文章编号:1001 - 9081(2008)06 - 1435 - 03

基于约简重要性的最佳约简求解算法

江敬之
(同济大学 电子与信息工程学院,上海 201804)
(oryukiko@ hotmail. com)

摘 要:为解决多约简决策表的约简选取问题,在综合考虑约简中属性的平均重要性以及属性个数的基础上,提出了约简重要性的概念,并对其进行了详细论证。以此概念为基础介绍了最佳约简求解算法,即以此概念为准则对多个约简进行比较,可选出一个最佳约简。最后以实例验证了算法的实用性。

关键词:粗糙集;属性约简;属性重要性;相对必要属性;人工神经网络
中图分类号: TP311.13 **文献标志码**:A

Algorithm of finding the best reduction based on reduction significance

JIANG Jing-zhi
(*College of Electronic and Information Engineering, Tongji University, Shanghai* 201804, *China*)

Abstract: To solve the problem of choosing a best reduction from several reductions of one decision table, after thinking about the average significance of attributes in reduction and the number of attributes, a new definition named attribute significance was proposed and proved in detail. Then the algorithm of finding the best reduction was presented based on this new definition. Finally, examples demonstrate the usefulness of the algorithm.

Key words: rough set; attribute reduction; attribute significance; relative indispensable attribute; artificial neural network

0 引言

粗糙集理论是一种处理不精确、不完备数据的数学工具[1],其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或规则,已经在机器学习、数据库知识发现、医学诊断、故障诊断及决策支持等领域获得成功应用。

属性约简是粗糙集理论中的一个重要的课题,一般说来,知识库中的知识属性并不是同等重要的,而且还存在冗余,这不利于作出正确而简洁的决策,属性约简要求在保持知识库的分类和决策能力不变的条件下,删除不相关或不重要的属性。伴随着知识库所包含的数据量越来越庞大,我们已经不能够将属性约简作为数据预处理的最后一步了,这是因为通常情况下,一个知识库通常存在多个约简,我们最终要使用的只有一个,如何快速有效地从多个约简当中找到一个最佳约简,是一个亟待解决的问题。针对这个问题,本文提出了一个基于约简重要性的最佳约简求解方法。

1 粗糙集理论的相关概念

1.1 约简及相对必要属性

约简定义为不含多余属性并保证分类正确的最小条件属性集。一个决策表可能同时存在几个约简,这些约简的交集定义为决策表的核,核中的属性是影响分类的重要属性。

定义 1 若 RE 满足 $R_{RE} = R_C$(即 RE 与 C 在知识的发现上作用等价),且 $\forall d \in D, R_{D-\{d\}} \neq R_C$,则称 RE 为决策表 $T = (U, C \cup D, F)$ 的约简(其中 U 为论域,C 表示条件属性集,D 表示决策属性集,F 为映射规则,RE 是 C 的子集)。所有 $T = (U, C \cup D, F)$ 约简的交集称为 $T = (U, C \cup D, F)$ 的核心[2]。

通过决策表发现知识,主要是用属性来表达知识的分类。各种属性在表达知识分类中的作用是不同的。因此,有以下三种不同类型属性的定义。

定义 2 设决策表 $T = (U, C \cup D, F)$ 的所有约简为:$\{RE_i \mid RE_i$ 是约简,$i \in \tau\}$(τ 为约简的个数)。可将属性集 C 分为以下三部分:

1) 绝对必要属性(核心中的元素)b:$b \in \bigcap_{i \in \tau} RE_i$。

2) 相对必要属性 c:$c \in \bigcup_{i \in \tau} RE_i - \bigcap_{i \in \tau} RE_i$。

3) 绝对不必要属性 d:$d \in C - \bigcup_{i \in \tau} RE_i$[3]。

1.2 决策表中属性的重要性

定义 3 设给定一个决策表 $T = (U, C \cup D, F)$,其中,对 $\forall X \subseteq U$ 和独立于属性 $R(R \in C)$ 的论域 U 的划分 $\pi(U) = \{X_1, X_2, \cdots, X_M\}$,其中 $X_m (1 \leqslant m \leqslant M)$ 表示论域中第 m 个类别的集合,则定义属性 R 相对于类别 X_m 的重要性为:

$$sig_R(X_m) = \frac{|U - bn_R(X_m)|}{|U|} \tag{1}$$

其中,$bn_R(X_m)$ 表示属性 R 相对于类别集合 X_m 的下近似[4]。

2 基于约简重要性的最佳约简求解

2.1 求解最佳约简的准则:约简重要性

最佳约简求解算法的核心思想是通过比较约简的重要性来选取最佳约简,因此算法的关键部分是约简重要性的求解。约简的重要性通过综合考虑相对必要属性的重要性以及其个数两个因素得到的。

