基于属性桶的约简算法

王 浩[1,2],胡学钢[1],黄晓梅[2]
(1.合肥工业大学,安徽 合肥 230009; 2.安徽建筑工业学院,安徽 合肥 230022)

摘 要:基于粗糙集理论的属性约简算法是机器学习和数据挖掘领域的研究热点之一。粗糙集理论是一种新型的处理模糊和不确定信息的数学工具,在保证分类能力不变的前提下,通过知识的约简导出概念的分类规则。文中提出了一种基于属性桶的约简算法,其约简过程类似基于属性频度函数的约简算法。该算法首先构造一组与决策表决策属性个数相同的属性桶,不同的属性桶划分了不同长度的区分矩阵项,避免了约简前的排序过程。通过构造属性桶时对核属性进行特殊处理,在一定程度上简化了属性约简过程。

关键词:粗糙集;属性桶;区分矩阵;属性约简
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2008)07-0018-03

Reduction Algorithm Based on Attribute Buckets

WANG Hao[1,2], HU Xue-gang[1], HUANG Xiao-mei[2]
(1. Hefei University of Technology, Hefei 230009, China;
2. Anhui Institute of Architecture & Industry, Hefei 230022, China)

Abstract: Reduction algorithm based on rough set theory is one of the main subjects in the field of machine learning and data mining. The rough set theory is a new mathematics tool which is used to process fuzzy and indetermination problem. This theory which's advantages lie in not requiring prior information when carries out the classicification is to derive classification rules of conception by knowledge reduction without changing the classification capacity of the information system. Presents a reduction algorithm based on attribute buckets which reduction progress is more similar than attribute reduction algorithm based on attributes frequency. At first, this algorithm construct a set of attribute buckets which have the same number of attribute in the decision table. Because the items of discernable matrixes with different length will put in the different attribute buckets, the sort process of attribute reduction can be avoided. By particularly manipulating the core of attributes when constructing the attribute buckets, the algorithm can simplify the process of attribute reduction to a certain extent.

Key words: rough set; attribute buckets; discernibility matrix; attributes reduction

0 引 言

粗糙集理论是一种研究不精确性和不确定性知识的数学工具,由波兰学者 Pawlak.Z 首先提出[1,2]。在粗糙集理论中,不同的知识对应着相关论域的不同划分,多个知识构成一个知识库。知识约简的实质是找出知识库中一些冗余的知识(划分)[3]。决策表就是一个知识库,其属性是知识库中的每个知识。如何高效地对决策表的属性进行约简是当前粗糙集理论的主要研究方向之一[4]。

基于可辨识矩阵的约简算法是基本的约简算法[5,6],除此之外有基于属性频度函数的约简算法、基于信息熵的属性约简算法、基于散列的属性约简算法等等[7~10]。Hu根据属性的重要性提出了以属性重要性为启发规则的属性约简算法[9]。该算法将核作为计算约简的开始,并按照属性的重要程度从大到小依次加入属性,直到该集合是一个约简为止。然后检查该集合中的每个属性,判断移走该属性是否会改变该集合的对决策属性依赖度。如果不影响,则将其删除。基于信息熵的属性约简算法也均为类似的启发式属性约简算法[11,12]。

文中讨论的基于属性桶的约简算法在约简时和基于属性频度函数的约简算法类似,该算法采用属性桶的方式存储区分矩阵项,使得求取区分矩阵每项时预先按照项长进行分类,桶号越小的桶先参与约简。由

收稿日期:2007-10-04
基金项目:安徽省自然科学基金资助项目(050420207)
作者简介:王 浩(1980-),男,硕士,实验师,研究方向为数据挖掘、计算机网络;胡学钢,博士,教授,硕士生导师,研究方向为数据挖掘、算法设计。

于在构造属性桶时一旦发现核属性的存在则立即在后续项的比较中停止对核属性项的比较,从一定程度上简化了区分矩阵项的构造过程,使得算法效率有所提高。

1　粗糙集基本定义

以下为粗糙集基本定义[13]:

定义1:设 $U \neq \varnothing$ 是研究对象组成的集合,称为论域。任何子集 $X \subseteq U$,称为 U 中的一个概念或范畴,U 的任何概念簇称为关于 U 的知识。U 上的一族划分称为关于 U 的一个知识库。

定义2:若 $P \subseteq R$,且 $P \neq \varnothing$,则 $\bigcap P$ 是一个等价关系,称为 P 上的不可区分关系,记为 ind(P),且有 $[x]_{\mathrm{ind}(P)} = \bigcap_{R \in P} [x]_R$

