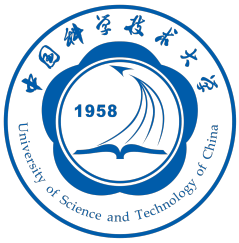


中国科学技术大学

University of Science and Technology of China

数据分析与实践 Lab4

频繁项集和关联规则挖掘

姓名：邱文韬

学号：JL22110003

专业：计算机科学与技术

学院：计算机科学与技术学院

2023 年 5 月 20 日

摘要

本文基于 Pisa 2018 数据集，在数据分析与实践 Lab3：**数据分析与特征处理**的基础上，最终选取了 5 个特征：**RepeatFactor**：复读评价指标、**SelfGrade**：年级指标 **EduIndex**：受教育水平指标、**ParentFactor**：父母地位及自我期望综合指标、**PowerIndex**：经济社会文化综合指标。具体含义以及特征的构造与处理，详见 section2。使用 *Python* 编程实现 *Apriori* 算法以及关联规则的计算，计算 $confidence$、$lift$、PS 、$\phi - coefficient$ 四个指标用于评价关联规则，最终得出了总体数据集上除 REPEAT 1.0 外的 7 个频繁项集，其中有两个频繁 2 项集，三个频繁 3 项集，一个频繁 4 项集，详见 section4。并继续分析了数据集中各个国家的频繁项集与关联规则，其中多米尼亚与哥斯达黎加数据量极少无法产生频繁项集，所以仅分析西班牙、墨西哥、智利、巴拿马的频繁项集与关联规则，并发现了与总体关联规则存在的异同，详见 section5。

关键字： Apriori，频繁项集，关联规则

目录

1 实验内容

1.1 关联规则

利用实验三的数据分析结果，选择与 **REPEAT** 列**最相关的 5 个特征**作为特征集，根据**关联规则算法**，挖掘这 5 个特征和 **REPEAT** 构成的**频繁项集**和**关联规则**

1.2 实验要求

- 数据集中与 REPEAT 列直接相关的三个特征在本次实验中仅允许合为一个特征使用。'ST127Q01TA', 'ST127Q02TA', 'ST127Q03TA'
- 需要挖掘的频繁项集的最小支持度 (min-support) 阈值为 0.6,最小置信度 (min-confidence) 阈值为 0.15
- 建议使用的关联规则算法：*Apriori* 或相关的改进算法 (不允许调包，需要自行实现)
- 选择的特征集可以是数据集里有的，也可以是自己在实验三中构建的特征

2 特征选取与特征处理

2.1 特征选取

首先，根据实验 3 的特征工程，初步选取以下特征

表 1: 弱相关特征选取

特征名称	特征含义
ISCEDL	ISCED level
FISCED	Fathers Education (ISCED)
MISCED	Mothers Education (ISCED)
PAREDINT	Index highest parental education
HISEI	Index highest parental occupational status
BMMJ1	ISEI of mother
BFMJ2	ISEI of father
BSMJ	Students expected occupational status (SEI)
ESCS	Index of economic, social and cultural status

它们与 REPEAT 列的相关系数如下：可以看出其中，它们和 REPEAT 都存在于一定相关性，其中 ISCEDL，HISEI，BSMJ，ESCS 相关性相对较高，超过了 0.2。