定义 4 若决策表最终需要被划分为 M 个类别,可知对于决策表中的属性 R,其相对于第 m 个类别的重要性为 $sig_R(X_m)$,则其相对于 M 个类别的重要性的平均值为:

$$\overline{SIG}(R) = \frac{\sum_{m=1}^{M} sig_R(X_m)}{M} \tag{2}$$

收稿日期:2007 - 12 - 05;**修回日期**:2008 - 02 - 01。

作者简介:江敬之(1984 -),女,河南新乡人,硕士研究生,主要研究方向:人工神经网络、数据挖掘、模式识别。

定理1 约简中存在两个影响分类性能的因素，分别是约简中属性的平均重要性以及约简中属性的个数。

1）约简中属性的平均重要性。这是由于约简是能够支持分类的总属性集的真子集，其中所包含的属性的平均质量（重要性）越高，分类的效果越好，因此约简中属性的平均重要性与约简的优劣成正比。

2）约简中属性的个数。属性的个数影响到数据的繁简，属性越多，数据量越大，规则也更难抽取，尤其是利用神经网络进行数据分类时，即使是一个属性的增减，也大大影响了网络结构的复杂程度，因此约简中属性的个数与约简的优劣成反比。

定义5 设给定一个决策表 $T=(U,A,F)$，在均等考虑约简所包含的属性平均重要性与属性个数两个因素的前提下，定义第 i 个约简的重要性为：

$$SIG(RE_i)=\frac{\sum_{n=1}^{N}(\overline{SIG}(n))}{N^2} \tag{3}$$

其中，$SIG(RE_i)$ 表示第 i 个约简的重要性，$\overline{SIG}(n)$ 表示属性 n 相对于 M 个类别的重要性的平均值，N 表示第 i 个约简中所包含的相对必要属性的个数。

证明 由定理1可以得到下面的公式：

$$SIG(RE_i)=k\cdot\frac{\sum_{n_1=1}^{N_1}(\overline{SIG}(n_1))+\sum_{n_2=1}^{N_2}(\overline{SIG}(n_2))}{N_1+N_2}\cdot\frac{1}{N_1+N_2} \tag{4}$$

其中，k 是一个常数系数，N_1 为约简 i 中核属性的个数，N_2 为约简 i 中相对必要属性的个数，$\overline{SIG}(n_1)$ 表示了约简中核心属性 n_1 相对于类别的重要性的均值，$\overline{SIG}(n_2)$ 表示了约简中相对必要属性 n_2 相对于类别的重要性的均值。

由于每个约简当中都包含了所有的核心属性以及各自的相对必要属性，因此，核心属性的相关部分是相同的，只有相对必要属性的相关部分是变化的。

又因为约简的优劣程度是通过比较约简重要性值的大小而体现的，所以可以去除式(4)中相对于其他约简的固定部分（即核心属性的相关部分），由此得到下面的公式：

$$SIG(RE_i)=k\cdot\frac{\sum_{n_2=1}^{N_2}(\overline{SIG}(n_2))}{N_2}\cdot\frac{1}{N_2} \tag{5}$$

1）当 $k>1$ 时，侧重考虑属性平均重要性的因素，即对分类结果的影响。

2）当 $k<1$ 时，侧重考虑属性个数的因素，即对数据量及分类过程的繁简程度的影响。

3）当 $k=1$ 时，平均考虑两个因素。因此这里选取 $k=1$，可以得到公式：

$$SIG(RE_i)=\frac{\sum_{n_2=1}^{N_2}(\overline{SIG}(n_2))}{N_2}\cdot\frac{1}{N_2} \tag{6}$$

以 N 替换 N_2，将分母进行合并，使表达更加简洁，即可得到式(3)。

2.2 基于约简重要性的最佳约简求解算法步骤

1）对决策表进行离散化[5]。

2）由差别矩阵约简算法求解决策表的约简及核。

3）计算出相对必要属性。所有约简所包含的相对必要属性为 $C=\left\{c\mid c\in\bigcup_{i\in\tau}D_i-\bigcap_{i\in\tau}D_i\right\}$，即约简的并集与核的差集。

