文章编号:1001 - 9081(2009)02 - 0419 - 02

基于层次聚类的差异化属性约简算法

汤周文,叶东毅
(福州大学 数学与计算机科学学院,福州 350002)
(tangzhouwen@ gmail. com)

摘 要:属性约简是粗糙集用于数据分析的一个重要概念,提出了一个计算差异化属性约简的算法。利用自底向上的聚合层次聚类方法对决策表的条件属性集进行聚类,得到条件属性集的 k 个划分,然后对这 k 个属性子集进行后处理操作而得到 k 个有较大差异的约简属性集。实验结果表明了算法的有效性。

关键词:属性约简;差异;层次聚类
中图分类号: TP311.13;TP182 **文献标志码**:A

Dissimilar attribute reductions based on hierarchical clustering method

TANG Zhou-wen, YE Dong-yi
(*College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian* 350002, *China*)

Abstract: Attribute reduction is an important concept in rough set data analysis. An algorithm for dissimilar reductions finding was presented. Method of bottom-up agglomerative hierarchical clustering was used to get k partitions of conditional attributes set, and then these k attribute sets were dealt with by post-processing and get k dissimilar attribute reductions. Result of experiments show that our method is effective.

Key words: attribute reduction; dissimilar; hierarchical clustering

0 引言

粗糙集理论自 1982 年由 Pawlak 教授提出以来,得到了广泛的应用。决策表的属性约简是粗糙集理论的一个重要应用,关于决策表的属性约简问题,人们提出了各种各样的约简算法及其改进[1-4]。但是他们要么求解单个属性约简,要么求解所有的属性约简。由单个属性约简得到的规则所包含的条件属性比较单一;而求解所有属性约简在计算上是 NP 难的,没有多大的实际应用价值。如果能够求出若干个属性集合有较大差异的约简,同时在计算上又不是 NP 难的,则既有代表性,又有一定的知识覆盖面[8]。为此,文献[8]提出了希望能够在非 NP 难的计算复杂度下求解差异化属性约简问题。本文就这个问题进一步研究分析,并提出了基于聚合层次聚类的求解算法。同时,利用 UCI 上的数据对本文算法进行实验与分析,效果良好。

1 基本概念

考虑一个信息系统:$L = (U, Q, V_q, F_q)$,$q \in Q$,其中,$U = \{x_1, \cdots, x_n\}$ 是论域,Q 是属性集,V_q 为属性取值的集合,F_q 是 $U \times Q \to V_q$ 的映射[5]。在决策分析时,通常把属性集 Q 分为条件属性集 C 与决策属性集 D,这时信息系统称为一个决策表。在本文中,记决策表的条件属性个数为 $|C|$,记录的条数为 $|U|$。条件属性子集 AS 中属性的个数为 $num(AS)$[8],两个条件属性集 a 和 b 的交集记为 $common\_element(a,b)$。

定义 1 如果属性集 $P \subseteq Q$ 满足 $IND(P) = IND(Q)$,并且 $\forall a \in P$, $IND(P - \{a\}) \neq IND(P)$。则称 P 为 Q 的一个约简,其中 $IND(P)$ 为属性集 P 导出的等价关系[5]。

聚类是一种比较常见的数据分析工具,其作用是把大量数据分成若干类,使得每个类中的数据之间最大限度地相似,而不同类中的数据最大限度地不同[6]。层次聚类算法通过将数据组织成若干组并形成一个相应的树状图来进行聚类,它又可以分为两类,即自底向上的聚合层次聚类和自顶向下的分解层次聚类。聚合聚类的策略是先将每个对象各自作为一个原子聚类,然后对这些原子聚类逐层进行聚合,直至满足一定的终止条件[6]。

聚合层次聚类方法的基本步骤如下[7]:

1) 将每个对象归为一组,共得到 N 组,每组仅包含一个对象。组与组之间的距离就是它们所包含的对象之间的距离。

2) 将最近的两个组合并成一组。

3) 重新计算新的组与所有旧组之间的距离。

重复第 2 步和第 3 步,直到最后合并成一个组为止(此组包含了 N 个对象)。

如果要利用聚合层次聚类方法得到有 k 个分类的聚类,只要在剩下 k 组时结束算法即可。

2 算法描述

为求 k 个有较大差异的约简属性集,我们可以考虑适当定义属性集之间的距离,先利用聚类方法把这些条件属性聚成 k 个有较大差异的属性集,然后对这 k 个属性集进行后处理而得到 k 个有较大差异的决策表约简。在本文中,考虑对条件属性进行 k 划分,把区分能力相近的属性放在同一个属性集中。两个属性集的区分相近程度定义为决策表的所有记录中依据它们同时属于正区域或者负区域的记录个数,个数越多则越相近。

