

# 基于属性拓展的数据挖掘预处理技术研究

张春生,李 艳,图 雅

(内蒙古民族大学 计算机科学与技术学院,内蒙古 通辽 028043)

**摘 要:**目前的数据挖掘技术基本上依据的是原始数据库中的信息,数据预处理技术要维护原始数据库的信息基本不变,只是进行简单的数据标准化、数据平整、异常点发现、缺失数据修补、数据离散等基本预处理工作,不能从根本上拓展原始数据库中的信息。同时,为保密起见,兴起的隐私保护数据挖掘技术对原始数据库中的敏感数据进行处理,隐藏了一些基本信息,进一步弱化了原始数据库中的信息含量。基于属性拓展的数据挖掘预处理技术,从原始数据库出发,通过属性拓展,拓展基础数据库所蕴含的信息,使数据挖掘能产生更深的隐藏关联规则。

**关键词:**属性;拓展;数据挖掘;预处理技术;关联规则

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2014)03-0079-03

doi:10.3969/j.issn.1673-629X.2014.03.020

## Research on Data Mining Preprocessing Technology Based on Attribute Extension

ZHANG Chun-sheng, LI Yan, TU Ya

(College of Computer Science and Technology, Inner Mongolia University for Nationalities,  
Tongliao 028043, China)

**Abstract:** Nowadays the data mining technology is basically based on the information of the original database. Data preprocessing technology must maintain the information of the original database unchanged, just for simple data standardization, data level, abnormal points found, missing data repair work, data discrete and other basic data pretreatment, not radically expanded the information in the original database. Meanwhile, for the sake of secrecy, the privacy preserving of data mining technology deals with the sensitive data in the original database, and hides some basic information, further weakening the information content in the original database. The data mining preprocessing technology based on the attribute extension starts from the original database, expands the information implicated in basic database by extending the attribute from the original database perspective, so that data mining can produce more deeply hidden rules.

**Key words:** attribute; extension; data mining; preprocessing technology; association rule

### 1 概 述

随着数据库技术和网络技术的发展,各行各业都积累了大量的有用数据。如何从这些数据中提取出对决策有价值的知识,成为当务之急。数据挖掘行为作为一个强有力的数据分析工具,可以发现数据中潜在的模式和规律,在商务决策、科学研究等领域都有广泛的应用前景<sup>[1-3]</sup>。

数据挖掘的理论自产生以来,基础理论研究和应用研究发展迅速,但都是基于原始数据库或数据仓库的基础上,经过数据预处理后通过数据挖掘完成的。

目前关于数据挖掘预处理技术研究成果较多,但

这些技术没有从根本上改变原始数据库所蕴含的信息,只是作基本变换,以适应数据挖掘的需要<sup>[4-6]</sup>。数据预处理包括数据的标准化(按一定比例把数据映射到 $[-1, 1]$ 或 $[0, 1]$ 上),数据的平整(通过对精度的限制,减少数据值个数),丢失数据修补,连续数据离散化等基本预处理工作。这些工作的目的是为数据挖掘作准备工作,基本上保持了原始数据信息内容。

另一方面,数据库中的原始信息有时人为地需要进行隐藏。很多情况下,数据由不同的组织持有,分布于不同的地理位置,而且持有者可能出于安全性和敏感性等原因不愿直接共享他们的数据,出现了一个新的数据挖掘方向——隐私保护数据挖掘技术<sup>[7-12]</sup>,隐

收稿日期:2013-05-13

修回日期:2013-08-19

网络出版时间:2014-01-07

基金项目:国家自然科学基金资助项目(61261025);内蒙古自然科学基金(2012MS0913)

作者简介:张春生(1965-),男,河北乐亭人,教授,硕士,研究方向为数据库技术、数据挖掘、软件理论。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140107.1726.052.html>

