计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(5):345−350 [doi: 10.15888/j.cnki.csa.008447]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 Apriori 算法的高校体测数据关联分析①

蒋茜茜, 杨风暴, 杨童瑶, 张锦荣

(中北大学 信息与通信工程学院, 太原 030051)
通信作者: 蒋茜茜, E-mail: 1145830856@qq.com

摘　要: 为了能有效地分析高校体能测试数据且快速地反馈影响学生体测成绩的因素, 本文以我校体能测试数据为样本, 先进行数据预处理转换成适用于数据挖掘的数据集, 考虑到体测数据特征有限并且长度一致的特点, 采用事务压缩技术与 hash 技术相结合的 Apriori 算法进行数据分析, 减少了遍历数据库的次数和生成的候选项集的规模, 在保证挖掘精度的同时提高算法的运行效率. 最后与 Apriori 算法、基于事务压缩的 Apriori 算法、基于 hash 技术的 Apriori 算法进行对比分析, 实验结果表明, 本文提出的事务压缩和 hash 技术相结合的改进 Apriori 算法, 能有效地分析出学生体测成绩间的关联规则, 对学生的体能训练具有更强的指导意义, 与 Apriori 算法相比, 运行效率提高了 85% 以上.

关键词: 事务压缩; hash; Apriori; 关联规则

引用格式: 蒋茜茜,杨风暴,杨童瑶,张锦荣.基于改进 Apriori 算法的高校体测数据关联分析.计算机系统应用,2022,31(5):345–350. http://www.c-s-a.org.cn/1003-3254/8447.html

Association Analysis of College Physical Fitness Test Data Based on Improved Apriori Algorithm

JIANG Xi-Xi, YANG Feng-Bao, YANG Tong-Yao, ZHANG Jin-Rong

(School of Information and Communication Engineering, North University of China, Taiyuan 030051, China)

Abstract: To effectively analyze college physical fitness test data and quickly feed back the factors that affect students' test results, this study takes the physical fitness test data of the North University of China as the sample and transforms preprocessed data into datasets suitable for data mining. Considering the limited features and consistent length of physical fitness test data, an Apriori algorithm that combines the transaction reduction technique with the hash technique is used to analyze data, which reduces the number of database traversal and the scale of candidate sets generated. It also improves the efficiency of the algorithm and ensures mining accuracy at the same time. Finally, comparison and analysis are made with the Apriori algorithm, the Apriori algorithm based on transaction reduction, and the Apriori algorithm based on the hash technique. The experimental results show that the proposed improved Apriori algorithm that combines transaction reduction and the hash technique can effectively analyze the association rules among students' physical fitness test results and therefore has a stronger guiding significance for students' physical fitness training. Compared with the Apriori algorithm, the proposed algorithm improves the operation efficiency by more than 85%.

Key words: transaction reduction; hash; Apriori; association rules

信息社会产生的大量数据急需强有力的数据分析工具, 将数据转换成有用的信息、知识, 因此, 数据挖掘技术[1] 应运而生. 高校体测数据的关联规则挖掘能直接反映项目集中数据的潜在关联, 对帮助学生提高

① 收稿时间: 2021-07-07; 修改时间: 2021-08-11; 采用时间: 2021-08-17; csa 在线出版时间: 2022-04-11

体能素质具有重要的意义.

张崇林等[2] 通过体质测试数据处理方法分析, 认为体质测试数据适合进行关联规则数据挖掘, 并且通过关联规则数据挖掘, 发现了原知识体系利用 BMI 和体脂率评价肥胖的矛盾现象, 可知 Apriori 算法利用逐层迭代搜索的方法挖掘数据间的关联关系, 是目前关联规则应用最广、最经典的算法. 赵常红等[3] 运用关联规则的 Apriori 算法并设置支持度、置信度, 分析男、女生的不同体测数据, 各测试指标等级数量分布情况及其之间的关联性, 但是忽略了算法的运行效率. 刘辛等[4] 提出一种基于数组的 Apriori 算法, 减少了候选频繁集冗余, 提高了 Apriori 算法效率, 找出了各体测项目间的关联关系, 但是忽略了遍历数据库的次数对运行效率的影响. 崔亮等[5] 提出一种基于动态散列和事务压缩的关联规则挖掘算法, 通过估计产生候选项集的大小来动态选择是否使用 hash 技术, 并且利用事物压缩技术删除不包含频繁项集的事务, 从而提高算法运行效率.

