基于一致性度量的属性约简的研究

陈 堃,李心科
(合肥工业大学 计算机与信息学院;安徽 合肥 230009)

摘 要:粗糙集理论作为一种处理不精确和不一致数据的数学工具被广泛应用于特征子集选择和属性约简中。在大多数现存的算法中,属性依赖度被用来度量特征子集的重要性,而依赖度在处理不一致信息系统时会出现找不到任何特征子集的问题。文中讨论了使用属性依赖性作为度量的缺点和不足,引入一种一致性度量,分析了其和依赖性之间的关系,重新定义了信息系统的多余属性和约简的概念,并构造了基于一致性度量的前向贪婪搜索算法。通过UCI数据集合验证了算法能够有效地处理不一致信息系统。

关键词:粗糙集;属性重要度;一致性;依赖度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)10-0064-04

Research of Attribute Reduction Based on Consistency Measure

CHEN Kun, LI Xin-ke
(School of Computer & Information, Hefei University of Technology, Hefei 230009, China)

Abstract: As a new mathematic method which analyses and treats the inexact, incomplete information and knowledge, rough sets has been widely used in feature subset selection and attribute reduction. In most of the existing algorithms, the dependency measure is employed to evaluate the quality of a feature subset. In this paper, discuss the disadvantages and problems of using dependency, and introduce the consistency measure to deal with the problems. The relationship between dependency and consistency is analyzed. Redefine the redundancy and reduction of rough sets, and construct a greed search algorithm to find the reduction based on consistency. The experimental results with UCI data set show that the new algorithm is effective and efficient.

Key words: rough set; attribute significance; consistency; dependency

0 引 言

粗糙集理论[1]是一种处理不一致和不精确数据的有力工具,它是由Pawlak教授于1982年提出来的。近年来它已经被广泛地应用于数据挖掘、人工智能、模式识别和故障诊断等领域。属性约简是粗糙集理论中的重要内容,而基于正区域的属性约简是其中被研究的较多的分支之一。人们通常希望找到具有最少属性的约简,即最优约简,而找到最优约简是一个NP-hard问题[2],所以通常采用添加启发式搜索信息来改进搜索策略,求出最优或次优约简。

以属性依赖度[3~5]作为启发信息的是一种比较常用的方法。粗糙集理论中的依赖度就是其正区域与样本空间大小的比值。笔者在研究基于正域的约简算法过程中发现,在某些应用中,以依赖度为度量的算法发现不了任何的特征子集。所以文中引入了一种一致性度量函数来代替属性依赖度作为度量的标准,分析了一致性和依赖度之间的关系,并且使用一致性度量来构造前向贪婪搜索约简算法。在几种离散化方法的处理下,分别对相应的数据集合用基于依赖度的算法和文中构造的算法进行研究,实验数据表明基于一致性度量比基于依赖性的度量有更好的区分能力。

1 粗糙集理论的基本概念

针对粗糙理论,这里给出主要概念的形式化定义:

定义1 决策信息系统是有序对 $K=(U,A\cup\{d\})$,其中 U 是非空有限集合,称为论域。A 是非空有限的条件属性集合。d 是一个决策属性。$A\cup\{d\}=\varnothing$。在任意子集 $B\subseteq A$ 上,有等价关系 IND(B),称为不可区分关系:

$$IND(B)=\{(x,y)\in U*U\mid x(a)=y(a),\forall a\in B\}$$

其中 x,y 为 U 中的元素。

收稿日期:2008-01-20
基金项目:安徽省科技计划项目(0012021A)
作者简介:陈 堃(1981-),男,安徽合肥人,硕士研究生,研究方向为软件工程、数据挖掘;李心科,副教授,硕士生导师,研究方向为软件工程、数据挖掘、神经网络。

