基于数据库系统的可变精度粗糙集模型*)

张东星[1] 苗夺谦[1] 李道国[1,2] 张红云[1]
(同济大学计算机科学与工程系 上海 200092)[1] (太原理工大学阳泉学院 阳泉 045001)[2]

摘 要 本文将粗糙集理论与关系数据库系统结合起来,使数据库的关系运算运用于粗糙集的集合操作,提出了基于SQL求可变精度粗糙集模型的近似核和近似约简的方法。由于数据库管理系统具有存取效率高,存储空间的利用率高,适合大规模数据的存储等优点,因此与普通文件的数据挖掘相比,基于SQL的可变精度粗糙集模型对于大规模数据库的处理更有效。

关键词 可变精度粗糙集,数据库系统,数据挖掘

A Variable Precision Rough Set Model Based on Database System

ZHANG Dong-Xing[1] MIAO Duo-Qian[1] LI Dao-Guo[1,2] ZHANG Hong-Yun[1]
(Department of Computer Science and Engineering, Tongji University, Shanghai 200092)[1]
(Yang Quan College of Tai Yuan University of Technology, Yangquan 045001)[2]

Abstract In this paper rough sets theory is integrated with relational database management systems and instead of the operations of original rough sets relational calculus is taken advantage of. The implementation of calculating approximate core and reduct of available precision rough set model is based on SQL. With the merits of high efficiency of accessing data, using rate of space and applicability for storage of large data sets, compared with data mining performed in flat files, our algorithms is very efficient.

Keywords Variable precision rough set, Database system, Data mining

1 引言

粗糙集理论[1](Rough Sets)是Z. Pawlak于20世纪80年代提出的,它是一种处理模糊的和不确定知识的数学工具,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,导出决策和分类规则。到现在为止,粗糙集理论已成功地应用于数据挖掘、机器学习、决策分析等众多领域之中。尽管目前提出了许多的粗糙集模型,例如Ziarko等提出的可变精度粗糙集模型(Variable Precision Rough Set Model)、Hu提出的广义粗糙集模型(Generalized Rough Set Model)等,其主要目的是为了消除噪音或者错误的数据。这些模型不能很好地和关系数据库系统结合起来,实现数据挖掘。由于当前基于粗糙集理论的数据挖掘都是基于普通文本的数据格式,为了计算核和约简,需要构造基于条件属性值和决策属性值的所有等价类,这些运算很耗时。可是,这在数据挖掘的过程中又是非常普遍的,在普通文件上运算效率的低下限制了粗糙集理论在大规模数据上的适用性。

数据挖掘的一个重要问题是数据挖掘系统与数据库或数据仓库系统是否能够很好地耦合。由于数据库管理系统(Database Management System,简称DBMS)具有管理方便、存贮占用空间小、检索速度快、修改效率高、安全性好、并发控制、存取控制、完整性检查、运行日志管理、事务管理、自动恢复等优点,因此数据仓库大都是基于DBMS构造的。但是原有粗糙集理论是基于集合之间的运算(包含、交、并等)或者逻辑运算(析取、合取等)的,对于这些运算结构化查询语言(SQL)不能很好地实现。因为SQL只能做一些基本的统计操作(求和、计数、最大、最小、标准差等),所以我们需要将原有的粗糙集的集合操作转化为SQL语言的统计操作。

本文将数据库的关系运算运用于粗糙集的集合操作,提出了基于SQL语言实现,实现了粗糙集理论与关系数据库系统的紧密结合,提高了算法的性能。由于数据库系统具有存取效率高,存储空间的利用率高,适合大规模数据的存储等优点,因此与普通文件上的数据挖掘相比,我们的基于数据库系统的粗糙集模型对大规模数据更有效。

2 可变精度粗糙集理论

表1 含有噪音的信息表

object	$a1$	$a2$	$a3$	$a4$	$a5$	d
$O_{1\text{-}20}$	1	2	1	2	1	1
O_{21}	1	2	1	2	1	2
$O_{22\text{-}45}$	1	2	1	2	2	1
$O_{46\text{-}70}$	1	2	1	6	1	2
$O_{71\text{-}97}$	1	2	3	4	3	1
$O_{98\text{-}126}$	1	3	1	4	2	2
$O_{127\text{-}150}$	1	3	3	4	1	2
$O_{151\text{-}180}$	2	3	1	2	2	3
$O_{181\text{-}182}$	2	3	1	2	2	4

