文章编号: 1001-0920(2005)10-1111-04

基于动态规划方法优化关联规则发现

陈细谦, 迟忠先, 曹秀坤
(大连理工大学 电子与信息工程学院, 辽宁 大连 116024)

摘 要: 为了得到准确可信任的关联规则, 将关联规则的发现归纳为多阶段决策问题, 利用动态规划方法对关联规则发现进行优化分析. 通过条件概率分析, 计算出了动态规划状态转移方程和最优期望代价方程, 并得到了关联规则发现的决策策略. 该策略不需要每一步计算条件概率, 其实现平稳方便. 最后给出了一个应用例子, 并通过模拟实验将该方法与增量关联规则挖掘进行了比较分析, 实验结果证明了该方法的有效性.

关键词: 关联规则; 动态规划; 决策策略
中图分类号: TP311 **文献标识码**: A

Optimal Strategy for Association Rules Mining Based on Dynamic Programming

CHEN Xi-qian, CHI Zhong-xian, CAO Xiu-kun
(Institute of Electronics and Information Engineering, Dalian University of Technology, Dalian 116024, China. Correspondent: CHEN Xi-qian, E-mail: cxqian@163.net)

Abstract: In order to improve the trustiness and accuracy of association rules mining, the association rules mining is reduced to the multi-step decision process, and an optimal strategy is proposed based on dynamic programming. The equation of state and optimal value function used to achieve the optimal strategy is figured out through the analysis of conditional probability of the process. The strategy is realized placidly and easily, without the computation of conditional probability in each step. The experimental results show that the mining based on dynamic programming is superior to traditional incremental updating algorithms for mining association rules.

Key words: Association rules; Dynamic programming; Decision strategy

1 引 言

关联规则是数据挖掘研究中的一个重要的研究课题. 关联规则挖掘的目的是发现大规模数据集中项集之间有趣的关联或相关关系[1]. 国内外对关联规则的发现方法进行了积极深入的研究, 提出了很多算法. 关联规则挖掘算法主要考虑两个描述用户兴趣程度的阈值: 可信度和支持度. 常见的算法包括基于候选集找频繁项集的Apriori算法[2]和基于频繁模式树的FP-growth算法[3]及其扩展.

大多研究都讨论如何在一个给定的固定数据集上挖掘有效的关联规则. 在很多情况下, 关联规则的发现是一个随时间推移的交互过程, 并需要根据环境变化情况调整相应的阈值, 因此研究者提出了关联规则增量更新算法, 如FUP[4], IUA[5]等, 以适应数据集和阈值不断变化的情况. 然而, 为了找到真正令用户感兴趣并信任的规则, 在对不同阶段的数据进行观测时, 需要在观测代价与以较高概率接受正确关联规则之间进行决策. 如在钢铁企业生产过程中, 从某阶段生产数据中挖掘出符合可信度和支持度阈值的关联规则: 当成分C含量达0.03%时, 材质各向异性较优. 显然, 生产决策者并不能根据这一次分析结果进行决策, 需要继续观察几个周期, 判断该关联规则是偶然性的还是必然性的.

本文利用动态规划在多阶段决策问题方面的解决能力[6], 结合条件概率理论, 对关联规则发现进行了优化分析.

收稿日期: 2004-11-03; **修回日期**: 2005-01-24

作者简介: 陈细谦(1976—), 男, 湖北黄石人, 博士生, 从事数据仓库和数据挖掘等研究; 迟忠先(1939—), 男, 山东牟平人, 教授, 博士生导师, 从事数据仓库、数据挖掘和面向对象建模技术等研究.

2 问题定义

将观测过程划分为一个序列 $z_0, z_1, \cdots, z_n$, 它们是独立恒定分布的.

设每一周期开始时, 两类数据的关联 x_i 可能存在两种可能的状态之一: x^1, 两类数据关联; x^2, 两类数据尚未看出关联. 每一周期开始时, 必须作出两个决策之一: u^1, 结束观测分析; u^2, 继续观测分析. 每一周期观测序列 z_i 有两种可能的结果: z^1, 该周期内关联规则成立; z^2, 该周期内关联规则不成立.

