智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于可决系数的自适应关联规则挖掘算法

王雪平， 林甲祥， 巫建伟， 高敏节

引用本文:
王雪平, 林甲祥, 巫建伟, 等. 基于可决系数的自适应关联规则挖掘算法[J]. 智能系统学报, 2020, 15(2): 352–359.
WANG Xueping, LIN Jiaxiang, WU Jianwei, et al. Adaptive-association-rule mining algorithm based on determination coefficient[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(2): 352–359.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201809030

DOI: 10.11992/tis.201809030
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190513.1210.002.html

基于可决系数的自适应关联规则挖掘算法

王雪平[1], 林甲祥[1], 巫建伟[2], 高敏节[1]
(1. 福建农林大学 计算机与信息学院, 福建 福州 350002; 2. 自然资源部第三海洋研究所, 福建 厦门 361001)

摘 要: 针对以频繁项集产生−规则产生为核心的两阶段关联规则挖掘, 存在需要人工以先验知识指定最小支持度和最小置信度阈值的缺陷。本文提出以支持数和置信度为依据, 采用曲线拟合技术, 根据可决系数自动确定曲线的次数及对应多项式的算法 AARM_BR(Adaptation Association Rule Mining Based on Determination Coefficient R^2), 从而确定支持度和置信度阈值。在标准数据集 Trolley 和 Groceries 上进行关联规则挖掘实验, 结果表明本算法更具有数据依赖性, 在用户不具备先验知识的情况下, 无须人为指定多项式阶次、支持度和置信度阈值的优点。

关键词: 关联规则; 阶次; 自适应; 可决系数; 规则; 支持度; 置信度; 曲线拟合; 多项式; 数据挖掘
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2020)02−0352−08

中文引用格式: 王雪平, 林甲祥, 巫建伟, 等. 基于可决系数的自适应关联规则挖掘算法 [J]. 智能系统学报, 2020, 15(2): 352–359.
英文引用格式: WANG Xueping, LIN Jiaxiang, WU Jianwei, et al. Adaptive-association-rule mining algorithm based on determination coefficient[J]. CAAI transactions on intelligent systems, 2020, 15(2): 352–359.

Adaptive-association-rule mining algorithm based on determination coefficient

WANG Xueping[1], LIN Jiaxiang[1], WU Jianwei[2], GAO Minjie[1]
(1. College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2. Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361001, China)

Abstract: The two-stage association-rule-mining algorithm based on the frequent item set generation and rule generation requires the manual assigning of minimum support and minimum confidence. To overcome this defect, this paper proposes a new method using the curve fitting technology based on the number of supports and confidence, in which the number of the order of curve and corresponding polynomial is automatically determined by a determination coefficient, which is called "adaptation association rule mining based on the determination coefficient R^2" (AARM_BR). As the proposed AARM_BR method is driven by data, the thresholds of support and confi-dence can be automatically obtained. The experiments on two standard datasets Trolley and Groceries show that compared with a recently published method, the proposed method is more data-dependent and automatically determines the number of order of polynomial and the threshold of support and confidence under the circumstance of not having a priori knowledge.

Keywords: association rule; order; adaptive; coefficient of determination; rule; support; confidence; curve fitting; polynomial; data mining

收稿日期: 2018−09−15. **网络出版日期**: 2019−05−14.
基金项目: 国家自然科学基金项目 (41401458); 福建省自然科学基金项目 (2018J01644, 2018J01645, 2016J01753); 中国−东盟海上合作基金项目 (2020399); 国家海洋局第三海洋研究所项目 (2016020); 福建省中青年教师教育科研项目 (JT180129).
通信作者: 王雪平, E-mail: gggfvgu@163.com.

关联规则挖掘是数据挖掘研究领域重要任务之一, 目标就是从事务数据集中发现隐藏的、有意义的联系, 目前已广泛应用于购物篮分析、网络入侵检测、关联规则分类、交通事故模式分析、药物成分关联分析、病人症型判断等领域[1-3]。常

用关联规则算法有 Apriori、FP-Growth、Magnum-Opus、Closet 等[1, 4], 其中最常用也是最经典的挖掘算法是 Apriori 算法。

为了挖取规模合适的规则, 大部分关联规则算法执行前需用户设置两个阈值: 最小支持度和最小置信度, 以期找到所有超过用户设定阈值的规则。因此, 用户必须具备一定的先验知识才能寻找到合适的最小支持度和最小置信度以便获得有应用价值的规则。但在实际应用过程中, 1) 不同领域数据差异较大, 导致算法在不同的数据集中设置的最小支持数和最小置信度存在较大差异, 没有一个统一的标准; 2) 存在许多非专业用户, 对算法参数的取值具有较大的随意性。因此如何利用数据集本身的特性自动确定阈值而无须先验知识是一个很有意义且亟待解决的问题。本文针对这一问题提出基于可决系数的自适应关联规则挖掘算法, 依据待挖掘数据集中所有项的支持数和所有规则的置信度的数据分布特性, 采用曲线拟合技术, 根据可决系数自动确定拟合多项式, 并在此基础上自动确定具有数据统计依赖意义的最小支持度和置信度, 使其关联规则挖掘的应用门槛降低。