收稿日期:2008 - 09 - 02;**修回日期**:2008 - 10 - 15。
基金项目:国家自然科学基金资助项目(60602052);福建省自然科学基金资助项目(A0710006)。
作者简介:汤周文(1984 -),男,福建莆田人,硕士研究生,主要研究方向:智能技术; 叶东毅(1964 -),男,福建南安人,教授,博士,主要研究方向:计算智能和最优化算法。

定义2 对于一个决策表 S 和它的两个条件属性子集 $A1$,$A2$,S 中依据 $A1$ 和 $A2$ 同时属于正区域或者负区域的记录个数称为这两个属性集的区分相似度。

定义3 对于一个由 n 个属性集组成的集合,如果有一个 $n \times n$ 的矩阵 $\boldsymbol{M}$,其中 M_{ij} 为第 i 个属性集和第 j 个属性集的区分相似度,则矩阵 $\boldsymbol{M}$ 称为这 n 个属性集的区分相似度矩阵。

算法1 基于聚合层次聚类求解决策表条件属性集 k 个划分算法。

输入:决策表 $S = \langle U, Q, V_q, F_q \rangle$,正整数 k,其中 $Q = C \cup D$,C、D 分别为条件和决策属性集。

输出:条件属性集的 k 个划分 $k - SA$。

第1步 把整个条件属性集 C 分解成 $|C|$ 个属性集,每个属性 $a \in C$ 组成一个属性集。

第2步 如果 $|C| \leqslant k$,则依照各个属性在决策表中出现的顺序依序循环复制已有的 $|C|$ 个属性集直到有 k 个属性集为止,转第5步。否则,计算这些属性集对决策表中各个记录被确定为正负区域的情况。并求出它们的区分相似度矩阵。

第3步 如果剩下 k 个属性集,转第5步。否则,根据区分相似度矩阵,找出两个区分相似度最大的属性集,把这两个属性集中的元素合并成一个新属性集,删掉这两个旧属性集。

第4步 计算对决策表中各个记录被这个新属性集确定为正负区域的情况,同时据此更新区分相似度矩阵。转第3步。

第5步 输出这 k 个属性集划分 $k - SA$,结束。

算法1不能保证最后得到的 k 个属性集都是决策表的约简,因此需要对这 k 个属性集进行后处理。在属性的后处理问题上,我们的目标是要得到属性集有尽量大差异的约简属性集,文献[8]提出了一个基于统计方法的属性集后处理算法,本文也采用这个属性集后处理算法,初始时,各个属性被使用过的次数置1;然后对这 k 个属性集都进行后处理。

算法2 属性集后处理算法[8]:

输入:决策表 $S = \langle U, Q, V_q, F_q \rangle$,其中 $Q = C \cup D$,C、D 分别为条件和决策属性集,一个属性集 $AS \subseteq C$。

输出:约简属性集 RED

第1步 求 C 相对于 D 的核属性集 $CORE$。

第2步 把核属性 $CORE$ 加入属性集 AS 得到属性集 RED,因为核属性是必须的。

第3步 把由 $a \in C - RED$ 的属性组成的属性集按照属性被用过的次数从小到大排序,得到有序属性集 R。

第4步 把属性 $a \in R$ 依序逐个加入到 RED 中,直到 RED 导出的等价关系和整个条件属性集 C 所导出的等价关系一样为止。

第5步 新加入的属性,属性被用过次数加1。

第6步 对 RED 进行冗余属性剔除。先判断第4步加入的各个属性是否是 RED 的冗余属性,如果是,则把它从 RED 中剔除,然后再判断其他的非核属性。

第7步 被剔除的属性,属性被用过的次数减1。

第8步 最后得到的 RED 就是决策表 S 的一个约简。

在算法1中,计算条件属性集 AS 相对 D 对决策表中的各个记录确定属于正区域还是负区域的时间复杂度为 $O(num(AS) * |U| \text{ lb } |U|)$,$num(AS)$ 的平均大小约 $|C|/k$,共计算了 $|C| + |C| - k$ 次。求区分相似度矩阵的时间复杂度为 $O(|C|^2 * |U|)$,所以算法1的时间复杂度为 $O(|C|^2/k * |U| \text{ lb } |U|)$,算法2的时间复杂度为 $O(|C|^2 * |U| \text{lb } |U|)$[8],所以基于聚合层次聚类方法求解 k 个差异化属性约简的总的时间复杂度为 $O(|C|^2 * k * |U| \text{lb } |U|)$。

3 算法实验与分析

为了测试算法效果,我们采用 UCI 数据集(http://mlearn.ics.uci.edu/databases/)进行实验。为表示上方便,实验之前对数据表进行预处理:把决策属性放在最后,把条件属性按照本来数据表中的先后顺序更名为 $F1 \sim Fm$(设有 m 个)[8]。

