支持度和置信度自适应的关联规则挖掘

林甲祥[1]，巫建伟[2]，陈崇成[3]，张泽均[1]，舒兆港[1]

（1. 福建农林大学 计算机与信息学院，福建 福州 350002；2. 国家海洋局第三海洋研究所海洋环境管理与发展战略研究中心，福建 厦门 361001；3. 福州大学 空间数据挖掘与信息共享教育部重点实验室，福建 福州 350116）

摘 要：以事务集中所有项的支持数和所有规则的置信度数据为依据，使用统计拟合技术，对支持度和置信度阈值的自动化确定进行研究，提出支持度和置信度自适应的关联规则挖掘算法 AdapARM（adaptive association rule mining），降低算法对先验知识的依赖性。在标准数据集 Trolley 和 Groceries 上进行实验研究，实验结果及其分析验证了 AdapARM 算法的有效性，其具有用户不必具备数据集的先验知识、不需人为设定支持度和置信度参数的优点。

关键词：参数自适应；关联规则；支持度；置信度；频繁项集

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2018）12-3746-09

doi：10.16208/j.issn1000-7024.2018.12.026

Association rule mining algorithm with adaptive support and confidence

LIN Jia-xiang[1], WU Jian-wei[2], CHEN Chong-cheng[3], ZHANG Ze-jun[1], SHU Zhao-gang[1]

(1. College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2. Research Center for Marine Environment Administration and Development Strategy, Third Institute of Oceanography, State Oceanic Administration, Xiamen 361001, China; 3. Key Lab of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou 350116, China)

Abstract: An association rule mining algorithm with adaptive support and confidence (adaptive association rule mining, AdapARM) was proposed. The mimimum support and the mimimum confidence values were acquired on the basis of statistical fitting technology, which was imposed on the support count data of all items in the transaction dataset and the confidence data of all the rules generated from frequent itemsets. Hence, the parameter values were changed and adaptive to the transaction dataset itself. Experimental research was performed on two standard datasets including Trolley and Groceries, to verify the effectiveness of AdapARM, and it is found that AdapARM has the advantage that users do not need to have any prior knowledge about the dataset, and users do not need to set the parameter values of mimimum support and mimimum confidence to conduct association rule mining tasks.

Key words: parameter adaptive; association rule; support; confidence; frequent itemset

0 引 言

关联规则是数据挖掘一个重要研究主题，在购物篮分析、医疗诊断、股票存货分析、Web 日志挖掘、客户市场分析、生物信息资讯等领域有着广泛的应用[1]。文献中提出的大多数关联规则挖掘方法采用两步骤的策略，首先从数据集中寻找所有频繁项集，然后由这些频繁项集生成强关联规则，算法的原理简单且易于实现，其中第一步频繁项集的生成是算法的关键，典型算法如 Apriori、FP-tree 等[2]。

收稿日期：2018-04-18；**修订日期**：2018-05-24

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、2016J05148、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基金项目（2016020）；福建省科技计划重点基金项目（2015H0015）

作者简介：林甲祥（1982-），男，福建泉州人，博士，讲师，CCF 会员，研究方向为数据挖掘和云计算；巫建伟（1984-），男，福建泉州人，博士，助理研究员，研究方向为专家系统和本体知识库；陈崇成（1968-），男，福建福州人，博士，教授，研究方向为空间数据挖掘和地理知识服务；张泽均（1984-），男，贵州遵义人，博士，讲师，CCF 会员，研究方向为图像分析与处理；舒兆港（1981-），男，江西上饶人，博士，副教授，研究方向为云计算与网络安全。E-mail：linjx@fafu.edu.cn

以频繁项集搜索为核心的关联规则挖掘算法，在支持度的约束下进行频繁项集搜索，在置信度的约束下进行强关联规则确认。目前，许多关联规则挖掘实践中，最小支持度（*minsup*）和最小置信度（*minconf*）参数的取值方法仍是由用户根据先验知识进行人为的指定，一方面，非专业关联规则算法用户，对算法参数的取值一般具有较大的主观随意性、往往很少考虑参数取值的科学性；另一方面，挖掘模型中支持度和置信度阈值的不同取值，对挖掘过程中频繁项集的大小、自连接产生候选项集的规模、挖掘结果中关联规则的数目有着显著的影响。从算法设计的角度实现关联规则支持度和置信度参数的科学化和自适应化取值，文献中还鲜有相关的研究成果报道。因此，拟在分析事务数据集中所有项的支持数和所有规则的置信度的基础上，使用统计拟合技术，对支持度和置信度阈值的自动化确定技术进行研究，提出基于支持度和置信度自适应技术的无参化关联规则挖掘解决方案，以期降低关联规则挖掘行业化应用的门槛。