假设每周期的分析匹配和观察的代价为 I, 如果数据关联, 则结束分析匹配的代价为 0; 如果数据无关联, 结束分析匹配的代价为 C ($C > 0$). 每一周期的"观测 - 结束"策略, 可表示为自然增长到该周期的观测信息的函数. 本文的目的是确定状态转移方程和衡量策略优劣的指标函数, 最终找到一个最优策略, 使整个观测过程总的期望代价较小.

3 动态规划优化方法

关联规则观测分析过程中, 从一个状态到另一个状态的转移概率为

$$\begin{cases} P(x_{k+1} = x^1 \mid x_k = x^1) = 1, \\ P(x_{k+1} = x^2 \mid x_k = x^1) = 0, \\ P(x_{k+1} = x^1 \mid x_k = x^2) = t, \\ P(x_{k+1} = x^2 \mid x_k = x^2) = 1 - t, \\ 0 < t < 1. \end{cases}$$

观察结果在概率上依赖于系统的如下状态:

$$\begin{cases} P(z_{k+1} = z^1 \mid x_k = x^1) = 1, \\ P(z_{k+1} = z^2 \mid x_k = x^1) = 0, \\ P(z_{k+1} = z^1 \mid x_k = x^2) = r, \\ P(z_{k+1} = z^2 \mid x_k = x^2) = 1 - r, \\ 0 < r < 1. \end{cases}$$

根据贝叶斯规则, 可以得到如下条件概率 P 的演变:

$$p_{k+1} = P(x_{k+1} = x^1 \mid z_0, \cdots, z_{k+1}) = \frac{P(x_{k+1} = x^1, z_{k+1} \mid z_0, \cdots, z_k)}{P(z_{k+1} \mid z_0, \cdots, z_k)} = \frac{P(x_{k+1} = x^1 \mid z_0, \cdots,}{\sum_{i=1}^{2} P(x_{k+1} = x^i \mid z_0, \cdots,} \rightarrow$$

$$\leftarrow \frac{z_k) P(z_{k+1} \mid z_0, \cdots, z_k, x_{k+1} = x^1)}{z_k) P(z_{k+1} \mid z_0, \cdots, z_k, x_{k+1} = x^i)}. \quad (1)$$

同时, 由给定的概率描述, 有

$$P(z_{k+1} \mid z_0, \cdots, z_k, x_{k+1} = x^1) = P(z_{k+1} \mid x_{k+1} = x^1) = \begin{cases} 1, z_{k+1} = z^1; \\ 0, z_{k+1} = z^2. \end{cases} \quad (2)$$

$$P(z_{k+1} \mid z_0, \cdots, z_k, x_{k+1} = x^2) = P(z_{k+1} \mid x_{k+1} = x^2) = \begin{cases} r, z_{k+1} = z^2; \\ 1 - r, z_{k+1} = z^2. \end{cases} \quad (3)$$

$$\begin{cases} P(x_{k+1} = x^1 \mid z_0, \cdots, z_k) = p_k + (1 - p_k) t, \\ P(x_{k+1} = x^2 \mid z_0, \cdots, z_k) = (1 - P_k)(1 - t). \end{cases} \quad (4)$$

将式 (1) ～ (4) 合并, 得到如下递归函数:

$$p_{k+1} = \Phi(p_k, z_{k+1}). \quad (5)$$

其中

$$\Phi(p_k, z_{k+1}) = \begin{cases} \dfrac{p_k + (1 - p_k) t}{p_k + (1 - p_k) t + (1 - p_k)(1 - t) r}, \\ z_{k+1} = z^1; \\ 0, z_{k+1} = z^2. \end{cases}$$

也等价于

$$\Phi(p_k, z_{k+1}) = \begin{cases} \dfrac{1 - (1 - p_k)(1 - t)}{1 - (1 - r)(1 - p_k)(1 - t)}, z_{k+1} = z^1; \\ 0, z_{k+1} = z^2. \end{cases} \quad (6)$$