1　相关研究

针对上述所提的参数阈值设置方面问题, 在过去的十多年中, 研究者们从不同角度提出了一些解决方法。一种角度是优化参数或减少参数的方法。例如, Scheffer 提出的预测 Apriori 算法[5], 它自动解决了最小支持度和最小置信度这两个参数之间的平衡问题, 最大限度地提高了对数据集进行精确预测的可能性。该算法利用贝叶斯方法计算了一个称为精确期望预测精度的参数, 以实现精确的预测, 从而提供规则的精确性信息。最后的结果表明, 预测 Apriori 算法的性能优于使用增量因子的 Apriori 型算法。AI-Maqaleh 等[6] 提出了一种有效的置信度综合算法, 在挖掘频繁项集的过程中生成了真正有用的规则。吴瑞华等[7] 提出了一种多重支持度关联规则挖掘算法, 根据不同数据项的特点定义多重支持度, 通过挖掘数据库中的最大频繁项目集, 计算最大频繁候选项目集在数据库中的支持度来发现关联规则, 可解决关联规则中经常出现的稀少数据项问题。陈柳等[8] 提出了一个结合项集相关性的两级置信度阈值设置方法 (PNMC-TWO), 该方法不仅可以更好地确保提取出的关联规则有效和有趣,还可以显著地降低可信度低的关联规则数。于海燕[9] 提出了一种最小相关度优化 PNARC 算法的审计数据关联规则挖掘模型, 提高负关联规则的程度, 减少不相关的关联规则, 然后对最小相关度进行概率分析, 降低无关规则的产生几率。董博等[10] 针对传统关联规则挖掘方法存在计算冗余度过高的问题, 提出一种后处理闭包算子最小单约束的关联规则算法有效降低算法冗余计算, 提高算法计算效率。Li 等[11] 提出了一种新的联想分类器 SigDirect, 利用 Fisher 的精确检验作为一种剪枝策略, 直接挖掘分类关联规则。在没有最小支持度和最小置信度等阈值设置的情况下, SigDirect 能够找到非冗余的分类关联规则, 这些规则在一组先前项和随后的类标签之间表现出统计上的显著依赖性。还有一些学者提出利用智能技术进行关联规则挖掘, 而无需设置最小支持度和最小置信度, 如 Qodmanan 等[12] 利用多目标遗传算法进行 FP-Tree 模式关联规则挖掘, Sarath 等[13] 提出使用二进制粒子群优化策略, 吴琼等[14] 针对量化关联规则的特点, 提出基于多目标烟花算法全面搜索关联规则, Anping 等[15] 提出了基于 Pareto 的多目标二进制 BAT 算法 (MBBA)。Can 等[16] 提出利用引力搜索算法 (GSA) 进行数值型关联规则挖掘。这些基于智能技术, 通过搜索全域空间, 无须设置最小支持度和最小置信度参数即可获取支持度最好、置信度最高的强关联规则, 存在着需要很大计算量的问题。另一角度是利用数据内在的某些特性确定参数的方法。如王志愿等[17] 提出根据项集内部的语义相关度动态确定该项集的最小支持度, 并采用了项集语义相关度的增量计算方法。实验结果表明, DS-Apriori 算法在很大程度上提高了关联规则挖掘算法的效率和效果。Saurav 等[18] 提出了一种基于动态阈值的 FP-生长规则挖掘算法, 该算法将基于加权最短距离的基因表达、甲基化和蛋白−蛋白质相互作用剖面结合起来, 在多视点数据集中寻找不同基因对之间的新关联, 该方法主要优点之一是它考虑了属于每个规则的所有成对基因之间的定量和交互意义。与以往的方法相比, 该算法生成的规则更少, 运行时间也更短, 并且为得到的顶级规则提供了更大的生物学意义, 但该方法是基于生物学基础的统计, 缺乏通用性。Jitendra 等[19] 提出了一种基于项集范围和相关系数的关联规则集粒子群算法 SARIC, 该算法能快速、客观地自动确定支持度和置信度。林甲祥等[20] 提出一种新的自动确定支持度和置信度阈值的方法, 该方法采用数据分布特性进行曲线拟合确定阈值, 为算法的更广泛合

理应用提供了可能。但该方法对数据拟合时采用固定阶次的多项式拟合方式，实际应用中不同的数据往往需要采用不同阶次的多项式拟合。

因此，在文献 [20] 基础上提出一种多项式曲线拟合的阶次自动确定算法 AARM_BR(Adaptation Association Rule Mining Based on Determination Coefficient R^2)，为进一步自动确定支持度和置信度阈值提供更具有数据统计依赖意义的基础。

