文章编号:1009－2269(2006)04－0001－04

基于粗糙集的数据集预处理研究*

肖爱斌[1],李向伟[2]

(1. 西北师范大学 图书馆,甘肃 兰州 730070;
2. 西北师范大学 数学与信息科学学院,甘肃 兰州 730070)

摘要:针对分类模型在构造过程中存在冗余属性的特点,提出了一种基于粗糙集理论的分类预处理技术,利用其理论的属性约简与分辨矩阵得到属性的约简集.经实例证明,此方法对提高分类模型构造效率有较好的效果.

关 键 词:粗糙集;约简与核;分辨矩阵
中图分类号:TP 18 **文献标识码**:A

数据预处理是数据挖掘过程中的第一步,也是非常重要的一步.预处理完的结果作为数据挖掘的直接数据,故预处理的效果直接决定了挖掘任务的成功与否,一个好的数据源,不仅体现准确的挖掘结果,而且可以极大地提高算法的效率.

通常意义上的数据预处理指的是在实施挖掘算法前的数据清理、数据集成和变换、数据归约、离散化的概念分层生成等处理.然而数据挖掘技术涉及领域广、分支多、解决的问题多,故针对具体的应用领域而实施的挖掘算法中所涉及的具体预处理技术也不尽相同,这样才使得数据预处理功能显示其作用.文献[1]提出了一种基于经验贝叶斯统计方法的预处理方法,可以对一类名词性属性进行变换,使之能预测建模.文献[2]提出了一种面向领域知识的存储结构和将领域知识应用于数据随机处理过程的算法,使之有效地减少了数据源的数量,最终提高了挖掘算法的质量与速度.文献[3]给出了一种基于冲突分析的特征提取算法,并有效地消除了对分类不起关键作用的属性.文献[4]针对数据集的特点,利用粗糙集理论对其进行离散化,达到为下一步数据挖掘做好基础.文献[5]以粗糙集、概念格及包含度为工具,论述了在不确定环境下的决策规则的提取方法与规则融合方法.总之,以上文献均在分析数据特点及不同应用目的的前提下对数据进行特定的预处理,达到提高算法效果的目的.

分类作为数据挖掘中最基本功能之一,在信息检索、医疗诊断、决策分析、机器学习、近似推理等领域已成功使用[2].然而,现有的一系列分类算法在实现上并不理想.一方面,分类的效率相对较低.如对于给定的信息系统,提炼分类模型常常需要对训练集进行很长时间的训练.另一方面,分类算法的伸缩性差.如对于属性较少、数据量不大的信息系统分类性能较好,但随着属性与数据量的增加,性能会大幅度下降,甚至是无法实现的.其主要原因是传统分类算法均未能在消除冗余属性或没有考虑冗余属性存在的前提下进行分类模型的构造.而粗糙集理论中可以利用不可分辨关系构造分辨矩阵,进而在得到分辨函数的基础上求得属性的核.而核在构造分类模型的属性中可理解为不可约简的最小属性集,在此基础上进行分类模型的提取可得到与原属性集等价的效果,从而大大降低了分类模型构造的复杂性.

* 收稿日期:2006－09－27
基金项目:甘肃省自然科学基金 (3ZS051－A25－047)
作者简介:肖爱斌(1962－),女,甘肃临洮人,馆员.

1 相关粗糙集概念与理论

1.1 分辨关系及等价类

定义1 如果任意两个对象 x_i, x_j 对所有条件属性其值相等,则称其为不可分辨对象.

定义2 令 R 为等价关系族,设 $P \subseteq R$,且这 $P \neq \emptyset$,则 P 中所有等价关系的交集称为P 上的不可分辨关系,记为IND(P) 即有:

$$[X]_{IND(P)} = I_{R \in P}[X]_R \tag{1}$$

显然IND(P) 也是等价关系.

不可分辨关系将所有对象分成不同等价类[4],实例见下.

不可分辨关系是粗糙集理论中最基本的概念,若 $< x, y > \in IND(P)$,则称对象 x 与 y 是 P 不可分辨的,即等价关系族 P 形成的分类知识不可区分 x 与 y.