4）求出各个约简与 C 的交集 LI_i，并统计 LI_i 含有的元素个数（其中 $i=1,2,\cdots,I$，I 为约简的个数），LI_i 即为第 i 个约简所包含的相对必要属性集。若存在第 j 个约简，满足 $LI_j=\varnothing(1<j<I)$，即此约简只包含核，则选取这个约简为最佳约简，算法结束，否则继续。

5）计算各个相对必要属性相对于每个类别的重要性。

6）计算各个约简的重要性，依据各个约简所对应的重要性来决定最佳约简。

7）输出最佳约简，算法结束。

3 实例应用及最佳性验证

3.1 实例步骤

实验所采用的数据是胜利油田临南地区一口油井的测井信息，选取12个典型层位的数据为样本集，其中SP、GR、RT、DEN、CNL、RWA分别代表自然电位、自然伽玛、电阻率、密度、补偿中子和地层水电阻率，原始数据表是未经离散化的，由于篇幅原因，这里直接列出离散化后的决策表T，其中条件属性集为：E={SP,GR,RT,DEN,CNL,RWA}，决策属性为：D。

1）对连续决策表进行离散化，得到新的决策表T。

表1 离散化后的决策表

No.	SP	GR	RT	DEN	CNL	RWA	D
1	0	0	2	3	1	2	0
2	0	0	1	3	1	3	0
3	0	0	3	3	1	2	0
4	2	2	0	2	0	3	1
5	1	2	0	2	0	2	1
6	1	2	0	2	0	3	1
7	1	0	2	2	1	1	2
8	3	0	2	3	1	3	2
9	3	1	1	3	2	3	2
10	2	1	3	1	1	0	3
11	2	3	3	2	1	0	3
12	1	0	3	2	1	2	3

2）按照差别矩阵的方法求解决策表的核与约简。

由差别矩阵的定义与性质可以得到决策表的核为：{SP,RT,RWA}。

约简有：{SP,GR,RT,DEN,RWA}，{SP,GR,RT,CNL,RWA}，{SP,RT,DEN,CNL,RWA}。

3）依据所得到的核与约简得到相对必要属性。

相对必要属性集为：{GR,DEN,CNL}。

4）各个约简所包含的相对必要属性集分别为：

$LI_1=\{GR,DEN\}$，$LI_2=\{GR,CNL\}$，$LI_3=\{DEN,CNL\}$。

5）依据1.3节中的式(1)计算各个相对必要属性相对于每个类别的重要性。

表2 相对必要属性相对于各个类别的重要性

属性	类别			
	W_1	W_2	W_3	W_4
GR	0.500	1.000	0.333	0.333
DEN	0.583	0.500	0.083	0.500
CNL	0.333	1.000	0.250	0.333

6）利用式(3)对各个约简的重要性进行计算。

$$SIG(RE_i)=SIG(\{SP,GR,RT,DEN,RWA\})=\sum_{n=1}^{2}\frac{\overline{SIG}(n)}{2^2}=0.220$$

易知其中 $n = 1$ 对应 GR,$n = 2$ 对应 DEN。

$SIG(RE_2) = 0.255$, $SIG(RE_3) = 0.224$

通过对于约简重要性的求解,比较各个约简的重要性,可以得知第二组约简,即{SP,GR,RT,CNL,RWA}为最佳约简。

3.2 最佳约简与其他约简相比较

以上面所求得的三个约简分别替代原决策表作为新决策表的属性集,在这里定义为 $T1$,$T2$,$T3$,分别以 $T1$,$T2$,$T3$ 的属性集作为网络输入,使用表中的 12 组数据对网络进行训练,直到训练自行结束。

3.2.1 使用 SOM 网络进行分类

对实验重复 1 000 次,将取得的 1 000 组实验结果求平均值,可以得到如表 3 所示的数据,E 为训练停止时所达到的误差,STEP 为训练停止时所进行的训练次数。