2　概念术语

定义 1　D 是一个事务数据集，其中每个事务 T 是由一系列具有唯一标识 TID 的项目组成的项集。令 $I=\{i_1, i_2, \cdots, i_m\}$是事务数据集 D 中所有项的集合，每个事务 T 对应 I 上的一个子集，即 $T\subseteq I$, T 可表示为 $T=\{t_1, t_2, \cdots, t_n\}$, $t_i\in I$, n 为自然数 ($n\leqslant m$)，表示事务 T 项集所含的项数。包含 k 个项的项集，称为 k-项集。

定义 2　项集 X 的支持度计数 ($n_{\sup}(X)$) 是指事务数据集 D 中包含项集 X 的事务个数。若项集 $X\subseteq I$, $Y\subseteq I$, 且 $X\cap Y=\Phi$, 则形如 $X\Rightarrow Y$ 的蕴涵式称为关联规则。令事务数据集 D 中事务总个数为 n_{total}。规则 $X\Rightarrow Y$ 的支持度 Support($X\Rightarrow Y$) 是指事务数据集 D 中同时出现 X, Y 的事务占事务数据库的百分比；规则 $X\Rightarrow Y$ 的置信度 Confidence($X\Rightarrow Y$) 是指事务数据集中同时出现 X, Y 的事务数与包含 X 的事务数之比。其计算表达式分别为

$$\text{Support}(X\Rightarrow Y)=\frac{n_{\sup}(X\ U\ Y)}{n_{\text{tatal}}}$$

$$\text{Confidence}(X\Rightarrow Y)=\frac{n_{\sup}(X\ U\ Y)}{n_{\text{tatal}}}$$

定义 3　最小支持度 minSup 和最小置信度 minConf 是用户或专家定义的分别衡量支持度和置信度的两个阈值。最小支持数 (minCount) 是指某个项集 X 为达到最小支持度在事务数据集中至少应出现的次数，等于最小支持度 minSup 乘于所有事务数；频繁项集是指支持度不小于最小支持度 minSup 的项集；所谓强关联规则就是指满足最小支持度 minSup 和最小置信度 minConf 的关联规则。

通常地，给定一个数据库 D，挖掘关联规则的过程可以转换成寻找满足最小支持度 minSup 和最小置信度 minConf 阈值的强关联规则过程，即分解成求解所有频繁项集和由频繁项集产生规则两阶段实现。

3　AARM_BR 的设计与实现

以最经典的关联规则挖掘算法 Apriori 为基础，在文献 [20] 所提基于自适应支持度和置信度的基础上进一步改进，提出一种基于可决系数的自适应关联规则挖掘算法 AARM_BR，该算法为进一步确定更具统计意义的支持度和置信度阈值提供可能。文献 [20] 提出一种支持度和置信度自适应的无参化关联规则挖掘算法 AdapARM，该算法是通过数学方法，拟合频繁项集支持数和规则置信度数据的曲线及其二阶导函数，确定数理意义上最适合的数值作为关联规则挖掘的 minCount 和 minConf 阈值。该文提出的自适应方法很有应用价值，但文中求取自适应支持度和自适应置信度时采用的拟合多项式是固定次数多项式。本文提出基于可决系数的自适应 k 次多项式拟合方法，多项式次数 k 根据自动拟合精度确定，最后以此自适应确定最小支持度和置信度阈值。

3.1　自适应 k 次多项式

首先，根据事务数据集 D 中各项的支持数或某个关联规则集的置信度，按从大到小顺序排序并建立“序−值”队列，如式 (1) 所示。其中，V_t 代表某个项在事务数据集 D 中的支持数或某个规则的置信度。对于支持数，序对的个数 t 等于事务数据集 D 中项的个数；而对于置信度，t 等于所产生的所有规则数目。

$$\{(\text{序列值},\text{序号})\}=\{(y,x)\}=\{(V_1,1),(V_2,2),\cdots,(V_t,t)\} \tag{1}$$

然后，以“序−值”队列中的序号为 x，序列值为 y 建立有序的平面坐标点序列 $(x_i, y_i)(i=1, 2, \cdots, t)$。并采用 k 次多项式进行曲线拟合。拟合的多项式曲线模型为

$$y=f(x)=\sum_{i=0}^{k}a_i x^j$$

式中：k 的取值根据自适应实现，从 2 次开始，每次递增 1，求取对应多项式并判断曲线拟合程度，本算法中选取可决系数R_k^2判断曲线的拟合程度。可决系数R_k^2衡量因变量与自变量关系密切程度的指标，取值范围在 [0, 1] 之间，R_k^2越大代表曲线拟合越好。

