一种基于粗糙集的分类数据挖掘算法*)

马君华　陈云开

(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　目前,粗糙集理论及数据挖掘的研究已经成为热点领域。本文提出一种基于粗糙集理论的分类数据挖掘算法,从实际数据出发,运用不同简化层次的算法,导出每个层次上的信息集,最后得到规则集。在进行推理和决策分析时,按照一定算法进行匹配,得出结论。算法分析表明,这一算法属性约简具有较好有效性,可减少未知样本参与分类的决策属性,适合模型相对稳定、更新不频繁且建模过程可以在后台进行的应用。

关键词　粗糙集,数据挖掘,知识表达

A Data Mining Algorithm Based on Rough Set Theory

MA Jun-hua　CHEN Yun-kai

(School of Computer Science and Technology, Huazhong University of Science &Technology, Wuhan 430074, China)

Abstract　At present, Rough Set Theory and Data Mining have become hot topics of computer research. This paper presents a model of data mining based on rough set, from applying various reductive level algorithms on practical data to elicit information set, and get a rules set eventually. Conclusions are acquired on matching according a certain algorithm when reasoning and decision are made. Lastly, a simulated example on how to create and apply this model is given.

Keywords　Rough sets, Data mining, Knowledge representation

1　引言

粗糙集理论是 20 世纪 80 年代初 Z. Pawlak[1] 针对 G. Firege 的边界域思想提出的,基于给定训练数据内部的等价类,用上下近似集合来逼近数据库中的不精确概念。用于分类,可以发现不准确数据或噪声数据内在的结构联系;用于特征归约,可以识别和删除无助于给定训练数据分类的属性;用于相关分析,可以根据分类任务评估每个属性的贡献或意义。其主要思想是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则[2]。

粗糙集理论具有这样的特点:能够处理各种数据,包括不完整的数据以及拥有众多变量的数据;能够处理数据的不精确性和模棱两可,包括确定性和非确定性的情况;能求知识的最小表达和知识的各种不同颗粒层次;能从数据中揭示出概念简单、易于操作的模式;能产生精确而又易于检查和证实的规则;因此特别适于智能控制中规则的自动生成。粗糙集的这些特点保证了它能够提供有效的技术用于数据挖掘的数据预处理、数据的约简、规则生成等方面,故该理论作为数据挖掘领域的一种主流方法,正受到越来越多研究者的关注[3]。

最初,关于粗糙集理论的研究主要集中在波兰,当时并没有引起国际计算机界和数学界的重视。直到 1990 年前后,由于该理论在数据的决策与分析、模式识别、机器学习与知识发现等方面的成功应用,才逐渐引起了世界各国学者的广泛关注。1991 年,Z. Pawlak 的专著《粗糙集——关于数据推理的理论》问世,标志着粗糙集理论及其应用的研究进入了活跃时期。1992 年,在波兰召开了关于粗糙集理论的第一届国际学术会议。1995 年,ACM Communication 将粗糙集列为新浮现的计算机科学研究课题。目前粗糙集理论已成为计算机科学最为活跃的研究领域之一,在许多应用领域如医疗数据分析、水泥窑生产控制算法、地理学、振动分析、飞行员技能评定、开关电路综合、语言识别、近似分类、故障诊断、成本预测等已得到发展。1998, 2000 和 2002 年,分别召开了三届 RSCTC (Rough Sets and Current Trends in Computing)国际会议,表明粗糙集的研究已步入发展期[4-6]。

与国外相比,国内研究稍晚,没有形成整体力量。国内对粗糙集理论的研究始于 90 年代中期,1993 年国家自然科学基金首次对数据库中知识发现领域的研究项目给予资助。目前,许多科研单位和高等院校竞相开展相关领域的基础理论及应用研究,取得了令人鼓舞的成果。2001 年 5 月,在重庆邮电学院举办了首届中国粗糙集和软计算学术研讨会;2002 年 10 月在苏州大学举办了第二届中国粗糙集和软计算学术研讨会;2003 年 5 月,在重庆邮电学院同时举办第三届中国粗糙集和软计算学术研讨会和第九届粗糙集、模糊集、数据挖掘与粒度计算国际学术会议(RSFDGrC' 2003),这些会议的举办,表明我国粗糙集理论和数据挖掘研究的队伍正在不断壮大,已经得到国际同行的重视和认可。粗糙集理论逐渐应用于数据挖掘(DM)领域中,并在对大型数据库中不完整数据进行分析和学习方面取得了显著的成果,使得粗糙集理论及数据挖掘的研究成为热点领域[7-10]。

