

# 改进的属性约简算法在数据挖掘中的应用研究

李智玲<sup>1</sup>, 胡 彧<sup>2</sup>

(1. 山西财经大学 实验教学中心, 山西 太原 030006;

2. 太原理工大学 测控技术研究所, 山西 太原 030000)

**摘要:**属性约简是应用粗糙集理论进行数据挖掘有效的方法之一, HORAFA 属性约简算法它的不足之处在于约简效率和完备性。应用粗糙集对知识分类的特点, 建立了新的数据挖掘模型。在模型的属性约简模块中, 详细分析了 HORAFA 算法, 提出了对其改进的 HORAFA-AFVDM 算法。该算法是在核中依次加入属性重要性最大的属性  $a$ , 对于  $\text{Red} = \text{Red} \cup \{a\}$ , 当  $\text{POS}_{\text{red}} - \text{ai}(D) = \text{POS}_c(D)$  时删除  $a$ , 直到不能再删为止, 保证了算法的完备性。实验在 MATLAB 环境下实现, 算法的测试数据来源于 UCI 数据集, 通过对改进前后两种算法的比较, 证实了改进后算法从属性约简效率和算法运行时间上均比之前的算法有显著的提高, 文中将该数据挖掘模型应用到短信数据挖掘系统中。

**关键词:**数据挖掘; 糙粗集; 区分矩阵; 属性约简; 属性频率

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2012)10-0047-04

## Application Research of Improved Attribute Reduction Algorithm in Data Mining

LI Zhi-ling<sup>1</sup>, HU Yu<sup>2</sup>

(1. Experimental Teaching Center, Shanxi University of Finance and Economics, Taiyuan 030006, China;

2. Institute of Measurement and Control Technology, Taiyuan University of Technology, Taiyuan 030000, China)

**Abstract:** Attribute reduction is an effective method of rough set theory for data mining, shortcomings of HORAFA algorithm are reduction efficiency and maturity. A new model of data mining was propounded according to the characteristic of rough sets classifying knowledge. HORAFA algorithm was analyzed in the attribute reduction of model and the improved attribute reduction algorithm, HORAFA-AFVDM was obtained. It joins the most important attributes  $a$  to core followed by, for  $\text{Red} = \text{Red} \cup \{a\}$ , if  $\text{POS}_{\text{red}} - \text{ai}(D) = \text{POS}_c(D)$ , then delete  $a$  from the core. so the maturity of the algorithm is ensured. The experiment is achieved using the tool of the MATLAB. The test data of algorithm comes from the UCI data sets. It proves validity of improved algorithm through the comparison of the efficiency of attribute reduction and running time of the two algorithms. Finally, the new data mining model was applied to SMS data mining system.

**Key words:** data mining; rough set; discernibility matrix; attribute reduction; attribute frequency

## 0 引言

粗糙集<sup>[1]</sup> (Rough Set, RS) 是波兰科学家 Pawlak 在 1982 年首先提出。它是一种刻画不完整性和不确定性的数学工具, 能有效地分析不精确、不一致、不完整等各种不完备的信息, 还可以对数据进行分析和推理, 从中发现隐含的知识, 揭示潜在的规律。

属性约简是粗糙集理论研究的核心内容之一, 是在保留基本知识, 同时保证对象的分类能力不变的基础上, 消除重复、冗余的属性和属性值, 实现对知识的

压缩和再提炼。目前, 属性约简大体分为两类, 即: 基于区分矩阵的属性约简算法<sup>[2]</sup> 和启发式属性约简算法<sup>[3]</sup>。典型的启发式属性约简算法有: 基于粒度的属性约简算法<sup>[4]</sup>、基于区分矩阵属性频度的属性约简算法<sup>[5]</sup>、基于正区域的属性约简算法<sup>[6]</sup>, 其优点是能快速找到一个约简, 缺点是约简算法是不完备的, 不一定能得到真正的属性约简结果。文中将文献[7]中的属性约简算法进行了改进, 保证了算法的完备性, 并将其应用到实际的数据挖掘系统中。

## 1 基本概念

定义 1: 等价类<sup>[8]</sup>。设  $R$  是  $A$  上的一个等价关系, 与  $A$  中的一个元素  $a$  相关的所有的元素的集合被称为  $a$  的一个等价类, 记作  $[a]_R$ 。

定义 2: 约简<sup>[9]</sup>。设  $R$  为一个等价关系族,  $r \in R$ ,

收稿日期: 2012-02-12; 修回日期: 2012-05-19

基金项目: 山西省自然科学基金资助项目 (2009011019-2)

