

一种基于粗糙集属性频度约简算法的改进

刘飞¹,孔媛媛¹,杨习贝²

(1.连云港职业技术学院,江苏连云港 222006;

2.南京理工大学计算机科学与技术学院,江苏南京 210094)

摘要:为了获得有效的属性最小相对约简,在基于属性频度的启发式约简算法的基础上,提出了一种同时满足属性重要性和频度改进的启发式约简算法。该算法的基本思想是:以属性的核为基础,以频度作为选择属性的启发信息,即把属性频度最大的属性添加到核属性中,这样就把分类能力较强的属性添加到约简集合中,从而能够获得较优的约简。

关键词:粗糙集;属性约简;属性重要性;属性频度;约简算法

中图分类号:TP18;TP301.6

文献标识码:A

文章编号:1673-629X(2008)12-0095-03

An Improvement of Reduct Algorithm Based on Rough Set of Attributes Frequency

LIU Fei¹, KONG Yuan-yuan¹, YANG Xi-bei²

(1. Lianyungang Technical College, Lianyungang 222006, China;

2. School of Computer Sci. and Techn., Nanjing Univ. of Science and Technology, Nanjing 210094, China)

Abstract: To obtain the minimal relative reducts of effective attributes, from the viewpoint of heuristic reduct algorithm based on attributes' frequency, propose a heuristic reduct algorithm, which satisfies both attributes' importance and amelioration of frequency. The main idea of algorithm is: the core of attributes is considered as the basis, the frequency is considered as the heuristic information for selecting attributes and then add the attributes with maximal frequency into the core attributes, from which the attributes with better ability for classification purpose can be joined the reducts, such reducts are preferable.

Key words: rough set; attributes' reduct; attributes' importance; attributes' frequency; reduct algorithm

0 引言

粗糙集(rough set)^[1]理论是由波兰科学家 Pawlak 提出的一种处理不精确和不确定性问题的新型数学工具。主要思想是在保持分类能力不变的前提下,通过属性约简,导出问题的决策或分类规则。应用粗糙集理论处理不确定性问题的最显著特点是不需要提供问题所需处理的数据集合之外的任何先验信息。属性集的约简(Attribute Reduct)是粗糙理论中关键的问题之一。所谓约简是属性集的子集,它与原属性集具有同样的分辨能力。约简反映了一个信息系统的本质信息,求解一个信息系统的全部约简或计算最佳约简都是 NP-hard 难题。因此在有限时间内求出尽可能好

的约简是个启发式搜寻问题,也成为一些学者的研究重点^[2-4],文中对前人提出的算法进行了分析和试验,做了一些算法改进,提出一种同时满足属性重要性和频度的启发式约简算法。

1 基本概念

定义1 给定一决策表 $T = (U, A)$, 其中属性集合 $A = C \cup D$, 且条件属性集 C 的个数 N , D 为决策属性集。

定义2 $R \subset C$, 对于任意属性 $a \in C - R$ 的重要性定义如下:

$$SGF(a, R, D) = k(R \cup a, D) - k(R, D)$$

$k(R, D) = \text{card}(\text{pos}_R(D)) / \text{card}(\text{pos}_C(D))$, 并且 $\text{card}(X)$ 表示 X 集合的基数, $\text{pos}_R(D)$ 是 D 的 R 正域。

定义3 设 U 为一个论域, P 和 Q 是 U 上的两个等价关系族。设 P 和 Q 在 U 上导出的划分分别为 X 和 Y , 则 P 和 Q 在 U 上的子集组成的 σ 代数上的概率分布为:

收稿日期:2008-04-14

基金项目:国家自然科学基金(60472060, 60572034);江苏省自然科学基金(BK2006081)

作者简介:刘飞(1971-),男,讲师,硕士,主要研究方向为智能信息处理。

$$(X:P) = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ P(X_1) & P(X_2) & \cdots & P(X_n) \end{bmatrix}$$

$$(Y:P) = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ P(Y_1) & P(Y_2) & \cdots & P(Y_m) \end{bmatrix}$$

其中: $P(X_i) = \frac{|X_i|}{|U|}, i = 1, 2, \dots, n; P(Y_j) = \frac{|Y_j|}{|U|}, j = 1, 2, \dots, m$ 。

定义 4 知识(属性集合) P 的熵 $H(P)$ 定义为:

$$H(P) = - \sum_{i=1}^n P(X_i) \log(P(X_i))$$

定义 5 设决策表 T 的差别矩阵为 M , 以属性在差别矩阵中出现的次数 $p(a)$ 作为该属性重要性的度量, 则有 $SGH(a, R, D) = p(a)$, 由上面的定义可知, $SGH(a, R, D)$ 的值越大, 则说明在已知 R 的条件下, 属性 a 对于 D 就越重要。