仿真实验系统使用VC++6.0开发实现。在实验过程中,选取几个约简属性集相对多样的数据集如表1所示,令 $k = 5$,求出的 k 个有较大差异的约简属性集如表2所示。

为评价约简属性集之间的相似程度,引入一个相似度评价函数 $similar(a,b)$ 来评价两个约简属性集 a 和 b 之间的属性相似程度,这个值越大越相似。同时定义 $k(k > 1)$ 个约简属性集的总的属性相似度 $total\_similar(reductions)$ 为这些属性集中两两之间的属性相似度的平均值。图1是所选的几个数据集在 k 从1到10时本文算法得到的约简属性集相似度的走势曲线图。

$$similar(a,b) = \frac{num(common\_element(a,b))}{maximal(num(a),num(b))} \tag{1}$$

$$total\_similar(reductions) = \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} similar(reductions_i, reductions_j)}{k*(k-1)/2} \tag{2}$$

其中,k 为约简属性集的个数。$reductions_i$ 为第 i 个约简属性集。

表1 实验选取的数据集信息

数据集编号	数据集名	样本个数	条件属性个数	核属性	类别个数
1	housing	506	13	无	229
2	lung-cancer	32	56	无	3
3	Lymphography	148	18	无	4
4	Soybean-small	47	25	无	4
5	Sponge	76	45	F28,F36,F39,F40	76

表2 本文算法实验结果

数据集号	本文算法得到的约简属性集	相似度/%
1	(1,5),(1,6),(5,6,7),(7,8),(3,12,13)	15.00
2	(34,35,42,44,49,53,55,56),(23,24,26,27,28,29,45),(20,22,30,31,36,52,54),(3,25,32,38,39,41,46),(2,6,15,21,37,51)	0.00
3	(1,11,12,13,14,16,18),(1,2,5,6,10,11,12,13,15),(2,5,6,8,10,15,16,17,18),(3,8,11,12,13,14,15),(1,3,6,8,10,14,16,17,18)	42.38
4	(2,5,20,22),(21,22),(1,3,4,35),(12,25,27,28),(8,9,10,23,24,35)	4.17
5	(1,15,19,28,29,32,36,39,40),(5,9,28,32,33,36,39,40),(7,10,16,18,28,31,35,36,39,40),(4,26,28,30,35,36,39,40,43),(4,26,28,33,35,36,39,40)	50.92

(下转第423页)

整 strip 单元大小,即条纹深度。

现有磁盘大多采用分区技术[7],靠近磁盘轴心的近区,扇区数较远区少,数据量小,N 个磁盘采用较小条纹布局,当数据块需要跨盘操作时,式(4)中的寻道时间、延迟时间以及同步时间均会降低,具有较短的响应时间,进而提高 IOPS,对于大块数据以及顺序访问,保证了多块磁盘的并发,可以提高系统带宽;在远区磁道中,扇区数较多,采用较大条纹布局,寻道、旋转等待时间都会相应降低,可以更好地提高响应速度,不会过于降低系统的 IOPS,由于模型仍然采用条纹布局方式,在顺序访问操作时,也不会降低吞吐量;而根据磁盘的具体磁道的扇区配置,可以将多个磁道组合为一个条纹单元,增加条纹大小,进一步增加 IOPS。

4　实验结果

采用广泛应用于磁盘阵列的仿真工具——RAIDframe[8]进行仿真实验。对该软件包进行了改进,修改条纹划分算法,改变布局方式,使其支持本文设计的粗粒度条纹布局模型;采用写穿策略,以随机写为主要访问方式,减小 Cache 对实验性能的影响。

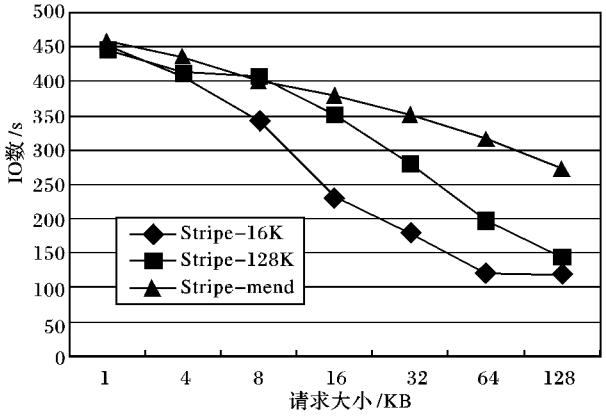

图 2　不同数据块请求下的性能比较

实验结果如图 2 所示。粗粒度条纹布局模型,针对随机的小块数据访问,可以有效提高 IOPS;随着访问数据不断增大,I/O 性能下降幅度不大,比现有布局模式有较好的 I/O 并发效果。