本算法拟合结束以相邻两阶可决系数R_k^2的差小于某个值 (设为 Rend) 为终止条件，Rend 由用户指定。多项式次数取拟合结束时的 k 值。自适应 k 次拟合多项式算法的具体流程如图 1 所示，流程中 Rend 取 0.05。

然后，根据确定的 k 次多项式求取拟合曲线的二阶导数 $f''(x)$：

$$y''=f''(x)=\sum_{i=2}^{t}i\cdot(i-1)\cdot a_i\cdot x^{i-2}$$

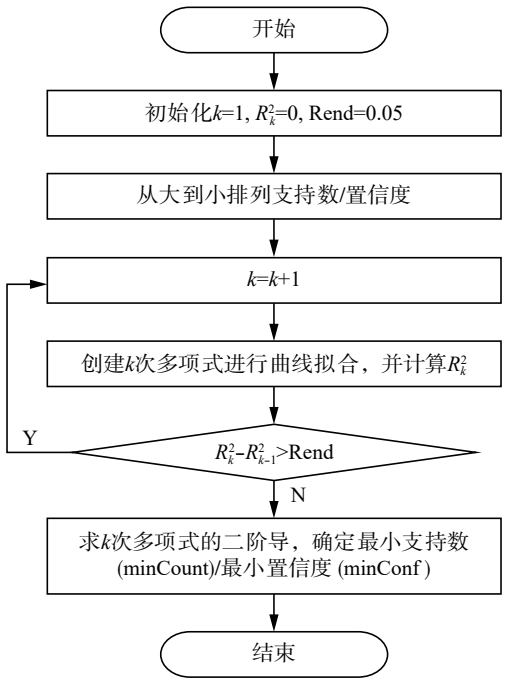

图 1　自适应 *k* 次拟合多项式流程

Fig. 1　The flow of adaptive *k*-order fitting polynomial

最后, 求解拟合曲线中二阶导数 $f''(x)=0$ 的点 (记 x_0) 及对应的函数值 $f(x_0)$, x_0 下取整作为最小支持数阈值 minCount, $f(x_0)$ 作为最小置信度阈值 minConf。

3.2　AARM_BR 算法实现

基于可决系数自适应阶次的多项式曲线拟合模型下确定最小支持数和置信度阈值的关联规则挖掘算法 AARM_BR 的核心流程如图 2 所示。

AARM_BR 算法描述如下:

输入　事务数据集 D, 拟合结束条件 Rend

输出　所有强规则 SR, SR={SR_1, SR_2, ⋯, SR_r}

1) C_1 = find_candidate_1-itemsets(D);

2) C'_1= sort_InDescOrder(C_1); //降序排序

//自适应确定多项式拟合曲线的次数 k;

3) k=find_k_polynomial_curve_fitting(C'_1); //产生支持数 k 次多项式;

4) $f(x)$ = polynomial_curve_fitting(C'_1,k);

5) find x_0, where $f'((x_0)=0$;

6) minCount =int(x_0); //x_0 下取整并赋给 minCount;

7) {L_1, L_2, ⋯, L_k}= find_all_frequent_k-itemsets(D, minCount);

8) {R_1, R_2, ⋯, R_t}= generateRule_from_frequent_k-itemsets(L_1, L_2, ⋯, L_k); //对所有规则按置信度降序排序;

9) R'= sort_InDescOrder(R_1, R_2, ⋯, R_t); //自适应确定多项式拟合曲线的次数 k

10) k=find_k_polynomial_curve_fitting(R'); //产生 k 次多项式

11) $h(x)$ = polynomial_curve_fitting(R',k);

12) find x_0, where h''(x_0)=0;

13) minConf=$h(x_0)$; //得到最小置信度;

14) {SR_1, SR_2, ⋯, SR_r} = find_strong_rules({R_1, R_2, ⋯, R_t}, minConf);

