文章编号: 1001-0920(2004)08-0957-04

动态决策系统中的几何挖掘算法——概念格

何友全[1], 肖 建[1], 黄碧霞[2], 雷 妍[3], 熊启军[4]
(1. 西南交通大学 电气工程学院, 四川 成都 610031; 2. 西南交通大学 土木工程学院, 四川 成都 610031;
3. 西南交通大学 经济管理学院, 四川 成都 610031; 4. 襄樊学院 电气信息工程系, 湖北 襄樊 441053)

摘 要: 探讨了一种基于概念格的几何数据挖掘算法, 根据不动点理论和伽罗瓦连接原理, 在数据库中寻找大于一定支持度的闭项目集, 分解闭项目集便可得到数据间的关联关系. 实验结果表明, 此方法适用于决策系统, 并且挖掘效率较高.

关键词: 动态决策; 数据仓库; 数据挖掘; 概念格
中图分类号: TM 76; TP18 **文献标识码**: A

Geometric data mining algorithm of dynamic decision system ——Conception lattice

HE You-quan[1], *XIAO Jian*[1], *HUANG Bi-xia*[2], *LEI Yan*[3], *XIONG Qi-jun*[4]
(1. School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China; 2. School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China; 3. School of Economic Management, Southwest Jiaotong University, Chengdu 610031, China; 4. Department of Electrical and Message Engineering, Xiangfan College, Xiangfan 441053, China. Correspondent: HE You-quan, E-mail: hyq123003@163.com)

Abstract: A new geometric data mining algorithm based on conception lattice is discussed. According to fixed point theory and Galois connection principle, frequent closed itemset that is larger than conditional support rate can be found. The association rule between data can be obtained by subdividing the itemset. Experiment shows that the method is useful and efficient in the decision support system.

Key words: dynamic decision; data warehouse; data mining; conception lattice

1 引 言

DSS 结构可分为两大类: 一类是基于两库(模型库和数据库)的框架结构, 另一类是基于知识集的框架结构[1]. 随着人工智能的不断发展, 两种结构相互渗透、相互联系, 它们的一个重要特点是依赖于具有知识集的数据源, 即在数据仓库和数据挖掘的基础上构建动态决策系统. 这一新的决策系统已成为计算机、人工智能、管理决策等领域研究的热点.

数据仓库是大容量数据的集合, 而知识集则由关联规则库组成, 从原始数据源中挖掘关联规则是构建动态决策系统的重要依据. 因此, 如何从大量的原始数据中挖掘出有用信息和关联规则, 是组建决策系统面临的重要课题. 数据挖掘的方法很多, 常见的有: 基于机器学习的诱导式学习算法[2], 多属性判别式矩阵分析方法[3], 粗糙集与神经网络混合模型[4], 模糊认知图(fuzzy cognitive map-FCM)[5], 演绎规则基推理(deductive rule-based reasoning-CBR)[6], 粗糙集和模糊规则的集合[7], 导出式规则

收稿日期: 2003-09-08; **修回日期**: 2003-12-23.
基金项目: 国家自然科学基金资助项目(69774024); 湖北省教育厅基金资助项目(2002D06001).
作者简介: 何友全(1964—), 男, 湖北监利人, 副教授, 博士生, 从事智能控制、数据挖掘的研究; 肖建(1952—), 男, 湖南长沙人, 教授, 博士生导师, 博士, 从事智能控制等研究.

推理生成法[8], 遗传算法[9]等.

对于决策支持系统, 传统的挖掘算法效率低、繁琐、周期长, 若采用一种几何数据挖掘算法——基于概念格的理论和方法, 则算法简单, 挖掘效率高, 能有效地挖掘出数据之间的相互关系, 为决策系统提供支持依据.

2 概念格与闭项目集的基本概念

概念格是数学中的一个重要分支, 它在数学的其他分支中(如泛函分析、拓扑学等)以及计算机科学和社会科学各领域有着广泛的应用. 闭项目集是关联关系中一个概念的内涵. 对于关系型事务数据库, 闭项目集保留了事务数据库中的所有支持度信息, 而概念格则保留了闭项目集的信息.

定义 1[10](格) 设 P 是一集合, $(P, \leq)$ 是偏序集, 若其中任意两元素都有上确界 $x \vee y$ 和下确界 $x \wedge y$, 则称 $(P, \leq)$(简称为 P)为一个格. 在格中, 元素 $x \vee y$ 与 $x \wedge y$ 分别称为 x 与 y 的并和交.