作者简介: 李智玲 (1978-), 女, 山西方山人, 讲师, 硕士, 主要研究方向为数据挖掘、知识工程; 胡 彧, 教授, 博士, 主要研究方向为智能信息处理与数据挖掘、智能控制理论与应用。

如果:  $\text{ind}(R) = \text{ind}(R - \{r\})$ , 则称  $r$  是  $R$  中可被约去的知识, 如果  $P = R - \{r\}$  是独立的, 则  $P$  是  $R$  中的一个约简。

定义 3: 核<sup>[9]</sup>。  $R$  中所有不可约去的关系称为核, 由它构成的集合称为  $R$  的核集, 记作  $\text{core}(R)$ 。

定义 4: 区分矩阵<sup>[10]</sup>。 设  $S = (U, A)$  是一个信息系统, 并置  $A = \{a_1, \dots, a_m\}$ ,  $M(S)$  表示  $n \times m$  阶矩阵, 称它为  $S$  的区分矩阵, 使得:

$$(c_{ij})_{n \times m} = \begin{cases} \{a \in C: a(x_i) \neq a(x_j)\} & D(x_i) \neq D(x_j) \\ \emptyset & D(x_i) = D(x_j) \\ -1 & a(x_i) = a(x_j) \text{ and } D(x_i) \neq D(x_j) \end{cases}$$

定义 5: 下近似集、上近似集<sup>[11]</sup>。 设集合  $X \subseteq U$ ,  $R \subseteq \text{ind}(K)$ , 定义两个子集:

$$RX = \cup \{Y \in U/R \mid Y \subseteq X\}$$
$$\bar{R}X = \cup \{Y \in U/R \mid Y \cap X \neq \emptyset\}$$

分别称它们为  $X$  的  $R$  下近似集和  $R$  上近似集。

集合  $bn_R(X) = \bar{R}X - RX$  称为  $X$  的  $R$  边界域;  
 $\text{POS}_R(X) = RX$  称为  $X$  的  $R$  正域;  $\text{neg}_R(X) = U - \bar{R}X$  称为  $X$  的  $R$  负域。

2 海量数据挖掘系统建构

数据挖掘的目的是得出隐藏在数据库中的有价值的信息。传统的数据分析工作量大, 主要是依赖工作人员的经验, 因此很难从数据库中获取准确的知识。文中应用数据分析与粗糙集的理论知识, 建构了海量数据的数据挖掘系统框架, 如图 1 所示。

2.1 属性约简

属性约简常用的方法是, 去掉信息系统中的某一条件属性, 判断信息系统仍是否相容, 若相容, 说明该属性是冗余的, 否则, 该属性是必要的。

2.2 HORAFA 算法

文献[7]中提出了 HORAFA 算法。它以属性在区分矩阵中出现的频率为启发, 对信息系统进行约简。因为只计算区分矩阵中属性出现的频率, 不计算粗糙集复杂的概念, 因此效率被提高了。

HORAFA 算法描述如下:  
Begin  
先给 Red, count 附初值为  $\text{Red} = \Phi$  and  $\text{count}(ai) = 0$ , 对于  $i = 1, \dots, n$   
得出区分矩阵  $M$  同时属性的加权频率值  $\text{count}(ai)$  也被计算出来  
区分矩阵中相同的项被合并, 同时按照矩阵中每项的长度和属性的频率对区分矩阵进行排序

repeat  
if (  $m \cap \text{Red} = \Phi$  )  
    选择  $m$  中的具有最大的  $\text{count}(a)$  的属性  $a$   
     $\text{Red} = \text{Red} \cup \{a\}$   
End  
until  $M$  中的每项  $m$  都计算完  
Return Red  
end