表1 信息系统模型

$P1$	$P2$	$P3$	$P4$	
$X1$	0	y	1	h
$X2$	0	y	1	h
$X3$	1	y	1	n
$X4$	0	n	0	1
$X5$	0	n	0	m

设一信息系统模型如表1所示:

由以上定义可知对于以上信息模型的等价类为:$E1 = \{1,2\}$; $E2 = \{3\}$; $E3 = \{4,5\}$.

粗糙集理论的不确定性是建立在上、下近似的概念之上.令 $X \subseteq U$ 是一个集合,R是定义在U 上的等价关系,则

$$R_-(X) = Y\{\gamma_i \in U/R : \gamma_i \in X\} \tag{2}$$

$$R^-(X) = Y\{\gamma_i \in U/R : \gamma_i \in X\} \tag{3}$$

分别称为 X 的 R 下近似与 R 上近似.

1.2 分辨矩阵

设 $S = (U, A)$ 为一信息系统,S 的分辨矩阵 $\boldsymbol{M}$ 定义为一个 n 阶对称矩阵,其 i 行,j 列的元素定义为:

$M_{ij} = \{a \in A \mid f(x_i, a) \neq f(x_j, a)\}$ $i, j = 1, \cdots, n$.

即 M_{ij} 是能够区别对象 x_i 和 x_j 的所有属性的集合.

1.3 约简与核

定义3 设 $Q \subseteq P$,若 Q 是独立的,且IND(Q) = IND(P),则称 Q 是等价关系族 P 的一个约简. P 中所有不可省关系的集合称为等价关系族 P 的核,记为CORE(P).

这两个概念是后面属性约简的基础,从定义可见:核是知识库中最重要的部分,是约简时不能消去的知识.

2 过程模型描述

2.1 分辨矩阵的构造

将粗糙集理论中的分辨矩阵理论引入到关系数据库或信息系统,可方便地建立属性集的分辨矩阵.分辨矩阵基于粗糙集理论中的不可分辨关系,也反应了知识或分类的粒度,而信息系统中的属性及子集可认为是知识粒度的体现,粒度的精度直接反应了元组的区分程度.以信息系统中的属性集为基础,构建基于属性的分辨矩阵,可以借助于数学理论分析各属性之间的关系,即各属性之间的独立性与依赖性.

以文献[6]中的信息表为例来说明,见表2.为表示方便,将表2的具体信息系统符号化可得到简化的信息系统,见表3.

表 2　一个具体的信息系统

	属性 1	属性 2	属性	属性 4	属性 5
客户 1	2	1	1	1	1
客户 2	1	1	2	1	2
客户 3	1	2	2	1	1
客户 4	1	2	2	1	2
客户 5	1	2	2	2	2
客户 6	1	1	2	1	1

表 3　一个简化的信息系统

	$P1$	$P2$	$P3$	$P4$	$P5$
$U1$	2	1	1	1	1
$U2$	1	1	2	1	2
$U3$	1	2	2	1	1
$U4$	1	2	2	1	2
$U5$	1	2	2	2	2
$U6$	1	1	2	1	1

根据分辨矩阵及等价类的定义可得到如下分辨矩阵，如图 1 所示.

$$\begin{bmatrix} 0 & p1p3p5 & p1p2p3 & p1p2p3p5 & p1p2p3p4p5 & p1p3 \\ & 0 & p2p5 & p2 & p2p4 & p5 \\ & & 0 & p5 & p4p5 & p2 \\ & & & 0 & p4 & p2p5 \\ & & & & 0 & p2p4p5 \\ & & & & & 0 \end{bmatrix}$$

图 1　分辨矩阵

由图 1 可看出分辨矩阵是一个以信息系统的属性集元素的个数为阶的一个对称矩阵，为研究方便，只写出其上三角，主对角线元素为 0.

2.2　求划分约简集及核

定义 4[5]　设(U, A, F)是一个信息系统，对于 $B \subseteq$ A，若 RB = RA，称 B 是划分协调集. 若 B 是划分协调集，而 B 的任何子集均不是划分协调集，则称 B 为划分约简集.

定义 5[5]　设 $\boldsymbol{D}$ 是信息系统(U, A, F)的分辨矩阵，则 B 是划分协调集当且仅当对于任意$[x_i]A \cap [x_j]A = \varnothing$，有 $B \cap \boldsymbol{D}([x_i]A, [x_j]A) \neq \varnothing$，依据定义 3，可知：

1) 若取掉属性集中的某个属性 P_i 后，元组仍是可分辨的，则称属性 P_i 是属性集中可约简的.

