数据挖掘中基于 Rough Set 方法研究

纪 滨
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:随着数据挖掘的兴起,有许多分类和预测的方法。数据挖掘研究的实施对象多为关系型数据库,这给粗糙集方法的应用带来了极大的方便。关系表可被看作为粗糙集理论中的决策表,而利用粗糙集理论来处理数据挖掘有着传统挖掘工具所不具有的优点。粗糙集理论是一种处理不确定和不精确问题的数学工具,文中通过实例介绍了粗糙集的基本理论,并通过实例详细介绍了在基于对决策表属性约简的基础上采用了可变精度粗糙模型实现规则的获取。该实例说明了对于不完备的信息系统,应用粗糙集理论进行数据挖掘是非常有效的。

关键词:粗糙集;数据挖掘;信息系统;分类规则;数据归约
中图分类号:TP301.6;TP392 **文献标识码**:A **文章编号**:1673-629X(2008)02-0126-03

Research of Data Mining Based on Rough Set

JI Bin
(School of Computer Science, Anhui University of Technology, Maanshan 243002, China)

Abstract: With the rise of data mining, there are many classification and prediction methods. What data mining researchs largely are relational databases. This has brought great convenience for rough set's application. The relational table may regard as the decision table in rough set theory, and using the rough set to deal with data mining is more than the traditional mining tools. The rough set theory is a new mathematical approach to data analysis which are indiscernible with respect to some features. In this paper, basic theory of rough set is introduced using an example, and the implementation of rule mining by variable precision rough set model based on reduction of decision form feature is illustrated using an example. It is effective for rule mining based on rough set by the example.

Key words: rough set; data mining; information system; classification rule; data reduction

0 引 言

数据挖掘[1]是从大量数据中提取或"挖掘"知识。从技术角度看,数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先并不知道的、但又是潜在有用的信息和知识的过程。随着数据挖掘的兴起,有许多分类和预测的方法,而利用粗糙集[2,3]理论来处理数据挖掘有着传统挖掘工具所不具有的优点。粗糙集(Rough Set)理论[4]是一种研究不精确、不确定性知识的数学工具,由波兰科学家 Z. Pawlak 在 1982 年首先提出。粗糙集理论受到数据挖掘研究者的重视进而受到研究界的广为注意。粗糙集和数据挖掘关系密切,它为数据挖掘提供了一种新的方法和工具。

第一,数据挖掘研究的实施对象[5,6]多为关系型数据库。关系表可被看作为粗糙集理论中的决策表,这给粗糙集方法的应用带来了极大的方便。

第二,现实世界中的规则有确定性的,也有不确定性的。从数据库中发现不确定性的知识,为粗糙集方法提供了用武之地。

第三,从数据库中发现异常,排除知识发现过程中的噪声干扰也是粗糙集方法的特长。

第四,运用粗糙集方法得到的知识发现算法有利于并行执行,这可极大地提高发现效率。

第五,粗糙集方法比模糊集方法或神经网络方法在得到的决策规则和推理过程方面更易于被证实和检测。

许多实验表明,对于同一个数据集,在粗糙集理论工具下进行处理,最终得到的所需的信息更简单、更准确、更易于被决策者接受和理解[1,5]。

1 信息系统

粗糙集把客观世界或对象世界抽象为一个信息系统,一个信息系统 K 被定义为:$K=(U,A,V,f)$,其

收稿日期:2007-05-03
基金项目:安徽省自然科学研究项目(2006KJ063B)
作者简介:纪 滨(1970-),男,江苏人,讲师,硕士,研究方向人工智能、信息安全。

中:U 为论域,是对象(或事例)的有限集合,$U = \{x_1, x_2, \cdots, x_n\}$;$A$ 为属性的全体,$A = \{a_1, a_2, \cdots, a_x\}$;$V$ 为属性的值域,$V = \{v_1, v_2, \cdots, v_n\}$,$f$ 为信息函数[3](information function)。