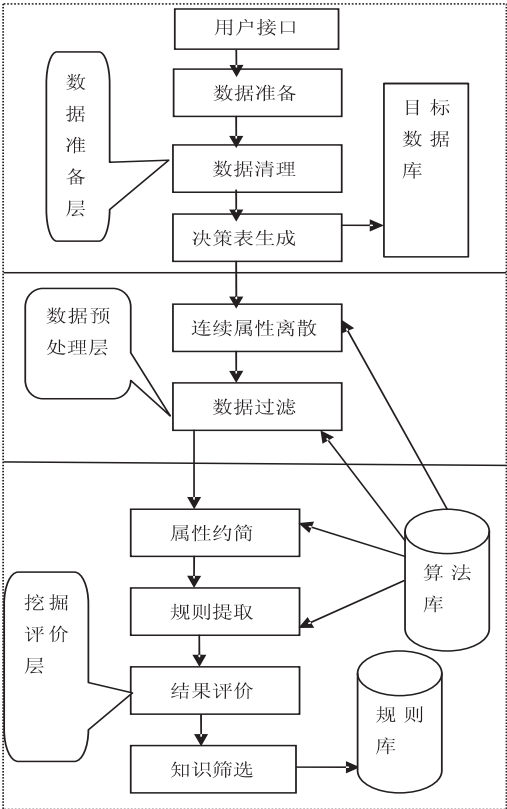


图 1 数据挖掘系统结构

2.3 HORAFA-AFVDM 算法

实验证明<sup>[7]</sup> HORAFA 算法不一定能够找到信息系统的较优约简, 它是不完备的。文中对其进行了改进, 提出了新的 HORAFA-AFVDM<sup>[12]</sup> (HORAFA based on attribute frequency value of discernibility matrix) 算法。它一定能找到信息系统的最优约简。

文中对 HORAFA 算法提出了如下两点改进:  
1) 属性频率函数更新为:  $f(a) = f(a) + |A| / |c'|$  ;  
2) 约简以核为基础, 依次加入属性重要性最大的属性  $a$ , 然后对于  $\text{Red} = \text{Red} \cup \{a\}$ , 当  $\text{POS}_{\text{red}} - ai(D) = \text{POS}_c(D)$  时删除  $a$ , 直到不能再删为止, 确保了算法的完备性。  
HORAFA-AFVDM 算法描述如下:  
Begin  
先给 Red, count 附初值为  $\text{Red} = \Phi$ ,  $\text{count}(ai) = 0$ , 对于  $i = 1, \dots, n$

得出区分矩阵  $M$  合并其中相同的元素,在 core 中加入只有一个元素的属性,  $n++$

将 core 附值给 Red,区分矩阵  $M$  中包含核的所有项被删除,计算  $\text{count}(ai)$  和  $|M|$

```
if  $|M| \neq 0$ 
    在 Red 中加入  $\text{count}(ai)$  最大的  $ai, n++$ , 区分矩阵  $M$  中包含  $ai$  的所有项被删除, 然后重新计算  $\text{count}(ai)$  和  $|M|$ 
end
if  $\text{POSred} - ai(D) \neq \text{POS}_c(D)$ 
    保留  $ai$ 
else 删除  $ai, n--$ 
end
Return Red
end
```

2.4 实验及结果分析

算法测试的硬件环境是运行于 windowsXP 的联想笔记本,双核 2.00GHz 的 CPU,1.0GBM 的 Memory。软件环境是 MATLAB7.0,SQL SERVER2000。

实验数据来源于 UCI<sup>[13]</sup> 数据集,实验结果如表 1,其中 Red1 为 HORAFA 算法的属性约简个数;Red2 为改进算法 HORAFA-AFVDM 的属性约简个数;Time1 为 HORAFA 算法的运行时间;Time2 为改进算法 HORAFA-AFVDM 的运行时间;h 为小时;m 为分钟;s 为秒。

从表 1 中看出除 tic-tac-toe、vehicle、glass 外,改进后算法的属性约简个数均小于改进前的算法,HORAFA-AFVDM 算法的约简能力总体上比 HORAFA 算法强。从算法的运行时间上看,改进后算法的运行时间都不同程度地得到了改善。从实验结果得出,HORAFA-AFVDM 算法从算法约简能力和运行时间上

都得到了大大的改进。

表 1 HORAFA 算法与 HORAFA-AFVDM 算法的属性约简和运行时间比较

数据集	实例数	条件属性	Red1	Red2	Time1	Time2
solar	333	10	9	7	6m0.078s	4m1.765s
tic-tac-toe	201	9	1	1	3m22.469s	1m2.109s
vehicle	150	18	2	2	3m7.853s	1m73.506s
zoo	101	16	7	5	41.75s	18.703s
ctr	21	9	7	4	0.07s	0.006s
flag	214	27	14	6	13m26.56s	9m40.375s
glass	232	9	2	2	11m25.678s	3m17.893s
auto_mpg	650	7	3	2	1h10m57.054s	24m3.456s
heart	66	16	6	4	8.83s	1.894s
house_votes	226	16	16	8	7m24.497s	3m8.656s

3 短信数据挖掘系统应用

短信数据挖掘系统采用 VisualC++和 SQLSERVER 2000 实现。测试数据来源于一个工作日忙时的数据,一共 20 列 20000 行。只显示部分数据如图 2 所示。

在该系统中,对属性进行约简之前,进行了数据预处理。图 3 为系统中数据约简后的结果,其中属性值全为零的属性是冗余属性,删除这些属性,并没有影响整个数据表的相容性,原有的分类能力也没有变。