2) 若属性集中的每个属性 P_i 性都是不可约简的，则称属性集是独立的. 即意味着属性集中的任意属性都是必不可少的. 它构成了信息系统的最小精简属性集，即所谓的“核”.

由定义 3、4、5 可知，求属性集的核可转化为求信息系统的划分约简集，而定义 4 给出了一个集合为划分约简集的充分必要条件，为我们求解划分约简集提供了一种捷径.

根据定义 4、5，我们得到图 1 分辨矩阵的划分约简集为 $B = \{P2, P4, P5\}$.

通过以上方法求得信息系统的划分约简集(属性集的核) 后，分类模型的构造将更为简化. 为便于说明，以上只是引用了一个非常简单的信息系统，可以认为是信息系统的一个原型，但这种方法可推广到任意复杂的信息系统，而且系统越复杂，越能体现这种方法的优越性.

2.3　算法描述

算法：generate - Lcore

输入：训练样本数据表.

输出：精减的属性集(核).

方法：

(1) for I = 1 to n //n 为数据表中元组的个数

for j = 1 to m//m 为数据表中属性的个数

begin

if $p_{ik} = p_{jk}$ then//$k = 1, 2, \cdots, n$

$m_{ij} = \varnothing$

else

$m_{ij} = \{a \in A \mid f(x_i, a) \neq f(x_j, a)\}$ $i, j = 1, \cdots, n$ // m_{ij} 是能够区别对象 x_i 和 x_j 的所有属//性的集合

end

(2) for $i = 1$ to n // n 为分辨矩阵的行数

for $j = 1$ to m // m 分辨矩阵的列数

begin

if $B \cap \boldsymbol{D}([x_i]A, [x_j]A) \neq \varnothing$ and $[x_i]A \cap [x_j]A = \varnothing$

then output B

end

以上算法在构造分辨矩阵和求约简集过程中分别扫描一次数据表和分辨矩阵,它们的时间复杂度均最大为 $0(n^2)$.

3 结束语

分类算法是数据挖掘中最重要的功能之一,通常称之为有指导的学习,基本思想是依训练集提炼分类模型,然后根据分类模型对未知数据进行分类.但对于现实中广泛使用的二维信息系统,现有的传统分类方法在没有经任何预处理的前提下建模,这样不仅潜在地增加了分类的复杂性,而且影响了分类的准确性.本文在引入粗糙集理论中的核心理论不可分辨矩阵和约简与求核对属性集在不影响分类结果的情况下对分类数据的属性进行了化简,结合实例说明了此理论在分类数据预处理中有很好的效果,对传统分类算法的效率与速度有极大的提高.

参考文献:

[1] 陆 勤.分类和预测问题中名词性属性的一种预处理方法[J].计算机工程,2004,30(3):92~95.

[2] 王大玲,于 戈,鲍玉斌,等.一种面向数据挖掘预处理过程的领域知识的分类与表示[J].小型微型计算机系统,2003,24(5):863~868.

[3] 杨 阳,刘 锋.分类器的数据预处理[J].计算机工程,1998,24(4):33~34.

[4] 施 伟,战守义,盛思源.基于粗糙集理论的数据预处理[J].计算机工程与应用,2003,(22):193~194.

[5] 张文修,仇国芳.基于粗糙集的不确定决策[M].北京:清华大学出版社,2005.

[6] 于 洪,杨大春.基于粗糙集理论的数据挖掘的应用[J].计算机与现代化,2001,(4):45~49.

Research on Data Set Preprocessing Based on Rough Set Theory

XIAO Ai-bin[1], LI Xiang-wei[2]

(1. Library of Northwest Normal University, Lanzhou 730070, China;

2. College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

Abstract: Aimed at the characteristic of redundancy in the process of constructing classification model, this paper offers a classification reprocessing technology based on the Rough Set Theory to make use of the attribute reduction and distinguish matrix of the theory to produce the core. It is tested by the concrete example that the technology has good effect on constructing classification model.

Key words: rough set; reduction and core; distinguish matrix