由上述文献可知关联规则 Apriori 算法可应用于高校体测数据场景, 其弊端在于数据量增加到一定程度时算法效率低下[6]. 与崔亮等[5] 提出的方法不同, 本文提出的基于事务压缩和 hash 技术的改进 Apriori 算法, 无需估计产生候选项集的大小, 是根据体测数据特征有限并且长度一致的特点, 先使用散列技术压缩要考察的候选项集, 再结合事务压缩技术减少遍历数据集的事务项数和长度, 将这两种算法优势完美结合, 从而大大提高算法的运行效率, 并且有效得到体测数据间的关联, 能较好地预测影响学生体能素质的因素, 从而辅助指导学生的体能训练.

1 基于事物压缩和 hash 技术的 Apriori 关联规则算法

关联规则是描述数据事务特征属性间的规律, 通过某些属性来预测其他属性的关联. 在体测数据中, 设项集 $itemset$={$item\_1$, $item\_2$, ⋯, $item\_n$}是高校学生体测数据特征属性的集合, n 是属性的个数[7]. 一条学生体测数据是一个事务 T, 一个事务 T 是一个项集, 每个事务均与一个唯一标识符 Tid 相联系, 体测数据中学籍号是唯一的标识符 Tid[8]. 不同学生的体测数据组成了事务集 D, 它构成了关联规则发现的事务数据库.

用支持度和置信度衡量关联规则的标准, 例如形如$A \Rightarrow B$的关联规则, 体测数据属性项 A 与属性项 B 同时存在的概率即为该数据关联规则的支持度 $Support(A \Rightarrow B)$; 体测数据属性 A 存在的条件下, 属性 B 也存在的概率即为体测数据关联规则的置信度 $Confidence(A \Rightarrow B)$, 其分别如式 (1) 和式 (2) 表示.

$$Support(A \Rightarrow B) = P(A \cup B) = \frac{Support\_count(A \cup B)}{Total\_count} \tag{1}$$

其中, $Support\_count(A \cup B)$ 表示 A 和 B 同时存在的事务个数, $Total\_count$ 表示所有事务的个数.

$$Confidence(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)} = \frac{Support\_count(A \cup B)}{Support\_count(A)} \tag{2}$$

其中, $Support\_count(A)$ 表示 A 存在的事务数.

如果$A \Rightarrow B$的支持度和置信度同时满足最小置信度和最小支持度阈值, 则$A \Rightarrow B$为强关联规则.

1.1 Apriori 算法

Apriori 算法是利用 k-项集来产生 (k+1)-项集迭代的方法[9], 其操作流程如图 1, 设定最小支持度 min_sup 和最小置信度 min_conf, 扫描整个数据库, 统计长度为 1 的每一项的个数, 筛选满足最小支持度的项, 得到频繁 1-项集 L_1; 由 L_1 自身连接生成候选 2-项集 C_2, 得到的 C_2 进行剪枝操作, 统计长度为 2 的每一项的个数, 删除不满足最小支持度的项, 得到频繁 2-项集 L_2, 如此迭代下去, 直到不能再找到频繁 k-项集[10], 最终得到频繁项集 L={L_1, L_2, ⋯, L_k}.

Apriori 算法的弊端在于数据量增加到一定程度时算法效率低下, 具体表现在以下 3 个方面:

1) 在 L_{k-1} 与自身连接生成 C_k 过程中, 可能需要产生大量候选项集. 例如, 假设 L_{k-1} 中有 m 个 k-项集, 则在进行自身连接需要进行比较的时间复杂度为 $O(k \times m^2)$.

2) 在对 C_k 剪枝的过程中, 判断 C_k 中任意一个 k 项候选集的 (k–1)-项候子集是否为 L_{k-1} 的子集, 最快只需要扫描一次得出结果, 最慢需要 k–1 次, 所以平均时间为$|C_k| \times |L_{k-1}| \times k/2$.

3) 在筛选满足最小支持度的候选项集过程中, 对 C_k 中的每个项进行计数, 需要扫描数据库$|C_k|$次.