私保护数据挖掘通过一些技术,对原始数据中的一些敏感数据实施保护。这样进一步降低了原始数据的信息量,降低了数据挖掘的规则的数量。

无论是数据预处理还是隐私保护处理,都不同程度地降低了原始数据所包含的信息量,从而降低了数据挖掘所产生的有效规则数量,这一点已经引起人们的关注。

文中提出的基于属性拓展的数据挖掘预处理技术不同于简单的数据预处理技术,基于属性的扩展是在原始数据基础上通过对单属性或属性间的运算而得到新属性,这些新属性参与数据挖掘,从而达到拓展原始数据所蕴含的信息量的目的。

综上所述,目前的数据挖掘技术没有在原始数据的基础上进行拓展,甚至压缩了原始数据信息,文中从原始数据出发,通过属性含义的拓展,扩展原始数据的隐含知识,并在此基础上实现数据挖掘,必将挖掘出更深层的隐含知识,提高数据挖掘的性能。

2 属性拓展方法

2.1 属性拓展知识库建立

根据属性拓展需求,构建如下属性拓展知识数据库:

单属性拓展知识库结构。

- (1)主表定义。  
(标识,原始属性,操作类型,拓展属性)  
(m\_id,A,op\_type,EA)
- 其中,m\_id为主表记录标识,不重复;A为待转换的原始属性;op\_type为操作符类型;EA为转换后的属性。

其中 op\_type 可以是算数运算、比较运算、逻辑运算、集合运算等。

- (2)子表定义。  
(子标识,主标识,转换表达式,拓展属性值)  
(s\_id,m\_id,expression,EA\_value)
- 其中,s\_id为子表记录标识,不重复;m\_id为子表记录对应的主表记录标识;expression为转换表达式;EA\_value为转换后的拓展属性值。

示例如表 1 和表 2 所示。

表 1 主表示例			
m_id	A	op_type	EA
1	年龄	集合	年龄段
2	购买日期	集合	季节

当某原始属性需转换时,先在主表中找到符合条件的记录,通过两表连接(通过 m\_id),在子表中找到相应的转换公式和转换后的拓展属性值,实现属性拓

展。

表 2 子表示例			
s_id	m_id	expression	EA_value
1	1	年龄 in [ 0,14]	少年
2	1	年龄 in [15,35]	青年
3	1	年龄 in [36,45]	中年
4	1	年龄 in [46,60]	老年
5	2	购买日期 in [2010-03-01,2010-05-31]	春季
6	2	购买日期 in [2010-06-01,2010-08-31]	夏季
7	2	购买日期 in [2010-09-01,2010-11-30]	秋季

多属性拓展知识库结构。  
构建如下知识表:  
(标识,属性串,转换表达式,拓展属性)  
(id,Attribute\_string,expression,EA\_value)  
其中,id为属性串标识;Attribute\_string为属性串;  
expression为转换表达式;EA\_value为拓展属性。  
示例见表 3。

表 3 多属性表实例			
id	Attribute_string	expression	EA_value
1	数量.单价	数量*单价	金额
2	应发工资.实发工资	应发工资-实发工资	扣款

注:属性串中出现的属性可以是原始属性,也可以是拓展的单属性。

2.2 属性拓展方案

文中提出 2 种属性拓展方案,从原始属性出发,拓展原始属性的知识。

- (1)单属性拓展。  
根据对单属性的计算进行知识变换,得到原始属性隐藏的知识。

定义 1:设某数据对象的原始属性为  $A_i$ ,构建用于变换的知识属性  $F_i$ , $F_i$ 由用户构建,包括属性名和域, $\theta$ 为变换操作,则单属性拓展表示为: $F_i\theta A_i \rightarrow A_i'$ , $A_i$ 与  $A_i'$ 的含义不同,属性  $A_i'$ 称为拓展属性。

算法流程如图 1 所示。

如图 1 所示,转换开始,首先遍历所有的原始属性,在主表中查找此属性是否可以拓展,若找到该属性,则根据该属性的 m\_id 在子表中查找相应的转换公式,通过转换公式得到拓展后的属性及属性值,否则失败。