对于上述函数 Φ, 数据关联的条件概率随着关联规则的不成立降低到零, 随着每一次正确结果而递增. 简单证明如下:

令 $y = p_{k+1} - p_k$, 即

$$y = \frac{1 - (1 - t)(1 - p_k) - [1 - 1 - (1 - r)(1 - \rightarrow}{\leftarrow \dfrac{(1 - r)(1 - t)(1 - p_k)]p_k}{t)(1 - p_k)}} = \frac{(1 - p_k)[(1 - r)p_k + rtp_k + (1 - p_k)t]}{1 - (1 - r)(1 - t)(1 - p_k)}.$$

显然对于上式, 分子、分母均大于零, 即 $p_{k+1} > p_k$. 这样便得到了动态规划算法中的状态转移方程 Φ.

下面考虑该问题的动态规划算法中最优期望代价方程的推导. 在第 N 个周期末, 假定数据关联已继续到该周期, 那么期望代价为

$$J_N(p_N) = (1 - p_N) C.$$

假定在第 N 个周期, 已计算出数据关联的条件概率 P_N. 此时决定是停止关联并达到期望代价 $(1 - p_N) C$, 或继续匹配关联并达到期望代价

$$J_N(p_N) = I + \underset{z_{N+1}}{E} \{J_{N+1}[\Phi(p_N, z_{N+1})]\}.$$

这样便可以得到最优期望代价方程

$$J_N(p_N) = \min[(1 - p_N) C, I + \underset{z_{N+1}}{E} \{J_{N+1}[\Phi(p_N, z_{N+1})]\}]. \quad (7)$$

根据概率定义, 有

$$\begin{cases} P(z_{k+1} = z^2 \mid p_k) = \\ (1 - t)(1 - r)(1 - p_k), \\ P(z_{k+1} = z^1 \mid p_k) = \\ 1 - (1 - t)(1 - r)(1 - p_k). \end{cases} \tag{8}$$

结合式(6), 函数 Φ 和式(8), 有

$$\begin{aligned} & \underset{z_{N+1}}{E}\{J_{N+1}[\Phi(p_N, z_{N+1})]\} = \\ & [1 - (1 - t)(1 - r)(1 - p_N)] \rightarrow \\ & \leftarrow J_{N+1}\left[\frac{1 - (1 - t)(1 - p_N)}{1 - (1 - t)(1 - r)(1 - p_N)}\right] + \\ & (1 - t)(1 - r)(1 - p_N) J_{N+1}(0) = \\ & (1 - t)(1 - p_N) C, \end{aligned}$$

即式(7) 可简化为

$$J_N(p_N) = \min[(1 - p_N)C, I + (1 - t)(1 - p_N)C].$$

当$(1 - p_N)C < [I + (1 - t)(1 - p_N)C]$时, 可以认定无需继续进行数据关联观测. 令

$$y = (1 - p_N)C - [I + (1 - t)(1 - p_N)C], \tag{9}$$

当 $y < 0$ 时, 可以停止数据关联观测.

进一步观察式(9), 得到

$$y = -tCp_N + tC - I,$$

即, 如果 $tC > I$, 令

$$\alpha = 1 - I/(tC), \tag{10}$$

则当 $P_N < \alpha$ 时, $y > 0$, 观测代价及正确接受关联规则的条件概率均未达到最优; $P_N > \alpha$ 时, $y < 0$, 观测代价逐步增大. 因此, 条件概率值 α 是观测期望代价的分界, 决定了最后周期的最优策略:

如果 $P_N \leq \alpha$, 继续关联观测;

如果 $P_N > \alpha$, 则停止关联观测.

可以看到, 该策略比较平稳, 实现方便, 不需要每一周期计算条件概率. 因为通过式(6) 可以得到, 若某一周期内, 关联规则不成立, P_k 降为 0; 若某一周期内, 关联规则成立, 则 $P_{k+1} > P_k$, 条件概率可由下式递归得到:

$$p_1 = \Phi(0, z^1), \cdots, p_{k+1} = \Phi(p_k, z^1). \tag{11}$$

取 N 为 $P_N > \alpha$ 的最小整数, 则可以得到如下简化的决策策略: 如果连续 N 个周期, 观测结果均正确, 则停止观测, 并判定关联规则成立; 否则继续观测.

