基于概率统计与德尔熵值法的隐私保护综合评价模型

史玉良[1),2)] 周 卫[3)] 臧淑娟[1)] 陈 玉[1)]

[1)](山东大学软件学院 济南 250101)

[2)](山大地纬软件股份有限公司 济南 250101)

[3)](中共山东省委党校图书馆 济南 250103)

摘 要 云计算环境中,租户将数据存储于 SaaS(Software as a Service)应用平台中,利用分块混淆的隐私保护技术将数据切分为多个数据块并存储到不同的存储模式中,实现明文状态下租户数据的分离与保护.但是这种隐私保护技术的隐私保护程度如何,用户是无法明确感知的,为此,针对分块混淆的隐私保护技术,该文提出一种基于概率统计与德尔熵值法的隐私保护综合评价模型.首先分析了基于分块混淆的隐私保护技术的当前状况,在此基础上定义了隐私保护评价指标,利用概率统计的知识,定义评价指标的计算规则,构建隐私保护层次分析模型,通过由数据块层到数据存储模式层(Data Storage Mode,DSM)层再到顶层回逆的方式,得到隐私保护后数据分布的评价指标值;然后分析了德尔菲法和熵值法在权重确定方面的优势,将德尔菲法的主观判别与熵值法的客观判断相结合,改进两种方法的计算过程,提出基于德尔熵值法的指标权重确定模型,得到隐私保护效果评价指标的权重;最后定义评价等级,建立基于概率统计和德尔熵值法的隐私保护综合评价模型,实现基于分块混淆的隐私保护技术的综合评价.实验结果证明本文提出的综合评价模型不仅可以客观地评价基于分块混淆的隐私保护技术的隐私保护效果,也证明了分块混淆隐私保护技术的有效性,为 SaaS 应用平台中的数据隐私保护提供了强有力的理论支撑.

关键词 隐私保护;概率统计;德尔熵值法;层次分析;模糊综合评价

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2019.00786

A Comprehensive Evaluation Model of Privacy Protection Based on Probability Statistics and Del-Entropy

SHI Yu-Liang[1),2)] ZHOU Wei[3)] ZANG Shu-Juan[1)] CHEN Yu[1)]

[1)](*School of Software, Shandong University, Jinan* 250101)

[2)](*Dareway Software Co., Ltd., Jinan* 250101)

[3)](*CPC Shandong Provincial Party School Library, Jinan* 250103)

Abstract In the environment of cloud computing, the data is stored on the SaaS application platform by users. On the cloud-based SaaS application support platform, using block confusing privacy protection technology, data is divided into multiple data blocks and stored in different storage patterns, which achieves the separation and protection of privacy data. At the same time, the correlation of the data slices is stored on the reliable third party's platform by introducing trusted third party, which achieves the separation and protecting of private data in clear text. However, users are not clear to what degree their private information is protected by using this privacy protection method. In order to solve this confusion, this paper proposes a comprehensive privacy

收稿日期:2017-10-18;在线出版日期:2018-05-27.本课题得到国家重点研发计划(2018YFC0114709)、山东省泰山产业领军人才工程专项经费(TSCY20150305)、山东省重点研发计划(2016GGX101008,2016ZDJS01A09)、山东省自然科学基金重大基础研究项目(ZR2017ZB0419)资助.**史玉良**,男,1978 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为云计算、数据库、隐私保护.E-mail:shiyuliang@sdu.edu.cn.**周 卫**,女,1978 年生,硕士,副研究馆员,主要研究领域为数据库、隐私保护.**臧淑娟**,女,1992 年生,硕士研究生,主要研究领域为云计算、隐私保护.**陈 玉**,男,1990 年生,硕士研究生,主要研究领域为云计算、隐私保护.

