基于粒度的知识粗糙性研究

王 刚[1,2],王 浩[1]
(1.合肥工业大学 计算机与信息学院,安徽 合肥 230009;
2.铜陵学院 计算机系,安徽 铜陵 244000)

摘 要:粒度计算是粗糙集理论研究的一种强有力的工具。在粗糙集理论中,提出知识是有粒度的并定义了知识粗糙度的概念。众所周知,知识获取是专家系统开发的"瓶颈"问题。文中主要从知识粒度、分辨度以及重要度等方面入手,着重研究了知识粗糙性的粒度原理与知识粗糙性的关系,提出了一种基于属性重要度的粗糙性知识获取算法,并通过理论与实验,证明该算法是正确的,行之有效的。

关键词:粗糙集;粒度;分辨度;约简;重要度;规则
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2008)01-0067-03

On the Roughness of Knowledge Based on Granularity

WANG Gang[1,2], WANG Hao[1]
(1. Computer and Information College, Hefei Technology University, Hefei 230009, China;
2. Department of Computer Science, Tongling College, Tongling 244000, China)

Abstract: Granular computing is emerging as a powerful tool for rough set theory. The theory of rough-set points out that knowledge has granularity and defines the concept of roughness of knowledge. It's well known that the difficulties of system-developing lie in the acquisition of knowledge. From the granularity, recognizability and importance of knowledge, studies the relationship between the roughness of knowledge and its priciple of granularity, puts forward a method of acquiring knowledge based on the importance of attributes, and proves the feasibility of this merthod through theory and experiments.

Key words: rough-set; granularity; recognizability; simplification; importance; rules

0 引 言

粗糙集理论(rough set theory)[1,2]是波兰数学家Z.Pawlak于1982年提出来的。这一理论为处理具有不精确和不完全信息的分类问题提供了一种新的框架,是一种新的软计算方法,已成为知识发现和诊断决策领域的一个研究热点。人们在解决、处理大量复杂信息问题时,由于能力限制,通常把大量复杂信息按其各自特征和性能对其划分成若干较简单的块,对于如此划分出来的块被看成一个粒,所以可以认为知识是具有粒度的。

在数学上,粗糙集理论把知识看作是关于论域的划分,论域上的一个划分与其上的一个等价关系是等价的,知识的这种颗粒状结构是通过等价关系的等价类来予以体现。正是由于知识的这种"颗粒"状,导致了知识的粗糙性。文中主要从知识粒度、分辨度以及重要度等方面入手,着重研究了知识粗糙性的粒度原理与知识粗糙性的关系,提出了一种基于属性重要度的粗糙性知识获取算法。

收稿日期:2007-03-16
基金项目:安徽省高等学校省级自然科学研究基金(KJ2007B382ZC)
作者简介:王 刚(1975-),男,安徽桐城人,硕士研究生,讲师,研究方向为数据挖掘、粗糙集理论、数据结构;王 浩,博士,教授,硕士生导师,研究方向为人工智能、数据挖掘、面向对象技术等。

1 粒度原理

对于粒度原理[3~5]的研究,更多学者从信息熵与条件信息熵作为切入点来进行研究,也有学者从信息量与条件信息量方面出发进行研究。为更好地研究有关粒度原理,对知识进行严密的分析和有效的操作,准确地完成数据挖掘任务。笔者先介绍与知识形式化定义相关的粗糙集理论。

定义1(信息系统):一个信息系统是一个对 $S=(U,A)$,其中 U 是一个非空、有穷、被称为全域的个体的集合,A 是非空、有穷的属性集合,即对于属性 $a\in A$,有 $a:U\rightarrow Va$,其中 Va 被称为属性 a 的值集;集合 $V=\bigcup_{a\in A}Va$ 被说成是属性集 A 的值区域。

定义 2(粒度):设 U 为一个论域,R 为集合 U 上的一个等价关系,以 R 的所有等价类作为元素的集合来划分 X 可得到 U/R(基本知识):$U/R = \{[x] \mid x \in U\} = \{X_1, \cdots, X_n\}$,$X$ 的"粒度"定义为:$p(X_i) = \text{card}(X_i)/\text{card}(U)$,它表示等价类 X_i 在 U 中的概率,其中 card(S) 表示 S 的基数。

定义 3(知识与知识粒度):知识库可定义为序对形式:$K = (U, R)$,R 为一等价关系,即知识。知识 R 的"粒度"定义为:$\text{GD}(R) = \text{card}(R)/\text{card}(U^2)$,其中,card($R$) 表示 $R \subseteq U \times U$ 的基数。

由上述定义可知,当知识 R 为相等关系时,知识 R 的"粒度"达到最小值 1/card(U);当知识 R 为全域关系时,知识 R 的"粒度"达到最大值 1。一般情况下,知识 R 的粒度取值在[1/card(U),1] 范围内,如果 GD(R) 取值越大,得到的知识的信息量就相对越小;反之,信息量就越大。正是由于知识的粒度可大可小,导致了知识的粗糙性。