1 相关研究进展

近年来，研究人员在提高关联规则计算效率、寻找更紧致数据结构、处理的数据类型扩展等方面做了大量的研究。典型工作如：增量式关联规则挖掘方面，Nath 等[3]对增量式关联规则挖掘算法进行了综述。基于特定数据结构的关联规则挖掘方面，Vo 等[4]对基于 N-list 和包含概念的频繁项集挖掘 PrePost 算法进行研究，在 N-lists 创建和交互中融合了 Hash 技术，对算法进行了改进。此后，Huong 等[5]对基于 N-list 和考虑权重的频繁项集挖掘算法进行了研究。Maria Luna 等[6]提出了一种称为倒排索引压缩（inverted index compression）的数据结构，能够用于很多现有的关联规则挖掘算法，用于提高算法的效率。Anand 等[7]为提高关联规则挖掘的效率，提出了基于横二叉树（Treap）数据结构的关联规则挖掘算法。牛新征等[8]提出了一种基于 FP-tree 的快速关联规则隐藏算法，避免了遍历原始数据集产生的大量 I/O 时间，减少了关联规则隐藏处理对原始数据集的影响。分布式 & 并行关联规则挖掘方面，Wang 等[9]使用微处理器技术对关联规则的并行化和效率提升进行了研究。Vu 等[10]针对频繁模式搜索，开展了多核共享存储器的并行关联规则挖掘研究。Liu 等[11]面向大数据，提出了一套基于 MapReduce 和最大频繁项集的启发式多流程关联挖掘解决方案。基于智能技术的关联规则挖掘方面，Sohrabi 等[12]对基于元胞自动机的频繁模式挖掘进行了研究，并与 Apriori、FP-Growth、BitTable 等知名算法的效率进行了对比分析。Zou 等[13]提出了一种基于模糊概念格的关联规则挖掘及其动态更新算法。Heraguemi 等[14]将使用多重合作策略的蜂窝算法用于关联规则挖掘。基于关联规则的应用研究方面，Tseng 等[15]对事务数据库中高效用项集的关联规则挖掘进行研究，提出了效用模式增长算法 UP-Growth 及其改进算法 UP-Growth(+)。Leung 等[16]针对大数据环境下的 Web 页面推荐，提出了一套按位（bitwise）并行关联规则挖掘方法。何明等[17]为了提高个性化推荐效率和推荐质量，平衡冷门与热门数据推荐权重，提出了优化 Apriori 算法且适合不同测评标准值的 k 前项频繁项集挖掘算法。

在关联规则算法参数取值研究方面，Sarath 等[18]使用二进制粒子群优化策略，对关联规则挖掘进行研究，无须指定最小支持度和最小置信度参数，只需指定要获得的规则的数目。虽然，一些基于智能技术的关联规则挖掘，通过问题解空间的全域搜索，无须设置最小支持度和最小置信度参数的值，即可获得支持度最高、置信度最高的若干强关联规则，但需要很大的计算量。对于关联规则挖掘算法参数的自适应取值研究，文献中未见相关成果报道。

2 核心概念与挖掘流程

2.1 相关概念

令交易事务数据库中所有商品种类的集合为 $I=\{i_1,i_2,\cdots,i_m\}$，其中 m 为自然数，表示数据集中不同商品的数目。D 为规模为 N 的事务集，即数据库中交易记录的集合，N 为交易的数量（记录数），其中每一个事务 t 是一个项集，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 中商品的数目（项数），即 $t\subseteq I$。对于项集 X，若 X 中含有的项数为 k，则称 X 为 k-项集；若 $X\subseteq t$，则称事务 t 包含 I 的一个子集 X。

事务集 D 中的关联规则是一种由支持度（*support*）和置信度（*confidence*）约束的蕴含式 $X\Rightarrow Y$，其中 $X\subset I$，$Y\subset I$ 且 $X\cap Y=\varnothing$，支持度表示规则的频度，置信度表示规则的强度。

在事务集 D 中，规则 $X\Rightarrow Y$ 的支持度是 D 中同时包含 X 和 Y 的交易数与所有交易数之比，记为 $support(X\Rightarrow Y)$，如式（1）所示，$support(X\Rightarrow Y)\in[0,1]$。其中符号 $\|S\|$ 为模，表示集合 S 中元素的个数

$$support(X\Rightarrow Y)=\frac{\|\{T\mid X\cup Y\subseteq T,\ T\in D\}\|}{\|D\|} \tag{1}$$

在事务集 D 中，项集 X 的支持数是 D 中包含 X 的交易数，记为 $support\_count(X)$，如式（2）所示，$support\_count(X)\in\{0,1,\cdots,N\}$，即 X 的支持数是不大于交易记录数 N 的一个自然数

$$support\_count(X)=\|\{T\mid X\subseteq T,\ T\in D\}\| \tag{2}$$

在事务集 D 中，规则 $X\Rightarrow Y$ 的置信度是 D 中同时包含 X 和 Y 的交易数与包含 X 的交易数之比，记为 $confidence(X\Rightarrow Y)$，如式（3）所示，$confidence(X\Rightarrow Y)\in[0,1]$

$$confidence(X\Rightarrow Y)=\frac{\|\{T\mid X\cup Y\subseteq T,\ T\in D\}\|}{\|\{T\mid X\subseteq T,\ T\in D\}\|}=\frac{support\_count(X\cup Y)}{support\_count(X)} \tag{3}$$

对给定的事务集 D，关联规则挖掘的任务是产生所有支持度和置信度大于用户给定的最小支持度（$minsup$）和最小置信度（$minconf$）的关联规则。最小支持数（$mincount$）是指某个项集 X 为达到最小支持度 $minsup$，在事务集 D 中至少应出现的次数（频数），即 $mincount=\lceil\|D\|\times minsup\rceil$，其中符号$\lceil\ \rceil$表示向上取整。关联规则挖掘实践中，那些支持度和置信度分别大于 $minsup$、$minconf$ 的规则称为强关联规则，而那些支持度大于 $minsup$ 的项集称为高频项集（或称频繁项集，简称频集），对规则的最小支持度约束，与最小支持数约束是等价的。

2.2 关联规则挖掘的一般流程

现有大多数关联规则挖掘算法采用两阶段频集思想，首先获得事务集 D 的所有频繁项集，然而基于频繁项集进行关联规则生成，挖掘的一般流程包括图 1 所示的 4 个核心步骤。