属性集 A 常常又划分为两个集合 C 和 D,$A = C \cup D$,C 表示条件属性集,D 表示决策属性集。D 一般只有一个属性。设 a 是任一属性,x_i 是任一对象,则 $f(x_i, a)$ 表示 x_i 在 a 属性的取值。一般来说,条件属性是描述事物特性的特征,往往是人们容易得到的;决策属性是事物的类别,是人们希望得到的(分类目标)。粗糙集理论就是发现并定量描述事物特征与分类之间对应关系的理论。

2　有关定义

定义 1:令 $S = \{U, A, V\}$ 为一知识表达系统,$B \subseteq A$,定义 B 不可分辨二元关系 IND(B) 为:IND(B) $= \{(x_1, x_2) \mid f(x_1, b) = f(x_2, b), \forall b \in B\}$,显然,不可分辨关系是一个等价关系,如果 $(x_1, x_2) \in$ IND(B),说明根据已有的信息不能将 x_1 和 x_2 区分开,包含元素 x 的等价类用 $[x]_{IND(B)}$ 表示。

例 1:见表 1,利用属性"头痛"将 U 分成 $\{x_1, x_2, x_3\}$ 和 $\{x_4, x_5, x_6\}$ 两个集合,说明根据属性"头痛"现有的信息再不能将实体对象 x_1, x_2, x_3 和 x_4, x_5, x_6 分开,属性"头痛"的不可分辨关系为:IND(头痛$_1$) $= \{x_1, x_2, x_3\}$,IND(头痛$_2$) $= \{x_4, x_5, x_6\}$。

表 1　信息系统示例

对象 U	条件属性 C			决策属性 D
	头痛	乏力	发烧	流感
x_1	是	是	正常	否
x_2	是	是	偏高	是
x_3	是	是	很高	是
x_4	否	是	正常	否
x_5	否	否	偏高	否
x_6	否	是	很高	是

定义 2:设 $S = \{U, A, V\}$ 为一知识表达系统,$B \subseteq A$,$X \subseteq U$,定义集合:X 的 B 下近似:$B_-(X) = \{x \in U \wedge [x]_{IND(B)} \subseteq X\}$。$X$ 的 B 上近似:$B^-(X) = \{x \in U \wedge [x]_{IND(B)} \cap X \neq \emptyset\}$。

X 的 B 正域:X 的 B 正域就等价于 X 的下近似,$POSB(X) = B_-(X)$。X 的 B 边界区间:$BN^-(X) = B^-(X) - B_-(X)$。

例 2:见表 1,取"流感 = 否"的集合 $X = \{x_1, x_4, x_5\}$,取 $B =$ {头痛,乏力},B 的不可分辨关系为 $\{x_1, x_2, x_3\}, \{x_4, x_6\}, \{x_5\}$,则 $B^*(X) = \emptyset \cup \emptyset \cup \{x_5\} = \{x_5\}$。$B_*(X) = \{x_1, x_2, x_3\} \cup \{x_4, x_6\} \cup \{x_5\} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$。

上、下近似的含义[4]是:上近似代表"可能",它可以解释为由那些根据现有知识判断出可能属于 X 的对象所组成的最小集合,对上例来说,有与 $x_1, x_2, x_3, x_4, x_5, x_6$ 任意一个相同症状的可能不是流感;下近似代表"一定",解释为由那些根据现有知识可以判断出肯定属于 X 的对象所组成的最大集合,对上例来说,有 x_5 症状的一定不是流感。

定义 3:设 $S = \{U, C \cup D, V\}$ 为一决策系统,$C = \{a_i \mid i = 1, \cdots, m\}$ 是属性集,$U = \{x_1, x_2, \cdots, x_n\}$ 是论域,$\forall a_i \in C, a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。$\boldsymbol{M}(i, j)$ 表示可辨识矩阵中第 i 行 j 列的元素,则可辨识矩阵 $\boldsymbol{M}$ 定义为:$\boldsymbol{M}(i, j) = \{a_k \mid a_k \in C \wedge a_k(x_i) \neq a_k(x_j)\}$,否则 $\boldsymbol{M}(i, j) = \emptyset$。