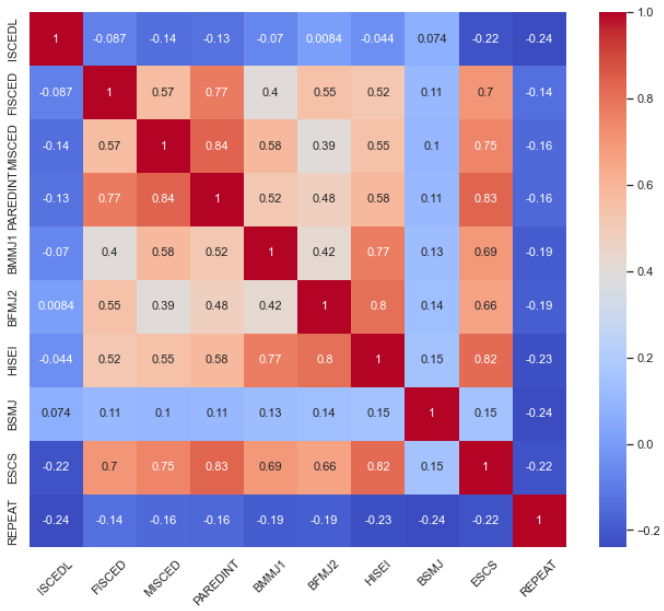


图 1: 弱相关特征相关系数热力图

而还有部分特征与 REPEAT 直接相关，如下所示

表 2: 强相关特征选取

特征名称	特征含义
ST001D01T	Student International Grade (Derived)
ST127Q01TA	Have you ever repeated a <grade>? At <ISCED 1>
ST127Q02TA	Have you ever repeated a <grade>? At <ISCED 2>
ST127Q03TA	Have you ever repeated a <grade>? At <ISCED 3>
GRADE	Grade compared to modal grade in country

根据特征间相关系数的热力图2，可以看出 REPEAT 和这几个特征的确具有明显的强相关性，其中'ST127Q01TA', 'ST127Q02TA', 'ST127Q03TA' 需要合并使用。

2.2 特征处理

2.2.1 RepeatFactor：复读评价指标

对于直接相关的三个特征：'ST127Q01TA', 'ST127Q02TA', 'ST127Q03TA'，因为 1 表示没有复读过，所以首先将 1 改为 0，然后将三列数值相加得到一个新的特征 RepeatFactor，

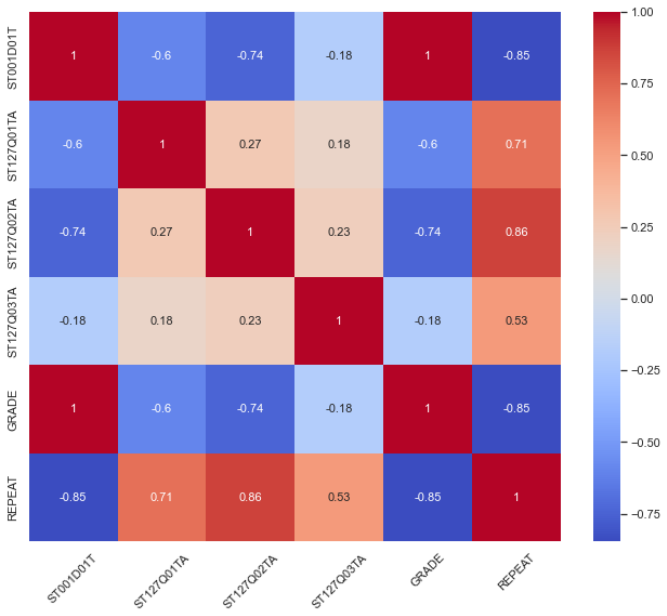


图 2: 强相关特征相关系数热力图

但在处理过程中遇到了一个问题：'ST127Q03TA' 的缺失率太高了，超过了 60%，所以选择不加入这个特征，只将'ST127Q01TA', 'ST127Q02TA' 相加得到新的特征 RepeatFactor。

2.2.2 SelfGrade：年级指标

将'GRADE' 特征根据其含义重新命名为：SelfGrade，其是离散变量，并且缺失值很少。

2.2.3 EduIndex：受教育水平指标

将'ISCEDL' 特征根据其含义重新命名为：EduIndex，该特征没有缺失值并且是离散变量。

2.2.4 ParentFactor：父母地位及自我期望综合指标

首先，'PAREDINT','BMMJ1','BFMJ2','HISEI','BSMJ' 的特征处理工作与实验 3 一致，这几列特征是衡量学生自我期望与父母地位的指标，这几类特征缺失率不高，在 0.2% 左右，使用均值填充缺失值，异常值也用均值替代，再进行归一化的操作，归一化至区间 [0,1]。再将这几列特征的值相加，得到一个新的特征 ParentFactor，取值区间为 [0,5]。并且因为这几列特征为连续型变量，将其离散化至 5 个等间距的区间，可以采用四舍五入或者分桶的操作，在这里选择四舍五入进行离散化。

2.2.5 PowerIndex：经济社会文化综合指标

首先，计算 ESCS 数据的平均值和标准差，根据大数定理 [1] 将异常值替换为平均值，同样使用平均值填充缺失值。再将数据归一化至 [-1,1] 区间上，ESCS 特征同样需要进行离散化，将其分桶为 [-1, -0.5, 0, 0.5, 1]，在映射至标签 [-1, 0, 1, 2] 上，因此最后 ESCS 的取值为 {-1, 0, 1, 2}。再根据其含义重新命名为 PowerIndex。