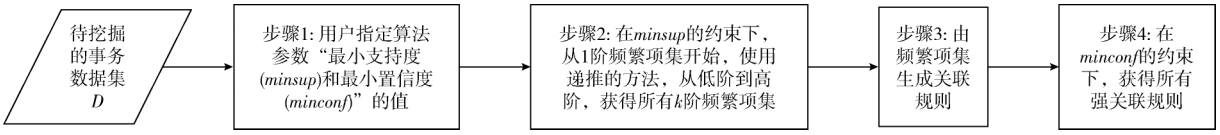

图 1 基于频繁项集的关联规则挖掘的一般流程

图 1 中，步骤 1 指定参数 $minsup$ 和 $minconf$ 的值是关联规则挖掘的前提；步骤 2 在事务集中寻找所有 k 阶频繁项集是关联规则挖掘的核心计算花销所在，往往需要对事务数据集进行多趟扫描；在步骤 2 得到所有 k 阶频繁项集后，步骤 3 和步骤 4 则较为容易。

3 自适应关联规则挖掘

3.1 算法思想

支持度和置信度自适应关联规则挖掘算法 AdapARM（adaptive association rule mining）的基本思想是：以事务集 D 中所有项的支持数和频繁项集能产生的所有规则的置信度数据为依据，用数据说话、由数据决定支持度和置信度参数的取值，在支持数和置信度数据从大到小有序排列的基础上，通过递减序列不低于 3 次的多项式曲线拟合，寻找拟合曲线切线斜率变化速度（二阶导数）首次为 0、即拟合曲线凹凸转换的拐点位置，作为最小支持数 $mincount$ 和最小置信度 $minconf$ 的取值。

由于支持数和置信度序列的单调递减性和多项式拟合曲线的连续性，二阶导数为 0 的地方是曲线变化由缓到急或由急到缓的转折点，将二阶导数首次为 0 的地方选取为支持数和置信度参数的阈值，具有较为科学的统计和数理分析依据。至于多项式拟合曲线的次数，由于 2 次多项式无法较好拟合支持数和置信度序列的单调递减性，而 1 次多项式退化为直线，也无法用于确定参数的科学取值，因此使用不低于 3 阶的多项式进行曲线拟合，从技术上确保了二阶导函数及其支持数和置信度参数取值的存在性。

总之，算法通过数学的方法，使用拟合支持数和置信度序列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，能够有效解决关联规则挖掘算法对先验知识的依赖问题。

3.2 核心步骤

传统关联规则挖掘通常要求算法用户根据先验知识事先指定最小支持度和最小置信度参数的值，而提出的算法 AdapARM 让数据说话、根据数据集自身的特性，自适应地确定最小支持数和最小置信度的值。AdapARM 的核心步骤如图 2 所示，其中步骤 2 和步骤 5：最小支持数和最小置信度的自动确定，是核心研究内容。

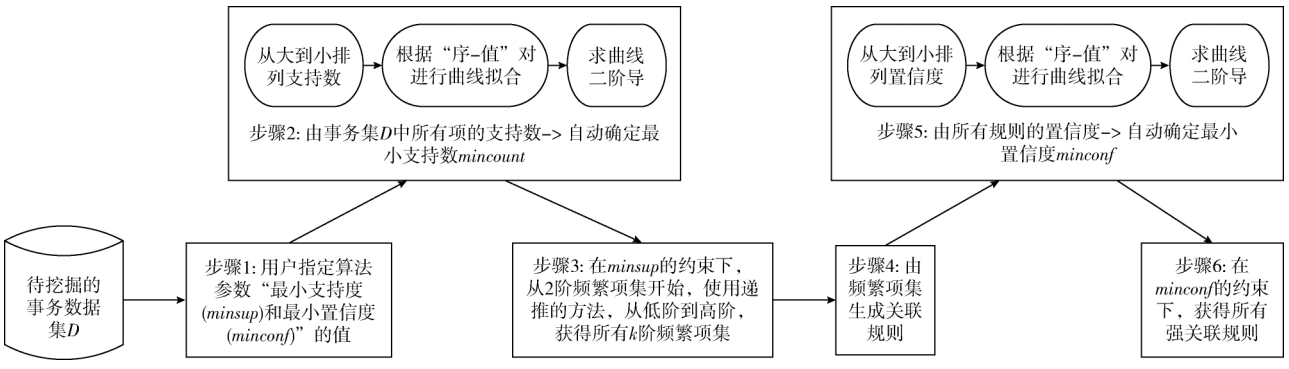

图 2 支持度和置信度自适应的关联规则挖掘的核心步骤

3.3　支持度和置信度自适应的实现

首先，将事务集 D 中各商品（项）的支持数或某个规则的置信度，按从大到小的顺序进行排序，建立“序-值”对序列 $\{(x_p,y_p)\mid p=1,2,\cdots,t\}$（如图 3 所示），其中序号值 $x_p=p$，序列值 y_p 随着序号值 x_p 的递增而递减，即当 $x_p<x_q$ 时 $y_p\geqslant y_q$。

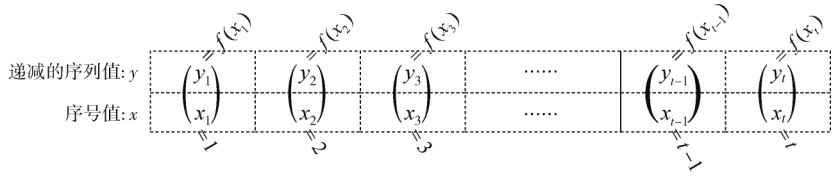

图 3　面向项支持数和规则置信度的“序-值”对

对于支持度，y_x 为某个项 i_x 在事务 D 中支持数（即 $y_x=support\_count(i_x)$），“序-值”对的数目 t 为事务数据集 D 中商品（项）的数目，即 $t=m$。

对于置信度，y_x 为频繁项集生成的某个规则的置信度（即 $y_x=confidence(X\Rightarrow Y)$），“序-值”对的数目 t 为所有 k 阶频繁项集能够产生的关联规则的数目。