15) Return {SR_1, SR_2, ⋯, SR_r} //得到所有强规则。

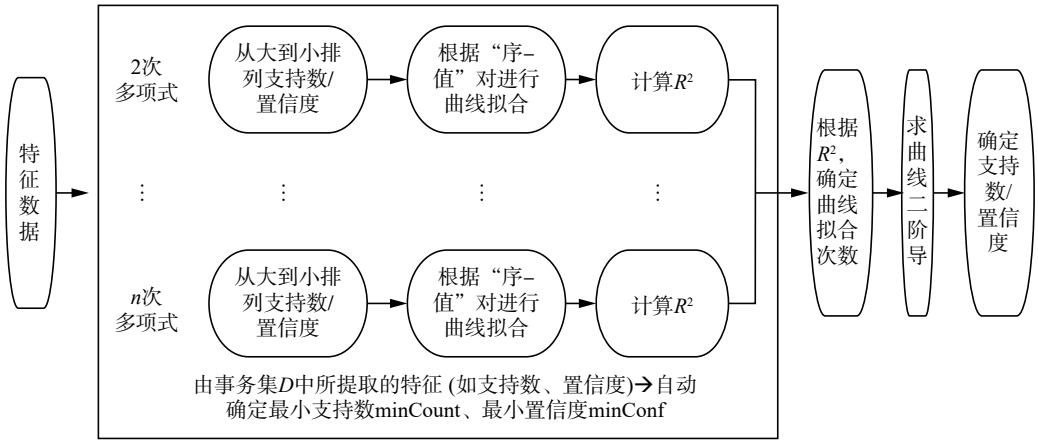

图 2　AARM_BR 核心流程

Fig. 2　Key process of AARM_BR

4　实验与分析

为了比较分析, 选取文献 [12] 使用的数据集进行实验。具体数据集为关联规则挖掘购物车 Trolley 数据集和开源软件 R GUI 里的 Groceries 数据集。Trolley 数据集总共有 9 条消费记录 (即 9 行), 包含 7 种不同商品; Groceries 数据集有 9 835 条消费记录 (即 9 835 行), 包含 169 种不同商品。下面对自适应 k 次多项式的挖掘流程进行

介绍，并对挖掘结果进行分析和讨论。

4.1　不同阶次拟合对比

首先，分析不同阶次多项式对自动确定支持度阈值的影响。下面给出 Trolley 数据集和 Groceries 数据集两个数据集在不同阶次下曲线拟合得到的 minCount，分别如表 1、2 所示。

表 1　Trolley 数据集中不同 k 下的 minCount

Table 1　MinCount of Trolley datasets under different k

k	拟合曲线多项式	R_k^2	minCount
3	$y=0.083\ 3x^3-1.142\ 9x^2+3.559\ 5x+4.285\ 7$	0.975 4	4
4	$y=0.041\ 7x^4-0.583\ 3x^3+2.458\ 3x^2-3.916\ 7x+9$	1	1

表 2　Groceries 数据集中不同 k 下的 minCount

Table 2　MinCount of Groceries datasets under different k

k	拟合曲线多项式	R_k^2	minCount
3	$y=-0.001\ 3x^3+0.413\ 5x^2-42.763x+1\ 486.1$	0.887	107
4	$y=2\times10^{-5}x^4-0.008\ 1x^3+1.163x^2-71.273x+1\ 734.8$	0.929 8	77
5	$y=-4\times10^{-7}x^5+0.000\ 2x^4-0.031\ 3x^3+2.645\ 7x^2-107.71x+1\ 949.8$	0.954 5	57

Groceries 数据集下确定 minCount 的不同阶次的多项式拟合曲线如图 3 所示。

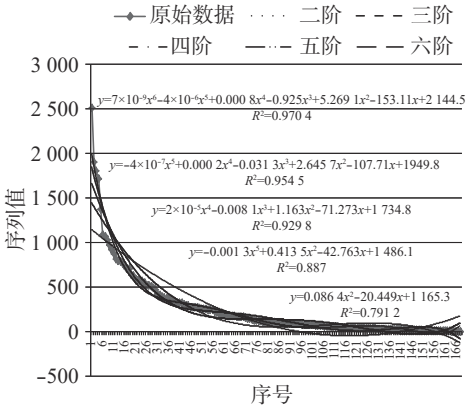

图 3　Groceries 中不同阶次多项式拟合曲线

Fig. 3　Polynomial fitting curve of Groceries under different order

从表 1 和表 2 及图 3 可以看出：1) 在相同的结束条件下，不同数据集拟合采用的多项式次数不一定相同，如本实验中数据集 Groceries 下确定最小支持度时 k 取 4 次，而 Trolley 数据集下 k 取 3 次；2) 同一数据集下，拟合选取不同次数的多项式，会影响 minCount 的值，从而也会影响关联规则挖掘的效果。这说明不同的数据集下，不能采取统一阶次的多项次拟合，多项式拟合的阶次应该根据数据拟合的实际情况自适应确定。

其次，分析不同阶次多项式拟合对自动确定置信度阈值的影响。对于 Trolley 数据集和 Groceries 数据集在上步 minCount 约束下所得规则的基础上，自适应拟合确定最小置信度 minConf，得到如表 3 和表 4 所示结果。

表 3　Trolley 数据集中不同 k 下的最小置信度

Table 3　MinConf of Trolley datasets under different k

k	多项式拟合曲线公式	R_k^2	minConf
3	$y=-6\times10^{-6}x^3+0.000\ 4x^2-0.019\ 9x+0.935\ 8$	0.933 2	0.612
4	$y=4\times10^{-6}x^4-0.000\ 3x^3+0.005\ 5x^2-0.056\ 5x+1.000\ 7$	0.952 7	0.762