- (2)多属性拓展。  
多属性拓展是通过原数据对象中的若干属性(原始属性或拓展属性)的相互操作变换出来的属性。

定义 2:设某数据对象的  $m$  个原始属性或拓展属性为  $\{A_i| i=1,2,\cdots,m\}$ , $\theta$ 为变换操作,则多属性拓展表示为: $A_1\theta A_2\cdots\theta A_m \rightarrow A_k$ ,其中  $k \notin \{1,2,\cdots,m\}$ , $A_k$ 与  $A_1,A_2,\cdots,A_m$ 的含义不同,则属性  $A_k$ 称为拓展属性。

算法流程如图 2 所示。

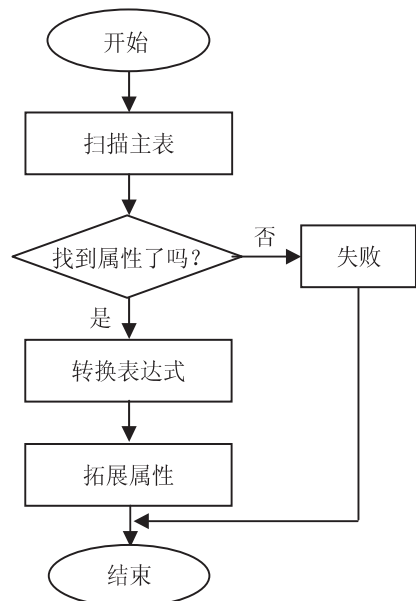


图 1 单属性拓展流程

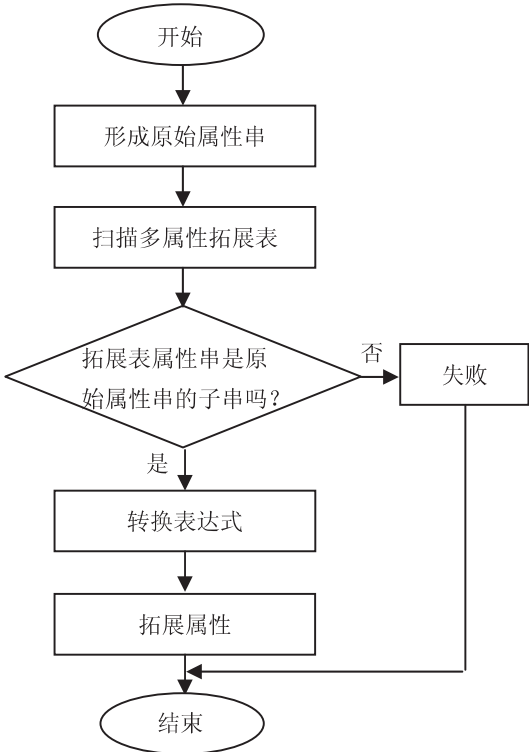


图 2 多属性拓展流程

如图 2 所示,转换开始,首先形成由全体原始属性组成的原始属性串,扫面所有属性表中的属性串,若属性串是原始属性串的子串,表示可以拓展,通过转换表达式得到拓展属性和属性值,否则失败。

3 属性拓展的实现实例

(1)单属性拓展的实现。

对于一个购物篮事务数据库,“购买时间”只能代表时间顺序,在规则挖掘中不起绝对作用,现构建一个

扩展属性。

首先构建一个用于变换的知识属性  $F_i, F_i = (开始日期,结束日期,季节)$ ,这样,通过 (“购买时间” in  $F_i$ ) 操作,得到“购买季节”属性,由于季节与购买的商品关系密切,所以“购买季节”属性在规则产生的过程中将可能产生有用的规则。

(2)多属性拓展的实现。

对于一个描述少年犯罪基本信息的数据库,存在三个属性,“罪犯出生日期”,“父亲出生日期”,“母亲出生日期”,这 3 个属性在规则挖掘构成中可能是非重要属性,但通过多属性拓展可拓展出 2 个新属性:

(犯罪年—date(父亲出生日期))—(犯罪年—date(罪犯出生日期))→罪犯出生时父亲年龄

(犯罪年—date(母亲出生日期))—(犯罪年—date(罪犯出生日期))→罪犯出生时母亲年龄

“罪犯出生时父亲年龄”与“罪犯出生时母亲年龄”某种程度上代表了父母对孩子的溺爱程度,将成为对少年犯罪的重要因素参与规则的产生。

4 结束语

文中针对目前的数据挖掘预处理技术只对原始数据库中的属性作无损原始信息的基本变换的缺点,从原始属性出发,给出了基于原始属性知识拓展的 2 种方法,即单属性拓展的实现、多属性拓展的实现。

通过原始属性的知识拓展,从原始数据角度拓宽了数据的知识含量,可挖掘出更深层的隐含知识,这些隐含知识是普通数据挖掘方法无法得到的。当然,文中只是抛砖引玉,若通过更有效的方法对原始属性进行拓展,将一定取得好的效果。

参考文献:

[1] 毛国君,段立娟,王 实,等.数据挖掘原理与算法[M].第 2 版.北京:清华大学出版社,2007.

[2] Richard R J,Geatz M W.数据挖掘教程[M].翁敬农,译.北京:清华大学出版社,2003.

[3] 于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社,2003.

[4] 袁 健,金 鑫.一种重构网站结构的 Web 日志挖掘数据预处理方法[J].小型微型计算机系统,2011,32(7):1427-1430.

[5] 王亚军,王传安.基于属性重要性的 WUM 数据预处理方式[J].计算机系统应用,2011,20(5):219-222.

[6] 据春华,梅 铮,刘东升.一种基于粗糙等价类的商业数据预处理方法[J].小型微型计算机系统,2009,30(5):955-958.

[7] Zhu Jianghua,Li Haibo,Pan Feng. Knowledge-reduction bas-

Leach 协议生命周期延长了 30%。这说明改进后的粒子群算法具有良好的多目标寻优能力,簇首选择优化后的 Leach 协议节点的能量消耗均衡,避免了 Leach 协议中簇首节点能量消耗大而提前死亡,从而缩短网络生命周期的现象。

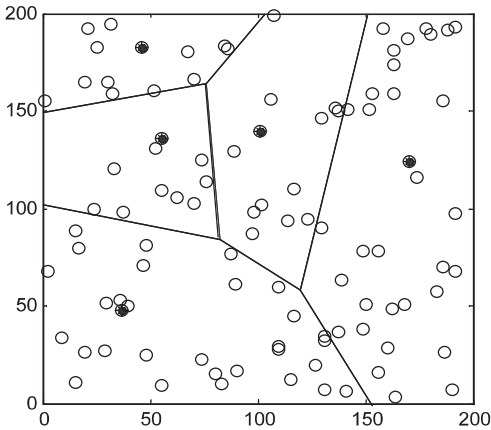


图2 AWSPSO-Leach 协议一轮的分簇示意图

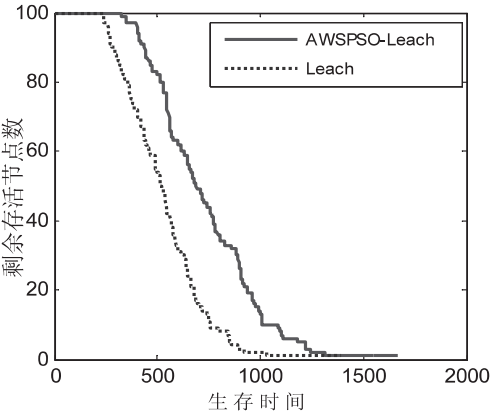


图3 生存周期比较

5 结束语

文中提出了一种改进的粒子群算法,在引入自适应惯性权重以及异变学习因子后,克服了基本粒子群算法的局部收敛问题,提高了算法的多目标寻优能力。针对基本 Leach 协议簇首选择的随机性问题,利用粒子群算法的多目标寻优能力对其进行优化,均衡了网络的能量消耗。仿真结果表明:经过改进粒子群算法优化后的 Leach 协议,分簇更加均匀有效,网络的生命

周期得到延长。

参考文献:

[1] Heinzelman W,Chandrakasan A,Balakrishnan H. Energy efficient communication protocol for wireless microsensor networks[C]//Proceedings of the 33rd annual Hawaii international conference on system sciences. [s.l.]:[s.n.],2000:1-10.