定义 2(概念) 一个关系是一个三元组 (G, M, I), 这里 G 和 M 是集合, $I \subseteq G \times M$, G 和 M 的元素相应称为对象和属性. 用 gIm 表示 $(g, m) \in I$, 含意是"对象 g 具有属性 m". 则关系 (G, M, I) 的概念定义为元素对 (A, B), 这里 $A \subseteq G, B \subseteq M$.

定义 3(概念格) 关系 (G, M, I) 的所有概念的集用 $R(G, M, I)$ 表示, 称其为关系 (G, M, I) 的概念格.

定义 4[11](闭项目集) 一个项目集是一个闭项目集, 若不存在这样的一个项目集 Y, 满足 Y 是 X 的真超集且每个含 X 的事务也含 Y.

定义 5(频繁闭项目集) 一个闭项目集是频繁的, 如果它的支持度不小于给定的最小支持度阈值.

定义 6[12](伽罗瓦连接) 设 (G, M, R) 是一个关系, 且 $O \subseteq G, I \subseteq M$, 定义

$$f(O): 2^G \to 2^M,$$

$$f(O) = \{i \in M \mid \forall o \in O, (o, i) \in R\};$$

$$g(I): 2^M \to 2^G,$$

$$g(I) = \{o \in G \mid \forall i \in I, (o, i) \in R\}.$$

则 (f, g) 是 G 与 M 幂集之间的一个伽罗瓦连接, $h = f \circ g: P(M) \to P(M)$ 和 $h' = g \circ f: P(G) \to P(G)$ 是伽罗瓦算子, 若对所有 $I, I_1, I_2 \subseteq M$ 和 $O, O_1, O_2 \subseteq G$, 根据文献[13, 14], 有如下结论:

1) $I_1 \subseteq I_2 \Rightarrow g(I_1) \supseteq g(I_2), O_1 \subseteq O_2 \Rightarrow f(O_1) \supseteq f(O_2)$;

2) $h(h(I)) = h(I), h'(h'(O)) = h'(O)$;

3) $O \subseteq g(I) \Leftrightarrow I \subseteq f(O)$.

3 概念格与闭项目集的挖掘方法

3.1 Apriori 算法(频繁项集的挖掘)及其特点

对于闭项目集的挖掘, 首先了解 Apriori 算法[15]. Apriori 算法以挖掘频繁项集为主要特点, 主要工作在于寻找大物品集. 它利用了大物品集向下封闭性, 即大物品集的子集必须是大物品集, 它是一种宽度优先算法. 该算法首先计算所有的 1- 项集(m - 项集是含有 m 个项的项集), 记为 C_1; 找出所有满足支持度的常用 1- 项集, 记为 L_1; 接着根据常用 1- 项集确定候选 2- 项集的集合, 记为 C_2; 再从 C_2 中找出所有的常用 2- 项集, 记为 L_2; 然后根据常用 2- 项集确定候选 3- 项集, 如此下去直到不再有候选项集.

Apriori 算法的特点是: 要产生所有频繁集, 需扫描数据库 m 遍, 并进行 m 次迭代, m 为项集格的高. 近来虽不断有一些新的算法对其进行改进(如抽样算法、DIC 算法), 试图解决数据库的搜索空间和搜索效率问题, 但效果并不明显.

3.2 闭项集格的挖掘

挖掘关联规则的过程, 其实质是寻找项目闭集的过程, 从拓扑学的角度看就是寻找不动点. 根据定义 6, 寻找一个闭项目集的过程也是一个伽罗瓦连接的过程[12]. 对于事务型数据库(见表 1, 最小支持度设为 40%), 其概念格如图 1 所示. 其中: $U = \{1, 2, 3, 4, 5\}$, $I = \{A, B, C, D, E\}$, 其挖掘过程如图 2 所示.

表 1 事务数据库

ID	项集
1	AC
2	BCE
3	$ABCE$
4	BE
5	$ABCDE$

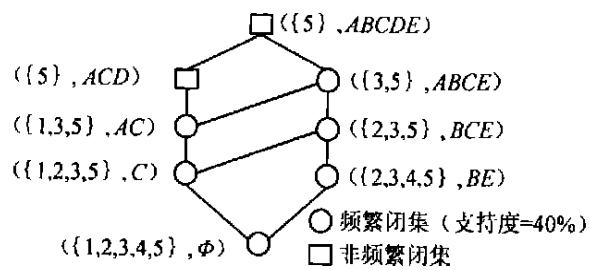

图 1 闭项集格

挖掘原理如下:

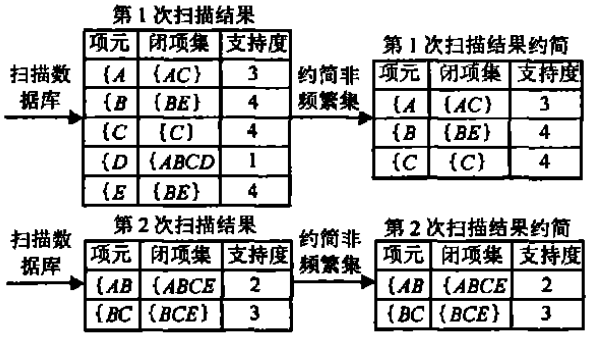

图 2　频繁闭项集的挖掘过程

1) 扫描原始数据库, 求各单项元的闭项集(如 A 的闭项集, 从 $g(A)$ 得到论域{1, 3, 5}, A closure = $f(1) \cap f(3) \cap f(5)$ = {AC}, 这实际上是一个伽罗瓦连接过程), 依据最小支持度约简非频繁项集;

2) 依次递增项元维数, 构造候选项集, 重复步骤 1), 如候选项集包含它的子集(或自身) 的闭集, 则应将其从候选项集中删除, 不再求其闭项集, 如{AC};

3) 当候选项集维数等于原始最大数据库项集维数时, 停止计算.

4　关联规则的生成及决策系统的构建

从表 2 可得到满足最小支持度 40% 的频繁闭项集{AC}, {BE}, {C}, {$ABCE$}, {BCE}, 则可生成如下关联规则:

1) 对于{AC}, 有:

$A \rightarrow C$, 支持度 60%, 可信度 1;

$C \rightarrow A$, 支持度 60%, 可信度 0.75.

2) 对于{$ABCE$}, 有:

$ABC \rightarrow E$, 支持度 40%, 可信度 1

$BCE \rightarrow A$, 支持度 40%, 可信度 0.67.

表 2　最终闭项集

闭项集	支持度
{AC}	3
{BE}	4
{C}	4
{$ABCE$}	2
{BCE}	3

本文根据系统的条件属性和决策属性, 确定采取哪些关联规则来构建决策系统

5　实验数据结果分析及结论

表1中的实验数据表示商场零售"货篮"数据, 对于具有数值特征的特征值数据(如电压、电流、温度等), 需先将其"泛化处理"(如将电压分为高、中、低等) 后才能进行频繁项集的挖掘

本文选取电力系统 SCADA (监控与数据获取) 作为原始数据库. 实验所用的软件开发平台如下: 前台 Delphi6.0, 后台 SQL 7.0, 主机主频 1.1G. 采集参数运行值如下: 短路保护处的电压、电流、自相关函数、互相关函数、付里叶变换系数、谐波分布因子(THD 分布)、0～6 次谐波. 建立了相应的数据库管理系统, 运用基于概念格的频繁闭项集数据挖掘方法对其进行挖掘, 共挖掘出相应关联规则 24 条, 例如:

If $1 * 10^5$V < 短路电压< $2 * 10^5$V, 3 000A < 短路电流< 4 000A or $3 * 10^5$V < 短路电压< $4 * 10^5$V, 1 000A < 短路电流< 2 000A, Then 发生单相近距离接地短路, 应切断主电源

这些关联规则构成了电力动态决策系统的智能部分. 对于具有 1 万条记录的大型数据库, 当最小支持度设定为 40%时, 挖掘出所有频繁闭集只需 1 min 左右. 几种挖掘算法所需时间比较如图 3 所示. 由图 3 可以看出, 基于概念格的频繁闭项集挖掘方法的效率高于其他方法

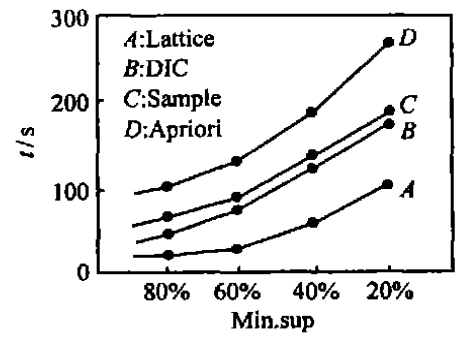

图 3　挖掘时间

关联规则是构建动态决策系统的前提和依据, 而关联规则的获取是通过数据挖掘得到的. 基于概念格的闭项目集几何挖掘算法, 具有几何特征形象化, 扫描数据库次数少, 运算简单, 运算速度快等优点, 适用于大型数据库关联规则的挖掘

参考文献(References):

[1] 黄梯云. 智能决策支持系统[M]. 北京: 电子工业出版社, 2001. 22-24

[2] Clark P, Boswell R. Rule induction with CN2: Some recent improvements[A]. *Lecture Notes in Artificial Intelligence*[C]. Berlin: Springer-Verlag, 1991. 151-163

[3] Altman E I, Haldeman R G, Narayanan P. Zeta analysis[J]. *J of Banking and Finance*, 1997, 6: 29-51.