2 基于属性频度的约简算法

2.1 属性频度约简算法

这里要讨论的启发式算法是基于属性频度的约简算法^[5], 目前为止比较常用的属性频度算法就是以决策表对应的差别矩阵为基础, 通过考虑其中的全体属性组合, 计算出所有非核属性在差别矩阵中出现的次数, 若某个非核属性的次数最大, 说明这个属性在辨别两个决策不同的对象中所起的作用比较大, 所以应首先考虑把它加入到约简集合中, 同时把包含这个属性的所有属性组合删除。给定一决策表 $T = (U, R, V, f)$, 其中, U 是论域, $R = C \cup D$, C, D 分别是条件属性集合和决策属性集合, 其中 $C = \{c_1, c_2, \dots, c_n\}$ 是 T 中所有条件属性。该决策表 T 所对应的差别矩阵为 M , M 中的元素表示为 A_{ij} , 称为其中的一项, 它代表着决策表中能区别第 i 个实例与第 j 个实例的所有属性的集合, 因此可以知道 $c_k \in A_{ij}, A_{ij} \in C$ 。另外, 可以假设 $P(c_k)$ 为属性 c_k 在 M 中的属性频度函数, 它就是前面提到的属性 c_k 在差别矩阵 M 中出现的次数, 因此 $p(c_k)$ 就可以定义为属性 c_k 的重要性, 在属性约简的过程中作为一种度量。下面的算法就是目前比较常用的属性频度算法^[5]。

具体描述如下:

输入: 给定一决策表 $T = (U, R, V, f)$, 其中, U 是论域, $R = C \cup D$, C, D 分别是条件属性集合和决策属性集合

输出: 决策表 T 的一个约简 B

Step1: 找出核属性 C_0 , 并初始化 $B: C_0 \rightarrow B$

Step2: 剔除 M 中所有与 B 的交集不为空的元素, 并且从条件属性集合中去掉 B 中所包含的元素, 即:

$M - Q \rightarrow M, C - B \rightarrow P$, 其中 $Q = \{A_{ij} | A_{ij} \cap B \neq \emptyset\}$

Step3: 对所有 $c_k \in P$, 计算在 M 中的属性频度函数 $p(c_k)$, 且从所有的 $p(c_k)$ 找出具有最大值的 c_q , 即 $p(c_q) = \max\{p(c_k)\}$ 。

Step4: 把 c_q 添加到约简集合中: $B \cup c_q \rightarrow B$ 。

Step5: 重复以上步骤直到 $M = \emptyset$ 。

上面的属性频度算法是按照属性的频度来决定将哪个属性添加到核属性中, 当某个属性在决策表对应的差别矩阵中出现的次数最多时, 就认为这个属性相对于决策是较重要的。但是在实际的应用当中会出现达到最大频度的属性不止一个, 对于这种情形上述算法从中任意取一个添加到核属性中。因而, 这样得到的约简有可能不是决策表的较优约简。

2.2 改进的属性频度约简算法

给定一决策表 $T = (U, R, V, f)$, 其中, U 是论域, $R = C \cup D$, C, D 分别是条件属性集合和决策属性集合。设 B 是要获得的一个约简, $POS_B(D)$ 是根据分类 U/R 的信息可以准确划分到决策属性的等价类中去的对象的集合, 而 $POS_B(D)/IND(B, D)$ 就是等价关系 $IND(B, D)$ 下的划分。改进后的算法的具体的思路就是: 当达到最大属性频度的属性存在多个时, 再对这些频度相等的属性进行一次处理, 即当选择把哪个属性频度最大的属性添加到核属性中时, 计算它们的 $POS_{B \cup \{c_k\}}(D)/IND(B, D)$ 中的元素个数, 然后从中选择包含元素个数最多的那个属性添加到核属性中, 这样就是把分类能力比较强的属性添加到了约简集合中, 从而能够获得较优的约简。

下面给出改进算法:

输入: 给定一决策表 $T = (U, R, V, f)$, 其中, U 是论域, $R = C \cup D$, C, D 分别是条件属性集合和决策属性集合

输出: 决策表的一个约简 B

Step1: 找出核属性 C_0 , 并初始化 $B: C_0 \rightarrow B$

Step2: 剔除 M 中所有与 B 的交集不为空的元素, 并且从条件属性集合中去掉 B 中所包含的元素, 即: $M - Q \rightarrow M, C - B \rightarrow P$, 其中 $Q = \{A_{ij} | A_{ij} \cap B \neq \emptyset\}$

Step3:

(1) 对所有 $c_k \in P$, 计算在 M 中的属性频度函数 $p(c_k)$;

(2) 从所有的 $p(c_k)$ 找出具有最大值的 c_q , 如果最大频度的属性只有一个, 则 $p(c_q) = \max\{p(c_k)\}$ 。把 c_q 添加到约简集合中: $B \cup c_q \rightarrow B$, 然后转 step4; 否则, 如具有最大频度属性有多个时转到(3);

