Advances in Social Science, Education and Humanities Research, volume 615
Proceedings of the 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021)

Analysis of Student Performance Based on Differential Privacy Protection

Zhao Lin[1,a], Sun Yingping[2,b*]

[1]*Department of Information Management and Information System, Shandong Normal University, Changqing, Jinan, China*
[2]*Department of Engineering Management, Shandong Normal University, Changqing, Jinan, China*
[a] *sdnuzhl@163.com*

[b*] *462124877@qq.com*

ABSTRACT
In the context of the information age, the use of data has greatly promoted the development of society. Education is inseparable from the development of the country， and education has become an area of close concern to society. How to maximize the useful value of mining data while protecting the privacy contained in the data is a problem we need to pay attention to. Based on the test scores of students in multiple courses, this paper analyzes the Apriori algorithm in the correlation analysis and mining, and obtains the correlation between the courses and the interval distribution of test scores. On this basis, we filter the obtained association rules, and use the Laplacian mechanism to increase noise interference to the filtered association rules, and protect the privacy of students through differential privacy. While using the maximum value of data as much as possible to promote the development of targeted teaching work, the privacy of students' performance is protected.

Keywords: *Differential privacy, association rule mining, student performance analysis*

基于差分隐私保护的学生成绩分析

赵琳 [1, a] 孙英平 [2, b*]

[1] *山东师范大学信息管理与信息系统系，长清，济南，中国*
[2] *山东师范大学工程管理系，长清，济南，中国*
[a] *sdnuzhl@163.com*

[b*] *462124877@qq.com*

摘要
在信息时代的背景下，数据的利用极大地推动了社会的发展。教育与国家的发展密不可分，教育也成为社会密切关注的领域。怎样在挖掘数据有用价值最大化的同时，尽可能地使数据中所包含的隐私不被泄露是我们需要关注的问题。本文针对学生多门课程的考试成绩，借助关联分析挖掘中的 Apriori 算法进行分析，得到各门课程间的关联性及考试成绩区间分布。在此基础上，我们对得到的关联规则进行筛选，采用拉普拉斯机制对筛算后的关联规则增加噪声干扰，实现差分隐私保护。在尽可能利用数据最大价值推动针对性教学工作发展的同时，使学生的成绩隐私得到了保护。

关键词：*差分隐私，关联规则挖掘，学生成绩分析*

1. 绪论

1. 1. 研究背景和意义

随着互联网时代数据信息的研究不断深入，数据间关联规则的挖掘迅速发展成为各个研究领域乃至各国政府的热门话题。挖掘数据相关性的算法有很多，其中关联规则挖掘算法得到了广泛的应用。同时，在关联规则挖掘算法的发展中首次提出了 Apriori 算法。由于其原理简单，结果更容易实现，

现已成为最实用的算法。尽管研究数据相关性的方法层出不穷，但在现有的数据相关性结论下，如何保护私有数据变得更加重要。差异化隐私保护可以有效保护数据:第一，无论隐私攻击者拥有多少相关的已知信息，都无法推断出信息所有人的其他重要隐私信息；其次，统计模型的严密性使得在使用差别隐私时有更好的准确性。

在教育领域，挖掘学生各学科成绩之间的相关性，有助于学校对课程之间的关系有更深入的了解，从而更加有针对性地为学生定制课程。同时，有利于学生检查和填补空白，加强优势学科，弥补薄弱学科。但是，在获取每门课程的相关结果时，存在学生成绩信息泄露的风险。从而我们可以采用差分隐私来保护学生的个人隐私。拉普拉斯机制中服从随机分布的噪声值干扰会使学生只获得自己的学习成绩信息，而不能推断出他人的学习成绩信息。

综上所述，挖掘教育领域中基于差异隐私保护的学生各学科成绩的相关规律具有重要意义。在本文中，首先利用关联规则找出课程之间的联系，使学校能够更好地优化课程安排，之后通过差分隐私引入噪声值干扰，保护学生成绩隐私不会在成绩信息发布的过程中被泄露。