2.3 最终特征

根据上述工作，最终得到以下五个特征：

表 3: 最终特征选取

特征名称	特征含义
RepeatFactor	复读评价指标
SelfGrade	年级指标
EduIndex	受教育水平指标
ParentFactor	父母地位及自我期望综合指标
PowerIndex	经济社会文化综合指标

下图是它们与 REPEAT 列的相关系数热力图，可以看出组合后得到的 ParentFactor 比之前任意单个特征的相关性都更高。

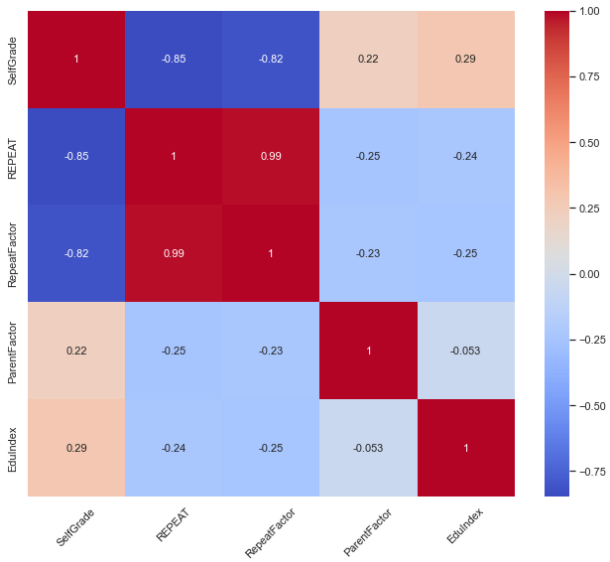


图 3: 最终特征相关系数热力图

3 频繁项集挖掘算法设计

3.1 *APriori* 算法

3.1.1 *APriori* 算法介绍

在本实验中选择使用 *APriori* 算法进行频繁项集的挖掘，*APriori* 算法是一个经典的频繁项集挖掘算法，在 1994 年由 IBM 研究员 Agrawal 提出 [2]，它的核心思想是: 广度优先搜索，自底而上遍历，逐步生成候选集与频繁项集，它基于一个重要的原理——反单调性原理 (Antimonotonicity)，其定义如下：

Definition 3.1 (反单调性)**.** 如果一个项集是频繁的，则它的所有子集一定也是频繁的。即

$$\forall X, Y : X \subseteq Y \rightarrow support(X) \geq support(Y)$$

其中 *support* 表示支持度，依据该性质，对于某 k+1 项集，只要存在一个 k 项子集不是频繁项集，则可以直接判定该项集不是频繁项集。算法步骤主要分为连接步和剪枝步。

连接步：从频繁 $K-1$ 项集生成候选 K 项集 ($K \geq 2$)。具体来说，对于两个频繁 $K-1$ 项集 I 和 J，如果它们的前 $K-2$ 个元素相同，那么可以将它们合并成一个候选 K 项集 $I \cup J$。

剪枝步：从候选 K 项集中筛选出频繁 K 项集。具体来说，对于每个候选 K 项集 C，检查它的所有 $K-1$ 子集是否都出现在频繁 $K-1$ 项集中，如果有任何一个子集不是频繁的，则可以将 C 剪枝掉。

算法伪代码如下：

Algorithm 1 Apriori Algorithm

Input: Dataset D, minimum support $minsup$

Output: Set of frequent itemsets L

$L_1 \leftarrow$ {frequent 1-itemsets};

$k \leftarrow 2$;

while $L_{k-1} \neq \emptyset$ **do**

  $C_k \leftarrow$ candidate k-itemsets generated from L_{k-1};

  prune C_k to obtain L_k;

  $k \leftarrow k+1$;

end

return $\bigcup_k L_k$;

3.2 *APriori* 算法设计实现

使用 python 语言进行编程，采用模块化函数设计实现 *Apriori* 算法及计算关联规则的功能。

频繁项集生成的各个函数及其对应功能如表4所示。

表 4: 频繁项集生成

函数名称	函数功能
SingleCandidateSet	生成初始一项候选集 singleSet
calculateSupport	计算一项候选集的支持度，得到频繁集及其支持度
generateCandidates	生成后续 k 项候选集
findFrequentSets	筛选满足 minsup 的频繁 k 项集
AprioriMain	得到最终频繁集及对应的支持度