表 3　SOM 网络进行分类结果对比

属性集	E	STEP
$T1$	0.000 20	21
$T2$	0.000 11	17
$T3$	0.000 15	19

3.2.2 使用粗糙 BP 网络进行分类

对实验重复 1 000 次,将取得的 1 000 组实验结果求平均值,可以得到如表 4 所示的数据。

表 4　粗糙 BP 网络进行分类结果对比

属性集	E	STEP
$T1$	0.000 12	18
$T2$	0.000 06	14
$T3$	0.000 11	17

然后使用另外选取的 20 个典型层位输入到训练好的粗糙 BP 网络进行分类,其识别率分别为:80%,95%,80%。

从表中数据可以知道:1)相对于其他两个约简而言,使用最佳约简可以更准确更快速地抽取规则,进行分类;2)约简的重要性越高,其应用的性能也越好。

4　结语

以选择最佳约简为目的,本文提出了一种衡量约简优劣的准则,即约简的重要性。通过比较计算出的各个约简的重要性,来决定最佳约简,使得在存在多个约简的情况下可以快速有效地求得一个最佳的约简。实验结果也证明了最佳约简用于决策分类时优于其他约简,并可以得知约简的优劣程度与其重要性成正比。

参考文献:

[1] PAWLAK Z. Rough set theory and its application to data analysis [J]. Cybernetics and System, 1998, 29(7): 661 - 688.

[2] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学, 2001.

[3] 魏玲, 张文修. 粗糙集约简的闭算子算法[J]. 计算机科学, 2007, 34(1): 159 - 162.

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

[5] 苗夺谦. Rough Set 理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296 - 302.

[6] 谢冲, 刘美玲, 徐章艳. 一种基于粗糙集的知识约简方法[J]. 广西民族学院学报, 2005, 11(1): 67 - 70.

[7] 周海岩, 杨汀. 基于二进制可辨识矩阵的属性约简算法的改进[J]. 计算机工程与设计, 2003, 12(24): 35 - 37

[8] PAWLAK Z. Rough sets [J]. International Journal of Information and Computer science, 1982, 11(5): 341 - 356.

[9] PAWLAK Z. Vagueness and uncertainty: A rough set prospective [J]. International Journal of Computer Intelligence, 1995, 11(2): 37 - 41.

[10] 张文修, 吴伟志, 梁吉业, 李德玉. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.

[11] 代建华, 李元香—种基于粗糙集的决策系统属性约简算法[J]. 小型微型计算机系统, 2003, 24(3): 523 - 526.

[12] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems [C]// Intelligent Decision Support-handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publisher, 1991: 331 - 362.

[13] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases [C]// Proceedings of the ACM SIGMOD Conference on Management of data. [S. l.]: ACM Press, 1993: 207 - 216.

(上接第 1434 页)

[11] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases [C]// Proceedings 4th International Conference on Foundations of Data Organizations and Algorithms (FODO), LNCS 730. Berlin: Springer-Verlag, 1993: 69 - 84.

[12] YI B K, FALOUTSOS C. Fast time sequence indexing for arbitrary Lp norms [C]// Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000). Cairo: Morgan Kaufmann Publishers, 2000: 385 - 394.

[13] KORN F, JAGADISH H, FALOUTSOS C. Efficiently supporting ad hoc queries in large datasets of time sequences [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Tucson: ACM Press, 1997: 289 - 300.

[14] LIN J, KEOGH E. Group SAX: Extending the notion of contrast sets to time series and multimedia data [C]// Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin: [s. n.], 2006: 284 - 296.

[15] LIN J, KEOGH M, LI WEI, *et al*. Experiencing SAX: A novel symbolic representation of time series [J]. Journal of Data Mining and Knowledge Discovery, 2007, 15(2): 107 - 144.

[16] KEOGH E, CHAKRABARTI K, PAZZANI M, *et al*. Dimensionality reduction for fast similarity search in large time series databases [J]. Journal of Knowledge and Information Systems, 2001, 3(3): 263 - 286.

[17] VLACHOS M, KOLLIOS G, GUNOPULOS D. Discovering similar multidimensional trajectories [C]// Proceedings of the 18th International Conference on Data Engineering. San Jose, CA: [s. n.], 2002: 673 - 684.

[18] LEE S, CHUN S, KIM D. Similarity search for multidimensional data sequences [C]// Proceedings of the 16th International Conference on Data Engineering. Washington: IEEE Computer Society, 2000: 599 - 608.

[19] KEOGH E J, KASETTY S. On the need for time series data mining benchmarks: A survey and empirical demonstration [J]. Journal of Data Mining and Knowledge Discovery, 2003, 7(4): 349 - 371.