定义 4(知识的分辨度):设知识 R 的分辨度为 Dis(R),$\text{Dis}(R) = 1 - \text{GD}(R)$。知识的粒度可以用于表示知识的分辨能力,GD($R$) 越大,得到的信息量就越小,对知识的分辨能力就越弱,分辨度的大小就直接反映了知识的分辨能力,这也为研究知识的粗糙性提供了一个方向。

关于熵值,它也是知识颗粒状的一种度量。对于知识粒度、分辨度与熵值的关系,给出表 1,对其深入理解、研究信息系统下知识粒度、信息熵和粗糙熵之间的关系,可以参考文献[3,4,6]。

表 1 粒度、分辨度与熵值的关系比较

R	全域关系	相等关系	一般的 R
$H(R)$	0	$\log_2(\text{card}(U))$	$0 \leqslant H(R) \leqslant \log_2(\text{card}(U))$
GD(R)	0	1/card(U)	$1/\text{card}(U) \leqslant \text{GD}(R) \leqslant 1$
$I(R)$	0	$1 - 1/\text{card}(U)$	$0 \leqslant \text{Dis}(R) \leqslant 1 - 1/\text{card}(U)$

2 属性重要度与数据约简

关于属性重要度的概念,用不同的粒度函数就有不同的定义方法,从而有不同的属性约简算法。

2.1 属性重要度

定义(属性重要度)[3]:设 $X \subseteq A$ 是一属性子集,$x \in A$ 是一属性,x 对于 X 的重要度为 $\text{Sig}(x) = 1 - |X \cup \{x\}| / |X|$,其中 $|X|$ 表示 $|\text{IND}(X)|$。

从上述定义中可知 x 对于 X 的重要度,即 X 中增加属性 x 之后分辨度的提高程度,提高程度越大,认为 x 对于 X 越重要。

2.2 属性约简与规则获取

对信息系统进行约简,可以从纵横两个方向来深入研究[2,5~8]。从横向上看,属性约简能从属性集合中去掉一些冗余属性,相当于对决策表进行横向压缩;从纵向上看,信息系统中存在相似或相同的对象,这些对象具有相似的条件属性和相同的决策值,在一定的抽象层次上将是难以区分的,因此可以去掉这些对象以达到约简目的,相当于对信息系统进行纵向压缩。在信息系统中,从个体的全域到粒的分类计算是产生较少的粒名数据表,它是第一层个体的属性约简;再从粒名的数据表对粒进行组合,产生粒组合数据表。在这种数据表中,可以进一步进行挖掘,寻找给定关系中具有一定形式的关联规则,其满足条件的规则数目应小于或等于商级中粒的数目,所以从商级到关联规则集也是数据的一种约简。下面完整给出基于这种思想获取关联规则集的算法:

S0:应用上述描述的粒度基本原理,刻划出对应域中合适的粒度,并对粒度中的粒进行恰当的描述,完成这一层相应的数据约简。

S1:基于属性的重要度,根据步骤 S0,计算出最重要的属性,如果所有属性的重要度相同,可以选择信息量最大的属性,完成相关属性的确定。

S2:根据步骤 S1,得到相应的属性取值,构成样本空间的划分粒度,构造出相应的多变量的树形结构图 T。

S3:对树形结构图 T 进行遍历,在访问过程中,如果每一个粒,它的所有个体元素属于同一个类,则转步骤 S4;否则,删除步骤 S1 中找出的属性,并重复 S1 和 S2,直到它们属于同一类或者无条件属性为止。

S4:对每一粒进行描述,输出相应规则,获取知识。

S5:基于提取的规则得到规则集,完成数据的约简。

根据此算法的思想,实质是寻求对样本空间进行划分,其划分过程是一个由粗到细的过程,在结构上形成了一棵多变量的树 T。在对树进行访问过程中,可以从树根到叶子结点提取所需要的规则,继而完成数据的约简。通过应用上述方法,提取粗糙性知识。实际上,在对多变量的树 T 进行遍历的过程中,可以应用现已比较成熟的技术,完成对树 T 的剪枝操作,以提高算法的效率,同时获取更加准确的知识。

3 实验及实验结果分析

为了验证算法的可行性及有效性,进行了下面实验,实验数据从文献[1]中获得,数据如表 2 所示。

从上述表格中,可以看出,U 中共有 3 种属性,分别为产品颜色、型号及价格,按属性进行分类,可以得到不同的商集。假设现在,得到如下关系:$U/\text{IND}(C)$