定义3:当 $K = (U, R)$ 为一个知识库,ind(K) 定义为 K 中所有等价关系的族,记作 $\mathrm{ind}(K) = \{\mathrm{ind}(P) \mid \varnothing \neq P \subseteq R\}$。

定义4:给定知识库 $K = (U, R)$,对于每个子集 $X \subseteq U$ 和一个等价关系 $R \in \mathrm{ind}(K)$,定义两个子集:

$$\underline{R}X = \bigcup \{Y \in U/P \mid Y \subseteq X\}$$

$$\overline{R}X = \bigcup \{Y \in U/P \mid X \cap Y \neq \varnothing\}$$

分别称为 X 的 R 下近似集和 R 上近似集。集合 $bn_R(X) = \overline{R}X - \underline{R}X$ 称为 X 的 R 边界域,$\mathrm{pos}_R(X) = \underline{R}X$ 称为 X 的 R 正域,$\mathrm{neg}_R(X) = U - \overline{R}X$ 称为 X 的 R 负域。

定义5:令 R 为一族等价关系,$P \in R$,如果 $\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$,则称 P 为 R 中不必要的;否则称 P 为 R 中必要的。如果每一个 $P \in R$ 都为 R 中必要的,则称 R 为独立的;否则称 R 为依赖的。

定义6:设 $Q \subseteq R$,如果 Q 是独立的,且 $\mathrm{ind}(Q) = \mathrm{ind}(R)$,则称 Q 为 R 的一个约简;R 中所有必要关系组成的集合称为 R 的核,记作 core(R)。

定义7:对于四元组 $S = (U, A, V, f)$,其中

U:对象的非空有限集合,称为论域;

A:属性的非空有限集合;

$V = \bigcup_{a \in A} V_a$,$V_a$ 是属性 a 的值域;

$f: U \times A \to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即 $\forall a \in A, x \in U, f(x, a) \in V_a$。

则 $S = (U, A, V, f)$ 是一个知识表达系统。

2　基于属性桶的约简算法

2.1　属性桶的定义

定义8:知识表达系统 $S = (U, A, V, f)$,其对应区分矩阵任一元素为 $a(x, y) = \{a \in A \mid f(x, a) \neq f(y, a)\}$,属性桶 $B_c = \{a(x, y) \mid \mathrm{Len}(a(x, y)) = c\}$,其中 Len($a(x, y)$) 为区分矩阵某元素的长度,$c$ 为常数。

2.2　算法基本思想

基于属性频度函数的约简算法效率较高,但该算法需要建立完整的区分矩阵并且将区分矩阵的每项按照项的长度进行排序后方可进行约简。基于属性桶的约简算法并不直接构造完整的区分矩阵,先建立一组与决策表决策属性个数相同数量的属性桶,编号为 x 的属性桶存储区分矩阵中对应长度为 x 的项。显然,编号越小的桶重要性越高,应先参与约简。编号为1的属性桶内存储的均为核属性。当1号桶求出属性项时,原决策表中每项在对比时将不考虑该属性的对比,从而在一定程度上减少了生成后续属性桶的比较次数。

基于属性桶的约简算法约简时首先从编号为2的桶开始,以出现频率的大小为优先级条件将属性加入到约简集中(如果桶中某项的属性已在约简集中出现,则该项不参与约简),直至遍历完所有的桶。1号桶中的属性与约简集的并集即为最终的约简结果。

2.3　算　法

算法名称:基于属性桶的约简算法

输入:决策信息系统 $(U, C \cup \{d\})$,其中 C 为条件属性集合,$a \in C$。

输出:约简 Re d。

Step1:建立一组空的属性桶,B_1 至 $B_{|C|}$;

Step2:对比决策信息系统中的每两项,全部非核属性比较完毕后将该两项不同的属性按照其个数多少存储到对应的属性桶中,每个桶记录该属性在桶中出现的频率。

Step3:决策信息系统中所有项是否全部比较完毕,如果是则转 Step4,否则转 Step2;

Step4:令 Re $d = \varnothing$,Re d = Re $d \cup B_1$;

Step5:从 B_2 至 $B_{|C|}$,对于桶内每一项所含属性 a,如 $a \notin$ Re d 且 a 在该项所有属性中频率最高,则 Re d = Re $d \cup \{a\}$;

Step6:所有桶的项均遍历完毕则输出约简 Re d,否则转 Step5。

2.4　算法分析

基于属性桶的约简算法需要对原决策表中的每项进行比较并将结果存入对应的属性桶中,所以最多一共有 $|U|(|U|-1)/2$ 次比较,时间复杂度为 $O(|U|^2)$,求 Re d 消耗的最大代价为 $O(|A \| U^*|) = O(|A \| |U|^2)$。因此,算法总的时间复杂度为 $O((|A|+1) |U|^2)$。由于生成属性桶时一旦求出核属性,则立