2　相关知识

2.1　粗糙集的定义

定义 1　U 为一对象集合,称为论域。R 为 U 上的一个由对象属性集 Ω 确定的等价关系,即 R 是 U 的一个划分,我

*)本课题获国家自然科学基金项目(60403027)、国家科技攻关项目(2001BA102A06-11)资助。马君华　博士生,主要研究领域为电力系统自动化;陈云开　博士生,研究方向为金融数据挖掘等。

们称$\langle U,R\rangle$为近似空间。对于任何 $P\subset R$ 且 $P\neq\varnothing$，则$\cap P$ 也是 U 上的一等价关系，我们称$\cap P$ 为 P 上不可分辨关系，记为 Ind (P)。

对于 $X\subseteq U$，我们称

$R_(X)=U\{Y_i\in U|\mathrm{Iud}(R):Y_i\subseteq X\}$是 X 的下近似集；

$R^-(X)=U\{Y_i\in U|\mathrm{Ind}(R):Y_i\cap X\neq\Phi\}$是 X 的上近似集；

$R^-(X)-R_(X)$是 X 的边界域。

定义 2 $\xi=\langle U,\Omega,Vq,fq\rangle(q\in\Omega)$为一知识系统。

U 为对象集合；Ω 为属性集合；Vq 为属性值集合；

对每一 $q\in\Omega$，有一个映射函数 $fq:U\to Vq$，$fq(O)$也可以写作 Oq，表示对象 O 的 q 属性的值。

定义 3 分辨矩阵定义为

$M(\Omega(=(m_{i,j}),(1\leqslant i,1\leqslant j\leqslant|U/\mathrm{Ind}(\Omega/|),(m_{i,j})=\{q\in\Omega|O_i^q\neq O_i^q\}$

分辨矩阵对象将粗糙集 Ind(Ω)转换为分辨矩阵 $M(\Omega)$。

2.2 基于粗糙集模型的知识表达

在数据挖掘中，知识由决策系统表示，决策系统是一个带有决定域的知识表达系统。

定义 4 假定 $\xi=\langle U,\Omega,Vq,fq\rangle(q\in\Omega)$为一个定义 2 中描述的知识系统，$\eta=\langle U,C,D,Vq,fq\rangle$为一个决策系统，$C$ 是 Ω 的一个子集，表示条件属性集合，D 是 Ω 的子集，表示决策属性集合。

定义 5 对于决策系统 η，分辨矩阵定义为

$M(C,D)=(m_{i,j}),(1\leqslant i,j\leqslant n,n=|U/\mathrm{Ind}(C)|)(m_{i,j})=\{q\in C\|O_i^q\neq O_i^q,O_i^d\neq O_i^d,d\in D\}$

定义 6 $\eta=\langle U,C,D,Vq,fq\rangle$为一个决策系统，Ind$(C,D)=\{(x,y)\in U\times U,\forall_r\in C,r(X)=r(y)\}$被称为 η 上的不可分辨关系。

定义 7 $\eta=\langle U,C,D,Vq,fq\rangle$为一个决策系统，条件属性 C 的约减集 C' 是一非空子集，如果

(1)Ind$(C',D)=$Ind(C,D)；

(2)不存在 $C''\subset C'$，使得 Ind$(C'',D)=$Ind(C,D)。C 的约减记为 Red$\eta(C)$；所有约减集的交集称为核，记为 Core$\eta(C)$。Core$\eta(C)=\cap$Red$\eta(C)$。