从定义可以看出,可辨识矩阵存放的是知识表达系统中实例记录之间的差别,如果辨识矩阵中第 i 行 j 列的元素 $\boldsymbol{M}(i, j)$ 为空,表示知识表达系统的第 i 行和第 j 行的实例记录完全相同,否则,$\boldsymbol{M}(i, j)$ 存放知识表达系统的第 i 行和第 j 行的实例记录中属性值不同的属性。

例 3:信息表 1 的属性列表中,令 $a =$ "头痛",$b =$ "乏力",$c =$ "发烧",则表 1 的可辨识矩阵如下:

	1	2	3	4	5	6
1		c	c			ac
2	c			ac	ab	
3	c			ac	abc	
4		ac	ac			c
5		ac	abc			bc
6	ac			c	bc	

定义 4:核可定义为识别矩阵中只有一个元素的矩阵项的集合,即:CORE(C) $= \{a \in C, \delta(u_1, u_2) = \{a\}, \exists u_1, u_2 \in U\}$,其中,$\delta(u_1, u_2)$ 表示决策表中对象 u_i 和对象 u_j 具有不同值的属性。

定义 5:每一个识别矩阵对应一个惟一的识别函数 $f_c(s)$,其定义:设 n 阶识别矩阵的任一非空元素 C_{ij} 为:$C_{ij} = \{a_1, a_2, \cdots, a_l\}$,其中对应的析取项 E_{ij} 为:$E_{ij} = a_1 \vee a_2 \vee \cdots \vee a_l$,则对应的属性化简 RED($C$) 为所有非空元素对应的析取项 E_{ij}:RED(C) $= E_1 \wedge E_2 \wedge \cdots \wedge E_K$。

例 4:根据定义 4,例 3 中识别矩阵的条件属性核为 CORE(C) $= \{c\}$,根据定义 5 相应地将属性化简为:RED(C) $= \{a, c\}$,因此表 1 可化简为表 2。

表 2 属性约简后决策表

对象 U	条件属性 C		决策属性 D
	头痛	发烧	流感
x_1	是	正常	否
x_2	是	偏高	是
x_3	是	很高	是
x_4	否	正常	否
x_5	否	偏高	否
x_6	否	很高	是

说明:条件属性“乏力”对分类不起作用,去掉该冗余属性后仍能保持分类精度不变。

3 数据挖掘中规则挖掘实例

以教师的课堂教学质量与哪些因素相关这个挖掘任务为例,详细介绍一个信息系统在基于对决策表属性约简的基础上采用了可变精度粗糙集模型实现规则的获取。选取分析目标,经过属性约简后,产生决策表(见表 3)。

表 3 决策表

U	C			D
	a_1	a_2	a_3	
1	40 ~ 50	高级	博士级	优
2	40 ~ 50	高级	硕士级	良
3	30 ~ 40	中级	硕士级	良
4	30 ~ 40	中级	硕士级	中
5	30 ~ 40	初级	硕士级	良
6	30 以下	初级	硕士级	中
7	50 以上	高级	本科级	良
8	50 以上	高级	本科级	良
9	30 ~ 40	中级	本科级	中
10	30 ~ 40	中级	本科级	差

所有分析对象的集合 $U = \{1,2,3,4,5,6,7,8,9,10\}$,属性集 $A = C \cup D$,其中 C 表示条件属性集,D 是决策属性。条件属性 $C = \{a_1, a_2, a_3\}$,其中 a_1 = 该参评教师所属年龄段;a_2 = 该参评教师的职称;a_3 = 该参评教师的学历。决策属性 D = 教师的教学效果,教师的教学效果分成 4 个等级{优、良、中、差}。

分析表 3,根据现有的 3 个条件属性不能够区分开的实例对象有{3,4},{7,8} 和{9,10}。针对这一问题采用可变精度粗糙集模型来实现规则提取。根据决策属性 D,将 U 划分成 4 个子集:$Y1 = \{1\}$,$Y2 = \{2,3,5,7,8\}$,$Y3 = \{4,6,9\}$,$Y4 = \{10\}$。根据条件属性,将 U 划分成 7 个子集:$C1 = \{1\}$,$C2 = \{2\}$,$C3 = \{3,4\}$,$C4 = \{5\}$,$C5 = \{6\}$,$C6 = \{7,8\}$,$C7 = \{9,10\}$。令 $\beta = 0.6$ 对于概念 $Y1 = \{1\}$ 有:

(1) $Y1$ 的 β 正域为 $C1$,即 U 中,以不大于 β 的分类误差,能分于集合 $Y1$ 的对象集 $C1 = \{1\}$;

(2) $Y1$ 的 β 负域为 $C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7$,即 U 中,以不大于 β 的分类误差,能分于集合 $Y1$ 的补集的对象集 $C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 = \{2,3,4,5,6,7,8,9,10\}$;

(3) $Y1$ 的 β 的边界区域为空集。

通过以上分析,可以得到表 4,在表中,$Y1$ 的值域 1 表示该实例记录属于 $Y1$ 的 β 正域,0 表示该实例记录属于 $Y1$ 的 β 负域。

表 4 决策表

U	C			$Y1$	$Y1$ 中的实例	非 $Y1$ 中的实例
	a_1	a_2	a_3			
1	40 ~ 50	高级	博士级	1	1	0
2	40 ~ 50	高级	硕士级	0	0	1
3,4	30 ~ 40	中级	硕士级	0	0	2
5	30 ~ 40	初级	硕士级	0	0	1
6	30 以下	初级	硕士级	0	0	1
7,8	50 以上	高级	本科级	0	0	2
9,10	30 ~ 40	中级	本科级	0	0	2

根据表 4,可以产生决策规则:(a_1 = “40 ~ 50”) $\wedge$ (a_2 = “高级”) $\wedge$ (a_3 = “博士级”) $\rightarrow D = 1$,可信度为 1 即是根据表 4 现有的数据分析,当满足年龄段在 40 ~ 50 岁之间,职称为“高级”,学历为“博士”的教师,教学评价结果为“优”。

对于概念 $Y2 = \{2,3,5,7,8\}$,有:

(1) $Y2$ 的 β 正域为:$C2 \cup C3 \cup C4 \cup C6 = \{2,3,4,5,7,8\}$;

(2) $Y2$ 的 β 负域为:$C1 \cup C3 \cup C5 \cup C7 = \{1,3,4,6,9,10\}$;

(3) $Y2$ 的 β 的边界区域为空集。

通过以上分析,可以得到表 5,在表 5 中,$Y2$ 的值域 1 表示该实例记录属于 $Y2$ 的 β 正域,0 表示该实例记录属于 $Y2$ 的 β 负域。

表 5 决策表

U	C			$Y1$	$Y2$ 中的实例	非 $Y2$ 中的实例
	a_1	a_2	a_3			
1	40 ~ 50	高级	博士级	0	0	1
2	40 ~ 50	高级	硕士级	1	1	0
3,4	30 ~ 40	中级	硕士级	1	1	1
5	30 ~ 40	初级	硕士级	1	1	0
6	30 以下	初级	硕士级	0	0	1
7,8	50 以上	高级	本科级	1	2	0
9,10	30 ~ 40	中级	本科级	0	0	2

根据表 5,可以产生决策规则:

(a_1 = “40 ~ 50”) $\wedge$ (a_2 = “高级”) $\wedge$ (a_3 = “硕士级”) $\rightarrow D = 1$,可信度为 1;

(a_1 = “30 ~ 40”) $\wedge$ (a_2 = “中级”) $\wedge$ (a_3 = “硕士级”) $\rightarrow D = 1$,可信度为 0.5;

(a_1 = “30 ~ 40”) $\wedge$ (a_2 = “初级”) $\wedge$ (a_3 = “硕士级”) $\rightarrow D = 1$,可信度为 1;

(a_1 = “50 以上”) $\wedge$ (a_2 = “高级”) $\wedge$ (a_3 = “本

(下转第 132 页)

成任务的总时间与效益最优算法相比，有较大的降低，提高了调度性能。

4 结 论

基于经济模型的网格资源调度算法的关键是构造效益函数。效益函数构造的方便性和它的性能决定资源调度算法的优劣。借助计算经济模型的思想，提出了一种基于遗传编程改进的网格资源调度算法，利用遗传编程技术来寻找和构造的有效的效益函数，一方面提高了效益函数构造的方便性，另一方面提高了资源调度算法的性能。

参考文献：

[1] Foster I, Kesselman C. The Grid: Blueprint for a Future Computing Infrastructure[M]. USA: Morgan Kaufmann Publishers, 1999.