1.2 基于散列技术的 Apriori 算法

散列技术又名 hash 技术[11], 基于散列技术的 Apriori 算法的优势主要体现在能将生成的 C_k 进行压缩, 大大缩小了 C_k 占用空间. 首先, 扫描数据库所有的事务得

到 C_1, 筛选满足最小支持度的候选项集得到 L_1; 然后利用散列技术来改进产生频繁 2-项集的方法, 具体如下: 将每条事务生成其所有的 2-项集并且运用散列技术构造散列函数把它散列到相应的散列表中, 即分配到不同的桶中并计数, 而且相同的数据项被分配到同一存储空间, 且对应着同一计数器, 只需计数加一即可, 且降低了所需的存储空间. 同样的生成频繁 k-项集也是用相同的方法, 并且在散列表中可以筛选出容器计数小于支持度阈值的 k-项集, 因此可以很好的对要检测的 k-项集执行压缩操作.

散列技术的缺点是其运行效率依赖于事务的平均长度, 也就是说如果数据集的特征少, 平均长度短, 其运行效率就高, 反之运行效果差. 而 Apriori 方法运行效率不依赖于数据集特征的长度和数量, 所以一定程度上弥补 hash 技术的缺点, 两者互补可以大大增加算法的运行效率.

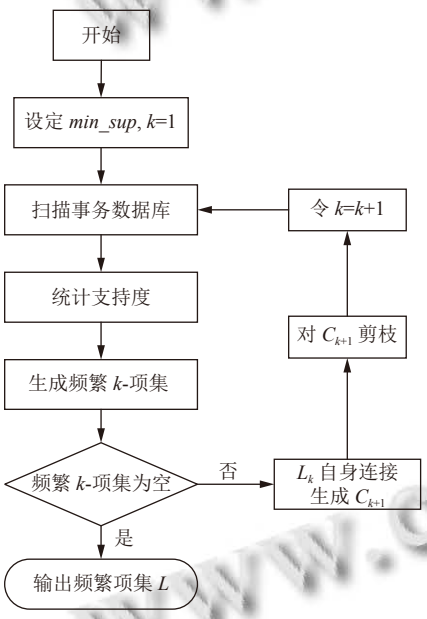

图 1 经典 Apriori 算法流程

1.3 基于事务压缩技术的 Apriori 算法

基于事务压缩技术的 Apriori 算法的优势主要体现在将进一步迭代扫描的数据库进行压缩, 进而影响扫描事务库的次数, 增加算法的运行效率[12]. 该技术的改进是根据经典 Apriori 算法的性质, 即如果一个事务不包含任何的频繁 k-项集, 则其也不包含任何的频繁 $(k+1)$-项集. 因此在生成 L_j $(j>k)$ 时, 遇到上述事务时, 可以加上标记或删除, 扫描数据库时不需要再扫描该事务, 减少扫描事务库的次数.

1.4 基于散列技术和事务压缩相结合的改进

根据以上了解可知, 基于散列技术的改进和基于事务压缩技术作用于 Apriori 算法过程中的不同部分, 基于散列技术的改进体现在减少候选项集的规模, 并更进一步减少数据库扫描的次数, 而事务压缩技术能够逐渐减小数据库中事务的长度和规模, 都主要作用在 Apriori 算法的剪枝部分[13], 并且两者会促进对方的效果, 而不产生负面影响.

改进算法流程如算法 1 所示, 根据高校体测数据的特征有限和并且长度一致的特点, 利用散列技术遍历一次事务集, 即可生成频繁项集 L_1, L_2, L_3, 再利用事务压缩技术由 L_3 生成 L_4, 如此迭代下去, 直到不能再找到频繁 k-项集. 因为迭代生成候选项时可能性太多, 数据量大时占用内存太大, 所以不继续使用 hash 技术继续生成 L_4, 而改用事务压缩技术, 有效地避免 hash 技术的短板.