3 基于粗糙集模型的数据挖掘算法

基于粗糙集的数据库中发现规则需经过以下步骤(图 1)：

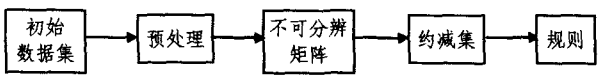

图 1 基于粗糙集的数据挖掘处理过程示意图

预处理。将数据库中的初始数据信息转换为粗糙集形式，明确条件属性和决策属性；

属性约减。生成不可分辨矩阵，并在分辨矩阵的基础上生成约减属性集；

发现规则。在约减的信息表中，根据可信度阈值发现规则。

(1)属性约减。属性约减是粗糙集理论中的一个重要的研究课题。通常，决策信息系统中的属性并非同等重要，且存在冗余，这不利于做出正确而简洁的决策。属性约减是在保持决策信息系统的分类和决策能力不变的前提下，删除不相关或不重要的属性。因此，在进行属性规约时，人们总希望能找到属性的最小规约，但这是一个 NP 难度的问题。幸运的是，在大多数情况下，无需找出属性的最小约减，因为用户只对与处理任务相关的最佳子集感兴趣。根据属性之间的依赖关系、重要性等，可以比较容易有效地找出一个最佳规约集。下面给出了计算最佳规约集的算法。该算法 1 核心作为计算的起点，每次选择重要性最大的属性加入属性集 Redη。在 Step4 的前向选择结束后，属性集 Redη 已经包含了一些重要作用的属性，且没有改变原始属性集与决策属性之间的依赖程度。最后的反馈过程是从属性集 Redη 中逐个去掉属性。如果去掉该属性会造成依赖度变化，则恢复该属性，否则删除该属性。最后剩下的属性集就是最佳规约集。删除决策信息系统中不属于最佳规约集的属性就可以得到最佳属性约减的决策信息系统。

算法 1 计算最佳规约集

```
输入:条件属性集 C,决策属性集 D,决策信息系统 Ind(C,D)
输出:生成分辨矩阵 M(Ω),最佳规约集 Redη(Ω)
Step1  置 Redη(Ω)=∅,Coreη(C) =∅,n=Ind(C,D),生成一个 n*
n 的空属性矩阵;
Step2  生成分辨矩阵
       for ( i=0 ; i<n ; i++ )
           for( j=i+1 ; j<n ; j++ )
       根据分辨矩阵的定义生成 m i,j
Step3  求核
for ( i=0 ; i<n ; i++)
              for ( j=i+1 ; j<n ; j++)
若|m_{i,j}| - 1 将 m_{i,j}加入 Coreη(Ω)
Step4  将含有 Coreη(Ω)中元素的矩阵元素置空;
Step5  将得矩阵中出现频率最高的属性 q,将 q 加入 Redη(Ω),且将
含 q 属性的矩阵元素置空;
Step6  若 M(Ω)≠∅(则转到 Step5;否则结束。
```

(2)元组合并。合并最佳属性规约的决策信息系统中的元组分两个步骤。首先，如果多个(两个或两个以上)元组的各个条件属性和决策属性都一一对应相同，则将这多个元组合并成一个元组；其次，如果多个元组的决策属性相同，条件属性只有一个不同，并且这多个元组在该条件属性上的取值覆盖了它所有可能的值，则将这个多个元组合并成一个元组，并且将该条件属性从元组中删除，最终便可得到最简决策信息系统。

(3)规则提取。尽管由上述步骤得到的最简决策信息系统已经可以为决策者提供正确而简洁的决策，但用这种方法表示知识的可读性很差，不容易被人理解，尤其是当生成的最简决策信息系统仍然较大时。因此，有必要提取出隐含在其中的决策规则。其具体方法是：对于简化的决策信息系统中的每一个元组，将它的每一个条件属性的属性-值对形成规则前件的一个合取项；将它的每一个决策属性的属性-值对形成规则后件的一个合取项。最终便可得到决策规则集。