然后，基于“序-值”对数据，以序号值为 x、序列值为 y，建立有序的平面坐标 (x_v,y_v) 点序列（$v=1,2,\cdots,t$），并采用多项式进行曲线拟合。拟合的多项式曲线如式（4）所示

$$y=f(x)=\sum_{i=0}^{t}a_i\cdot x^i \tag{4}$$

紧接着，求取以获得拟合曲线的二阶导函数 $f''(x)$，如式（5）所示

$$y''=f''(x)=\sum_{i=2}^{t}i\cdot(i-1)\cdot a_i\cdot x^{i-2} \tag{5}$$

最后，求解拟合曲线二阶导函数 $f''(x)$ 在区间（1，m）内首次出现 $f''(x)=0$ 的 x 值（记 x_0），并求取对应的拟合曲线函数值（记 $f(x_0)$），m 为事务集 D 中项的数目。对于支持数拟合曲线，取 $\lfloor x_0\rfloor$ 作为最小支持数阈值 $mincount$；而对于置信度拟合曲线，取 $f(x_0)$ 作为最小置信度阈值 $minconf$，其中符号 $\lfloor v\rfloor$ 表示向下取整，即不大于 v 的最大整数值。

从 $mincount$ 和 $minconf$ 的取值方法可见，其数值取决于所拟合的曲线 $f(x)$，而 $f(x)$ 主要由有序支持数或置信度构造而来的平面点序列 (x_v,y_v) 所刻画。因此，可以说按提出方法自动获得的 $mincount$ 和 $minconf$ 的取值，与事务集 D 中项的支持数与产生的规则的置信度的分布情况密切相关，能够自适应地贴合数据集自身的特性。

3.4　AdapARM 算法实现

支持度和置信度自适应的关联规则挖掘算法 AdapARM 的伪代码如下：

(1)C_1 = find_candidate_1-itemsets(D); //get all items's support count

(2)$\boldsymbol{C}_1'$=**sort_InDescOrder**($\boldsymbol{C}_1$); //sort C_1 in descending order according the support count

(3)$\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})$ =**polynomial_curve_fitting**($\boldsymbol{C}_1'$); //(x=order id, y=support count)

(4)**find mini** $\boldsymbol{x}_0$, **where** $\boldsymbol{f''}(\boldsymbol{x}_0)=\boldsymbol{0}$; //compute second derivative and get the minimal resolution

(5)***mincount*** $=\lfloor x_0\rfloor$; //*mincount* is defined as the rounded down integer value of x_0

(6)$\{L_1, L_2, \cdots, L_k\}$=find_all_frequent_k-itemsets(D, *mincount*); //according *mincount*

(7)$\{R_1, R_2, \cdots, R_t\}$=generateRule_from_frequent_k-itemsets($\{L_1, L_2, \cdots, L_k\}$); //generate rules

(8)$\boldsymbol{R}'$=**sort_InDescOrder**($\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_t\}$); //sort rules in descending order according the rule confidence

(9)$\boldsymbol{y}=\boldsymbol{h}(\boldsymbol{x})$=**polynomial_curve_fitting**($\boldsymbol{R}'$); //(x=order id, y=rule confidence)

(10)**find mini** $\boldsymbol{x}_0$, **where** $\boldsymbol{h''}(\boldsymbol{x}_0)=\boldsymbol{0}$; //compute second derivative and get the minimal resolution

(11)***minconf***$=\boldsymbol{h}(\boldsymbol{x}_0)$; //*minconf* is defined as the value $h(x_0)$

(12)$\{SR_1, SR_2, \cdots, SR_r\}$=find_strong_rules($\{R_1, R_2, \cdots, R_t\}$, *minconf*); //find strong rules with *minconf*

首先，扫描一次数据库，产生候选 1 项集 C_1；

然后，根据支持数从大到小的方式对项进行降序排列，对序号与支持数值进行多项式曲线拟合，寻找二阶导数为 0 的点 x_0，并将 x_0 向下取整的值 $\lfloor x_0\rfloor$ 作为最小支持数阈值 $mincount$。

紧接着，根据最小支持数阈值 $mincount$，寻找所有 k 阶频繁项集。

再接着，通过所有 k 阶频繁项集，生成关联规则，并根据规则置信度值从大到小的方式对规则进行降序排列，对序号与规则置信度值进行多项式曲线拟合，寻找二阶导数为 0 的点 x_0，并将 x_0 对应的拟合曲线函数值 $h(x_0)$ 作为最小置信度阈值 $minconf$。

最后，依据最小置信度阈值 $minconf$，筛选并获得所有强关联规则。

4　实验及分析

本节使用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，对无须用户指定支持度和置信度参数的 AdapARM 算法的挖掘流程进行介绍，并对挖掘结果进行分析与讨论，从而验证提出算法在自动确定参数上的有效性和实用性。

4.1 数据集说明

Trolley 数据集是很多数据挖掘教材中用于讲解 Apriori 和 FP_Growth 算法的某面包店仿真交易记录，总共有 7 个不同商品，有 9 条消费记录（即 9 行），如表 1 第 2 列所示。

Groceries 数据集记录了某个杂货店一个月的真实交易记录，总共有 169 个不同商品，有 9835 条消费记录（即 9835 行），如表 1 第 3 列所示（仅列举前 9 条事务记录）。