依据数据表中所选择的字段名称和生活常识,可以确定数据挖掘系统的约简结果和实际情况是一致的。

4 结束语

文中使用 MATLAB 工具实现了算法,实验数据来源于 UCI 数据集,实验比较了改进前后算法的属性约简效率和运行时间,并将新算法应用到短信数据挖掘系统中,发挥了粗糙集模式识别与分类的长处,提高了数据挖掘的效率,并在系统中得到了充分的验证。

FAreaCode	FCPID	FDeliverTime	FDestAddr	FMsgType	FO
010	0	2005/02/28 02:12:38	03513406583	3	011
0471	0	2005/02/28 02:13:43	03545995927	14	13
010	0	2005/02/28 02:13:50	03513406593	3	011
0349	0	2005/02/28 07:24:42	106980142	5	03
010	0	2005/02/28 02:11:25	03546189545	3	011
0349	0	2005/02/28 02:11:28	106980142	5	03
0471	0	2005/02/28 02:11:30	0358859331	14	13
010	0	2005/02/28 02:11:31	03513406576	3	011
010	0	2005/02/28 02:16:10	03513406611	3	011
010	0	2005/02/28 02:15:01	03513406603	3	011
0351	11	2005/02/28 02:15:43	03518692069	7	99
010	0	2005/02/28 02:15:45	03513406608	3	011
	11	2005/02/28 00:00:13	03512877393		99
0358	11	2005/02/28 23:31:15	03587236542	7	99
0359	0	2005/02/28 23:29:46	1069757	5	03
0471	0	2005/02/28 23:29:46	03518521972	14	13
	0	2005/02/28 23:29:44	03513565789		07
0471	0	2005/02/28 23:29:42	03542628497	14	13
010	0	2005/02/28 23:29:37	03566851899	3	011
010	0	2005/02/28 02:24:02	03513406671	3	011
010	0	2005/02/28 02:24:43	03513406676	3	011
010	0	2005/02/28 02:24:50	03546186770	3	011
010	0	2005/02/28 02:22:17	03513406657	3	011
0351	9	2005/02/28 02:22:16	03513316243	7	99
010	0	2005/02/28 02:23:05	03546580634	3	011
010	0	2005/02/28 02:23:06	03546189724	3	011
010	0	2005/02/28 02:23:07	03513406663	3	011
0358	1	2005/02/28 08:04:01	03588906362	1	12
0354	1	2005/02/28 08:04:01	03542626670	1	12
0354	1	2005/02/28 08:04:01	03544882545	1	12
0358	1	2005/02/28 08:04:01	03588909255	1	12
0358	1	2005/02/28 08:04:01	03588902198	1	12
0351	0	2005/02/28 08:04:02	1383510	13	03
	1	2005/02/28 08:04:02	03554129579		12
	-	2005/02/28 08:04:02	03556688855		-

图 2 部分源数据

数据挖掘系统									
DataGrid1									
fscndaccoun	frcvaccount	fsubmittime	fdelivertime	forqaddr	fdestaddr	fareacode	fcpid	fmsgtype	fsmtypr
3	5	2	0	4	0	1	0	3	5
1	5	2	0	1	0	0	0	14	3
3	5	2	0	4	0	1	0	3	5
6	3	4	0	0	0	3	0	5	0
3	5	2	0	4	0	1	0	3	5
6	3	2	0	0	0	3	0	5	0
2	5	2	0	3	0	0	0	14	3
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	0	0	3	0	7	2
3	5	2	0	4	0	1	0	3	5
3	0	1	0	0	0	0	0	0	2
3	5	12	0	0	0	3	0	7	2
4	0	5	0	0	0	0	0	0	5
4	0	5	0	0	0	0	0	0	5
4	0	5	0	0	0	0	0	0	5
6	3	12	0	5	0	0	0	5	0
2	3	12	0	3	0	0	0	14	3
3	0	12	0	0	0	0	0	0	3
1	5	12	0	2	0	0	0	14	3
3	5	12	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
3	5	2	0	4	0	1	0	3	5
4	5	5	0	0	0	3	0	1	2
4	0	5	0	0	0	0	0	0	5
4	0	5	0	0	0	0	0	0	5

属性约简

剩余行数 50

返回首页

图 3 属性约简数据

参考文献:

[1] Pawlak Z. Rough set; theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] 王立宏, 吴耿锋. 基于并行协同进化的属性约简[J]. 模式识别与人工智能, 2003, 26(5): 630-635.