算法 2 规则提取

```
输入:条件属性、可信度阈值 ū0
输出:规则集
Step1  输入条件属性 C1;
Step2  对于 Ind(C,D)
找出与 C1 属性相同的元素个数 N;
找出与属性 C1,D 都相同的元素个数 M;
Step3  若 ū=该划分大小/N≥ū0,且该规则不存在于规则表中,则输
出该规则。
```

(4)规则评估。对于特定的决策者来说，他不一定对上述生成的所有决策规则都感兴趣。因此，决策者需要根据自己的目标进一步限制挖掘过程产生的不感兴趣的决策规则的数量。这可以通过设定规则兴趣度度量方法来实现。规则兴趣度度量的方法很多，其中最常见的方法有支持度度量和置信度度量。只有那些支持度和置信度都大于或等于决策者事先设定的最小支持度阈值和最小置信度阈值的决策规则才被认

为是有效的。对于形如"A→B"的决策规则，其支持度和置信度的定义分别为：支持度(A→B)＝包含 A 和 B 的元组数/元组总数；置信度(A→B)＝包含 A 和 B 的元组数/包含 A 的元组数。

4 算法分析

4.1 复杂度分析

算法 1 最坏情况下的时间复杂度为 $O(|C||U|2)$。分辨矩阵中的项数最多是 $|U|(|U|-1)/2$，因此算法 1 最佳规约集计算中，Step2 最坏情况下的分辨矩阵计算，其时间复杂度是 $O(|C||U|2)$；Step3 求核的代价等同步骤 2，因为该步骤实质上是对分辨矩阵进行了一遍扫描；Step5 求约简是通过遍历核完成，而表示核的矩阵大小同分辨矩阵一样，因此最坏情况的时间复杂度仍是 $O(|C||U|2)$，但由于在求核时，已经从分辨矩阵去掉了一些项，因此这一步的运行时间应当比步骤 2 和步骤 3 要快。

规则提取的时间复杂度为 $O(|C||U|)$。上文列出的算法 2 仅针对一个决策属性进行计算，该算法 Step2 中对于 M 和 N 的计算可通过扫描决策系统一次完成，因此该算法的时间复杂度为 $O(|U|)$，那么对 $|C|$ 个决策属性进行规则提取的时间复杂度就为 $O(|C||U|)$。

综合以上步骤，基于粗糙集的分类算法最坏情况下的时间复杂度是 $O(|C||U|2)$。与用于分类模型生成的决策树算法 ID3 相比，基于粗糙集的分类算法在模型建立的时间耗费上略逊(ID3 算法的时间复杂度是 $O(|C||U|\log|U|)$)。但与 ID3 相比，由于实现了属性约简，生成的规则集中决策属性要少于 ID3，因此在对未知样本进行分类时，平均时间耗费要小于 ID3。这一特性使得该算法适合模型相对稳定、更新不频繁且建模过程可以在后台进行的应用，而金融、电力等领域的数据挖掘系统正需要这样的特性。

在金融、电力等领域的数据挖掘系统中，需要处理的交易量特别多，给存储以及处理带来问题。基于粗糙集的分类算法虽然耗时较多，但它处理的是数据量较小的训练集，进行属性约简后，为庞大的未知交易集指定了参与分类的决策属性，从而增加分类处理的效率。因此，算法属性约简的有效性成为算法的关键，接下来的实验分析主要围绕这一问题进行。

4.2 实验对比分析

对基于粗糙集的分类算法的实验分析主要集中于属性约简的有效性上，即属性约简后，与原决策系统相比，决策属性有没有减少。另外，这种减少是否影响对未知样本的分类结果。以此为原则，我们将基于粗糙集的分类算法与 ID3 算法生成的规则相比较，分析约简的有效性。在给出实验结果之前，给出实验所用的数据集。

算法程序采用 Java 编写，Java 运行环境为 JRE 1.5.1，编程平台采用了 Eclipse 3.1.2，以及 Hibernate 的开发工具包，系统测试程序运行于 Windows XP Professonal with SP2。测试采用的机器硬件环境包括：CPU 为 Intel Pentium(R)D 2.80GHz，内存 1G，硬盘 160GB(7200 转)8MB 缓存。