综上所述, 本文将关联规则的发现归纳为多阶段决策问题, 在对不同阶段的数据进行观测中, 得到了一个有效地在观测代价与以较高概率接受正确关联规则之间进行决策的方法. 其中不同的阶段划分策略对于挖掘结果有着重要的影响. 主要有两种划分策略: 按时间周期进行划分, 如小时、周、月、季度等; 按企业生产周期或事务特征划分阶段. 前者操作简单、容易实施, 但会导致增量数据特征不明显, 延长关联规则观测周期. 而后者通常结合具体的业务环境, 可以使增量数据挖掘更具有针对性.

4　应用例子

在某超市的交易数据集挖掘中发现关联规则:

规则 1: buys(X, "面包") ≥ buys(X, "牛奶");

规则 2: buys(X, "无公害蔬菜") ≥ pays(X, "信用卡").

下面考虑用上述动态规划方法观测这两个规则是否可以被用户接受. 将每一个观测周期定为 7 d, 最大观测周期为 10. 根据经验数据, 设定每周期分析关联规则和观测的代价为 $I = 5$, $t = 0.05$, $r = 0.4$. 如果关联规则不成立, 则结束分析的代价为 $C = 500$.

根据式(10), 可以取 $\alpha = 1 - I/(tC) = 0.8$.

由式(6) 和(11), 可以得到不同周期的条件概率序列, 如表 1 所示.

表 1　条件概率序列

N	P_N
1	0.116 279
2	0.323 336
3	0.581 423
4	0.791 099
5	0.909 888

从计算结果可以看到, P_k 是递增的, 这与前面的证明结果吻合. 当 $k = 5$ 时, 计算得到 $P_k = 0.909\,888 > \alpha = 0.8$. 因此, 如果在观测过程中连续得到 5 个相继正确结果, 就可以停止关联规则观测, 认为关联规则成立; 否则继续观测.

在实际观测过程中, 连续 5 个周期, 规则 2 都成立, 所以判定该关联规则可以信任. 而对于规则 1, 在 10 个观测周期全部结束时, 仍然没有得到连续 5 个正确结果, 因此判定规则 1 不成立. 该结果与超市领导决策层的实践经验吻合, 得到了认可. 因为牛奶销售商不定期举行优惠活动, 购物者此时会批量购进牛奶, 因此规则 1 并不稳定. 而规则 2 则相对稳定, 由于无公害蔬菜价格是普通蔬菜的 5～6 倍, 购物者大多为崇尚健康、生活时尚的白领阶层.

5　比较与分析

增量关联规则的挖掘也讨论了关联规则随时间变化的研究, 本文选择与较为常用的增量关联规则挖掘方法 FUP[4] 进行比较. 动态规划优化过程中

每阶段挖掘采用Apriori算法[2]. 通过模拟实验, 在时间效率和挖掘结果两方面, 对本文所提出的动态规划优化方法与传统的增量关联规则挖掘方法进行了比较

本文采用与文献[2]同样的生成程序得到所需的测试数据(http://www.almaden.ibm.com/software/quest/Resources/datasets/data/assocgen.tar.Z), 其中的有关参数使用文献[2]中同样的记号表示. 这里不考虑其他参数, 主要考虑D(数据记录的数目)和d(增量记录的数目), 测试数据库实例记为Dm.dn, 表示初始数据记录的数目为m, 增量记录的数目为n. 由于最小可信度阈值对性能影响不大, 可不予考虑. 在下面的实验中考虑最小支持度阈值和增量记录数对性能的影响

采用增量挖掘方法与本文方法执行时间之比作为评价指标. 图1给出了分别在测试数据库实例D10k.dx和D100k.dx上使用上述两种方法, 执行时间之比随增量记录数变化的实验结果, 其中最小支持度阈值取固定值2%. 可以看出, 增量挖掘方法执行时间是本文方法的2～7倍. 当增量数据远大于初始数据时, 两者执行时间虽然趋于接近, 但本文方法仍有一定优势.