[4] Ahn, B S, Cho S S, Kim C Y. The integrated methodology of rough set theory and artificial neural network for business failure prediction[J]. *Expert Systems with Application*, 2000, 18: 65-74.

[5] Kun Chang Lee, Jin Sung Kim, Nam Ho Chung, et al. Fuzzy cognitive map approach to web-mining inference amplification [J]. *Expert Systems with Applications*, 2002, 22: 197-211.

[6] Lynn Ling X Li. Knowledge-based problem solving: An approach to health assessment[J]. *Expert Systems with Application*, 1999, 16: 33-42.

[7] Hong T -P, Wang T -T Wang S -L., et al. Learning a coverage set of maximally general fuzzy rules by rough sets[J]. *Expert Systems with Application*, 2000, 19: 97-103.

[8] Krone A, Kiendl H. An evolutionary concept for generating relevant fuzzy rules from data [J]. *Int J of Knowledge Based Intelligent Engineering Systems*, 1997, 1(4): 207-213.

[9] Wang Ching-hung, Hong Tzung-pei, Chang Ming-bao, et al. A coverage-based genetic knowledge-integration strategy [J]. *Expert Systems with Applications*, 2000, 19: 9-17.

[10] 黄天民. 格、序引论及其应用[M]. 成都: 西南交通大学出版社, 1998. 36-48.

[11] 李天瑞. 数据库中的关联规则及挖掘算法研究[D]. 成都: 西南交通大学, 2001. 33-41.

[12] Nicolas Pasquier, Yves Bastide, Rafik Taouil, et al. Efficient mining of association rules using closed itemset lattices[J]. *Information Systems*, 1999, 24(1): 25-46.

[13] Davey B A, Priestley H A. *Introduction to Lattices and Order*[M]. 4th Edition. Cambridge: Cambridge University Press, 1994.

[14] Wille R. Concept lattices and conceptual knowledge systems [J]. *Computers and Mathematics with Applications*, 1992, 23: 493-515.

[15] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. *Proc of the 20th Int Conf on Very Large Data Bases*[C]. Santuago, 1994. 9: 487-499.

(上接第 956 页)

5 结 论

本文针对拥塞控制方案实现中网络时延不能准确预估的具体特点, 结合 Smith 预估原理, 在单瓶颈多通道的网络模型下, 提出一种鲁棒拥塞控制器设计方案. 通过将网络时延误差视为不确定因素, 运用小增益定理来解决鲁棒稳定性问题, 利用网络传输时延最大误差调整参数, 使得在估计误差最差情况下, 保证网络的稳定运行. 同时队列长度能够快速逼近期望值, 从而改善了网络的QoS.

参考文献(References):

[1] 韩兵, 曲润涛, 席裕庚. ATM 网络虚通道资源动态分配的分散控制[J]. 控制与决策, 2000, 15(4): 415-418.
(Han B, Qu R T, Xi Y G. Decentralized control of dynamic resource allocation problem of virtual path in ATM networks[J]. *Control and Decision*, 2000, 15(4): 415-418.)

[2] Mascolo S. Smith's principle for congestion control in high-speed data networks[J]. *IEEE Trans on Automatic Control*, 2000, 45(2): 358-364.

[3] Gomez S F, Fornes J M, Rubio F R. Dead-time compensation for ABR traffic control over ATM networks[J]. *Control Engineering Practice*, 2002, 10(5): 481-491.

[4] 汪小帆, 孙金生, 王执铨. 控制理论在 Internet 拥塞控制中的应用[J]. 控制与决策, 2002, 17(2): 129-134.
(Wang X F, Sun J S, Wang Z Q. Application of control theory to internet congestion control [J]. *Control and Decision*, 2002, 17(2): 129-134.)

[5] Natalie Giroux. The ATM forum technical committee: Traffic management specification version 4.1[EB/OL]. http://www.atmforum.com/standards/approved.html, 1999-03-19.

[6] Raj Jain. Congestion control and traffic management in ATM networks: Recent advances and a survey [J]. *Computer Networks and ISDN System*, 1996, 28(13): 1723-1738.