表 4　Groceries 数据集中不同 k 下的最小置信度

Table 4　MinConf of Groceries datasets under different k

k	多项式拟合曲线公式	R_k^2	minConf
3	$y=-7\times10^{-10}x^3+2\times10^{-6}x^2-0.001\ 4x+0.543\ 9$	0.995 4	0.094
4	$y=8\times10^{-13}x^4-3\times10^{-9}x^3+3\times10^{-6}x^2-0.001\ 8x+0.564\ 7$	0.998 1	0.116
5	$y=-1\times10^{-15}x^5+4\times10^{-12}x^4-6\times10^{-9}x^3+4\times10^{-6}x^2-0.002x+0.573\ 5$	0.998 5	0.065

同理，不同数据集中根据自动确定多项式次数是不完全相同的，如本实验拟合精度差 Rend 取 0.05 时，数据集 Trolley 和数据集 Groceries 的多项式次数都是 3 次；而若拟合精度差 Rend 取 0.02 时，数据集 Trolley 下多项式次数为 4 次，而数据集 Groceries 的多项式次数为 3 次。同一数据集下不同阶次的多项式下得到的 minConf 也不尽相同，如 Trolley 数据集下 k 取 3 次时得到 min-

Conf 为 0.612, 而 k 取 4 次时得到 minConf 为 0.762; Groceries 数据集下 k 取 3 次时得到 minConf 为 0.094, 而 k 取 4 次时得到 minConf 为 0.116。综合上述两部分分析说明多项式的阶次会影响阈值, 从而也会影响挖掘结果。如何自适应确定多项式阶次是很有必要的。

4.2　结果对比与分析

根据挖掘的流程, 首先选取待挖掘数据集中各事务项支持数作为特征数据并按支持数大小进行降序排序建立“序–值”对序列。

其次, 采用文中所提的自动阶次拟合方法建立“序–值”对序列的 k 次多项式曲线拟合模型, 自动确定 k 的次数及对应的多项式。k 从 2 开始, 以相邻两阶精度差 R^2 小于 Rend 为结束条件, 本实验以 Rend 取 0.05 为例。本步实验结果与文献 [12] 所提的 AdapARM 算法比较如表 5 所示。

然后, 根据 k 阶多项式二阶导函数求取最小支持数, 得到对应的二阶导函数及最小支持数 minCount 如表 6 所示。

按照 Apriori 算法思想, 在上步求得的最小支持数基础上, 从一阶频繁项开始逐层向上, 获取所有 k 阶频繁项集, 并根据频繁项集产生关联规则, 得到如表 7 结果。

表 5　k 次拟合多项式比较 (支持数)

Table 5　Comparison of k order polynomial fitting curves (support number)

数据集	算法	k	确定的拟合多项式(支持数)
Trolley	AARM_BR	3	$f_T(x)=0.083\ 3x^3-1.142\ 9x^2+3.559\ 5x+4.285\ 7$
	AdapARM	3	$f_T(x)=0.083\ 3x^3-1.142\ 9x^2+3.559\ 5x+4.285\ 7$
Groceries	AARM_BR	4	$f_G(x)=2\times10^{-5}x^4-0.008\ 1x^3+1.163x^2-71.273x+1\ 734.8$
	AdapARM	3	$f_G(x)=-7\times10^{-10}x^3+2\times10^{-6}x^2-0.001\ 4x+0.543\ 9$

表 6　最小支持数比较

Table 6　Comparison of Minimum support number

数据集	算法	二阶导函数(支持数)	minCount
Trolley	AARM_BR	$f''_T(x)=-2.285\ 7+0.5x$	4
	AdapARM	$f''_T(x)=-2.285\ 7+0.5x$	4
Groceries	AARM_BR	$f''_G(x)=2.326\ 1-0.048\ 8\ x+2.415\ 8\times10^{-4}x^2$	77
	AdapARM	$f''_G(x)=0.827\ 1-0.007\ 7\ x$	107

表 7　产生的规则数目比较

Table 7　Comparison of the produced rules number

数据集	算法	求得的关联规则数目
Trolley	AARM_BR	30
	AdapARM	30
Groceries	AARM_BR	1 146
	AdapARM	498

根据数据集 Trolley 和数据集 Groceries 产生的关联规则的置信度从大到小排序并进行 k 次多项式曲线拟合, k 从 2 开始, 以相邻两阶精度差 R^2 小于 Rend 为结束条件, 本实验以 Rend 取 0.05 为例。在本步中, 两个方法得到的阶次 k 均为 3, 确定的拟合多项式如表 8 所示。

根据 k 次多项式二阶导函数求取最小置信度, $h_T(x)$ 和 $h_G(x)$ 的二阶导函数 $h''_T(x)$ 和 $h''_G(x)$ 公式及对应的最小置信度分别如表 9 所示。