4.2 挖掘流程

步骤 1 计算各数据项的支持数

遍历数据集，获得每个事务项的支持数，即获得候选 1 项集 C_1 中各项的支持数，并按支持数从大到小对项进行排序，见表 2（仅列举支持数最大的前 7 个项）。

表 1 Trolley 数据集和 Groceries 数据集

ID	Trolley 的事务项(全部)	Groceries 的事务项(部分:前 9 条)
T1	牛奶,鸡蛋,面包,薯片	citrus fruit,semi-finished bread,margarine,ready soups
T2	鸡蛋,爆米花,薯片,啤酒	tropical fruit,yogurt,coffee
T3	鸡蛋,面包,薯片	whole milk
T4	牛奶,鸡蛋,面包,爆米花,薯片,啤酒	pip fruit,yogurt,cream cheese,meat spreads
T5	牛奶,面包,啤酒	other vegetables,whole milk,condensed milk,long life bakery product
T6	鸡蛋,面包,啤酒	whole milk,butter,yogurt,rice,abrasive cleaner
T7	牛奶,面包,薯片	rolls/buns
T8	牛奶,鸡蛋,面包,黄油,薯片	other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)
T9	牛奶,鸡蛋,黄油,薯片	pot plants

表 2 数据集 Trolley 和 Groceries 中 C_1 各项的支持数和排序号信息

<Trolley 的 1 项名,支持数,排序号>(全部)	<Groceries 的 1 项名,支持数,排序号>(部分)
<面包, 7, 1>	<whole milk, 2513, 1>
<薯片, 7, 2>	<other vegetables, 1903, 2>
<鸡蛋, 7, 3>	<rolls/buns, 1809, 3>
<牛奶, 6, 4>	<soda, 1715, 4>
<啤酒, 4, 5>	<yogurt, 1372, 5>
<黄油, 2, 6>	<bottled water, 1087, 6>
<爆米花, 2, 7>	<root vegetables, 1072, 7>……(共 169 项)

步骤 2 根据支持数从大到小排序数据项并进行曲线拟合

将 Trolley 数据集和 Groceries 数据集中各数据项的支持数，按从大到小的方式进行排序并使用 3 次多项式对数据进行曲线拟合，支持数与序值对应关系及其拟合曲线如图 4 所示。

Trolley 数据项的支持数拟合曲线如式（6）所示

$$f_T(x)=4.2857142857+3.5595238095\cdot x-1.1428571429\cdot x^2+0.0833333333\cdot x^3 \quad (6)$$

Groceries 数据项的支持数拟合曲线如式（7）所示

$$f_G(x)=1486.065650975-42.7627470864\cdot x+0.4135366641\cdot x^2-0.001283055089\cdot x^3 \quad (7)$$

步骤 3 根据拟合曲线的二阶导函数求取最小支持数

$f_T(x)$ 的二阶导函数 $f''_T(x)$ 如式（8）所示。$f_G(x)$ 的二阶导函数 $f''_G(x)$ 如式（9）所示

$$f''_T(x)=-2.2857142858+0.4999999999\cdot x \quad (8)$$

$$f''_G(x)=0.8270733282-0.007698330534\cdot x \quad (9)$$

对于二阶导函数 $f''_T(x)$，当 $x_0=4.5714285714$ 时，$f''_T(x)=0$。根据算法思想，将 x_0 向下取整（即 $\lfloor x_0 \rfloor=4$）作为 Trolley 数据集关联规则挖掘最小支持数。

对于二阶导函数 $f''_G(x)$，当 $x_0=107.4354140279$ 时，$f''_G(x)=0$。根据算法思想，将 x_0 向下取整（即 $\lfloor x_0 \rfloor=107$）作为 Groceries 数据集关联规则挖掘的最小支持数。

步骤 4 根据最小支持数获得所有 k 阶频繁项集

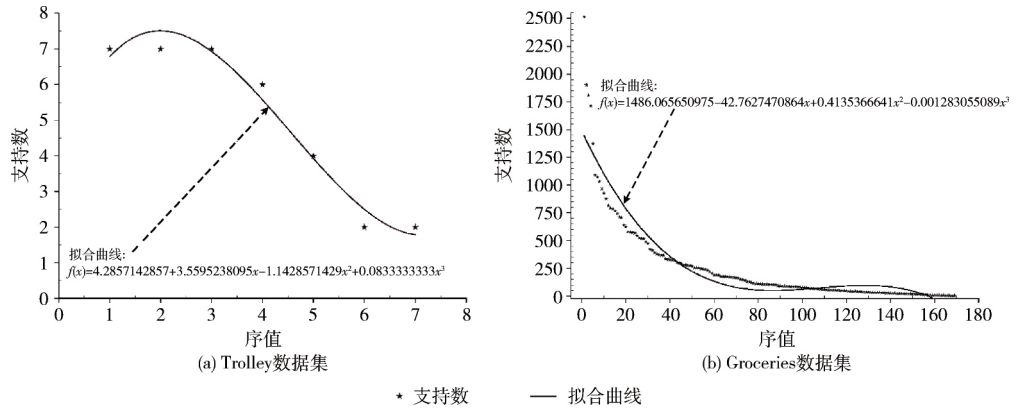

图 4　从大到小有序的数据项支持数及其拟合曲线

根据获得的最小支持数（Trolley 数据集为 4、Groceries 数据集为 107），按照经典 Apriori 算法的思想，从一阶频繁项开始逐阶向上，获得数据集对应的所有 k 阶频繁项集。