(3) 计算属性频度最大的各个属性的 $POS_{B \cup \{c_k\}}(D)/IND(B, D)$ 中包含的元素个数, 然后把

$POS_{B \cup c_q}(D)/IND(B, D)$ 含有元素个数最多的属性 c_q 添加到约简集合中： $B \cup c_q \rightarrow B$ 。

Step4:重复以上步骤直到 $M \neq \emptyset$ 。

3 实验分析

为了验证改进后的算法的有效性,本节通过 UCI^[6]中的一个信息系统来进行测试分析。

实验总共分为三部分:

第一部分是,将数据集用传统的属性频度算法进行约简,即当遇到有多个属性频度同时达到最大时,采用的方式是随机选择一个添加到约简集合中,其中这里随机选择的策略是利用系统时间作为随机种子的随机数实现的;

第二部分是,用改进后的属性频度算法对所选择的数据集进行约简处理。即当遇到有多个属性频度同时达到最大时,再次对这些频度最大的属性进行处理,从中找出分类能力最强的一个添加到约简集合中。

第三部分:根据前两部分实验结果进行比较分析,最终得出结论。

采用的是一组医疗数据作为实验数据集,共有 700 条,这组数据是用来判断病人的肿瘤是良性的还是恶性的。其中条件属性和决策属性(class)如表 1 所示^[5]。

表 1 一组医疗数据的属性列表

属性编号	属性名称	属性取值
A1	Clump- Thickness	int[1,10]
A2	Cell- Size- Uniformity	int[1,10]
A3	Cell- Shape- Uniformity	int[1,10]
A4	Marginal- Adhesion	int[1,10]
A5	Singl- Epi- Cell- Size	int[1,10]
A6	Bare- Nuclei	int[1,10]
A7	Bland- Chromatin	int[1,10]
A8	Normal- Nucleoli	int[1,10]
A9	Mitoses	int[1,10]
A10	Class	{benign,malignant}

运用两种算法分别对该测试数据集进行试验得到的结果如表 2 所示。

表 2 两种算法得到的约简集合

算法名称	约简结果
原属性频度算法	{A1,A3,A5,A6} {A3,A5,A6,A7}
改进后算法	{A3,A5,A6,A7}

根据约简集合 {A1,A3,A5,A6} 和 {A3,A5,A6,A7} 对原始数据进行约简,最后的约简对比如表 3 所示。

表 3 两种约简集合下约简结果对比

约简集合	属性约简率	约简后对象个数	数据对象约简率
{A1,A3,A5,A6}	55.6%	317	54.1%
{A3,A5,A6,A7}	55.6%	279	61.4%

很显然,集合 {A3,A5,A6,A7} 要比集合 {A1,A3,A5,A6} 约简效率高,所以改进后的算法在一定程度上提高了约简率,得到数据集的较优约简。

4 结束语

在分析原有算法(基于属性频度的约简算法)的基础上,提出一种改进算法,该算法在用属性频度作为启发信息的同时,又把其中多个属性频度一样的属性进行再次分析处理,把分类能力强的属性添加到约简集合中,使约简效率提高。目前粗糙集的应用越来越广泛,通过实践证明把粗糙集理论运用在属性约简上是十分有效的,但是目前粗糙集理论还有一些待解决的问题、待提高的方面:

1)首先粗糙集还不能处理连续属性,只有将连续属性离散化后才能用粗糙集的知识进行属性约简。

2)其次现在在知识发现中还缺乏高效的知识约简算法,在存储空间和运行时间上都有待提高。因此,寻求快速的约简算法仍然是粗糙集的主要研究方向之一。

参考文献:

[1] Pawlak Z. Rough Sets[J]. International Journal of Computer Information Sciences,1982,11(5):341-356.
[2] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.
[3] 史忠植.知识发现[M].北京:清华大学出版社,2002.
[4] 曾黄磷.粗集理论及其应用——关于数据推理的新方法[M].重庆:重庆大学出版社,1996.
[5] 姚明臣.基于粗糙集的属性约简算法研究与实现[D].哈尔滨:哈尔滨工业大学,2002.
[6] Merz C J, Murphy P. UCI repository of machine learning database[EB/OL]. 2002. <http://www.cs.uci.edu/2002>.

(上接第 94 页)

[J].计算机工程,2006,28(7):104-105.
[5] 杨 凯,张小平,马 垣.基于属性分组的高效挖掘关联规则算法[J].计算机工程与应用,2005(31):157-159.
[6] 章志明,黄龙军,余 敏,等.一种动态的频繁项集挖掘算法[J].计算机工程,2006,32(24):78-80.
[7] 张梅峰,张建伟,张新敬,等.基于 Apriori 的有效关联规则挖掘算法的研究[J].计算机工程与应用,2003(19):196-198.