[2] Heinzelman W B,Chandrakasan A P,Balakrishnan H. An application specific protocol architecture for wireless micro-sensor networks[J]. IEEE transactions on wireless communications,2002,1(4):660-670.

[3] 邢云冰,史浩山,赵洪钢. 基于备用节点的无线传感器网络 LEACH 协议的改进[J]. 传感技术学报,2007,20(7):1592-1596.

[4] Lindsey S,Raghavendra C,Sivaligam K M. Data gathering algorithms in sensor networks using energy metrics[J]. IEEE transactions on parallel and distributed systems,2002,13(9):924-935.

[5] Kennedy J,Eberhart R C. Particle swarm optimization[C]//Proc of IEEE international conference on neural networks. Piscataway:IEEE Service Center,1995:1942-1948.

[6] Latiff N M A,Tsimenidis C C,Sharif B S. Energy-aware clustering for wireless sensor networks using particle swarm optimization[C]//Proceedings of IEEE 18th international symposium on personal, indoor and mobile radio communications. [s.l.]:[s.n.],2007:1-3.

[7] 吴昌友,王福林,马力. 一种新的改进粒子群优化算法[J]. 控制工程,2010,17(3):359-362.

[8] Liu B,Wang L,Jin Y H. An effective PSO-Based memetic algorithm for flow shop scheduling[J]. IEEE transactions on systems,man and cybernetics,2007,37(1):18-27.

[9] 李洪亮,侯朝桢,周绍生. 一种高效的改进粒子群优化算法[J]. 计算机工程与应用,2008,44(1):14-16.

[10] 张雪东,赵传信,季一木. 一种混合粒子群算法及其在 Job Shop 问题中的应用[J]. 计算机技术与发展,2006,16(9):109-111.

[11] 牛小娇,吕程林. 一种基于 LEACH 协议的分簇路由算法[J]. 计算机技术与发展,2011,21(7):13-16.

[12] 朱丽莉,杨志鹏,袁华. 粒子群优化算法分析及研究进展[J]. 计算机工程与应用,2007,43(5):24-27.

(上接第 81 页)

ed on GA and fuzzy-rough set[J]. Computer simulation, 2007,24(1):68-70.

[8] Miao Duoqian,Hu Guirong. A heuristic algorithm for reduction of knowledge[J]. Computer research and development,1999,36(6):681-684.

[9] Shi Huaji,Qin Chuan,Chen Huijun. Heuristic algorithm of attribute reduction in condition entropy[J]. Computer engineer-

ing and design,2008,29(19):5014-5016.

[10] 陈晓明,李军怀,彭军. 等. 隐私保护数据挖掘算法综述[J]. 计算机科学,2007,34(6):183-186.

[11] 华蓓,钟诚. 数据挖掘中的隐私保护技术进展分析[J]. 微电子学与计算机,2009,26(8):38-41.

[12] 韩建民,于娟,虞慧群. 等. 面向敏感值的个性化隐私保护[J]. 电子学报,2010,38(7):1723-1728.

基于属性拓展的数据挖掘预处理技术研究

作者：[张春生](#)，[李艳](#)，[图雅](#)，[ZHANG Chun-sheng](#)，[LI Yan](#)，[TU Ya](#)

作者单位：[内蒙古民族大学 计算机科学与技术学院, 内蒙古 通辽, 028043](#)

刊名：[计算机技术与发展](#)



英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(3)

本文链接：[http://d.g.wanfangdata.com.cn/Periodical\\_wjfz201403020.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjfz201403020.aspx)