定义2　给定信息系统 $K=(U,A\cup\{d\})$,设 $X\subseteq U$ 是一组对象,$B\subseteq A$ 是一组属性。X 的相对于 B 的下近似是 $B_-(X)=\{x\in U\mid[x]_B\subseteq X\}$。$X$ 相对于 B 的上近似为 $B^-(X)=\{x\in U\mid[x]_B\cap X\neq\varnothing\}$。$X$ 相对于 B 的正区域 $\mathrm{Pos}_B=B_-(X)$,负区域 $\mathrm{Neg}_B(X)=U-B^-(X)$,边界域为 $\mathrm{Bn}_B(X)=B^-(X)-B_-(X)$。则决策属性 d 相对于 B 的正区域为 $\mathrm{Pos}_B(d)=\bigcup\{B_-(X)\mid X\in U/\mathrm{IND}(d)\}$。

定义3　属性 D 以程度 $r_p(D)$ 依赖于 K 中的 A 的子集 P,若 $r_p(D)=|\mathrm{Pos}_p(d)|/|U|$。

定义4　属性 $a\in B\subseteq A$ 是可缺少的,如果 $\mathrm{Pos}_B(d)=\mathrm{Pos}_{B-\{a\}}(d)$。属性子集 $C\subseteq B\subseteq A$ 是 B 的一个约简,当(1)$\mathrm{Pos}_B(d)=\mathrm{Pos}_C(d)$;(2)对于任意 $C'\subseteq C$(1)中所述不成立。其中 B 的约简不唯一,所有约简的交集称为 B 的核,记做 $\mathrm{Core}(B)=\bigcap\mathrm{Red}(B)$。

定义5　属性 a 相对于 P 对于 D 的依赖程度的重要性为:$\mathrm{SGF}(a,P,D)=r_{p+\{a\}}(D)-r_p(D)$

2　基于一致性度量属性约简

2.1　一致性的相关概念

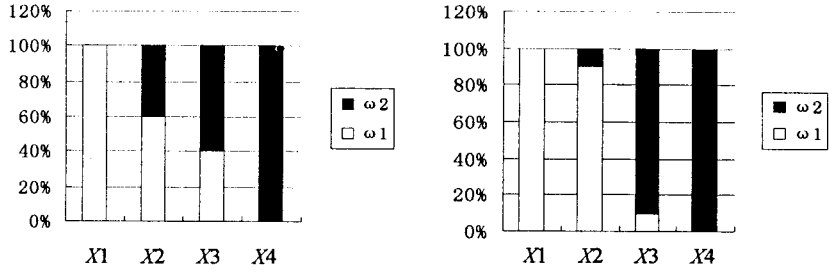
经典粗糙集理论中的属性依赖度函数是指样本中所有正域的大小与样本本身的比值。正域就是样本集合中根据现有属性可以毫无疑问的正确分类的样本子集。在前向搜索算法中,经常是从一个空的集合开始进行的,然后一次次地从候选属性集合中向约简集合中添加。在第一轮中,要计算每个属性的依赖度,并且选择依赖度最大的属性。有时,会发现在某些应用中计算出来的最大的依赖度为零,这也就意味着通过候选属性不能够区分样本集合中的任何元素。因为根据依赖度函数的定义,依赖度应该是大于零的,所有就没有任何的属性被选出来,特征选择算法也就发现不了任何东西。但事实是,虽然单个属性不能够区分样本,但某些属性的组合却可以区分样本集合中所有的元素。

表1所示的决策表,若按照属性依赖度进行前向贪婪属性约简的话,那么第一轮是不会发现任何属性的依赖度大于零,算法中止,且发现不了任何的特征子集,但其实 $\{a,d\}$ 是它的一个约简。

图1中展示的是一个离散空间上的二元分类问题,其中样本元素依据它们的特征值被分成一个等价类的非空集合 $\{X_1,X_2,\cdots,X_k\}$,有相同特征值的元素被分到同一个组中去。其中有些等价类是一致的 X_1 和 X_4,因为其中的元素属于同一个决策类,同时也会存在一些不一致的等价类。比如图中的 X_2,X_3。根据经典粗糙集理论,它们是属于决策边界域中的,而一致性等价类则是属于决策正域。