关联规则计算的各个函数及其对应功能如表5所示。

表 5: 关联规则计算

函数名称	函数功能
calculate_confidence	计算置信度
generate_subsets	获取集合的子集
generate_association_rules	生成对应的关联规则

3.3 关联规则评价指标

3.3.1 $confidence$

$confidence$(置信度)[3] 是指在包含一个前提项集的事务中同时包含另一个后继项的比例。置信度越高，表示这个关联规则所描述的现象越可信。通常情况下，置信度的阈值也需要根据具体数据集和应用场景来确定。置信度的计算公式为:

$$\frac{P(X,Y)}{P(Y)} \tag{1}$$

3.3.2 $lift$

$lift$(提升度)[4] 是指规则中后继项集的出现概率与其在前提项集中出现独立的概率之比。提升度越大，表示后继项集的出现与前提项集的出现之间的关联性越强。当提升度大于

1 时，表示后继项集与前提项集之间存在正相关关系；当提升度小于 1 时，表示后继项集与前提项集之间存在负相关关系；当提升度等于 1 时，表示后继项集与前提项集之间不存在关联关系。提升度的计算公式为：

$$\frac{P(X,Y)}{P(X)P(Y)} \tag{2}$$

3.3.3 PS

PS(协方差)[5, 6] 是用来度量两个随机变量之间线性关系强度的一种统计量。协方差的取值范围是正无穷到负无穷，取值为正数表示两个变量呈正相关关系，取值为负数表示两个变量呈负相关关系，取值为零表示两个变量之间没有线性关系。协方差的计算公式为：

$$P(X,Y) - P(X)P(Y) \tag{3}$$

3.3.4 $\phi - coefficient$

$\phi - coefficient$(Phi 相关系数)[5, 6] 是测量两个二元变数之间相关性的工具，如果 $\phi - coefficient$ 为 1 或-1，则表示两个二分类变量完全相关或完全不相关；如果 $\phi - coefficient$ 为 0，则表示两个二分类变量之间没有线性关系。$\phi - coefficient$ 的计算公式：

$$\frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)P(Y)(1-P(X))(1-P(Y))}} \tag{4}$$

3.4 数据变换

由于原始数据的每个取值都是一个数字，在生成频繁项集时不便于分辨，所以将每一列的取值都改为” 列名 + 取值” 的字符串，变换后的数据示例如6所示。

表 6: 变换后数据示例

SelfGrade	**EduIndex**	**RepeatFactor**	**ParentFactor**	**PowerIndex**	**REPEAT**
SelfGrade 9.0	EduIndex 2.0	RepeatFactor 2.0	ParentFactor 3.0	PowerIndex 0.0	REPEAT 1.0

4 频繁项集与关联规则计算结果

4.1 频繁项集

将前述数据作为输入，只取 REPEAT=1 的数据集并取支持度阈值 $minsup = 0.6$，得到的频繁项集及其支持度如表7所示。

表 7: 总体频繁项集与支持度

频繁项集	支持度
{'REPEAT 1.0'}	1.000000
{'REPEAT 1.0', 'SelfGrade 9.0'}	0.753535
{'REPEAT 1.0', 'RepeatFactor 2.0'}	0.753535
{'REPEAT 1.0', 'EduIndex 2.0'}	0.921558
{'REPEAT 1.0', 'CompareGrade -1.0', 'RepeatFactor 2.0'}	0.678584
{'EduIndex 2.0', 'REPEAT 1.0', 'RepeatFactor 2.0'}	0.700462
{'EduIndex 2.0', 'REPEAT 1.0', 'SelfGrade 9.0'}	0.678406
{'EduIndex 2.0', 'REPEAT 1.0', 'RepeatFactor 2.0', 'SelfGrade 9.0'}	0.650836