即停止后续项对核属性的对比，即只比较非核属性，所以算法实际复杂度低于 $O((|A|+1)|U|^2)$。基于属性桶的约简算法在求约简时无需先对区分矩阵的每项按照长度进行排序，在一定程度上优于基于属性频度函数的约简算法。

3 算 例

表1是文献[13]24页中所给出的一个知识表达系统，文献[13]使用区分矩阵并求出区分函数的方法求出该知识表达系统的两个约简 $\{a,b\}$、$\{b,d\}$，核是 $\{b\}$。

表1 决策表

U	a	b	c	d
1	0	1	2	0
2	1	2	0	2
3	1	0	1	0
4	2	1	0	1
5	1	1	0	2

根据基于属性桶的约简算法，首先构造出4个空的属性桶(见图1)，接着对比决策表中的每项，由不同属性的个数决定比较结果填入哪个属性桶，当进行到决策表第2项和第5项进行比较时，仅有属性 b 不同，即该属性是核属性，则后续的所有比较均不在查看 b 属性的值是否相同，在一定程度上提高了效率(图中 * 表示未参与比较的属性)。算法执行约简时，先从1号桶中取出属性 b 做为初始约简集，然后从2号桶中选择出现频率最高的非核属性 d，此时的约简为 $\{b,d\}$，在后续桶中，由于每项均出现了 b 和 d 属性，故均不用作约简，最后的约简结果为 $\{b,d\}$。显然，与基于属性频度函数的约简算法一样，基于属性桶的约简算法也是不完备的，只能求出某个约简或是该约简的超集。

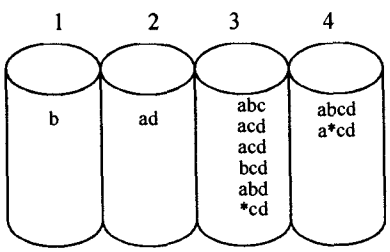

图1 属性桶

4 结束语

基于属性桶的约简算法在求约简时和基于属性频度的约简算法类似，该算法的核心是在构建属性桶的过程中对核属性进行了特别处理，一旦发现核属性的存在则立即在后续项的比较中停止对核属性项的比较。由于在构造属性桶的同时，自动区分了不同长度的区分矩阵项，省略了约简前的排序工作，从一定程度上简化了算法。

参考文献：

[1] Pawlak Z. Rough Sets, Theoretical Aspects of Reasoning about Data[M]. Dordrecht, Boston, London: Kluwer Academic Publishers, 1991.

[2] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341－356.

[3] 瞿彬彬，卢炎生. 基于粗糙集的属性约简算法研究[J]. 华中科技大学学报：自然科学版，2005，33(8)：30－33.

[4] 张文修，吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学，2000，14(4)：1－12.

[5] 叶东毅. 粗糙集属性约简的一个贪心算法[J]. 系统工程与电子技术，2000，22(9)：63－65.

[6] 王国胤. 决策表核属性的计算方法[J]. 计算机学报，2003，26(5)：611－615.

[7] 程玉胜，胡学钢. 不完备信息系统的属性约简方法研究[J]. 计算机工程与应用，2004(1)：68－70.

[8] 王加阳，陈松乔，罗 安. 粗集动态约简研究[J]. 小型微型计算机系统，2006，27(11)：2056－2060.

[9] Hu Xiao－hua, Cercone N. Learning in relational database: a rough set approach[J]. Computational Intelligence, 1995, 11(2)：323－337.

[10] 杨 静. 基于粗糙集合和信息熵的分类模型研究[D]. 合肥：合肥工业大学，2004.05.

[11] 苗夺谦. 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.

[12] 张建军，张静波. 一种新的基于粗糙集理论的决策表离散化算法[J]. 西安电子科技大学学报：自然科学版，2004，31(3)：469－472.

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

(上接第17页)

l.]: Research Studies Press LTD, John Wiley & Sons Ins, 1989.

[3] Berndt J D, Clifford J. Finding patterns in time series: a dynamic programming approach[M]//Fayyad U, Gpiatetsky－Shspiro, Smythp. Advances in Knowledge Discovery and Data Mining. [s.l.]: MIT Press, 1996.

[4] Anders T B. Mining Time Series Using Rough Set－A Case Study[C]//In: Proceeding of the First European Symposium. PKDD'97. Trondheim, Norway: [s.n.], 1997: 256－263.

[5] 尹旭日，商 琳，何佳洲，等. Rough集挖掘时间序列的研究[J]. 南京大学学报，2001，37(2)：87－91.