表1　一个决策表

	a	b	c	d	D
1	1	1	1	1	0
2	2	2	2	1	1
3	1	1	1	1	0
4	2	3	2	2	0
5	2	2	2	1	1
6	3	1	2	1	0
7	1	2	3	2	2
8	2	3	1	2	3
9	3	1	2	1	1
10	1	2	3	2	2
11	3	1	2	1	1
12	2	3	1	2	3
13	3	3	1	2	1
14	1	2	3	2	3
15	2	3	2	2	2



(a) 分类情况1　　(b) 分类情况2

图1　离散空间中的分类复杂度的两种情况

从图1中a,b的比较,可以发现不一致样本出现的概率是相同的,但其错误分类的概率[6]却是不同的。经典粗糙集理论的属性依赖度并不能真实反映这样的不同。因此在特征选择中给定一个度量来反映不一致区域的大小是十分必要的。在文献[6]中 Dash 和 Liu 提出了一种能够度量这种区别的一致性函数。下面就介绍一下一致性的基本定义,并重新定义决策信息系统中的多余和属性约简的概念。

定义1:给定信息系统 $K=(U,A\cup\{d\})$,$\exists x,y\in U$ 使得 $x(A)=y(A)$,而 $x(d)\neq y(d)$,则称信息系统是不一致的。

定义2:给定信息系统 $K=(U,A\cup\{d\})$,$P\subseteq A$ 是条件属性的一个子集,论域在 P 上的分类为 $\{P_1,P_2,\cdots,P_n\}$,那么一个分类 P_i 中的不一致数量 ε_i 就定义为样本中分类 P_i 出现的次数与分类 P_i 中不同决策类中个数最大的类的元素数量之差。

定义3:给定信息系统 $K=(U,A\cup\{d\})$,其上的不一致率定义为论域中所有分类的不一致数量之和

与论域元素总数的比值，即 $\sum \varepsilon_i / |U|$，相对的，一致率就是 $\delta = (|U| - \sum \varepsilon_i)/|U|$。

基于上面的分析就可以理解依赖度反映的是样本能够正确分类的比率，而一致性则反映的是样本可能正确分类的比率。那么现在就存在着两种样本 $Pos_B(D) \cup Q$。其中 $Pos_B(D)$ 是一致性样本集合，而 Q 则是那些边界域中有着最大数量分类标志的样本，称之为伪一致样本。称 $Pos_B(D) \cup Q$ 为伪正域。下面来分析一致性和依赖性之间的关系。

性质 1：给定一个决策信息系统，$K = (U, A \cup \{d\})$，$B \subseteq A$，有：

$0 \leqslant \delta_B(D) \leqslant 1$，并且有 $r_B(D) \leqslant \delta_B(D)$

性质 2：给定一个决策信息系统，$K = (U, A \cup \{d\})$，如果有 $B_1 \subseteq B_2 \subseteq A$，则有：$\delta_{B_1}(D) \leqslant \delta_{B_2}(D)$

性质 3：给定一个决策信息系统，$K = (U, A \cup \{d\})$，当且仅当 $U/C = U/D$ 系统是一致的，那么有：$r_C(D) = \delta_C(D) = 1$

定义 4：给定一个决策信息系统，$K = (U, A \cup \{d\})$，$B \subseteq A$，$a \in B$，就说条件属性 a 相对于 B 对于决策属性 D 是不可缺少的，如果有 $\delta_{B-a}(D) \leqslant \delta_B(D)$，否则称 a 是多余的。称 $B \subseteq A$ 是不可缺少的，当 B 中所有属性对于 B 都是不可缺少的。

$\delta_B(D)$ 反映的不仅仅是正域的大小，同时也包含了边界样本。当一个属性被删除时，一致性不发生变化，则称这个属性是多余的。这里“多余”有两层意思，一是表示属性是相关的，同时也是多余的；二则表示属性是不相关的。所以一致性能够发现这两种“多余”的属性。