本实验使用的数据集来源于金融领域的进出口核销数据，其中出口收汇数据来自核销单总帐表。顺序选取前 65536 条记录，并从中选取贸易方式为来料加工、对口进料加工、进料非对口，以及三资进料加工的记录；进口付汇数据来自付汇数据表，随机选取 37336 条数据；将以上两表合并，并根据逃套汇可疑交易的特点进行分类标记，经格式转换形成具有 116 条记录的训练集，用该训练集形成决策系统，则该系统包含 4 个决策属性(a, b, c, d)和 1 个分类属性 e，各属性及其取值代表的含义如表 1 所示。

表 1 决策系统属性说明

属性名称	属性说明	取值说明
a	国外直接来料加工进口额/总来料加工额度	0：0%～33% 1：34%～70%2：71%～100%
b	进料加工出口收汇额/进料加工进口付汇额	0：0%～33% 1：34%～70%2：71%～100%
c	来料加工出口收汇额	0：0～70%1：71～100%
d	具备何种合法加工业务	0：同时具备来料加工和进料加工业务 1：来料加工 2：进料加工
e	交易判定	0：正常 1：待查 2：可疑

表 2 则给出了决策系统 $\eta=\langle U,\Omega,V_q,f_q\rangle$ 的片断。

经计算得 η 的分辨矩阵，如表 3 所示。由分辨矩阵得 $\mathrm{Red}\eta(\Omega)=\{\{a,b\},\{b,c,d\}\}$，从这些约简出发，衍生出如图 2 所示的一系列节点。再经过每个节点的决策系统计算 $u\geqslant\bar{u}0$，将规则记入与前件相匹配节点的规则集中。仅以节点 Nbc 上的规则为例，如表 4 所示。

表 2 决策系统 η 片断图

	Attributes	Decision
n	a b c d	e
1	0 0 1 1	0
2	0 1 0 2	1
3	0 2 1 2	1
4	1 0 1 2	0
5	1 1 0 1	0
6	1 2 0 1	2
7	2 0 0 1	2
8	2 1 1 0	1

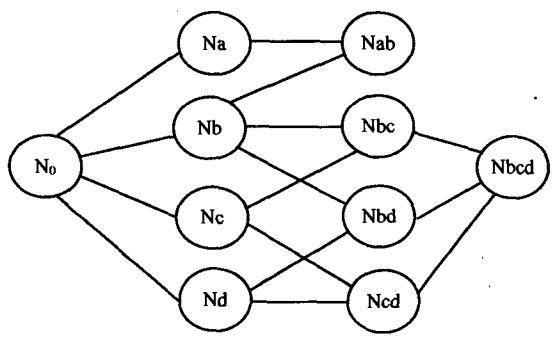

图 2 模型中节点关系示意图

表 4 节点 Nbc 上的规则集合

Attributes	Decision
a b	e
2 2	0
0 2	1
1 1	1
0 1	2
0 0	2
1 0	2

表5 ID3算法生成的规则

ttributes	Decision
a b c d	e
2 2 0 0	0
2 2 0 1	0
0 2 1 0	1
0 2 0 0	1
1 1 1 2	1
0 1 1 1	1
0 1 0 0	2
0 0 1 2	2
1 0 1 2	2
0 0 0 1	2

针对决策系统 η，基于粗糙集的分类算法得到的规则可理解为，当交易具有以下特征：(1)有较大的进料加工进口付汇，但没有或只有较少的进料加工出口及收汇；(2)有较大的来料加工出口，但没有或只有较少的来料加工国外直接进口(来料大部分由国内通过深加工转厂转入)，则该交易为可疑交易。

表5给出了ID3算法在决策系统 η 上生成的规则。对比这两个算法得出的规则，基于粗糙集的分类算法得到规则的前件基本为ID3算法得到规则前件的子集，且决策结论也基本一致。因此，从实验结果看，该算法的属性约简功能具有较好的效果，与粗糙集理论描述一致。