算法 1. 改进 Apriori 算法

输入: 数据库 D, 最小支持度 min_sup, 最小置信度 min_conf
过程:
1) 扫描数据库, 利用散列技术生成频繁 1-项集 L_1, 频繁 2-项集 L_2, 频繁 3-项集 L_3;
2) 通过频繁项集 L_{k-1} 自身连接生成 C_k 候选项集, 再进行剪枝;
3) 利用事务压缩技术减少扫描事务集的次数, 删除不满足最小支持度的项集, 生成频繁 k-项集 L_k;
4) 对 $k>3$, 重复执行步骤 2) 和 3), 判断频繁 k-项集的长度, 如果为 0, 则跳出循环, 得到最终结果频繁项集 L, 否则返回步骤 2);
5) 筛选出满足 min_sup 和 min_conf 的频繁项集即强关联规则, 结束.
输出: 强关联规则

2 实验过程及分析

2.1 数据准备

以我校 2016 级 7 709 名大学生第一学年的体测成绩为实验对象, 从数据库中提取与体测成绩有关的数据, 如: 学籍号、性别、身高、体重、肺活量、50 m、立定跳远、1 000 m (男生)、800 m (女生)、引体向上 (男生)、仰卧起坐 (女生) 共 11 个属性, 表 1 是以学籍号为主键的体测成绩.

2.2 数据清洗及转换

实验数据来源于我校体测中心, 所以得到的数据较为规范, 但是数据中存在部分学生体测成绩缺项, 又考虑到部分特殊学生免测的情况, 所以要对原始数据先进行预处理, 经过数据清洗、数据转换和数据消减, 实现算法在高校学生体测中的应用.

表 1 学生 2016 年的体测成绩

学籍号	性别	身高 (cm)	体重 (kg)	肺活量 (mL)	50 m (s)	…
****124	女	156	58	2 674	10.1	…
****127	男	179	64.7	4 110	6.87	…
****129	男	172	57.7	5 112	6.4	…
****137	男	166	78.9	4 437	7.78	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮
****306	女	160	54	2 416	8.84	…

数据清洗的实际操作主要有:

(1) 针对缺项数据, 如果缺项数据较多则删除该学生的体测成绩[12]; 如果缺项较少, 则可以通过求该项目成绩的平均值、众数或者中位数填充;

(2) 针对免考学生的数据, 由于各项成绩为缺项, 可以直接删除, 使得清洗后的数据标准、干净且连续.

数据转换将不同标准的连续型数据转换成标准相同的离散型数据, 便于后续的统一处理. 虽然体测项目间的评判标准不同, 但是都可以划分为 A、B、C、D 四个等级, 即优秀、良好、及格、不及格, 其次, 为了方便统计以及提高数据挖掘的效率, 用数字 1 和 2 分别代替性别的男和女, 用小写字母开头代替各项目的名称, 如: 肺活量用 fhl 代替.

数据转换结果如表 2 所示, 经过数据清洗及转换得到的标准数据最终有 7 337 条.

表 2 数据转换结果

xjh	xb	BMI (kg/m^2)	fhl (mL)	50 m (s)	…
*****124	2	A	C	C	…
*****127	1	A	C	A	…
*****129	1	A	A	A	…
*****137	1	C	B	C	…
⋮	⋮	⋮	⋮	⋮	⋮
*****306	2	A	C	C	…

2.3 数据消减

数据消减是在保持数据库的完整性的情况下, 从上述数据集中获得精简的数据集, 能够保证不影响最终的算法结果. 男生和女生体测项目和评价标准有所不同, 所以本文将在数据挖掘时对男生和女生的体测成绩数据分别进行处理. 数据消减就是针对数据库中筛选出的男生、女生的体测数据, 消除与体测项目属性无关的项, 例如学籍号, 性别.

2.4 实验验证

试验环境为 Intel(R) Core(TM) i5-5200U 2.20 GHz 处理器, 4 GB 内存, 操作系统为 Windows 10, 算法在 Python 3.8 下实现.

设定 *min_conf* = 0.9, 设置不同的最小支持度, 利用 Apriori 算法以及本文提出的事务压缩和 hash 结合的算法得到的结果如表 3 所示, 在相同的最小支持度下得到的关联规则数目相同, 表示改进后的算法确保了 Apriori 算法的准确性, 由此可以得到改进算法的可靠性; 随着最小支持度的增加, 改进后的算法执行效率大大提高. 改进后的算法在保证挖掘精度的同时提高算法效率.