Trolley 数据集和 Groceries 数据集的所有 k 阶频繁项集如表 3 所示，其中 NULL 表示“空”。

步骤 5　根据频繁项集产生关联规则

根据获得的所有 k 阶频繁项集（见表 3），为 Trolley 数据集和 Groceries 数据集产生关联规则，见表 4。

步骤 6　根据置信度从大到小排序频繁项集产生的规则

表 3　Trolley 数据集和 Groceries 数据集的所有频繁项集

阶数	Trolley 的频繁项及支持数(全部)	Groceries 的频繁项及支持数(部分)
1	<[面包，7>；<[牛奶]，6>；<[啤酒]，4>；<[薯片，7>；<[鸡蛋]，7>	<[hard cheese],241>；<[whole milk],2513>；<[beverages],256>；<[canned beer],764>；<[baking powder],174>；<[root vegetables],1072>；<[frozen meals],279>；……(共 81 项)
2	<[牛奶,鸡蛋]，4>；<[薯片,鸡蛋]，6>；<[薯片,面包]，5>；<[面包,鸡蛋]，5>；<[薯片,牛奶]，5>；<[面包,牛奶]，5>；	<[tropical fruit,pastry],130>；<[root vegetables,curd],107>；<[root vegetables,beef],171>；<[yogurt,whole milk],551>；<[pork,other vegetables],213>；<[pastry,rolls/buns],206>；……(共 183 项)
3	<[薯片,面包,鸡蛋]，4>；<[薯片,面包,牛奶]，4>；<[薯片,牛奶,鸡蛋]，4>；	<[root vegetables,yogurt,other vegetables],127>；<[domestic eggs,other vegetables,whole milk],121>；<[root vegetables,other vegetables,whole milk],228>；<[tropical fruit,rolls/buns,whole milk],108>；……(共 22 项)
4	NULL	NULL

表 4　Trolley 数据集和 Groceries 数据集频繁项集产生的关联规则

ID	Trolley 的规则及置信度(共 30 条)	ID	Groceries 的规则及置信度(共 498 条)
1	[牛奶,鸡蛋]−>[薯片]，conf=1.0	1	[root vegetables,tropical fruit]−>[other vegetables],conf=0.58454106
2	[鸡蛋]−>[薯片]，conf=0.857142857	2	[butter,other vegetables]−>[whole milk],conf=0.57360406
3	[薯片]−>[鸡蛋]，conf=0.857142857	3	[root vegetables,tropical fruit]−>[whole milk],conf=0.57004831
4	[牛奶]−>[薯片]，conf=0.833333333	4	[root vegetables,yogurt]−>[whole milk],conf=0.56299213
5	[牛奶]−>[面包]，conf=0.833333333	5	[domestic eggs,other vegetables]−>[whole milk],conf=0.55251142
6	[薯片,牛奶]−>[面包]，conf=0.8	6	[yogurt,whipped/sour cream]−>[whole milk],conf=0.52450980
7	[面包,牛奶]−>[薯片]，conf=0.8	7	[root vegetables,rolls/buns]−>[whole milk],conf=0.52301255
…	…	…	…
29	[鸡蛋]−>[薯片,面包]，conf=0.57142857	497	[whole milk]−>[tropical fruit,rolls/buns],conf=0.04297652
30	[鸡蛋]−>[牛奶]，conf=0.57142857	498	[whole milk]−>[yogurt,whipped/sour cream]，conf=0.04257859

并进行曲线拟合

根据 Trolley 数据集和 Groceries 数据集频繁项集产生的关联规则的置信度，按从大到小的方式进行排序并使用 3 次多项式对数据进行曲线拟合，置信度与序值对应关系及其拟合曲线如图 5 所示。

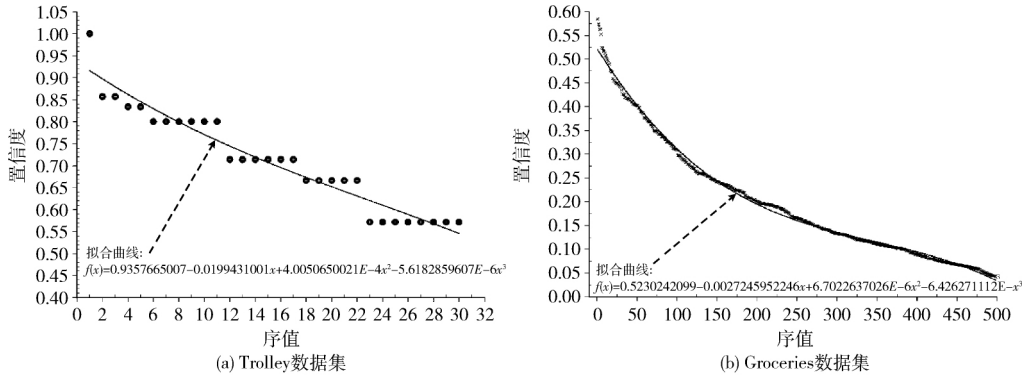

图 5 从大到小有序的规则置信度及其拟合曲线

Groceries 数据集产生的规则的置信度拟合曲线如式（11）所示

$$h_G(x) = 0.5230242099 - 0.0027245952246 \cdot x + 6.7022637026E-6 \cdot x^2 - 6.426271112E-9 \cdot x^3 \quad (11)$$

步骤 7 根据拟合曲线的二阶导函数求取最小置信度

$h_T(x)$ 的二阶导函数 $h''_T(x)$ 如式（12）所示。$h_G(x)$ 的二阶导函数 $h''_G(x)$ 如式（13）所示

$$h''_T(x) = 8.0101300042E-4 - 3.3709715764E-5 \cdot x \quad (12)$$

$$h''_G(x) = 1.3404527405E-5 - 3.8557626672E-8 \cdot x \quad (13)$$

对于二阶导函数 $h''_T(x)$，当 $x_0 = 23.7620811171$ 时，$h''_T(x) = 0$。根据算法思想，将 $h_T(x_0) = 0.6126373314$ 作为 Trolley 数据集进行强关联规则挖掘的置信度阈值。

对于二阶导函数 $h''_G(x)$，当 $x_0 = 347.6491828518$ 时，$h''_G(x) = 0$。根据算法思想，将 $h_G(x_0) = 0.1158444298$ 作为 Groceries 数据集进行强关联规则挖掘的置信度阈值。