preservation evaluation model to present an intuitive impression on block confusing privacy protection technology to users based on probability and Del entropy method. First of all, this paper defines the evaluation index of privacy protection by analyzing the current status of privacy protection technologies. Based on the knowledge of probability and statistics, the calculation rules of evaluation index are defined and the model of privacy protection levels composed of data blocks, Data Storage Mode (DSM), and privacy protection is constructed. By the method of backtracking and calculating the privacy protection evaluation index of each data blocks, the DSM layer privacy protection evaluation index, and the top privacy protection index, the evaluation index value of the data distribution after each tenant adopting the privacy protection policy is obtained. Then, using the weighting advantages of Delphi method and entropy method, the subjective judgments of Delphi method and the objective judgments of entropy method is combined. On the basis of improving the calculation process of these two methods, this paper proposes an index weight determination model based on Del-entropy method. By this model, the weights of the privacy protection effectiveness evaluation indicators are obtained. At last, the levels of evaluation are defined, the privacy preservation comprehensive evaluation model of privacy protection based on probabilistic statistics and Del-entropy method is set up to realize the intuitionistic evaluation of blocking confusion privacy protection methods, and a single-factor evaluation algorithm, evaluation index weight distribution algorithm, and privacy protection fuzzy comprehensive evaluation algorithm are proposed to calculate the privacy protection level of tenant data. In order to verify the correctness of the privacy protection evaluation index proposed in this paper and the effectiveness of the privacy protection evaluation model based on probability statistics and Del-entropy method, experiments are performed by using the privacy protection calculation method. The experimental results show that the comprehensive evaluation model proposed in this paper can not only objectively demonstrate the privacy protection effects of privacy protection technology based on block obfuscation, but also prove the effectiveness of blocking confidential privacy protection methods and pointed out the direction of improvement. The most important is that it provides a good theory for data privacy protection in SaaS application platform.

Keywords　privacy protection; probability statistics; Del-entropy method; analytic hierarchy; fuzzy comprehensive evaluation

1　引　言

随着信息技术和社会的发展，以及移动互联网、物联网、云计算应用的进一步丰富，数据已呈指数级增长，大数据时代悄然到来. 近几年，数据存储技术越来越成熟，应用也越来越多，人们在关注数据的存储、处理和融合的同时，也逐渐开始重视数据的隐私保护问题.

数据隐私保护一直深受各界学者的追捧，许多成熟的隐私保护方法已经被提出. 针对 SaaS(Software as a Service)应用面临的租户隐私泄露的风险，我们在文献[1-3]中分别提出了一种面向 SaaS 的隐私保护机制. 文献[1]利用租户自定义的隐私约束，将租户的隐私属性划分到不同的数据块中，并利用第三方实现数据切片之间的关联关系的混淆和重构，保证数据的统一；文献[2]利用键能算法实现属性的聚类，将关联关系密切的属性尽可能分到同一分块中，减少了数据分块之间的连接次数从而提高了应用性能；文献[3]针对租户的个性化隐私保护需求和事务处理响应要求，提出了一种基于策略的个性化隐私保护机制. 上述文献均基于非加密技术，在利用加密技术进行隐私保护方面，学者们也提出了很多方法. 文献[4]提出了一种面向矩阵乘法的数据混淆方法，通过随机添加噪声和重新洗牌的方法，实现矩阵的行列分裂，从而保护数据的隐私性. 文献

[5]在数据发布过程中利用 k-匿名的技术，实现数据隐私保护. 文献[6]针对大数据的隐私保护，基于 MapReduce 框架提出一种可扩展的两阶段自上而下的专业化匿名数据方法，该方法给借助 MapReduce 框架的多维度匿名化，极大地提高了数据的处理效率. MapReduce 的开源性可能造成恶意代码的嵌入，从而在计算过程中泄露用户的敏感信息，文献[7]通过限制用户访问系统资源的权限和向 Reduce 输出结果添加噪声的方式，实现敏感数据的隐私保护.