5　结语

本文针对 RAID 条纹布局在随机访问的不同数据块大小下的性能做了整体分析,总结出条纹的增大对 IOPS 性能的影响会逐渐降低,根据磁盘的磁道分布思想提出了粗粒度条纹布局模型。通过实验仿真,验证了模型在小块随机数据请求下具有良好性能,可以减缓 IOPS 的下降趋势。在仿真实验中没有采用缓存,如果为模型设计一种适合的缓存方式,可以使模型的整体性能得以更加显著的提高。

参考文献:

[1] PATTERSON D A, GIBSON G, KATZ R H. A case for redundant arrays of inexpensive Disks(RAID) [C]// Proceedings of ACM SIGMOD. New York: ACM, 1988: 109 - 116.

[2] RUEMMLER C, WILKES J. An introduction to disk drive modeling [J]. IEEE Computer, 1994, 27(3): 17 - 29.

[3] 王芳,张江陵,冯丹. RAID 的并行 I/O 调度算法分析[J]. 计算机工程与科学,2003,25(3):3 - 4,30.

[4] 骆新国,张江陵. 一种 RAID 评价新方法[J]. 华中理工大学学报,1994,25(10):98 - 100.

[5] 周可,张江陵,冯丹. 带 Cache 的磁盘阵列 I/O 响应时间及吞吐量分析[J]. 微电子学与计算机,2003,20(8):66 - 68.

[6] 谢长生,刘艳,李怀阳,等. 基于分条单元的 RAID 数据分布优化[J]. 计算机科学, 2006,33(3): 275 - 278.

[7] METER R V. Observing the effects of multi-zone disks[C]// Proceedings of the annual conference on USENIX Annual Technical Conference. Berkeley: USENIX Association, 1997: 19 - 30.

[8] COURTRIGHT II W V, GIBSON G, HOLLAND M, *et al*. RAIDframe: Rapid Prototyping for Disk Arrays[EB/OL]. [2008 - 06 - 10]. http://www.pdl.cmu.edu/PDL-FTP/RAID/Sigmetrics96.pdf.

(上接第 420 页)

其中,数字 i 表示 Fi 属性。为使得结果更直观,属性按 Fi 中的 i 的大小从小到大排好序。

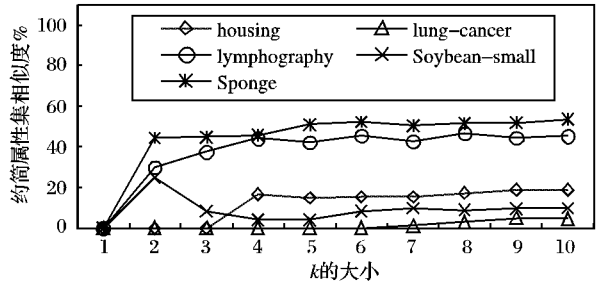

图 1　本文算法对表 1 选取的数据表的实验效果

通过实验发现,本文的算法在 k 较大时,得到的约简属性集的相似度趋于稳定。在条件属性集的个数比较多而各个约简所用到的属性个数比较少的情况下,得到的效果较佳。在决策表有较多核属性或者约简属性集的属性个数比较多的情况下效果较差,而这种情况下,各个约简之间的相似度本来就有比较大的相似度。总体上说,本文的算法在计算差异化属性约简问题上可以得到一个较好的效果。

4　结语

本文针对文献[8]提出的如何计算差异化属性约简问题,提出了一个计算方法并实现了它。把层次聚类的思想应用到差异化属性约简的计算过程中。先对决策表的条件属性集按定义好的距离进行聚类,得到条件属性集的若干个划分;接着,对这些属性子集进行后处理操作,从而得到若干个有较大差异的属性约简集。最后,为说明算法的效果,利用 UCI 上的数据集对本文算法进行实验,效果良好。

参考文献:

[1] HU XIAOHUA , CERCONE N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323 - 337.

[2] 王珏. Rough sets 约简与数据浓缩[J]. 高技术通讯, 1997, 7(11) : 40 - 45.

[3] 叶东毅. 属性约简算法的一个改进[J]. 电子学报, 2000, 28(12) : 81 - 82.

[4] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 - 766.

[5] ZIARKO J. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993 , 46(1): 39 - 59.

[6] 贺玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述[J] 计算机应用研究, 2007, 24(1): 10 - 13.

[7] TAN PANG-NING, STEINBACH M, KUMAR V. Introduction to data mining[M]. New Jersey : Addison-Wesley. 2006.

[8] 汤周文,叶东毅. 差异化属性约简计算的两个算法[J]. 计算机科学, 2008, 35(8A): 37 - 40.