步骤 8 根据置信度阈值判断并获得强关联规则

根据设定的置信度阈值（Trolley 数据集为 0.6126373314、Groceries 数据集为 0.1158444298），获得 Trolley 数据集和 Groceries 数据集对应的强关联规则，见表 5。

表 5 Trolley 数据集和 Groceries 数据集中的强关联规则

ID	Trolley 的强规则及置信度（共 22 条）	ID	Groceries 的强规则及置信度（共 339 条）
1	[牛奶,鸡蛋]->[薯片], conf=1.0	1	[root vegetables,tropical fruit]->[other vegetables],conf=0.58454106
2	[鸡蛋]->[薯片], conf=0.857142857	2	[butter,other vegetables]->[whole milk],conf=0.57360406
…	…	…	…
21	[牛奶]->[鸡蛋], conf=0.66666667	338	[other vegetables]->[pastry],conf=0.11665791
22	[牛奶]->[薯片,面包], conf=0.66666667	339	[root vegetables]->[rolls/buns,whole milk],conf=0.11660448

4.3 结果分析与讨论

从上述数据挖掘流程可见，与传统算法需要用户预先指定最小支持度 *minsup* 和最小置信度 *minconf* 的具体数值不同，譬如最小支持度取值为 50%、最小置信度取值为 80%，提出的 AdapARM 算法不需要用户指定任何算法参数值，算法将根据不同的数据集，自动确定 *mincount* 和 *minconf* 的值。对 Trolley 数据集，自动确定的 *mincount* 值为 4，*minconf* 值为 61.26373314%；对 Groceries 数据集，自动确定的 *mincount* 值为 107，*minconf* 值为 11.58444298%。可见，AdapARM 算法能够在用户不具备先验知识的前提下，自动确定参数 *minsup* 和 *minconf* 的值，且参数值与具体的数据集相适应，而不需要由用户进行算法参数值的人为指定，从而为关联规则挖掘任务的无参化运行提供了一种统计学意义上具有较高可信性的解

决方案。

虽然提出的 AdapARM 算法能够挖掘并获得统计意义上支持度和置信度较高的关联规则，然而对于具体的关联规则挖掘实践，合理的支持度和置信度阈值选取，主要依赖于用户的主观应用需求。以上述实验为例，Trolley 数据集中有 7 件不同的商品（项），自适应确定的支持数阈值为 4（支持度$\frac{4}{7}$(57.14%)），有 5 件商品为频繁 1 项，占比$\frac{5}{7}$(71.43%)；Groceries 数据集中有 169 个不同的商品（项），自适应确定的支持数阈值为 107（支持度$\frac{107}{169}$(63.31%)），有 81 件商品为频繁 1 项，占比$\frac{81}{169}$(49.93%)。较高的频繁 1 项集占比，在后续从低阶到高阶进行频繁项搜索时，可能增加一些无用的候选项生成和剪枝操作，从而一定程度上抑制了算法的高效性。与此相类似的问题，也出现于规则的置信度确定中，譬如，Trolley 数据集产生的关联规则有 30 条，具有 22 条强关联规则，占比$\frac{22}{30}$(73.33%)；Groceries 数据集产生的关联规则有 498 条，具有 339 条强关联规则，占比$\frac{339}{498}$(68.07%)。

尽管如此，提出的方案经过进一步处理，仍能为实际关联规则挖掘任务的参数确定提供一定的借鉴。有效的阈值二次确定技术如：在最高支持数（或置信度）和自适应支持数（或置信度）之间，采用第一四分位数（Q1）、第二四分位数（Q2）、第三四分位数（Q3）作为最终选定的支持数（或置信度）的值。

总之，算法在确定支持度和置信度阈值时，通过计算所有数据项的支持数和所有规则的置信度，经从大到小排序后进行多项式曲线拟合，用数据说话、凭数据进行算法参数决策，使得算法在实际行业应用中走向科学，对于算法的推广应用乃至普及具有重要的实践意义。

5 结束语

本文提出了一个支持度和置信度自适应的关联规则挖掘算法 AdapARM，能够在用户不具备先验知识、不指定支持度和置信度阈值的情况下，通过数据集自身的数据特征自适应地确定关联规则挖掘的支持数和置信度。支持度和置信度参数的自适应确定策略：是在数据集所有项的支持数、所有规则的置信度的有序序列的多项式曲线拟合的基础上，通过求取拟合曲线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，进而挖掘统计意义上支持度和置信度较高的强关联规则。在关联规则挖掘标准数据集 Trolley 和标准数据集 Groceries 上的实验结果及分析表明，提出的方法对于关联规则挖掘算法的实施和推广具有一定的实际应用价值。

参考文献：

[1] Solanki Surbhi K, Patel Jalpa T. A survey on association rule mining [C] //5th International Conference on Advanced Computing & Communication Technologies, 2015: 212-216.

[2] ZHAO Xuejian, SUN Zhixin, YUAN Yuan. An efficient association rule mining algorithm based on prejudging and screening [J]. Journal of Electronics & Information Technology, 2016, 38 (7): 1654-1659 (in Chinese). [赵学健，孙知信，袁源. 基于预判筛选的高效关联规则挖掘算法 [J]. 电子与信息学报，2016，38 (7)：1654-1659.]

[3] Nath B, Bhattacharyya D K, Ghosh A. Incremental association rule mining: A survey [J]. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2013, 3 (3): 157-169.

[4] Vo Bay, Le Tuong, Coenen Frans, et al. Mining frequent itemsets using the N-list and subsume concepts [J]. International Journal of Machine Learning and Cybernetics, 2016, 7 (2): 253-265.