可变精度粗糙集模型[2]是在基本粗糙集模型的基础上引入了误差因子$\beta(0\leqslant\beta<0.5)$,即允许有一定程度的错误分类率存在,这有利于利用粗糙集理论从不相关的数据中发现相

*)基金项目:国家自然科学基金项目(60175016,60475019)。张东星 硕士生,主要研究方向:粗糙集理论、粒度计算;苗夺谦 教授,博士生师,主要研究方向:人工智能、模式识别、数据挖掘、粗糙集理论、主曲线。

关的数据关联。变精度粗糙集模型主要用于解决属性间无严格意义上的函数关系，或者存在概率上的不确定关系时的分类问题，它适用于含有少量错误规则的信息表(例如表1，其中 $C=\{a1,a2,a3,a4,a5\}$，$D=\{d\}$)。特别地，当 $\beta=0$ 时，可变精度粗糙集模型退化为基本粗糙集模型。

由粗糙集理论可知，对于已知的信息系统 $IS(U,C\cup D)$，条件属性 C 导出的划分为 $U/C=\{C_1,C_2,\cdots,C_m\}$，决策属性 D 导出的划分 $U/D=\{D_1,D_2,\cdots,D_n\}$。有如下定义：

定义 2.1[2] 对于 $\forall C_i\in U/C, D_j\in U/D$，其中 $i=1,\cdots,m, j=1,\cdots,n$，定义 D_j 相对于 C_i 的条件概率

$$P(D_j|C_i)=\frac{|D_j\cap C_i|}{|C_i|}$$

定义 2.2[2] 令 $\beta(0\leqslant\beta<0.5)$ 是依赖于数据中噪音程度的一个误差因子，则

(1)β 正域为 $POS_C^\beta(D)=\bigcup_{D_j\in U/D}\{\bigcup_{P(D_j|C_i)\geqslant 1-\beta}C_i\}$

(2)β 边界为 $BN_C^\beta(D)=\bigcup_{D_j\in U/D}\{\bigcup_{\beta<P(D_j|C_i)<1-\beta}C_i\}$

(3)β 负域为 $NEG_C^\beta(D)=\bigcup_{D_j\in U/D}\{\bigcup_{P(D_j|C_i)\leqslant\beta}C_i\}$

定义 2.3[2] 条件属性集 C 与决策属性集 D 之间的相关程度为

$$K_\beta(C,D)=\frac{|POS_C^\beta(D)\cup NEG_C^\beta(D)|}{|U|}$$

3 可变精度粗糙集模型在数据库系统上的实现

由文[1]，对 $a\in C$，若 $POS_{C-\{a\}}(D)=POS_C(D)$，那么属性 a 是 C 中相对于 D 不必要的，否则 a 是 C 中相对于 D 必要的。由此可见，要检验属性 a 是否是属性集合 C 中的不必要属性，需要比较 $POS_C(D)$ 和 $POS_{C-\{a\}}(D)$ 这两个集合。如果它们含有的元素相同，则可以断定 a 是不必要的，否则 a 是必要的。但是 SQL 语言并不擅长集合的比较操作，但它适合进行关系运算[5]。为了解决这个问题，文[3]提出了在数据库系统上检验不必要属性的方法，即对 $a\in C$，如果 $Card(\pi_{C\cup D-\{a\}}(IS))\neq Card(\pi_{C-\{a\}}(IS))$，那么属性 a 为必要属性，否则就是不必要的，其中 Card 是统计信息系统的总记录数，π_B 表示信息系统 IS 在属性集合 B 上的投影操作。基于上面的不等式，文[3]也提出了相应的计算核、约简的方法，这种方法适用于在数据库系统上的 SQL 操作。但是它只适用严格意义下的集合包含关系，对于带有噪音的数据集却无法正确地实现知识约简。为此，我们提出可变精度的粗糙集在数据库系统上的实现。由于可变精度模型只在原有的粗糙集模型下增加了一个误差因子 β，因此当 $\beta=0$ 时，它便适合于原有粗糙集模型上严格包含关系。因此，我们的方法比文[3]中的方法更具有灵活性。