[3] 沈 玮, 赵佳宝. 一种新的启发式粗集决策表属性约简算法[J]. 计算机技术与发展, 2010, 20(10): 16-20.

[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): 59-66.

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

[6] 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D]. 北京: 中国科学院自动化研究所, 1997.

[7] 胡可云. 基于概念格和粗糙集的数据挖掘方法研究[D]. 北京: 清华大学, 2001.

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

[9] 刘 清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.

[10] 王国胤. Rough 集理论与属性获取[M]. 西安: 西安交通大学出版社, 2001.

[11] 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, 2005.

[12] 李智玲. 基于区分矩阵的粗糙集属性约简算法在数据挖掘中的应用研究[D]. 太原: 太原理工大学, 2007.

[13] Murphy K P. The Bayes net toolbox for matlab[J]. Computing Science and Statistics, 2001, 33(2): 1024-1034.

(上接第 46 页)

[5] 张 昱, 严洪森, 张 平. 基于 B/S 结构的知识化制造自重构子系统的实现[J]. 计算机技术与发展, 2008, 18(1): 25-29.

[6] 杨人子, 严洪森. 知识化制造系统中知识网的结构研究[J]. 计算机集成制造系统, 2008, 14(3): 595-601.

[7] 张 平, 严洪森, 余晓光. 基于混合算法的知识网运算表达式优化[J]. 计算机技术与发展, 2009, 19(3): 32-35.

[8] 杨人子, 严洪森. 基于知识网的知识表达度量方法及其应用[J]. 系统工程理论与实践, 2010, 30(6): 1067-1076.

[9] 王艳斌, 严洪森, 马力伟, 等. 知识化制造系统的知识网数据库设计[J]. 东南大学学报(自然科学版), 2004, 34(增刊): 24-29.

[10] 李金坚, 严洪森, 胡建悦. 基于知识网最简约生成的面向组件软件开发系统[J]. 计算机技术与发展, 2011, 21(1): 125-128.

[11] 殷乾坤, 严洪森, 王方顺. 知识化制造软件系统自动生成的实现[J]. 计算机技术与发展, 2011, 21(1): 16-18.

[12] 倪 兴, 张韵华. 动态测量数据的合理性检验方法研究[J]. 运筹与管理, 2011, 20(4): 113-115.

[13] 何 秋, 桂寿平, 朱 强. 区域物流系统动态学模型的建立与合理性检验[J]. 交通与计算机, 2002, 20(3): 30-33.

[14] 黄家贵, 詹武平. 合理性检验方法的研究[J]. 装备指挥技术学院学报, 2002, 13(4): 93-97.

[15] 张 洁, 刘世平, 李培根. 基于多 Agent 的车间重构模型[J]. 中国机械工程, 2000, 11(4): 432-434.

[16] Yan Hongsen. A new complicated-knowledge representation approach based on knowledge meshes[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 47-62.

[17] 李晓喆, 张晓辉, 李祥胜. SQL Server 2000 管理及应用系统开发[M]. 北京: 人民邮电出版社, 2002.

[18] 麦中凡, 陆永宁. C#编程语言[M]. 北京: 北京航空航天大学出版社, 2001.

作者: 李智玲, 胡戡  
作者单位: 李智玲(山西财经大学 实验教学中心, 山西 太原 030006), 胡戡(太原理工大学 测控技术研究所, 山西 太原 030000)  
刊名: 计算机技术与发展  
英文刊名: Computer Technology and Development  
年, 卷(期): 2012(10)

参考文献(13条)

1. Pawlak Z Rough set: theoretical aspects of reasoning about data 1991  
2. 王立宏; 吴耿锋 基于并行协同进化的属性约简 2003(05)  
3. 沈玮; 赵佳宝 一种新的启发式粗集决策表属性约简算法 2010(10)  
4. 王国胤; 于洪; 杨大春 基于条件信息熵的决策表约简 2002(07)  
5. 叶东毅 Jelonek属性约简算法的一个改进 2000(12)  
6. 苗夺谦 Rough Set 理论及其在机器学习中的应用研究 1997  
7. 胡可云 基于概念格和粗糙集的数据挖掘方法研究 2001  
8. Pawlak Z Rough sets 1982(05)  
9. 刘清 Rough集及Rough推理 2001  
10. 王国胤 Rough集理论与属性获取 2001  
11. 张文修; 吴伟志; 梁吉业 粗糙集理论与方法 2005  
12. 李智玲 基于区分矩阵的粗糙集属性约简算法在数据挖掘中的应用研究 2007  
13. Murphy K P The Bayes net toolbox for matlab 2001(02)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjfz201210014.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjfz201210014.aspx)