1.2.研究内容

然而，在学生所学的众多课程中，有一些课程的相关性较强，直接影响了学生的学习情况。因此如何有效挖掘学生学习成绩之间的关联规则，成为一个值得研究的课题。基于以上讨论，本文决定采用关联分析中的 Apriori 算法来挖掘学校课程之间的相关性，从而为学校课程优化提供便利。此外，随着越来越多的数据之间的相关性被挖掘出来，由于学生们对未知的事物的好奇心加重，在获得自己的成绩信息后，他们更容易推算出相应学科其他学生的分数。这样一来，学生成绩的隐私泄露的风险也就显现出来。因此，我们需要一种正确且合乎实际的方法来保护学生的学科成绩数据不被泄露。基于此，本文采用差分隐私保护机制引入噪声值，使学生无法推算出其他学生的分数以保护好学生的成绩隐私。

2.理论基础

2.1.关联规则

项集：令 $I=\{i_1,i_2,.....,i_n\}$ 是 n 个不同特征项组成的集合。其中，每个 $i_k(k=1,2,.....,n)$ 称为项。含有 k 个项的项集称为 k—项集。例如，{啤酒，面包，牛奶} 是一个 3—项集。

事务集：D 是包含若干事务 T 组成的事务集，每个事务都有唯一的标识符 Tid，每一个事务 T 是一个项集的子集。

支持度（support）:事务集 T 中包含 X 项集的事务数与总事务数的比值，即为项集 X 在事务集 T 中出现的概率。

$$s(X \to Y)=\frac{\sigma(X \cup Y)}{N} \quad (1)$$

其中 $\sigma(X \cup Y)$ 表示同时包含 X 和 Y 项集的支持度计数，X 与 Y 是两个不相交的项集，N 为事务总数。

置信度（confidence）：$X \to Y$ 的支持度与 Y 支持度比值，表示 Y 项集在包含 X 项集的事务集中出现的条件频率。

$$c(X \to Y)=P(X|Y)=\frac{\sigma(X \cup Y)}{\sigma(Y)} \quad (2)$$

频繁项集：频繁项集挖掘就是找到满足用户指定最小支持度阈值的所有项集，我们把这些项集称为频繁项集。

关联分析[1]是用于发现隐藏在已知事务集中各项间的联系的一种方法数据挖掘分析方法。用关联规则或频繁项集的形式来表示其所发现的联系。我们所熟知的经典应用沃尔玛超市里“啤酒与尿布”就是通过提取关联规则，发掘出啤酒与尿布的最佳销售组合[2]。此外，关联分析的应用领域非常多，本文主要研究在学生成绩分析方面的应用。

2.2.关联规则的算法实现原理

Apriori 算法：运用了以下两个关键的原理，提高了算法的效率。

先验原理：频繁项的所有非空子集一定都是频繁的。

原理：非频繁项集的所有超集一定都是非频繁的。

关联分析主要分为两个子任务：

（1）找到所有频繁项集

（2）提取关联规则在置信度高的关联规则

首先生成频繁 1— 项集 L_1，然后为频繁 2— 项集 L_2，……，直到项集 L_r 为空集时算法停止。在第 i 次迭代中 $(i>1)$，先产生候选择项集 C_i，其中的每个项只是对频繁 $i-1$— 项集中两个包含一个不同项的项集进行连接。筛选出 C_i 中满足给定的支持度最小阈值的候选项集形成频繁 i— 项候选项集，

得到频繁项集 L_i。

计算出频繁项集，我们就可以通过公式得到强关联规则。在上述过程中，可以明显的看出 Apriori 算法需要产生大量候选项集并经过多次重复地扫描数据库，因此它的运行时间及开销会随着事务数的增大而增大。