4.2 关联规则

继续计算其关联规则，并计算置信度，lift，PS，$\phi - coefficient$ 评价指标，结果如表8所示。

表 8: 总体关联规则及其度量指标

前项	后项	置信度	lift	PS	ϕ 系数
SelfGrade 9.0	REPEAT 1.0	0.787	5.411	0.085	0.713
RepeatFactor 2.0	REPEAT 1.0	1.000	6.872	0.093	0.844
EduIndex 2.0	REPEAT 1.0	0.206	1.413	0.039	0.234
SelfGrade 9.0, RepeatFactor 2.0	REPEAT 1.0	1.000	6.872	0.084	0.802
EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	0.836	6.872	0.087	0.816
SelfGrade 9.0, EduIndex 2.0	REPEAT 1.0	1.000	5.743	0.082	0.716
SelfGrade 9.0, EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000	6.872	0.081	0.783

4.3 结果分析

Grade9.0 表示学生为 9 年级，而 9 年级的学生与复读的关联比较强，而 9 年级应该是初中升高中的阶段，所以具有一定的合理性。并且初始选取的特征中在支持度阈值 0.6 的条件下，ParentFactor 与 PowerIndex 没有出现在频繁项集中，说明 ParentFactor 与 PowerIndex 的影响相对更小，而 EduIndex 2.0 代表的是数据集中学生教育水平最低的档次，所以根据常识判断，这些频繁项集应该是合理的，而关联规则中 EduIndex 的 $confidence$、$lift$、PS、以及 $\phi - coefficient$ 相对其他特征都更低一些，这是因为 EduIndex 本身与 REPEAT 特征的相关性就比较低，而其他的前项都有着较高的置信度等评估指数，符合对数据的直观理解。

5 单个国家的频繁项集关联规则

5.1 各个国家与总体的差异

用同样的支持度阈值与置信度阈值分别计算西班牙、墨西哥、智利、巴拿马这四个国家的频繁项集与关联规则。得到的各个国家的频繁项集与关联规则计算结果在5.2中以表格形式给出。

图4是数据集中学生所属国家的分布，按照学生数量排序，依次是西班牙、墨西哥、智利、巴拿马、多米尼亚共和国、哥斯达黎加。因为多米尼亚共和国和哥斯达黎加占比极小，两者相加还不到 0.25%，无法产生频繁项集与关联规则，所以只关心西班牙、墨西哥、智利、巴拿马四个国家的情况。

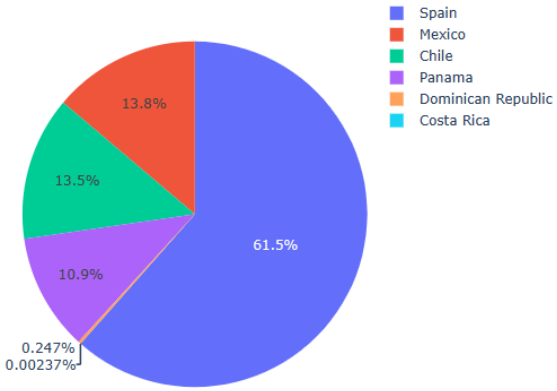


图 4: 各个国家人数占比

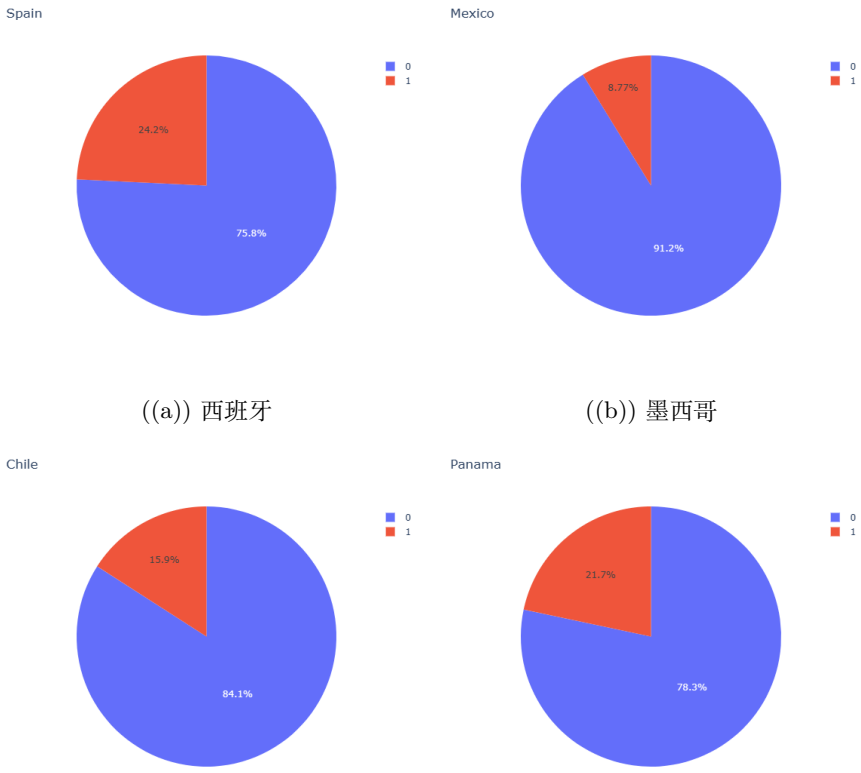
发现其中最明显差异的国家为巴拿马，在支持度阈值为 0.6，并且只在 REPEAT=1 的数据集上寻找频繁项集时，最多只找到了频繁二项集 (REPEAT 1.0, EduIndex 2.0)，而与 REPEAT 明显强相关的特征 RepeatFactor 2.0(支持度:0.58) 以及 SelfGrade 9.0(支持度:0.52) 在阈值为 0.6 时却没有出现，这与总体数据集的情况存在差异。