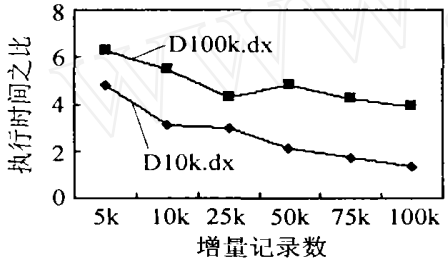

图1 执行时间之比随增量记录数变化

图2给出了在测试数据库实例D10k.d1k和D100k.d10k上执行时间之比随最小支持度阈值变化的实验结果. 本文方法在执行时间上同样具有较大的优势, 差异随着最小支持度阈值的增大而逐步缩小.

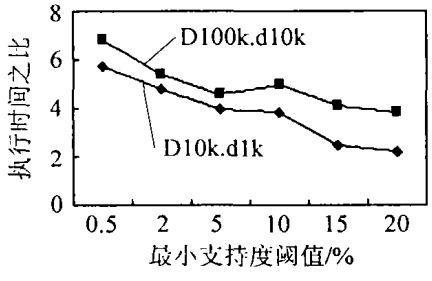

图2 执行时间之比随最小支持度变化

在实验过程中, 作者考察了某一关联规则的挖掘结果随时间变化的情况, 如表2所示. 为了使得描述简洁, 沿用了第4节实例中计算得到的α值.

表2 挖掘结果

阶 段	1	2	3	4	5	6	7
增量挖掘方法	√	-	-	√	√	-	-
动态规划方法	√	-	√	√	√	√	√

可以看到, 基于增量关联规则的挖掘方法受某阶段异常数据(如第2阶段)的影响, 关联规则的发现结果是不稳定的; 而根据动态规划优化方法, 异常阶段将使得观测回到起点, 此阶段后, 连续得到了5个正确的观测结果, 从而可以判断该关联规则可以接受.

通过对以上实验结果进行分析可以得到:

1) 关联规则的增量更新方法, 主要考虑通过充分利用前阶段的挖掘结果高效地生成较小的候选项目集, 但不可避免地要重新计算部分数据, 尤其在最小支持度阈值调整小后, 需要重新计算的记录数更大. 而通过动态规划方法对关联规则进行优化分析过程中, 只关心前一阶段的挖掘结果, 不关心具体数据, 因此在挖掘时间性能上具有较大的优势.

2) 通过将关联规则的发现过程视为多阶段决策问题, 可有效地规避某些阶段内的异常数据, 从而向决策者提供了以较低观测代价和较高概率接受稳定、正确关联规则挖掘结果的有效手段.

本文提出的方法实施方便, 可充分利用现有成熟的挖掘算法, 在各个观测阶段可以根据具体分析环境选用现有合适的挖掘算法. 由于在各个观测阶段执行独立的挖掘, 增量数据需要达到一定的规模, 因此该方法更适用于企业生产数据分析、Web日志分析等.

6 结 语

由于在大多数应用情况下关联规则的发现是一个随时间推移的交互过程, 本文将关联规则的发现归纳为多阶段决策问题. 利用动态规划方法, 通过条件概率分析计算出了状态转移方程和最优期望代价方程, 最终得到了关联规则发现的决策策略. 该策略实现平稳, 不需要每一步计算条件概率, 方便决策者以较低观测代价和较高概率接受正确关联规则.

参考文献(References)

[1] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[A]. *Proc 1993 ACM-SIGMOD Int Conf Management of Data*[C]. Washington, 1993: 207-216.

[2] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[A]. *Proc 1994 Int Conf Very Large Data Base*[C]. Santiago, 1994: 487-499.