[5] Huong Bui, Bay Vo, Ham Nguyen, et al. A weighted N-list-based method for mining frequent weighted itemsets [J]. Expert Systems with Applications, 2018, 96: 388-405.

[6] Maria Luna Jose, Cano Alberto, Pechenizkiy Mykola, et al. Speeding-up association rule mining with inverted index compression [J]. IEEE Transactions on Cybernetics, 2016, 46 (12): 3059-3072.

[7] Anand HS, Vinodchandra SS. Association rule mining using treap [J]. International Journal of Machine Learning and Cybernetics, 2018, 9 (4): 589-597.

[8] NIU Xinzheng, WANG Chongyi, YE Zhijia, et al. An efficient association rule hiding algorithm based on cluster and threshold interval [J]. Journal of Computer Research and Development, 2017, 54 (12): 2785-2796 (in Chinese). [牛新征，王崇屹，叶志佳，等. 基于簇和阈值区间的高效关联规则隐藏算法 [J]. 计算机研究与发展，2017，54 (12)：2785-2796.]

[9] Wang Ke, Qi Yanjun, Fox Jeffrey J, et al. Association rule mining with the micron automata processor [C] //IEEE 29th International Parallel and Distributed Processing Symposium, 2015: 689-699.

[10] Vu Lan, Alaghband Gita. Novel parallel method for association rule mining on multi-core shared memory systems [J]. Parallel Computing, 2014, 40 (10): 768-785.

[11] Liu Zelei, Hu Liang, Wu Chunyi, et al. A novel process-based association rule approach through maximal frequent itemsets for big data processing [J]. Future Generation Computer Systems-the International Journal of Escience, 2018, 81: 414-424.

[12] Sohrabi Mohammad Karim, Roshani Reza. Frequent itemset

mining using cellular learning automata [J]. Computers in Human Behavior, 2017, 68: 244-253.

[13] Zou Caifeng, Deng Huifang, Wan Jiafu, et al. Mining and updating association rules based on fuzzy concept lattice [J]. Future Generation Computer Systems-the International Journal of Escience, 2018, 82: 698-706.

[14] Heraguemi Kamel Eddine, Kamel Nadjet, Drias Habiba. Multi-swarm bat algorithm for association rule mining using multiple cooperative strategies [J]. Applied Intelligence, 2016, 45 (4): 1021-1033.

[15] Tseng Vincent S, Shie Bai-En, Wu Cheng-Wei, et al. Efficient algorithms for mining high utility itemsets from transactional databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25 (8): 1772-1786.

[16] Leung Carson K, Jiang Fan, Pazdor Adam G M. Bitwise parallel association rule mining for web page recommendation [C] //IEEE/WIC/ACM International Conference on Web Intelligence, 2017: 662-669.

[17] HE Ming, LIU Weishi, ZHANG Jiang. Association rules recommendation algorithm supporting recommendation nonempty [J]. Journal on Communications, 2017, 38 (10): 18-25 (in Chinese). [**何明, 刘伟世, 张江. 支持推荐非空率的关联规则推荐算法 [J]. 通信学报**, 2017, 38 (10): 18-25.]

[18] Sarath K N V D, Ravi Vadlamani. Association rule mining using binary particle swarm optimization [J]. Engineering Applications of Artificial Intelligence, 2013, 26 (8): 1832-1840.

(**上接第** 3721 **页**)

[7] Khajehnejad MA, Xu W, Avestimehr AS, et al. Improving the thresholds of sparse recovery: An analysis of a two-step reweighted basis pursuit algorithm [J]. IEEE Transactions on Information Theory, 2015, 61 (9): 5116-5128.

[8] Sajjad M, Mehmood I, Abbas N, et al. Basis pursuit denoi-sing-based image superresolution using a redundant set of atoms [J]. Signal, Image and Video Processing, 2016, 10 (1): 181-188.

[9] Fornasier M, Peter S, Rauhut H, et al. Conjugate gradient acceleration of iteratively re-weighted least squares methods [J]. Computational Optimization & Applications, 2016, 65 (1): 205-259.

[10] Saadat SA, Safari A, Needell D. Sparse reconstruction of regional gravity signal based on stabilized orthogonal matching pursuit (SOMP) [J]. Pure & Applied Geophysics, 2016, 173 (6): 2087-2099.

[11] Chen J, Zhou Y, Jin L, et al. An adaptive regularized smoothed norm algorithm for sparse signal recovery in noisy environments [J]. Signal Processing, 2017, 135 (C): 153-157.

[12] Ramli DA, Tan WC. Fast kernel sparse representation classifier using improved smoothed-l_0 norm [J]. Procedia Computer Science, 2017, 112: 494-503.

[13] Zhang Y, Yu J, Bai H, et al. Improved sparse signal reconstruction based on approximate hyperbolic tangent function with smoothed l_0 norm [J]. International Journal of Science, 2017, 4 (2): 201-210.

[14] Feng JJ, Zhang G, Wen FQ. MIMO radar imaging based on smoothed l_0 norm [J]. Mathematical Problems in Engineering, 2015 (8): 1-10.

[15] Gerdts M, Horn S, Kimmerle SJ. Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with applications in optimal control [J]. Journal of Industrial & Management Optimization, 2017, 13 (1): 47-62.

[16] WU Feiyun, ZHOU Yuehai, TONG Feng. A fast sparse signal recovery algorith-m based on approximate l_0 norm and hybrid optimization [J]. Acta Automatica Sinica, 2014, 40 (10): 2145-2150 (in Chinese). [**伍飞云, 周跃海, 童峰. 基于似零范数和混合优化的压缩感知信号快速重构算法 [J]. 自动化学报**, 2014, 40 (10): 2145-2150.]