而剩余的西班牙、墨西哥、智利国家寻找出的频繁项集和总体的差异不大，基本保持一致，比如根据表9，在支持度阈值 0.6 的条件下，西班牙的频繁项集与总体是一致的，这应该是因为总体数据中西班牙的学生占比是最高的，所以总体的数据分布受西班牙的数据影响较大，从而二者的频繁项集与关联规则是一致的。并且对于每个国家的关联规则，评价指标 $\phi - coefficient$ 的差异也不大，说明找出的这些前项在各个国家对 REPEAT 1.0 都是有影响的。

但也存在一些细微的差异，比如，西班牙的每一个关联规则的 $lift$ 提升度都有所降低。而其他国家的关联规则提升度都有较大幅度的提高，这应该也与西班牙的数据集占比最大有关。在西班牙的关联规则中 EduIndex 2.0 与 REPEAT 1.0 的协方差 PS 降低至 0，说明 EduIndex 2.0 与西班牙学生是否复读关联不大，在墨西哥的关联规则中，SelfGrade 9.0 的置

信度为 0.496，而总体数据集 SelfGrade 9.0 的置信度为 0.787，说明对于墨西哥的学生，他们的年级与是否复读的关联要相对于整体更弱一些，而对于智利的学生前项 EduIndex 2.0 的置信度为 0.534，总体数据集 EduIndex 2.0 的置信度仅为 0.206，说明 EduIndex 对于智利学生是否复读影响更明显一些。

西班牙、墨西哥、智利、巴拿马四个国家各自学生复读的情况分布如图5所示。



((a)) 西班牙　　((b)) 墨西哥

((c)) 智利　　((d)) 巴拿马

图 5: 各个国家复读人数分布

5.2 西班牙、墨西哥、智利、巴拿马的频繁项集与关联规则

表 9: 西班牙频繁项集与支持度

频繁项集	支持度
{'REPEAT 1.0'}	1.000000
{'REPEAT 1.0', 'EduIndex 2.0'}	1.000000
{'REPEAT 1.0', 'RepeatFactor 2.0'}	0.766667
{'REPEAT 1.0', 'SelfGrade 9.0'}	0.746171
{'REPEAT 1.0', 'EduIndex 2.0', 'RepeatFactor 2.0'}	0.766667
{'REPEAT 1.0', 'SelfGrade 9.0', 'EduIndex 2.0'}	0.746171
{'REPEAT 1.0', 'SelfGrade 9.0', 'RepeatFactor 2.0'}	0.725225
{'REPEAT 1.0', 'SelfGrade 9.0', 'EduIndex 2.0', 'RepeatFactor 2.0'}	0.725225

表 10: 西班牙关联规则及其度量指标

前项	后项	置信度	lift	PS	ϕ 系数
EduIndex 2.0	REPEAT 1.0	0.185	1.001	0.000	0.012
RepeatFactor 2.0	REPEAT 1.0	1.000	5.401	0.116	0.853
SelfGrade 9.0	REPEAT 1.0	0.911	4.920	0.110	0.790
EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000	5.401	0.116	0.853
EduIndex 2.0, SelfGrade 9.0	REPEAT 1.0	0.911	4.920	0.110	0.790
SelfGrade 9.0, RepeatFactor 2.0	REPEAT 1.0	1.000	5.401	0.109	0.826
SelfGrade 9.0, EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000	5.401	0.109	0.826