2.3. 差分隐私

定义：根据 Dwork 在 2006 年针对统计数据库的隐私泄露问题提出的一种新的隐私定义[3]，我们可以得到差分隐私的定义：数据集的计算处理结果对于特定记录的变化是不敏感的，数据集中的单个记录的增加或删除对计算结果影响很小，因此，向数据集中添加一条记录所造成的隐私泄露风险被控制在很小的可接受范围内，攻击者无法通过观察计算结果获得准确的个人信息。也就是说，通过在数据中添加噪声干扰来模糊数据集的计算结果，攻击者不能通过观察计算结果来推断出准确的个人信息，从而保护了在公开发布的数据中存在泄露风险的用户隐私。

简单来说，就是在保留统计特征的基础上，删除差异化的个体特征以保护用户隐私。采用一种更形式化的方法来理解差异隐私:首先，定义相似数据集。有两个数据集 D，它们具有相同的属性结构，但它们之间只有一个数据不同，称之为相似数据集。使用随机算法 F 对查询结果进行处理，如果查询结果中相邻数据处理的结果满足公式[4]（3），说明随机算法 F 提供了差分隐私保护。

$$Pr\{F(D)=O\} \le e^{\varepsilon} * Pr\{F(D')=O)\} \quad (3)$$

其中假设 Po 为用数据处理时运用随机算法 F 得到的所有可能结果的集合，则 O 为 Po 的任意子集，Pr{}为随机算法 F 得到某个特定结果的概率，ε 越为隐私保护预算（或者称隐私保护程度调节参数）。通常而言，ε 越小，数据的保密度越高。ε 越大，数据的可用度越高，数据的保密程度越低。因此，确定合适的 ε 是很关键的。若将该算法应用于任意两个相邻的数据集中，得到特定输出公式的概率应该是近似的。仅仅通过观察输出结果，观察者很难察觉到数据集的细微变化，从而达到隐私保护的目的。

2.4. 差分隐私的 Laplace 实现机制

基于本文中数值型的学生成绩数据，本文主要采用的是 Laplace 机制，实际输出值受到其分布产生的噪声的干扰影响，实现了差分隐私保护。

定义[4]：对于任意一函数 $f: D \to R^d$，函数的全局敏感性为

$$\Delta f = MAX_{D,D'} \| f(D) - f(D') \|_p \quad (4)$$

其中，D 和 D' 至多相差一条记录，p 表示度量使用的 L_p 距离，通常使用 L_1 来度量。

对于任意一个映射函数 $f: D \to R^d$，若有一随机算法 A 满足以下等式，我们就称算法 A 对数据集 D 提供了 $\varepsilon-$ 差分隐私保护[4]。

$$A(D) = f(D) + \left(\frac{Lap_1}{\Delta\varepsilon} + ,\ldots\ldots, + \frac{Lap_d}{\Delta\varepsilon}\right) \quad (5)$$

其中，$Lap_i / \Delta\varepsilon (1 < i < d)$ 是相互独立的拉普拉斯变量，R 表示所映射的实数空间，d 表示函数 f 的查询维度。噪声大小与 Δf 成正比，与 ε 成反比。算法 A 的全局敏感性越大，所需噪音就越大。

3. 基于差分隐私的学生成绩关联分析应用

3.1. 数据来源

首先，本文收集了本校商学院信息管理与信息系统专业 2017 级 68 名学生在校期间学习《计算机网络与应用》等十八门课程的成绩信息，主要包括十八门学科的学期末测试成绩。研究使用的信息包括学生序号，学科成绩，学号，课程序号等几部分，其中部分数据如图 1 所示：