可变精度粗糙集模型主要在消除错误对象集，又不损失有用信息的基础上达到知识约简的目的。我们所求的分类应该是在一定的误差水平 $\beta(0\leqslant\beta<0.5)$ 上计算的。

由定义 2.1、2.2，对于 $\forall o\in U$，属性集合 $B\subseteq C$，可以计算 $P([o]_D|[o]_B)=Card([o]_{B\cup D})/Card([o]_D)$，那么我们得到如下的定义：

定义 3.1 令 $\beta(0\leqslant\beta<0.5)$ 是依赖于数据中噪音程度的一个误差因子，则

(1)β 正域为：

$POS_B^\beta(D)=\{o\in U|Card([o]_{B\cup D})/Card([o]_B)\geqslant 1-\beta\}$

(2)β 边界为：

$BN_B^\beta(D)=\{o\in U|\beta<Card([o]_{B\cup D})/Card([o]_B)<1-\beta\}$

(3)β 负域为：

$NEG_B^\beta(D)=\{o\in U|Card([o]_{B\cup D})/Card([o]_B)\leqslant\beta\}$

对于同一个信息表，由于 U 是不变的，因此为了方便计算，我们有：

定义 3.2 条件属性集 $B(B\subseteq C)$ 与决策属性集 D 之间的相关程度为

$$K_\beta(B,D)=Card(POS_B^\beta(D))+Card(NEG_B^\beta(D))$$

由于 β 正域 $POS_B^\beta(D_j)$ 表示在误差水平为 β 的情况下，我们可以根据条件属性集合 B 的值而能够确定其能归入集合 D_j 的元素的集合。负域 $NEG_B^\beta(D_j)$ 表示在误差水平为 β 的情况下，我们可以根据条件属性集合 B 的值而能够确定其不能归入集合 D_j 的元素的集合。而我们不能确定边界部分 $BN_B^\beta(D_j)$ 的归入集合是哪个，因此我们用集合 $POS_B^\beta(D)\cup NEG_B^\beta(D)$ 来度量在误差水平为 β 下，条件属性集合 B 的分类能力。

为了便于理解，以表1为例给出误差因子为 0.10 时，对于 $B=\{a2,a3\}$ 对于 $D=\{d\}$ 求 $K_\beta(B,D)$ 的 SQL 语句如下：

```
Select Sum(count)
From 表 1 t
Where
    (select cast(sum(count) as float) from IS where a2=t. a2 and a3
    =t. a3 and d=t. d) / (select cast(sum(count) as float) from IS
    where a2=t. a2 and a3=t. a3) >=1-0. 10 OR
    (select cast(sum(count) as float) from IS where a2=t. a2 and a3
    =t. a3 and d=t. d) / (select cast(sum(count) as float) from IS
    where a2=t. a2 and a3=t. a3) <=0. 10
```

定义 3.3 已知信息系统 $IS(U,C\cup D)$，对于 β 是误差因子$(0\leqslant\beta<0.5)$，对 $\forall a\in C$，如果 $K_\beta(C,D)\neq K_\beta(C-\{a\},D)$，那么属性 a 是在误差因子 β 下属性集合 C 中的必要属性，否则 a 是不必要属性。

因此，为了求信息系统 $IS(U,C\cup D)$ 近似核 $Core_C^\beta(C)$，我们首先计算 $K_\beta(C,D)$，然后对于 $\forall a\in C$，计算 $K_\beta(C-\{a\},D)$，只需依次比较 $K_\beta(C,D)$ 和 $K_\beta(C-\{a\},D)$，若相等则 a 是核属性，否则在误差因子为 β 的情况下去掉属性 a，其分类能力不变。由表1可得，对于误差因子 $\beta=0.10$，可以使用类似上面的 SQL 语句计算，可得 $K_{0.10}(\{a1,a2,a3,a4,a5\},\{d\})=182$，而 $K_{0.10}(\{a1,a2,a3,a5\},\{d\})=136$，分类能力发生了改变，因此 $a4$ 相对于 $\{a1,a2,a3,a4,a5\}$ 是必要的，$a4$ 是核属性；$K_{0.10}(\{a2,a3,a4,a5\},\{d\})=182$，$a1$ 相对于 $\{a1,a2,a3,a4,a5\}$ 是不必要的，$a1$ 不是核属性。根据以上的算法依次判断其他条件属性，只需要执行 6 次 SQL 操作，然后比较，就可以求得表1的近似核 $\{a4\}$。