文献[1-7]从明文、加密以及加入噪声等方面均提出了适合不同领域的隐私保护方法，但这些隐私保护方法的效果如何评价目前业界并没有统一的标准. 不过，不同领域的学者在各自领域也已提出一些卓有成效的评价方法：文献[8]提出了一种对图像和视频隐私保护方法的客观评价框架，通过聚合和融合帧的两种新方法来评价现有的隐私保护方法及其实用性. 在分块混淆的隐私保护技术中由于数据分布不均衡，在数据应用过程中可能导致隐私泄露，针对这一问题，文献[9]提出了一种隐私保护效果评价机制及隐私泄露检测技术，并提出了处理隐私泄露的方法，但是该方法没有给出明确的隐私保护效果等级. 文献[10]通过一种无解释的评估方法，从隐私保护效果和性能两个方面进行评价，该方法既不依赖于主观判断，又不假定图像数据的数据类型，在图像隐私保护方面较具通用性. 在软件评估方面，通常的方法是通过范围缩减、多项式评估和重建三步进行评估，并在过程中尽量减少数据源误差对结果的影响，文献[11]针对这一问题，通过消除正弦和余弦评估的第二范围缩减中使用的列表值的舍入误差来消除三角函数评估过程中的误差来源，提高隐私评价准确性.

文献[8-11]针对各个领域的隐私保护技术进行评价，但是经过分析可知，这些评价方法与基于分块混淆的隐私保护技术不是非常匹配，而目前关于分块混淆的隐私保护技术评价的研究少之又少，因此本文基于分块混淆的隐私保护技术，提出一种针对该隐私保护技术的评价模型. 首先通过分析隐私保护效果的影响因素，确定评价指标；然后根据概率统计的知识提出一种基于评价指标的隐私保护层次分析模型，通过底层回逆的方式，得到租户数据在实施隐私保护策略之后的评价指标值；最后综合德尔菲法和熵值法的优势，提出一种基于德尔熵值法计算隐私保护评价指标权重的方法，得到评价指标的权重，提高评价指标权重的准确性，建立隐私保护综合评价模型，得到隐私保护技术的保护效果评价分数，从而实现对分块混淆的隐私保护技术的可视化评价.

本文第 2 节展示相关工作；第 3 节描述隐私评价指标的定义和基于分块混淆的隐私保护技术；第 4 节介绍基于概率统计的隐私保护层次分析模型；第 5 节介绍基于德尔熵值法的评价指标权重分配方法以及隐私保护综合模型；第 6 节是相关实验和分析；最后对本文进行简要的总结.

2 相关工作

在 SaaS 应用中，文献[1-3]基于分块混淆的隐私保护技术将原始数据进行分割并混淆数据之间的关联关系，引入可信任第三方，实现了明文状态下的数据隐私保护，但是目前缺少对这种隐私保护技术的效果评价，用户无法直观地感受到该技术的隐私保护程度.

目前，隐私保护效果评价倍受人们关注，且很多研究成果已应用于实践中. 文献[12]提出了一种基于可信性和隐私感知的云服务评价模型，该模型引用了时间衰减因子来解决信任动态随时间变化的问题，并且根据不同的交易金额给予不同的权重，该评价模型能够保证实际服务质量，对恶意实体的欺诈和恶意评估具有一定的检测和抵御能力，但并不能评估数据的隐私性. 文献[13]分析了防火墙、入侵检测和入侵容忍之间的关系，利用博弈论，均衡三者之间的关系，得到影响信息安全技术正负值的关键因素，提出针对于信息安全技术的综合评价模型，有效保护了用户隐私. 在虚拟培训过程中，现有方法在解决动态评估问题上稍有逊色，为此，文献[14]将时序因子引入，来构建动态评估模型，该模型采用时间步长和事件步长法组合的方法建立评价对象、评价时间和评价指标之间的三维评价矩阵，并使用加权平均函数来解决动态评估问题，实现动态评估过程的可视化. 但文献[13-14]都是针对特定问题的评价，在 SaaS 应用平台中，这些评价方法不能适应分块混淆方法的隐私评价. 文献[15]采用信息论来求解多目标评价问题，与本文的研究有一定的关联性，为本文使用的概率统计、信息熵等提供理论支持. 利用网络追踪技术可以取得客户的浏览记录和个人信息，如果这些信息被不法分子恶意截取，可能泄露用户的隐私，对此文献[16]提出一种动态的隐私评分模型，该模型通过追踪组件的隐私泄露程度和相关性评估与用户隐私相关的隐私泄露风险，从而保护用