学号	计算机网络与应	企业管理	马原	管理学	营销学	概率论	数据库设计
201824040101	93	88	74	89	85	90	75
201824040102	97	81	69	83	83	61	67
201824040103	91	88	68	84	83	75	76
201824040104	92	82	72	78	85	81	83
201824040105	76	87	65	93	81	78	71
201824040106	99	90	74	93	88	86	91
201824040107	98	88	66	81	78	75	71
201824040108	93	91	74	86	84	90	81
201824040109	99	88	79	93	90	88	84
201824040110	99	89	73	90	81	78	73
201824040111	99	92	83	93	85	79	94
201824040112	98	90	74	92	90	87	76
201824040113	100	85	71	94	86	71	70
201824040114	98	88	76	88	87	60	76
201824040115	96	88	70	86	82	85	72
201824040116	97	88	68	80	78	62	72
201824040117	90	82	81	87	81	64	82
201824040118	100	89	66	82	79	60	50

图 1 部分原始学生成绩数据

3.2.数据处理

首先对原始数据集中的数据进行删除重复值、处理缺省值并对数据进行标准化处理的操作收集的学生学科成绩中，实现数据清理。仍存在因学生兴趣差异而导致部分学生未选修课程的残缺值，因此，本文以各科课程测试成绩的平均成绩进行填充。

由于分数的数值区间较小，用学生分数的单一维度作为测试数据挖掘学科间的关联规则，不仅会导致算法执行效率低下且无法获得准确的相关性。因此我们在学生成绩为离散化百分制数值型数据的基础上，将各科成绩按照表 1 所示规则划分为 A~E 的五个区间，最终得到处理汇总之后的数据如图 2 所示。

表 1　学生成绩区间划分规则

学生成绩	区间等级
[90,100]	A
[80,90)	B
[70,80)	C
[60,70)	D
[0,60)	E

学号	序号	计算机网	企业运营	马原	管理学	营销学	概率论	数据库设
201824040101	1	A1	B2	C3	B4	B5	A6	C7
201824040102	2	A1	B2	D3	B4	B5	D6	D7
201824040103	3	A1	B2	D3	B4	B5	C6	C7
201824040104	4	A1	B2	C3	C4	B5	B6	B7
201824040105	5	C1	B2	D3	A4	B5	C6	C7
201824040106	6	A1	A2	C3	A4	B5	B6	A7
201824040107	7	B1	B2	D3	B4	C5	C6	C7
201824040108	8	A1	A2	C3	B4	B5	A6	B7
201824040109	9	A1	B2	C3	A4	A5	B6	B7
201824040110	10	A1	B2	C3	A4	B5	C6	C7
201824040111	11	A1	A2	B3	A4	B5	C6	A7
201824040112	12	A1	A2	C3	A4	A5	B6	C7
201824040113	13	B1	B2	C3	A4	B5	C6	C7
201824040114	14	B1	B2	C3	B4	B5	D6	C7
201824040115	15	A1	B2	C3	B4	B5	B6	C7
201824040116	16	A1	B2	C3	B4	C5	D6	C7
201824040117	17	A1	B2	B3	B4	B5	D6	B7
201824040118	18	A1	B2	D3	B4	C5	D6	E7

图 2 处理后的学生成绩数据

3.3.实验结果及分析

3.3.1.学生成绩关联分析

我们首先得到包含一个事务的候选项及集，即单门课程成绩所处区间类别及对应的支持度，如表 4-2 所示。所统计的十八门课程中，B(分数区间 80~90)是 9 个学科中占比最高的区间等级，A(分数区间 90~100)是 6 个学科中所占比最高的区间等级。分类的最高比例中没有 D 或 E 类，这说明学生在 18 个学科中的期末成绩大多在 80 分以上。81%的学生计算机网络及应用的成绩在 90 分以上，表明学生对该学科的掌握程度较好。概率论与数理统计、运筹学以及数据结构三门课程学生成绩占比最高的都为 B 区间，但其比例都接近 30%，说明该门课程对应的学生成绩分布较为均匀。