表 3 两种算法比较

算法	支持度	挖掘时间 (s)	关联规则个数
Apriori算法	0.01	22.03	940
	0.1	15.58	100
	0.2	14.56	39
	0.3	14.22	18
改进的Apriori算法	0.01	4.01	940
	0.1	1.55	100
	0.2	1.44	39
	0.3	1.39	18

为了能直观地反映出改进后的算法运行效率更高, 在设置不同最小置信度、最小支持度情况下, 将得出的实现数据进行分析对比, 结果如图 2 和图 3 所示. 图中结果可知, 本文提出的改进算法效率明显优于经典 Apriori 算法、基于 hash 的 Apriori 算法以及基于事物压缩的 Apriori 算法. Apriori 算法和基于事物压缩的 Apriori 算法随着支持度和置信度的增加执行效率比较接近, 其中基于事物压缩的 Apriori 算法效率稍高; 基于 hash 改进算法相较于 Apriori 算法和基于事物压缩的 Apriori 算法本身具有一定的优越性, 本文将 hash 技术与事务压缩技术相结合得到了更好的效果.

2.5 结果分析

原始的体测数据经过数据预处理后得到 2 266 条女生体测数据和 5 071 条男生体测数据, 设置最小置信度 *min_conf* = 0.9, 最小支持度 *min_sup* = 0.3, 将这两类数据分别用文中提到的 4 种算法进行关联规则挖掘, 将各运行效率进行比较, 如表 4 所示, 可以看出本文提出的基于事物压缩和 hash 结合的算法执行效率更高, 与经典 Apriori 算法执行效率相比, 女生体测数据算法执行效率提高了 86.12%, 男生提高了 90.63%; 与基于事物压缩的 Apriori 算法相比, 女生体测数据算法执行效率提高了 80.57%, 男生提高了 89.32%; 与基于

hash 的 Apriori 算法相比, 女生体测数据算法执行效率提高了 48.1%, 男生提高了 55.69%; 可见, 该 4 种算法应用于男生的体测数据的执行效率比女生更高, 说明数据集量越大改善效果越明显.

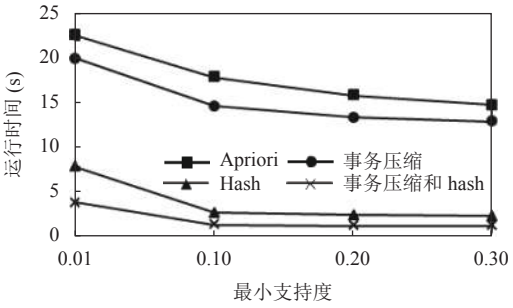

图 2 不同支持度性能对比

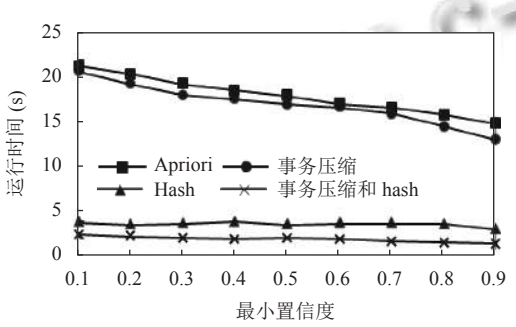

图 3 不同置信度性能对比

表 4 4 种算法执行效率比较 (s)

性别	Apriori	事物压缩 +Apriori	Hash技术 +Apriori	事物压缩 +hash+Apriori
男生	14.83	13.01	2.46	1.09
女生	4.9	3.5	1.31	0.68

表 5 和表 6 分别是男生和女生体测成绩的关联规则挖掘结果, 由表 5 可以看出挖掘出的男生第一条关联规则 50 m:C⇒ytxs:D, 支持度为 72.1%, 置信度为 93.5%, 说明男生 50 m 成绩如果是 C 等级, 那么他的引体向上的成绩为 D 等级的可能性为 93.5%, 而这种事件发生的可能性为 72.1%, 可知此学生的 50 m 成绩可以预测其引体向上的成绩, 以此类推. 可通过某学生的历史数据来预测其他成绩的及格情况, 达到预警的效果, 使该学生加强某方面的锻炼, 进行有针对性的体能训练, 帮助老师安排训练计划有着指导性的意义.