户上网过程中的隐私安全.对于车载物联网中的位置信息隐私保护问题,目前的文献都集中在预防方法,以实现位置信息的隐私保护,文献[17]引入了一种新的风险评测方法评测车载网络中基于攻击树的隐私泄露情况.该方法提供了一个通用的分析框架评估特定的威胁源可能带来的隐私泄露程度,同时还构建了攻击树用以识别攻击者可能攻击隐私系统的情况.文献[16-17]通过隐私泄露风险的评估,有效地提高了数据的安全性,同时评估的方法对于本文的研究有一定的参考价值.对于变电站设计方案的评估,现有的方法没有考虑到故障次数、运行成本等因素的随机性,同时也缺乏对安全性和有效性评估,因此文献[18]综合考虑上述不确定变量、安全性和效率,提出一种不确定性的评估模型.在模型中,利用模糊层次分析法确定各指标的权重,然后采用蒙特卡洛模拟方法对模型进行求解.文献[19]基于感官数据和模糊集理论提出一种模糊层次分析综合评价模型,通过该模型可以对感知数据之间的关系进行评价,从而更好地分析感官数据,获取更有价值的信息.文献[18-19]均为将层次分析运用在模糊评价方面的研究,模糊评价是在多影响因素的状况下综合所有因素给出一种全面评价的决策方法,而层次分析法又是评价问题的常用方法,与本文要评价的基于分块混淆方法的隐私保护技术目标契合,因此本文基于此方法从评价指标筛选和权重确定两个方面进行深入研究.

在指标筛选和权重确定方面,德尔菲法和熵值法一直倍受关注,文献[20]综合分析了德尔菲法的优势和局限性,针对该方法时间周期长、易受专家主观影响等问题对德尔菲法的实现形式做出改良,重新定义了德尔菲法的研究步骤,对本文评价指标的选定具有指引作用;熵值法的基本思想是系统中的信息量越大,可变性就越小,熵也就越小,权重越大,反之信息量越小,可变性越大,熵值越大而权重越小,文献[21]介绍了熵值法的基本原理和求解过程,并将层次分析法引入到评价模型中,构建主观评价体系,但是熵值法完全以客观数据为依据,只能从客观上确定各评价指标的权重系数.

综上可知,文献[12-19]从不同领域选取不同评价指标采用不同的方法对隐私保护技术进行评价,为本文解决评价问题提供了方法参考,文献[20-21]针对指标筛选和权重确定问题改进了德尔菲法和熵值法,为解决基于分块混淆的隐私保护评价提供了有效的借鉴.本文基于上述研究成果,提出了一种基于概率统计与德尔熵值法的隐私保护综合评价模型,实现了分块混淆下的数据隐私保护的等级评价,同时也很好地证明了分块混淆方法对数据的隐私保护效果,为隐私保护的深入研究提供了理论支撑.

3　基础知识

大数据时代,由于企业的管理制度和权限分配不完善,经常会发生云计算服务商或数据库管理员通过各种技术手段获取用户数据及其变更记录恶意泄露用户隐私的情况,虽然文献[1]中提出的面向分块混淆的隐私保护机制能够隐藏用户关键属性的关联关系,将数据分片存储,实现数据隐藏,但是,用户在应用数据的过程中,数据库管理员等无背景知识的攻击者仍可以查看到云端数据的物理存储操作日志、数据变更记录关联关系存在数据分片分布不均衡等问题,有一定概率会造成隐私泄露,现有隐私保护技术在上述三个方面的表现决定了其隐私保护效果.基于此结论,本章首先分析基于分块的隐私保护技术的基本情况,然后从上述三个方面分析并定义影响基于分块混淆的隐私保护技术效果的因素,并将影响因素作为隐私保护效果的评价指标.

3.1　基于分块混淆的隐私保护技术

在前期工作文献[1-3]中,我们提出了一种基于分块混淆的隐私保护技术,该技术通过增加可信第三方构造安全交互模型,实现明文状态下的隐私保护,模型架构图如图 1 所示.