表3 分辨矩阵

Ø	{b,c,d}	{b,d}	Ø	Ø	{a,b,c}	{a,c}	{a,b,c,d}
{b,c,d}	Ø	Ø{a,b,c}	{a,d}	{a,b,d}	{a,b,d}	{a,c,d}	
{b,d}	Ø	Ø	{a,b}	{a,b,c,d}	{a,c,d}	{a,b,c,d}	{a,b,c,d}
Ø	{a,b,c}	{a,b}	Ø	Ø	{b,c,d}	{a,c,d}	{a,b,c,d}
Ø	{a,d}	{a,b,c,d}	Ø	Ø	{b}	{a,b}	{a,c,d}
{a,b,c}	{a,b,d}	{a,b,d}	{b,c,d}	{b}	Ø	Ø	Ø
{a,c}	{a,b,d}	{a,b,c,d}	{a,c,d}	{a,b}	Ø	Ø	Ø
{a,b,c,d}	{a,c,d}	{a,b,c,d}	{a,b,c,d}	{a,c,d}	Ø	Ø	Ø

结束语 本文提出一个基于粗糙集的挖掘算法，该算法用于生成决策模型，该决策模型由一组规则组成，首先实现属性约简，形成分辨矩阵，然后从中发现规则。这一算法属性约简具有较好有效性，可减少未知样本参与分类的决策属性，适合模型相对稳定、更新不频繁且建模过程可以在后台进行的应用。

参考文献

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrech: Kluwer Academic Publishers, 1991:10-34

[2] Han B, Wu T J. Data Mining in Multisensor System Based on Rough Set Theory// Proceedings of the American Control Conference. Arlington, 2001:4427-4431

[3] Rough集理论与知识获取. 西安：西安交通大学出版社，2001：156-176

[4] 张文修，吴志伟. 粗糙集理论和研究综述. 模糊系统与数学，2000，14(4)：1-12

[5] 杨涛，李龙澎. 一种基于粗糙集聚类的数据约简算法. 系统仿真学报，2004，16(10)：2195-2200

[6] 徐泉清，朱玉文，李亮. 一种结合粗糙集和COBWEB的聚类器. 计算机应用，2005，25(6)：1350-1352

[7] 干庆东，马昕，戴华平. 基于粗糙集属性量度的数据库分解方法. 浙江大学学报(工学版)，2004，38(9)：1196-1199

[8] 刘业政，杨善林. 基于粗集理论的Null值估算方法研究. 计算机工程，2001，27(10)：41-42

[9] 苗夺谦，王压. 粗糙集理论中概念与运算的信息表示. 软件学报，1999，10(2)：113-116

[10] 苗夺谦，胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展，1999，36(6)：681-684

(上接第212页)

导致结论不同有两方面原因：首先，衡量的粒度不同，定义不同，使得结论不同；另一方面，Parssian乘积过程可以简化为量化向量的乘积，而本模型笛卡尔乘积可以看成是键属性、非键属性分别乘积然后归并合成的结果。

参考文献

[1] Kon H B, Madnick E, Siegel M D. Good Answers From Bad Data: A Data Management Strategy[C]. Massachusetts Institute of Technology (MIT), Sloan School of Management, 1995: 1-16

[2] Reddy M P, Wang R Y. A Data Quality Algebra for Estimating Query Result Quality[C]// CISMOD Conference. Bombay, 1996

[3] Yang W, Wang L, Richard Y, et al. Data Quality[M]. Kluwer, 2001

[4] Parssian A, Sumit S, Varghese J S. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product[J]. Management Science, 2004, 50(7): 967-982

[5] Parssian A, Sumit S, Varghese J S. Assessing Information Quality for the Composite Relational Operation Join[C]. 2002: 225-237

[6] Motro A, Igor R. Not All Answers Are Equally Good: Estimating the Quality of Database Answers[M]. Kluwer Academic Publishers, 1997: 1-21

[7] Motro A, Igor R. Estimating The Quality of Databases[C]// The 3rd International Conference on Flexible Query Answering Systems (FQAS). Cambridge, MA, 1998: 298-307

[8] Scannapieco M, Carlo B. Completeness in the Relational Model: a Comprehensive Framework[C]. 2004: 333-345