表 11: 墨西哥频繁项集与支持度

频繁项集	支持度
{'REPEAT 1.0'}	1.000000
{'REPEAT 1.0', 'EduIndex 2.0'}	0.845730
{'REPEAT 1.0', 'SelfGrade 9.0'}	0.66667
{'REPEAT 1.0', 'RepeatFactor 2.0'}	0.804407
{'SelfGrade 9.0', 'REPEAT 1.0', 'EduIndex 2.0'}	0.66667
{'REPEAT 1.0', 'EduIndex 2.0', 'RepeatFactor 2.0'}	0.768595
{'REPEAT 1.0', 'SelfGrade 9.0', 'RepeatFactor 2.0'}	0.650137
{'REPEAT 1.0', 'EduIndex 2.0', 'RepeatFactor 2.0', 'SelfGrade 9.0'}	0.650137

表 12: 墨西哥关联规则及其度量指标

前项	后项	置信度	lift	PS	ϕ 系数
EduIndex 2.0	REPEAT 1.0	0.534	8.312	0.048	0.645
SelfGrade 9.0	REPEAT 1.0	0.496	7.720	0.037	0.541
RepeatFactor 2.0	REPEAT 1.0	1.000	15.567	0.048	0.891
EduIndex 2.0, SelfGrade 9.0	REPEAT 1.0	0.496	7.720	0.037	0.541
EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000	15.567	0.046	0.870
SelfGrade 9.0, RepeatFactor 2.0	REPEAT 1.0	1.000	15.567	0.039	0.797
SelfGrade 9.0, EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000	15.567	0.039	0.797

表 13: 巴拿马频繁项集与支持度

频繁项集	支持度
{'REPEAT 1.0'}	1.000000
{'REPEAT 1.0', 'EduIndex 2.0'}	0.788617

表 14: 巴拿马关联规则及其度量指标

前项	后项	置信度	lift	PS	ϕ 系数
EduIndex 2.0	REPEAT 1.0	0.656514	5.152037	0.080987	0.674548

表 15: 智利频繁项集与支持度

频繁项集	支持度
{'REPEAT 1.0'}	1.000000
{'REPEAT 1.0', 'EduIndex 2.0'}	0.845730
{'REPEAT 1.0', 'SelfGrade 9.0'}	0.666667
{'REPEAT 1.0', 'RepeatFactor 2.0'}	0.804408
{'SelfGrade 9.0', 'REPEAT 1.0', 'EduIndex 2.0'}	0.666667
{'REPEAT 1.0', 'EduIndex 2.0', 'RepeatFactor 2.0'}	0.768595
{'REPEAT 1.0', 'SelfGrade 9.0', 'RepeatFactor 2.0'}	0.650138
{'REPEAT 1.0', 'EduIndex 2.0', 'RepeatFactor 2.0', 'SelfGrade 9.0'}	0.650138

表 16: 智利关联规则及其度量指标

前项	后项	置信度	lift	PS	ϕ 系数
EduIndex 2.0	REPEAT 1.0	0.533913	8.311688	0.04779	0.64476
SelfGrade 9.0	REPEAT 1.0	0.495902	7.719945	0.037277	0.541293
RepeatFactor 2.0	REPEAT 1.0	1.000000	15.567493	0.048353	0.890927
EduIndex 2.0, SelfGrade 9.0	REPEAT 1.0	0.495902	7.719945	0.037277	0.541293
EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000000	15.567493	0.0462	0.869814
SelfGrade 9.0, RepeatFactor 2.0	REPEAT 1.0	1.000000	15.567493	0.03908	0.7968
SelfGrade 9.0, EduIndex 2.0, RepeatFactor 2.0	REPEAT 1.0	1.000000	15.567493	0.03908	0.7968

参考文献

[1] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley & Sons, 1994.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. ACM, 1993, pp. 207–216.

[3] A. Agresti, Categorical Data Analysis. John Wiley & Sons, 2013.

[4] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011.

[5] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis. John Wiley & Sons, 2012.

[6] R. Rosenthal and R. L. Rosnow, Essentials of Behavioral Research: Methods and Data Analysis. McGraw-Hill, 1991.