[2] Oram A. Peer - to - Peer: Harnessing the Power of Disruptive Technologies[M]. USA: O'Reilly Press, 2001.

[3] Parastatidis S, Watson P, Webber J. Grid Resource Specification[R]. North East Regional e - Science Centre, School of Computing Science University of Newcastle, Newcastle - upon - Tyne, NE1 7RU, United Kingdom, 2003: 2 - 3.

[4] Steiglitz K, Honig M L. A computational market model based on individual action[M]//Market - based control: a paradigm for distributed resource allocation table of contents. [s. l.]: [s. n.], 1996: 1 - 27.

[5] Buyya R, Abramson D, Giddy J. A Case for Economy Grid Architecture for Service - Oriented Grid Computing[C]//Proceedings of the International Parallel and Distributed Processing Symposium: 10th IEEE International Heterogeneous Computing Workshop (HCW 2001). San Francisco, California, USA: IEEE CS Press, 2001.

[6] Buyya R, Abramson D, Giddy J. An Economy Driven Resource Management Architecture for Global Computational Power Grids[C]//Proceedings of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000). Las Vegas, USA: CSREA Press, 2000.

[7] 刘一萌，舒 勤. 基于 Bargain 的经济模型的网格资源交易管理算法[J]. 计算机工程与应用，2004(17)：191 - 193.

[8] 陈冬娥，杨 扬，刘 丽. 基于效用最优的网格计算资源调度算法[J]. 计算机工程与应用，2006(2)：191 - 193.

[9] 胡自林，徐 云，毛 涛. 基于效益最优的网格资源调度[J]. 计算机工程与应用，2005(7)：69 - 70.

[10] Koza I J R. Genetic Programming: On the Programming of Computers by Means of Natural Selection[M]. Cambridge, MA: MIT Press, 1992.

[11] Buyya R, Murshed M. GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing[J]. The Journal of Concurrency and Computation: Practice and Experience (CCPE), 2002, 14(13 - 15): 312 - 317.

(上接第 128 页)

科级") → $D = 1$，可信度为 1。

即根据表 5 现有的数据分析，当满足年龄段在 40～50 岁之间，职称为"高级"，学历为"硕士"的教师，教学评价结果为"良"。当满足年龄段在 30～40 岁之间，学历为"硕士"，职称为"初级"或"中级"的教师，教学评价结果为"良"。当满足年龄段在 50 岁以上，职称为"高级"，学历为"本科"的教师，教学评价结果为"良"。继续对表 5 进行分析，可以得到所有的规则。

4 结 论

信息系统经过粗糙集方法[7]处理后，可以实现以下目标：

1)从信息系统中去除冗余对象，即值约简；

2)从信息系统中去除冗余属性，即属性约简；

3)得到独立属性的最小子集，同时保证与原属性的分类质量相同，即约简的属性集；

4)得到各约简集的交集——核(core)，这是最优约简集的必要元素；

5) 得到分类规则。粗糙集方法得到的分类规则一般是符号形式的显示规则，正是数据挖掘所追求的，近年来得到越来越广泛的应用。

对于不完备的信息系统，应用粗糙集理论进行数据挖掘是非常有效的，可广泛的应用多个领域[7]。

参考文献：

[1] Dunham M H. 数据挖掘教程[M]. 北京：清华大学出版社，2005.

[2] 曾黄麟. 粗集理论及其应用[M]. 重庆：重庆大学出版社，1998.

[3] 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.

[4] Pawlak Z. Rough Set Theoretical Aspects of Reasoning about Data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[5] 温有奎. 知识元挖掘[M]. 西安：西安电子科技大学出版社，2004.

[6] Pyle D. 业务建模与数据挖掘[M]. 北京：机械工业出版社，2005.

[7] 纪 滨. 粗糙集理论及进展的研究[J]. 计算机技术与发展，2007，17(3)：69 - 72.