$=\{[红],[黄],[蓝],[白],[黑]\}$,$[红]=\{u_1,u_3,u_8,u_9,u_{12}\}$,$[黄]=\{u_2,u_7,u_{10}\}$,$[蓝]=\{u_4,u_6\}$,$[白]=\{u_5\}$,$[黑]=\{u_{11}\}$,其中[红]为粒名,$\{u_1,u_3,u_8,u_9,u_{12}\}$为相应的粒,如果按其它属性分类,依据上述描述,以此类推。

表 2　多属性关系表

U	产品颜色 C	产品型号 M	产品价格 P
u_1	红	B_{100}	中
u_2	黄	B_{200}	贵
u_3	红	B_{300}	贵
u_4	蓝	B_{400}	中
u_5	白	B_{100}	便宜
u_6	蓝	B_{100}	贵
u_7	黄	B_{500}	中
u_8	红	B_{200}	贵
u_9	红	B_{300}	贵
u_{10}	黄	B_{500}	中
u_{11}	黑	B_{300}	中
u_{12}	红	B_{200}	贵

在实际应用中,可以根据实际情况,计算属性重要度。在这里,为方便,假设直接以产品价格为重要属性,构造出一棵多变量的树 T,如图 1 所示。

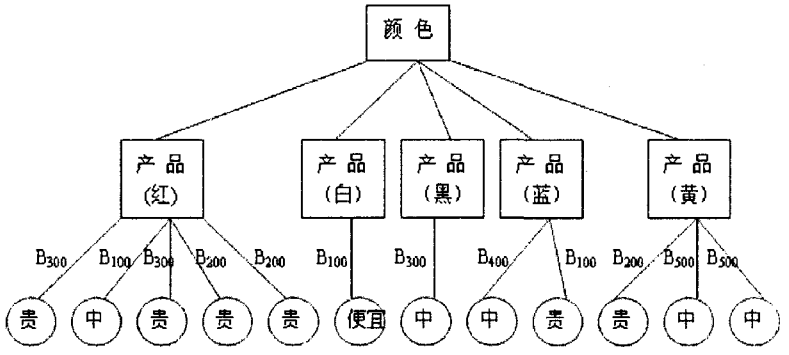

图 1　树 T

可以应用剪枝技术,对树 T 进行剪枝,在这里就不再进行赘述,可以参考相关文献。对粒的描述可以按照此树结构及粒度形成的先后过程,由连接树的根结点至叶子结点的路径分别给出,对树 T 按照某一确定原则,对其进行遍历,获取相应规则,如 if 颜色="红色"and 产品型号="B300" then 产品价格="贵"等。采用的硬件配置是 AMD Athlon64 处理器,2400MHz,512M 内存,软件是 Windows XP 环境,在软硬件环境相同的情况下,应用 VC++6.0 进行编程,进行有关对比实验,与应用 ID3 算法进行相应实验,实验结果一致,另外,应用多次属性约简,数值计算相对变小,说明该算法是行之有效的;运算过程也相对简单。

4　结束语

基于粒度的知识获取方式是根据人的思维特性而提出的,它与人认识客观世界的过程尽可能地保持一致,是思维过程的自然反映。从粗糙集理论出发,分析了知识粗糙性和知识粒度的关系,即知识的粗糙度愈大对应的知识的粒度也愈大,而得到的知识的信息量就愈小,并提出了基于粒度原理获取粗糙性知识的一种算法。如何进一步提高粗糙性知识获取的效率,还有待进一步深入研究。

参考文献:

[1] 刘　清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.

[2] 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.

[3] 苗夺谦,范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践,2002(1):48-56.

[4] 苗夺谦,王　珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论[J]. 模式识别与人工智能,1998,11(1):23-25.

[5] 林镇飚. 知识粗糙性的粒度原理及其属性约简[J]. 武汉科技学院学报,2006,19(9):31-34.

[6] Liang J Y, Shi Z Z. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004, 12(1): 37-46.

[7] 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,2001,21(12):76-80.

[8] 顿毅杰,张小峰,孙　昊,等. 一种基于粒度的规则挖掘方法[J]. 兰州理工大学学报,2006,32(1):105-108.

(上接第 66 页)

Services Allocation[C]//The Proceedings of the International Conference on Systems, Man and Cybernetics. Hague, Netherlands: [s.n.], 2004.

[3] Byoung-Dar L, Jon B W. Dynamic Replica Management in the Service Grid on Hi Performance Distributed Computing [C]//10th IEEE International Symposium. San Francisco, CA, USA: [s.n.], 2001.

[4] Meng L, Krithivasan R, Sue S Y W, et al. Flexible Inter-enterprise Workflow Management Using E-services[C]//on Advanced Issues of E-Commerce and Web-Based Information System, Proceedings of the 4th IEEE International Workshop, WECWIS. [s.l.]: [s.n.], 2002.

[5] Azzedin F, Maheswatan M. Towards Trust-Aware Resource Management in Grid Computing Systems, on Security and Grid Computing[C]//Proceedings of First IEEE International Workshop. [s.l.]: Springer-Verlag, 2002.