表 2　单门课程成绩所处区间分布

学科	占比最高区间	支持度
计算机网络与应用	A	0.81
企业运营管理	A	0.66
马克思主义原理	C	0.69
管理学	A	0.54
营销学	B	0.82
概率论	B	0.29
数据库设计	C	0.37
大学体育	B	0.5
应用统计学	B	0.43
机器学习	C	0.37
运筹学	B	0.29
毛概	B	0.49
Python	A	0.54
信息检索	B	0.51
数据结构	B	0.32
高等数学	A	0.76
线性代数	A	0.25
综合英语	B	0.54

进一步我们可以得到包含两个事务的频繁项集（给定支持度最小阈值为 0.6），及两门课程为同一成绩区间等级的分布情况及对应支持度，如表 3 所示。分析得出：计算机网络与应用最高区间均为 A，且所占比例高达 81%，表明大多数学生在学习此两门课程中，较易理解，得分都较高。而营销学与计算机网络与应用的组合中，营销学分类级别为 B 与计算机网络分类为 A 的所占比例最高，表明大多数学生在学习两门课程的最终得分都高于 80 分。

表 3　两门学科为同一成绩区间等级的分布情况

处于同一成绩区间的两门学科	区间	支持度
计算机网络，营销学	A，B	0.66
计算网络，高等数学	A，A	0.63
马原，营销学	C，B	0.6
营销学，高等数学	B，A	0.65

可获得到包含三个事务的频繁项集（给定支持度最小阈值为 0.45），即三门课程为同一成绩区间等级的分布情况及对应支持度，如表 4 所示。在计算机网络、营销学与高等数学的组合中，有 52%的学生高等数学与计算机网络同时得分在 90 分以上，营销得分在 80-90 分之间。

表 4　三门课程为同一成绩区间等级的分布情况

学科	区间	支持度
计算机网络，营销学，高等数学	A，B，A	0.52
计算机网络，马原，营销学	A，C，B	0.47
马原，营销学，高等数学	C，B，A	0.48

由此，我们通过分析挖掘得到隐藏在 18 门课程学生成绩中的强关联规则，如表 5 所示。可以看出计算机网络与大部分课程之间都存在成绩的强关联性，在已知该名学生计算机网络期末成绩为九十分上的条件下，我们可以推断出他高等数学成绩在九十分以上，营销学成绩在 80 到 90 之间，马原成绩在 70 到 80 之间的概率分别为 83%，80%，81%。另外，营销学分别与高等数学、马原的课程成绩组合之间的也存在着强关联性，置信度均大于 0.8。

表 5 课程成绩之间的强关联规则

强关联规则	置信度
A1,C3	0.81
A1,B5	0.8
A1,A16	0.83
B5,C3	0.83
A16,C3	0.79
B5,A16	0.83

3.3.2. 学生成绩隐私保护

从上的分析我们可以得出，以关联规则 $A11 \Rightarrow A16$ 为例，如果已知计算机网络和高等数学的成绩分布间的关系规则，那么在一个学生计算机网络成绩是 A 的前提下，很简单地就可以推断出他的学生计算网络成绩有 83%的概率在 90 分以上。因此我们需要对发布的学生成绩进行隐私保护，本文以高等数学成绩数据的差分隐私保护为例，对学生的期末成绩采用不同的隐私保护参数，进而观察不同结果中隐私安全度的影响。图 3、图 4 和图 5 所示分别为三个不同隐私参数值所对应数据处理结果。