定义 5：一个属性子集 B 是决策系统的基于一致性的属性约简，当：

1) $\delta_B(D) \leqslant \delta_C(D)$

2) $\forall a \in B, \delta_{B-a}(D) < \delta_B(D)$

定义 6：属性 a 相对于子集 B 对于决策属性的重要性则定义为：

$$SGF(a, B, D) = \delta_{B+\{a\}}(D) - \delta_B(D)$$

2.2 约简算法

因为 $SGF(a, B, D)$ 是随着新属性的增加而线性增长的[6]，基于此构造了一个前向贪婪搜索算法。

算法：前向贪婪算法

输入：决策信息系统 $K = (U, A \cup \{d\})$

输出：约简 Red

Step1：初始化 $Red = \Phi$，$\delta_\Phi(D) = 0$

Step2：对于任意的每一个 $a_i \in A - Red$

计算 $SGF(a_i, B, D) = \delta_{B+\{a_i\}}(D) - \delta_B(D)$

Step3：选出其中 $SGF(a_i, B, D)$ 最大的值，以及相应的属性 a_i

Step4：如果 $SGF(a_i, B, D)$ 不为 0

$Red = Red \cup a_i$

返回 Step2

否则 Return Red

Step5：结束

3 实验分析

实验的主要目的有两个，一是以属性依赖为度量的方法和文中方法在属性选择上的比较；二是对两种方法的分类精度进行比较。选用 UCI 机器学习数据库中数据集合进行分析，数据集合结构如表 1 所示。这里数据集合有些是连续型数字属性的，这里使用 FCM、等频率和信息熵的方法分别对数字属性进行离散化处理。然后在处理后的数据集合上验证两种度量方法的算法。

表 2 数据集合描述

数据集	实例个数	属性个数	决策类别
Ionosphere	351	34	2
Ecoli	336	7	7
Sonar	208	60	2
WPBC	198	33	2

表 2 中可以发现基于依赖性为度量的算法的严重的问题（其中 D 代表依赖性度量，C 代表一致性度量），算法在某些数据集合中只能选取 2 个或者更少的属性，而相比较而言基于一致性度量的算法却能选取中等数量的属性。在一些集合中，基于依赖性度量的算法甚至找不到任何的属性，原因就是依赖度关注的仅仅是正区域的比率，而贪婪算法是从空集合开始，每次将一个依赖度最大的属性放入其中，直到依赖度达到最大值，形成约简。而对于某些集合来说，其每个属性的依赖度为零，因此在第一轮的搜索中就没有任何的属性会被选出，算法停止且没有选择出任何的特征。一致性度量克服了这个问题，因为它能够反映出边界域中元素对象的作用。

表 3 不同离散方法下的特征选择比较

数据集	原始属性个数	FCM		等频率		信息熵	
		D	C	D	C	D	C
Ionosphere	34	10	9	1	7	10	8
Ecoli	7	1	6	7	7	1	7
Sonar	60	6	6	0	6	0	14
WPBC	33	7	7	0	6	11	7

用 Cart 学习算法在经过约简的数据集合上训练分类器。并且通过十折交叉验证来测试约简的分类精

度。表3中给出了Cart算法下的平均分类精度。从表中可以看出,当属性集合中的大部分属性被消除后,保留合适数量的属性可以保持甚至提高分类精度。而对于某些数据集合,由于依赖性度量的局限性,使得分类精度有较大的下降。

表4　平均分类精度

数据集	原始分类精度	FCM		等频率		信息熵	
		D	*C*	*D*	*C*	*D*	*C*
Ionosphere	0.88	0.91	0.91	0.75	0.90	0.93	0.89
Ecoli	0.82	0.42	0.81	0.82	0.81	0.42	0.82
Sonar	0.72	0.70	0.70	0	0.74	0	0.74
WPBC	0.70	0.70	0.70	0	0.71	0.68	0.69

