mathrm{lb}\frac{1}{2}+\frac{1}{2}\mathrm{lb}\frac{1}{2})=0.346\ 6$$

而 $H(R_C|\{a,b\})=-\sum_{i=1}^{5}p(X_i)\sum_{j=1}^{5}p(Y_j|X_i)\log(p(Y_j|X_i))=$

$$\frac{4}{8}\times(\frac{1}{4}\mathrm{lb}\frac{1}{4}+\frac{1}{4}\mathrm{lb}\frac{1}{4}+\frac{1}{4}\mathrm{lb}\frac{1}{4}+\frac{1}{4}\mathrm{lb}\frac{1}{4})=0.693\ 1$$

这里 $H(R_C|\{a,b\})>H(R_C|C)$,可见不一致决策表中属性约简的代数表示和新的条件信息熵的表示是不等价的。

定理 3[3] 设论域为 U,某个等价关系在 U 上形成的划分为 $A_1=\{X_1,X_2,\cdots X_n\}$,而 $A_2=\{X_1,X_2,\cdots,X_{i-1},X_{i+1},\cdots X_{j-1},X_{j+1},\cdots X_n\}$是将划分 A_1 中的某两个等价块 X_i 与 X_j 合并为 $X_i\cup X_j$ 得到的新划分 $B=\{Y_1,Y_2,\cdots Y_m\}$也是 U 上的一个划分,则 $H(B|A_1)\geqslant H(B|A_2)$。

推论 2 设决策表 $S=(U,C\cup D,V,f)$,任意 $a_i\in C,i=1,2,\cdots n(n=|C|)$,则有 $H(B|\{a_1\})\geqslant H(B|\{a_1\}\cup\{a_2\})\geqslant H(B|\{a_1\}\cup\{a_2\}\cup\cdots\{a_n\})=H(B|C)$。

推论 2 说明条件信息量的变化呈现单调增加的

推论 3 设决策表 $S=(U,C\cup D,V,f)$,A 是约简 C_0 为核,如果 $a_i\in A-C_0$ 是任意一个不能约简的属性,则有 $H(R_C|C_0)>H(R_C|C_0\cup\{a_1\})>\cdots>H(R_C|C_0\cup\{a_1\}\cup\cdots\cup\{a_i\})>\cdots>H(R_C|A)$。

推论 3 说明如果属性约简以核为起点,那么在约简过程中,条件信息熵的变化是单调递减的。

5　知识约简算法

以 $Sig_2(a,A,D)$为启发信息的约简算法中,每次循环时条件属性子集 A 均不变,这使得 $Sig_2(a,A,D)$最大的 a 就是 $H(R_C|A\cup a)$最小的 a。因此只需计算 $H(R_C|A\cup a)$,这样可避免计算 $H(R_C|A)$,减少了计算量。

属性约简算法:

输入:决策表 $S=(U,C\cup D,V,f)$,其中 U 为论域,C 为条件属性,D 为决策属性。

输出:决策表 S 的一个相对约简 A。

步骤 1 由覆盖算法得出 R_C,计算 $H(R_C|C)$。

步骤 2 计算条件属性集 C 相对于决策属性集 D 的核属性集 $CORE_D(C)$,并令 $B=C-CORE_D(C)$。

步骤 3 令 $A=CORE_D(C)$:

(1)如果$|A|\neq 0$,则计算条件熵 $H(R_C|A)$,转(4);

(2)对每个属性 $a\in B$,计算 Rc 相对条件属性集 $A\cup\{a\}$的条件熵 $H(Rc|A\cup\{a\})$;

(3)选择使 $H(Rc|A\cup\{a\})$最小的属性 a,把 a 从 B 中删除,并把 a 增加到 A 的尾部;

(4)如果 $H(R_C|A)=H(R_C|C)$,则终止,否则转(2)。

例 2 在表 2 的决策表中,条件属性 $C=\{a,b,c,e,f\}$,决策属性 $D=\{d\}$。

表 2　决策表

U	a	b	c	e	f	d
1	0	0	0	0	1	0
2	0	1	1	1	0	1
3	1	1	0	1	1	1
4	0	1	1	1	0	0
5	0	0	1	0	1	0
6	1	1	0	1	0	1
7	0	1	1	1	1	1
8	1	1	1	0	1	1
9	1	1	0	1	1	0
10	0	1	1	1	1	0

(下转 124 页)