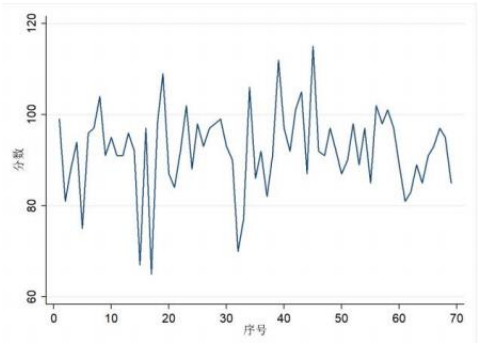

图 3　$\varepsilon = 0.2$

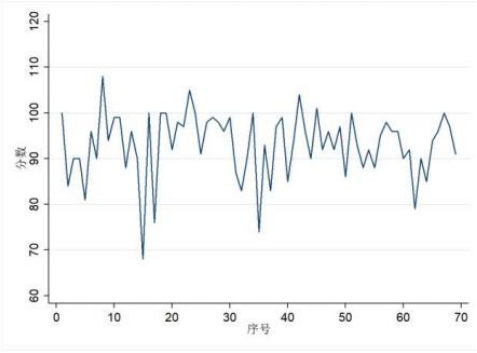

图 4 $\varepsilon = 0.5$

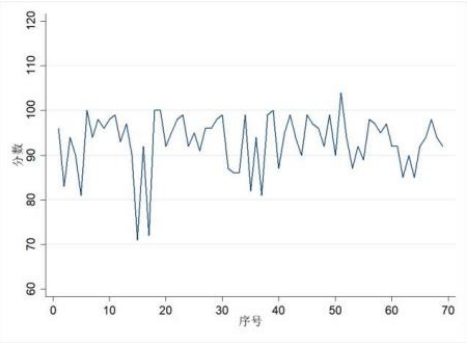

图 5 $\varepsilon = 1$

由以上三个图中的不同结果可以看出，当隐私参数设置不同时，噪声值的波动范围会不断变化。ε 值越小，越有可能偏离原来的数值，这样一来学生的分数范围就会波动很大，最大值也会扩变大，处理后数据的可利用价值也随之下降。但这将带来更高程度的隐私保护。因此，在未来的研究中，我们可以通过设置不同的 ε 值来实现隐私保护级别之间的划分，找到隐私保护级别与实现数据可利用价值之间的合理分配。

4. 结论

本文主要基于学生成绩数据进行了相关联分析与信息安全保护方面的研究，首先，明确了学生成绩数据分析对于高校课程优化的重要意义。其次，通过运用 Apriori 算法挖掘各学科学生成绩的相互关联规律，并分析各学科间存在的相互影响关联，为未来学校课程设置提供了较为直接的依据。第三，学生隐私泄漏的几率随着发布个人信息的数量的上升而提高，且我们根据上述数据分析得到的相关结论，也具有可能导致学生隐私的泄漏的风险。为了克服此问题隐患，人们可使用差分隐私的 Laplace 机制，调整隐私参数 ε 以控制隐私保护程度，在尽可能利用数据最大价值推动针对性教学工作发展的同时，使学生的成绩隐私得到了保护。在大数据时代，个人、组织之间的数据共享越来越多，数据共享的趋势越来越明显[5]。针对不同领域的个性化需求，我们可以在精细化的数据挖掘算法的基础上研究针对性的隐私保护算法。

项目基金

本文为 2020 年度大学生创新创业项目《大数据抗疫下患者隐私泄露机理及保护技术研究》的阶段性成果之一。

REFERENCES

[1] Pang-Ning Tan,Michael Steinbach(2011) Correlation Analysis. In Yang Hailing,Y.H.L(Eds.),Introduction to Data Mining.Post&Telecom Press, Beingjing.201-250.

[2] Beer and diapers: a magical shopping basket analysis [J]. Guangcai, 2009(01):63.

[3] Dwork, C. (2008) Differential Privacy:A Survey of Results. In: Agrawal, M., Du, D.Z., Duan, Z.H. and Li, A.S.(Eds)., Theory and Applications of Models of Computation, Springer, Berlin, 1-19.

[4] Dwork, C., F. McSherry, K. Nissim & A. Smith. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), Theory of Cryptography, Proceedings (Vol. 3876, pp. 265-284).

[5] Fang Yuejian, Zhu Jinzhong, Zhou Wen, Li Tongliang. A Survey of Data Mining Privacy Protection Algorithms[J].Information Network Security,2017(02):6-11.