3 结论与展望

现在数据挖掘技术已经非常成熟, 然而数据挖掘和体能分析结合的应用却很少, 本文通过对比 4 种算法实现的结果和执行效率, 表明改进后的算法在高校学生体测中的应用具有一定的可靠性和有效性, 对于比较大的数据集效率改善更明显. 并且将本文所提出的算法, 与经典 Apriori 算法相比, 可知执行效率提高了 85% 以上, 并且挖掘出体测成绩间的潜在关联, 分析身体素质的不足之处, 对学生重视体育锻炼, 辅导老师安排学生的训练计划具有指导性的意义.

表 5 关联规则 (男生)

序号	支持度	置信度	规则
1	0.721	0.935	50 m:C⇒ytxs:D
2	0.687	0.921	1000 m:C⇒ytxs:D
3	0.573	0.939	1000 m:C, 50 m:C⇒ytxs:D
⋮	⋮	⋮	⋮
18	0.303	0.910	1000 m:C, ldty:D⇒50 m:C

表 6 关联规则 (女生)

序号	支持度	置信度	规则
1	0.574	0.913	ldty:C⇒50 m:C
2	0.499	0.903	800 m:C⇒50 m:C
3	0.458	0.910	BMI:A, ldty:C⇒50 m:C
⋮	⋮	⋮	⋮
7	0.306	0.908	800 m:C, fhl:C⇒50 m:C

参考文献

1 Casado R, Younas M. Emerging trends and technologies in big data processing. Concurrency and Computation: Practice and Experience, 2015, 27(8): 2078–2091. [doi: 10.1002/cpe.3398]

2 张崇林, 虞丽娟, 吴卫兵. 关联规则数据挖掘技术在体质测试分析中的应用. 上海体育学院学报, 2012, 36(2): 42–44. [doi: 10.3969/j.issn.1000-5498.2012.02.010]

3 赵常红, 王琳, 冯刚, 等. 民族院校大学生体质测试数据的分析与应用——以西北民族大学为例. 兰州文理学院学报 (自然科学版), 2017, 31(3): 108–112.

4 刘辛, 杨素锦. 基于数组的 Apriori 算法在体质测试数据分析中的应用. 山东理工大学学报 (自然科学版), 2011, 25(5): 55–58. [doi: 10.3969/j.issn.1672-6197.2011.05.016]

5 崔亮, 郭静, 吴玲达. 一种基于动态散列和事务压缩的关联规则挖掘算法. 计算机科学, 2015, 42(9): 41–44. [doi: 10.11896/j.issn.1002-137X.2015.9.009]

6 Chen Z, Cai SB, Song QL, *et al*. An improved Apriori algorithm based on pruning optimization and transaction reduction. 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic

Commerce. Deng Feng: IEEE, 2011. 1908–1911.
7 纪文璐, 王海龙, 苏贵斌, 等. 基于关联规则算法的推荐方法研究综述. 计算机工程与应用, 2020, 56(22): 33–41. [doi: 10.3778/j.issn.1002-8331.2006-0158]
8 杜永兴, 高迪, 李宝山, 等. 改进 Apriori 算法在荒漠草原的应用. 计算机工程与设计, 2019, 40(7): 2082–2086, 2093. [doi: 10.16208/j.issn1000-7024.2019.07.046]
9 Jin R, Lin ZJ, Xue CM, *et al*. An improved association-mining research for exploring Chinese herbal property theory: Based on data of the Shennong's Classic of Materia Medica. Journal of Integrative Medicine, 2013, 11(5): 352–365. [doi: 10.3736/jintegrmed2013051]
10 Zhang CS, Li Y. Extension of local association rules mining algorithm based on Apriori algorithm. 2014 5th IEEE International Conference on Software Engineering and Service Science. Beijing: IEEE, 2014. 340–343.
11 彭永供, 王靓明, 朱敏, 等. 基于散列技术的高效剪枝关联规则挖掘算法. 南昌大学学报 (理科版), 2009, 33(5): 494–498. [doi: 10.3969/j.issn.1006-0464.2009.05.020]
12 孙金华, 谢彦麒. 基于事务压缩的关联规则挖掘算法改进. 微计算机信息, 2010, 26(27): 223–225. [doi: 10.3969/j.issn.2095-6835.2010.27.090]
13 Wang P, An CH, Wang L. An improved algorithm for Mining Association Rule in relational database. 2014 International Conference on Machine Learning and Cybernetics. Lanzhou: IEEE, 2014. 247–252.