根据基本粗糙集求约简的算法，有属性重要性 $M_\beta(a,B,D)=K_\beta(B\cup\{a\},D)-K_\beta(B,D)(\forall a\in C-B)$。我们发现，对于属性集合 B，$K_\beta(B,D)$ 是不变的，那么衡量某一属性 a 的重要性，只需要使用 SQL 语句计算 $K_\beta(B\cup\{a\},D)$，再求出 $a_m=\arg\max_{a\in C-B}(K_\beta(B\cup\{a\},D))$，将 a_m 并入集合 B，如此继续增加最重要属性，直到 $K_\beta(B,D)=K_\beta(C,D)$，那么属性集合 B 就成为该信息系统的一个近似约简。

由于已经求得表1的近似核是 $\{a4\}$，令 $B=\{a4\}$，为了求得近似约简，我们需要依次计算 $K_{0.1}(\{a4,a1\},\{d\})=102$，$K_{0.1}(\{a4,a2\},\{d\})=182$，$K_{0.1}(\{a4,a3\},\{d\})=57$，$K_{0.1}(\{a4,a5\},\{d\})=128$，因此我们可以得到 $a_m=a2$，令 $B=\{a4,a2\}$，又因为 $K_{0.1}(\{a4,a2\},\{d\})=K_{0.10}(\{a1,a2,a3,a4,a5\},\{d\})=182$，所以 $\{a4,a2\}$ 是表1的近似约简。

同理，取不同的误差因子，对于表1我们可以依次得到表2。

表 2 取不同的误差因子得到不同的近似核和近似约简

误差因子 β	近似核	近似约简
0.00	a4,a5	a1,a4,a5
0.10	a4	a2,a4
0.43	a4	a4
0.45	∅	a2

由表 2 可知，β 取值不同，我们得到的近似约简也不相同，最后所得到的决策规则也不相同。如果 β 太大，可能会丢失一些信息，得不到准确的规则；如果 β 太小，约简结果中就有可能包含错误信息，得到包含有错误的决策规则，导致规则的不正确。因此在设定 β 值的时候，我们应该尽量减少或者避免不确定信息的存在，即要使 $\mathrm{Card}(BN_{\beta}(D))$ 尽量地小。

以上求得近似核和近似约简的结果表明，我们可以使用 SQL 语句实现在数据库系统中信息表的知识约简。此外，由于我们使用的可变精度模型，当 $\beta=0$ 时，它同样适用于严格包含关系下的粗糙集模型。因此，与文[3]中的方法相比，我们的方法更灵活。

4 实验结果与分析

选用一个含有 36 个条件属性、1 个决策属性的对象集（包含重复记录），然后依次取其中的 400，2000，4000，10000，20000，24000，30000，50000，60000 个对象作为测试集。这些测试集经过预处理后，可以得到使用基于数据库系统的约简方法和 Rosetta[5] 中基于普通文本的约简方法的运行效率的对比如图 1。

由图 1 可知，小数据集在文本文件上约简速度较快。但是随着对象数的增多，文本文件上的约简速度越来越比数据库系统上的约简慢。这是因为我们运用了数据库系统的高存取效率、适合大数据集的优点，而普通文本面临大规模数据，其构造等价类的操作越来越耗时。因此我们基于数据库系统上的粗糙集模型在大数据集的知识约简上更有效。

总结 本文使用了基于数据库系统的 SQL 语言来实现可变精度粗糙集理论的求近似核、近似约简的操作。由于大多数粗糙集模型的这些操作都是使用基于普通文件上的，很少和关系数据库系统结合起来，这就限制了粗糙集理论在数据挖掘中的应用。我们基于传统的粗糙集理论的主要思想，结合关系数据库提出了一系列的方案，使用 SQL 语言采用高效的存储过程，实现了可变精度的粗糙集模型。由于当误差因子 $\beta=0$ 时，可变精度粗糙集模型又可退化为基本粗糙集模型，因此基于数据库系统的可变精度粗糙集模型更具有灵活性。实验证明，该方法在处理大规模数据上效率更高，有一定的应用价值。