(下转第1119页)

表 3　两种算法聚类中心及对应收率

丙烯腈反应器参数和收率	PSO-SOM 算法		基本 SOM 算法	
	优类聚类中心	分类聚类中心	优类聚类中心	分类聚类中心
反应压力/MPa	0.757 9	0.631 3	0.760 4	0.736 6
中段温度/℃	434.040 7	429.553 3	433.674 3	430.145 9
纯丙烯/(Kmol/h)	2 564.1	2 216	2 547.7	2 352.2
空比	9.450 4	7.898	9.456 1	8.993 9
氨比	1.152 1	1.158 5	1.676 7	1.151
触媒量/kg	57.835	54.513 5	56.773	54.493 3
反应线速/(m/s)	0.687 0	0.513 7	0.672 8	0.605 1
收率/%	79.035 5	77.437 9	78.390 7	78.322 0

6 结　论

本文采用近邻粒子群优化算法对权重失真指数进行直接优化, 并在优化过程中对获胜神经元进行直接更新, 网络质量得到提高, 所得聚类效果较好. 因为工业数据往往是非线性的, 所以本文引进了核函数进行数据转换. 转换的方法只适用于少量数据, 对工业的大数据集是不适用的, 但对网络进行训练时, 只需少量可以表征丙稀睛反应器运行状态的样本数据即可, 因此可以采用该转换方法. 虽然采用该方法会增加计算复杂性, 但由于粒子群算法收敛快, 升维所造成的复杂性并不明显. 采用该算法训练后的网络不仅可以根据优类聚类中心指导反应器参数调整, 而且可以监测反应器收率的高低.

参考文献(References)

[1] Guha S, Rastogi R, Shim K. An Efficient Clustering for Large Database[A]. *Proc of the ACM SIGMOD Int Conf on Management of Data*[C]. New York: ACM Prees, 1998: 73-84.

[2] Kohonen T. *Self-organizing Maps*[M]. New York: Springer-Verlag, 1987.

[3] Kohonen T. The Self-organizing Map [J]. *Neurocomputing*, 1998, 21(1): 1-6.

[4] Curry B, Morgan P H. Evaluating Kohonen's Learning Rule: An Approach Through Genetic Algorithms[J]. *European J of Operational Research*, 2004, 154(1): 191-205.

[5] Kennedy J, Eberhart R. Particle Swarm Optimization [A]. *IEEE Int'l Conf on Neural Networks*[C]. Perth Australia, 1995: 1942-1948.

[6] Ioan Cristian Trelea. The Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection [J]. *Information Processing Letters*, 2003, 85(6): 317-325.

[7] Kalyan Veeramacaneni, Thanmaya Peram, Chilukuri Mohan, et al. Optimization Using Particle Swarms with Near Neighbor Interactions[A]. *Proc of Genetic and Evolutionary Computation, Lecture Notes in Computer Science*[C]. Berlin Heidelberg: Springer, 2003, 2723: 110-122.

[8] Kaski S, Logus K. Comparing Self-organizing Maps [A]. *Proc of ICANN 96, Int Conf on Artificial Neural Networks*, Lecture Notes in Computer Science [C]. Berlin, Heidelberg: Springer, 1996, 1112: 809-814.

(上接第 1114 页)

[3] Han J, Pei J. Mining Access Patterns Efficiently from Web Logs [A]. *Proc of Pacific-asia Conf on Knowledge Discovery and Data Mining*[C]. Kyoto, 2000: 396-407.

[4] Cheung D W, Han J, Ng Vincent, et al. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique[A]. *Proc of the 12th Int Conf on Data Engineering*[C]. New Orleans, 1996: 106-114.

[5] 冯玉才, 冯剑琳. 关联规则的增量式更新方法[J]. *软件学报*, 1998, 9(4): 301-306.
(Feng Y C, Feng J L. Incremental Updating Algorithms for Mining Association Rules [J]. *J of Software*, 1998, 9(4): 301-306.)

[6] 麦永浩. *数据仓库和数据挖掘方法研究及其在公安信息建设中的应用*[D]. 上海: 华东理工大学, 2000.
(Mai Y H. *The Research and Application of Data Warehouse and Data Mining Technology in Police Information* [D]. Shanghai: East China University of Science and Technology; 2000.)