表 8　k 次拟合多项式比较 (置信度)

Table 8　Comparison of k order polynomial fitting curves (confidence)

数据集	算法	k	确定的拟合多项式(置信度)
Trolley	AARM_BR	3	$h_T(x)=-6\times10^{-6}x^3+0.000\ 4x^2-0.019\ 9x+0.935\ 8$
	AdapARM	3	$h_T(x)=-6\times10^{-6}x^3+0.000\ 4x^2-0.019\ 9x+0.935\ 8$
Groceries	AARM_BR	3	$h_G(x)=6.774\ 4\times10^{-10}x^3+1.598\ 8\times10^{-6}x^2-0.001\ 4x+0.543\ 9$
	AdapARM	3	$h_G(x)=-6.426\ 3\times10^{-9}x^3+6.702\ 3\times10^{-6}x^2-0.002\ 7x+0.523$

表 9　最小置信度比较

Table 9　Comparison of minimum confidence

数据集	算法	二阶导函数(置信度)	minConf
Trolley	AARM_BR	$h''_T(x)=8.010\ 1\times10^{-4}-3.371\ 0\times10^{-5}x$	0.612 637 3
	AdapARM	$h''_T(x)=8.010\ 1\times10^{-4}-3.371\ 0\times10^{-5}x$	0.612 637 3
Groceries	AARM_BR	$h''_G(x)=3.197\ 6\times10^{-6}-4.064\ 6\times10^{-9}x$	0.099 430 9
	AdapARM	$h''_G(x)=1.340\ 5\times10^{-5}-3.855\ 8\times10^{-8}x$	0.115 844 4

最后，根据最小置信度确定强关联规则，得到结果如表 10 所示。

上述算法挖掘结果比较可以看出，自适应多项式挖掘的结果与人为确定多项式次数的挖掘结果不一定相同，如 Groceries 数据集下人为确定次数有可能遗漏一些重要的规则。根据数据本身的特性确定多项式的拟合次数算法，不需要用户具备先验知识，在不指定多项式阶数、不指定最小支持度和最小置信度阈值的情况下，自动获取数据统计意义下的强关联规则。自适应多项式曲线拟合方法为支持度和置信度的自动确定提供了更具数据依赖意义的解决方案。

表 10　强关联规则数目比较

Table 10　Comparison of the strong association rules number

数据集	算法	强关联规则数目
Trolley	AARM_BR	22
	AdapARM	22
Groceries	AARM_BR	786
	AdapARM	339

从时间复杂度角度分析，本算法只在经典算法 Aprori 算法的基础上增加了两个步骤：排序和多项式阶次自动确定，其中排序的时间复杂度为 $O(n\log n)$，多项式阶次的自动确定时间复杂度为 $O(kn)$；与 AdapARM 算法比较，只多了一步自动确定多项式阶次的单层循环。这里增加的时间在整个算法中占用很小，可忽略不计，同时对自动确定最小支持度和最小置信度具有指导意义。

5　结束语

文中提出一种基于可决系数的数据自适应多项式拟合曲线确定支持度和置信度阈值的关联规则挖掘算法 AARM_BR。以曲线拟合的精确程度 R^2 为判断依据，自适应确定多项式拟合曲线的次数及多项式，在此基础上求取 k 次拟合多项式的二阶导函数为零的点 x_0 及其函数值 $f(x_0)$，作为支持数和置信度阈值，进而获取数据依赖意义下最小支持数和最小置信度及其强关联规则。该方法根据数据自身特点，在用户不具备经验知识、不指定支持度和置信度阈值的情况下，自动确定拟合曲线、最小支持度和最小置信度阈值。在两个标准数据集 Trolley 和 Groceries 上的实验结果和分析表明，该方法对关联规则的进一步推广应用具有一定价值。

参考文献：

[1] MALIK M, MAMTA, AGARWAL R P. A survey on association rule mining[J]. International journal of research in engineering and applied sciences, 2015, 5(6): 48–56.

[2] XI Jianfeng, ZHAO Zhonghao, LI Wei, et al. A traffic accident causation analysis method based on AHP-apriori[J]. Procedia engineering, 2016, 137: 680–687.

[3] ALWIDIAN J, HAMMO B H, OBEID N. WCBa: weighted classification based on association rules algorithm for breast cancer disease[J]. Applied soft computing, 2018, 62: 536–549.

[4] 张良均, 杨坦, 肖刚, 等. MATLAB 数据分析与挖掘实战 [M]. 北京: 机械工业出版社, 2016.

[5] SCHEFFER T. Finding association rules that trade support optimally against confidence[C]//European Conference on Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg, 2001: 424–435.

[6] AL-MAQALEH B M, SHAAB S K. Efficient algorithm for mining association rules using confident frequent itemsets[C]//Third International Conference on Advanced Computing and Communication Technologies. Rohtak, India, 2013.