由于基于一致性度量的方法将论域元素中那些边界元素也考虑进去,所以在实际过程中论域中的大部分样本会由约简中少数属性区分开来,而剩下的属性则用来区分那些少数样本。这样就会导致过度拟合问题的出现,所以需要合适的剪枝算法来避免过度拟合的出现。这里使用十折交叉验证来测试属性集合,选择最佳精度和最佳约简作为最终的输出。表4给出了十折交叉验证后选择的最佳约简和对应精度。

表5　十折校验前后分类精度比较

数据集	原始数据集合(Cart)		分类精度(Cart)	
	特征属性	分类精度	特征属性	分类精度
Ionosphere	34	0.88	3	0.94
Sonar	60	0.72	3	0.76
WPBC	33	0.78	5	0.82

4　结束语

介绍了一种基于一致性度量的方法来克服使用依赖性为度量的方法所出现的问题,讨论了一致性和依赖性之间的关系,分析了一致性的概念和相关性质。基于此重新定义了粗糙集中的多余和约简的概念,并且构造了一个前向贪婪搜索算法。实验分析证明构造的方法是有效的。

与依赖性相比,一致性反映的不仅仅是正域中的元素,而且还包括了边界域中的元素,因此一致性函数比依赖性函数有着更好的区分能力。但同时由于一致性度量的特点,其最终的约简集合中也许包含着只是区分极少量样本的属性,这样就会造成数据的过度拟合。因此有效的剪枝算法是十分必要的,通过十折交叉验证测试实验结果以求获得更有效的特征子集。

参考文献:

[1] Pawlak Z. Rough Set[J]. International Journal of Information and Computer Sciences,1982,11(5):341－356.

[2] Wong S K, Ziarko W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences,1985,33:693－696.

[3] Jensen R, Shen Q. Semantics－prserving dimensionality reduction: Rough and fuzzy－rough－based approaches[J]. IEEE transactions of knowledge and data engineering,2004,16:1457－1471.

[4] Bhatt R B, Gopal M. On fuzzy－rough sets approach to feature selction[J]. Pattern Recognition Letters,2005,26:965－975.

[5] Kwak N, Choi C －H. Input feature selection for classification problems[J]. IEEE Trans. on Neural Networks,2002,13:143－159.

[6] Dash M, Liu H. Consistency－based search in feature selection[J]. Artificial Intelligence,2003,151:155－176.

(上接第63页)

5　结束语

文中将内分泌思想引入对粒子群算法的改进,从改变粒子的更新方法入手改善粒子群的性能,设计了一种内分泌激素更新方案,将内分泌系统对粒子的行为调节转变成对粒子更新方法的调整。通过机器人全局路径规划实验,表明本方法优于传统的粒子群方法。

参考文献:

[1] Kennedy J, Eberhart R C. Particle Swarm optimization. proc [C]//IEEE international Conference on Neural Network. USA: IEEE press,1995:1942－1948.

[2] Shi Y, Eberhart R C. A modified Swarm Optimizer[C]// IEEE International Conference of Evolutionary Computation. Anchorage, Alaska: IEEE press,1998.

[3] 李　磊,叶　涛,谭　民,等.移动机器人技术研究现状与未来[J].机器人,2002,24(5):475－480.

[4] Lovbjerg M. Rasmussen T K, Krink T. Hybrid particle swarm optimization with breeding and subpopulations[C]//In: proc of the third Genetic and Evolutionary computation conference. San Francisco: Morgan Kaufmann Publishers,2001.

[5] Higasshi N, Iba H. Particle swarm optimization with Gaussian mutation[C]//In: proc of the congress on Evolutionary Computation. Indianapolis, Indiana: IEEE,2003:72－79.

[6] Habib M K, Asama H. Efficient method to generate collision free paths for antonomous mobile robot based on new free space structuring approach [C]//IEEE/RSJ International workshop on Intelligent Robots and Systems. Osaka, Japan Habib M K,1991:563－567.