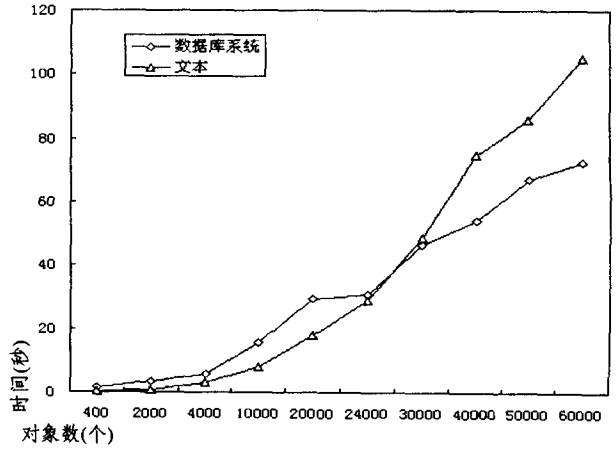

图 1 数据库系统与文本文件上的知识约简效率对比

参 考 文 献

1 Pawlak Z. Rough Sets . Theoretical Aspects of Reasoning about Data. Amsterdam: Kluwer Academic Publishers, 1991. 6～42

2 Ziarko W. Variable precision rough set model [J]. Journal of Computer and System Sciences, 1993, 46: 39～59

3 Hu X, Lin T Y. A New Rough Sets Model Based on Database Systems[A]. RSFDGrC, 2003. 114～121

4 Ullman J D. Principal of Database Systems[M]. Computer Science Press, 1980

5 Rosetta GUI V1. 4. 40. http://www. idi. ntnu. no/～ aleks/ rosetta/

（上接第 171 页）

2 Li DeYu, Zhang Bo, Yee Leung. On Knowledge reduction in inconsistent decision information systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004, 12 (5): 651～672

3 Kuroki N. Rough ideals in semigroups. Information Sciences, 1997, 100: 139～163

4 Davvaz B. Lower and upper approximations in Hv-groups. Ratio Math, 1999, 13: 71～86

5 Vakarelov D. A modal logic for similarity relations in Pawlak knowledge representation systems. Fundamenta Informaticae, 1991, 15: 61～79

6 Yao Y Y, Lin T Y. Generalization of rough sets using model logic. Intelligent Automation and Soft Computing, 1996, 2: 103～120

7 Yao Y Y. A comparative study of fuzzy sets and rough sets. Information Sciences, 1998, 109: 227～242

8 Wu Wei-Zhi, Zhang Wen-Xiu. Connections between rough set theory and Dempster-Shafer theory of evidence. International Journal of General Systems, 2002, 31(4): 405～430

9 Liang Ji-Ye, Shi Zhong-Zhi. The information entropy, rough entropy and knowledge granulation in rough set theory. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004, 12(1): 37～46

10 梁吉业，徐宗本，李月香. 包含度与粗糙集数据分析中的度量. 计算机学报, 2001, 21(12): 76～80

11 Andrzej C. Speaker-independent recognition of isolated words using rough sets. Information Sciences, 1998, 104: 3～14

12 Hamid M, Jerzy W G, James A B. Entropy of English text: experiments with humans and machine learning systems based on rough sets. Information Sciences, 1998, 104: 31～47

13 Ahn B S, Cho S S, Kim C Y. The integrated methodology of rough set theory and artificial neural network for business prediction. Expert Systems with application, 2000, 18(2): 65～74

14 Tsumoto S. Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Information Sciences, 2000, 124: 125～137

15 Hong T P, Wang T T, Wang S L, et al. Learning a coverage set of mximally general fuzzy rules by rough sets. Expert Systems with Application, 2000, 19: 97～103

16 Pawlak Z. Rough Sets. Theoretical Aspects of Reasoning about Data. Dordecht: Kluwer Academic Publishers, 1991

17 Grzymala-Busse J W. Algebraic properties of knowledge representation systems. In: Proc. of the ACM SIGART International Symposium on Methodologies for Intelligent Systems, Knoxville, 1986. 432～440

18 Grzymala-Busse J W, Sedelow W A Jr. On rough sets, and information system homomorphisms. Bulletin of the Polish Academy of Sciences Technical Sciences, 1988, 36(3-4): 233～239

19 Li De-Yu, Ma Yi-Cheng. Invariant characters of information systems under some homomorphism. Information Sciences: An International Journal, 2000, 129(1-4): 211～220