[7] 吴华瑞, 张凤霞, 赵春江. 一种多重最小支持度关联规则挖掘算法 [J]. 哈尔滨工业大学学报, 2008, 40(9): 1447–1451.
WU Huarui, ZHANG Fengxia, ZHAO Chunjiang. An algorithm for mining association rules with multiple minimum supports[J]. Journal of Harbin Institute of Technology, 2008, 40(9): 1447–1451.

[8] 陈柳, 冯山. 正负关联规则两级置信度阈值设置方法 [J].

计算机应用, 2018, 38(5): 1315–1319, 1338.
CHEN Liu, FENG Shan. Two-level confidence threshold setting method for positive and negative association rules[J]. Journal of computer applications, 2018, 38(5): 1315–1319, 1338.

[9] 于海燕. 最小相关度优化 PNARC 算法的审计数据关联规则挖掘模型 [J]. 科技通报, 2017, 33(12): 158–161.
YU Haiyan. Research on audit data association rule mining model with minimal relevance optimized PNARC algorithm[J]. Bulletin of science and technology, 2017, 33(12): 158–161.

[10] 董博, 王雪. 关联规则算法的计算效率优化研究 [J]. 计算机仿真, 2017, 34(9): 247–253.
DONG Bo, WANG Xue. Closure operator based post processing minimum single constraint association[J]. Computer simulation, 2017, 34(9): 247–253.

[11] LI Jundong, ZAIANE O. Exploiting statistically significant dependent rules for associative classification[J]. Intelligent data analysis, 2017, 21(5): 1155–1172.

[12] Qodmanan H R, Nasiri M, Minaei-Bidgoli B. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence[J]. Expert systems with applications, 2011, 38(1): 288–298.

[13] SARATH K N V D, RAVI V. Association rule mining using binary particle swarm optimization[J]. Engineering applications of artificial intelligence, 2013, 26(8): 1832–1840.

[14] 吴琼, 曾庆鹏. 基于多目标烟花算法的关联规则挖掘 [J]. 模式识别与人工智能, 2017, 30(4): 365–376.
WU Qiong, ZENG Qingpeng. Association rules mining based on multi-objective fireworks optimization algorithm[J]. Pattern recognition and artificial intelligence, 2017, 30(4): 365–376.

[15] A. S. 1, X. D. 1, J. C. 2, et al. Multi-objective associati-on rule mining with binary bat algorithm[M]. School of Computer Engineering and Science, Shanghai University, Shanghai, China. Yale Stem Cell Center and Department of Cell Biology, Yale University School of Medicine, New Haven, USA, 2016: 105–128.

[16] CAN U, ALATAS B. Automatic mining of quantitative association rules with gravitational search algorithm[J]. International journal of software engineering and knowledge engineering, 2017, 27(3): 343–372.

[17] 王志愿, 夏士雄, 张磊, 等. 语义驱动的关联规则挖掘算法研究 [J]. 计算机工程与设计, 2011, 32(3): 936–939, 944.
WANG Zhiyuan, XIA Shixiong, ZHANG Lei, et al. Study on semantic-driven association rule mining algorithm[J]. Computer engineering and design, 2011, 32(3): 936–939, 944.

[18] MALLIK S, BHADRA T, MUKHERJI A. DTFP-growth: dynamic threshold-based FP-growth rule mining algorithm through integrating gene expression, methylation, and protein-protein interaction profiles[J]. IEEE transactions on nanobioscience, 2018, 17(2): 117–125.

[19] AGRAWAL J, AGRAWAL S, SINGHAI A, et al. SET-PSO-based approach for mining positive and negative association rules[J]. Knowledge and information systems, 2015, 45(2): 453–471.

[20] 林甲祥, 巫建伟, 陈崇成, 等. 支持度和置信度自适应的关联规则挖掘 [J]. 计算机工程与设计, 2018, 39(12): 3746–3754.
LIN Jiaxiang, WU Jianwei, CHEN Chongcheng, et al. Association rule mining algorithm with adaptive support and confidence[J]. Computer engineering and design, 2018, 39(12): 3746–3754.

作者简介:

王雪平, 讲师, 主要研究方向为数据挖掘、模式识别。主持省级科研项目 1 项, 参与省级科研项目 10 余项。

林甲祥, 博士。主要研究方向为空间数据挖掘、人工智能和大数据。主持国家级和省部级科研项目 4 项, 参与省部级科研项目 20 余项; 获福建省科学技术奖二等奖 1 项, 获国家发明专利授权 2 项, 获国家计算机软件著作权登记 5 项。发表学术论文 40 余篇。

巫建伟, 工程师, 博士。主要研究方向为海洋环境管理信息系统、空间数据挖掘、海洋大数据分析。主持或参与国家级和省部级科研项目 10 余项。发表学术论文 10 余篇。