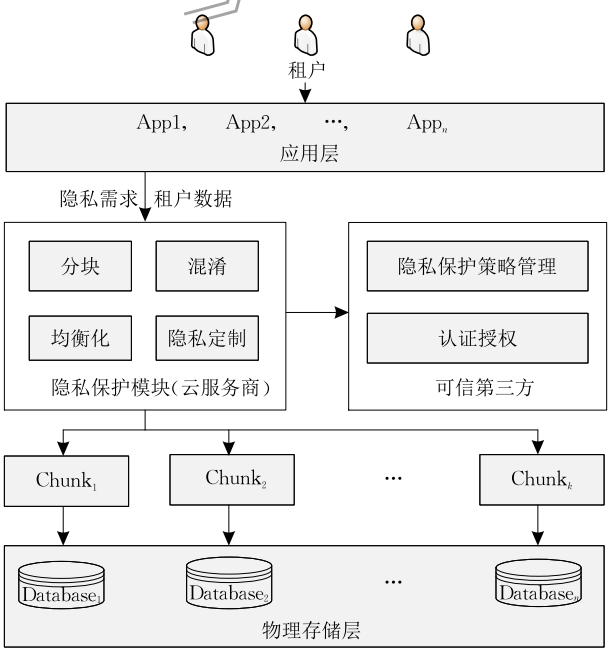


图 1　基于分块混淆的隐私保护架构图

首先,基于租户的敏感信息制定隐私约束,根据制定的隐私约束分割租户的数据,将每个隐私约束的隐私属性分配到不同的数据分块中,从而实现隐私属性之间的关联关系隐藏,同时对单隐私属性进行转换和拆分,使租户的每条数据切分到不同的数据分块中,并对数据切片进行标识,在明文状态下实现租户数据的隐私保护;然后通过建立可信第三方,将同一条租户数据的数据切片标识的关系存储到第三方,同时采用同态加密的方式实现分块数据的重构;最后,由于数据分块不均匀,可能会导致隐私泄露,提出了α、β和γ这3种均衡化机制[1],通过在数据分块中添加噪音数据,使每个数据分块满足均衡化要求,即保证数据隐私泄露的概率低于$\max(1/n, \beta^k, \gamma^k)$,平衡各个数据节点的存储能力、计算能力以及负载能力,将分块数据存储到各存储节点.

3.2 相关定义

由3.1节可以看出,影响隐私保护效果的因素主要存在于数据属性的泄露、数据切片分布不均、数据分块之间的关联关系泄露等方面,本节主要从这三个方面分析并定义相关影响因素.

定义1. 隐私相容元组(Fusion Tuple,FT),一条数据记录$DR=\{DA_1, DA_2, \cdots, DA_n\}$,$DA$表示租户隐私属性,根据不相容隐私约束,$DR$被分为了多个切片,即$DR=\{(DA_1, DA_3, DA_6), (DA_5, DA_8, DA_n), \cdots, (DA_4, DA_9, DA_{11})\}$,每个片段中的隐私属性组合都不会造成隐私泄露,即为一个隐私相容元组FT,对于任意的子元组FT_i和FT_j,有$FT_i \cap FT_j = \varnothing$.

在数据属性泄露方面,属性泄露的概率和比率是衡量隐私保护效果的重要参数,因此本文选用隐私属性泄露概率和隐私属性泄露比率两个评价指标来衡量隐私保护技术在防止数据属性泄露方面的效果,定义如下.

定义2. 隐私属性泄露概率(Privacy Attribute Leakage Probability,$PALP$),从经过隐私保护后的数据块集合中获得一条能唯一确定用户信息的元组标识或隐私属性的概率,定义为隐私属性泄露概率.例如,获得包含隐私属性"姓名"的隐私相容元组的概率.

假设某租户T_1经过隐私保护策略后,数据被分到k个不同的数据块中,每个数据块中的数据记录有$N_{R_i}(0<i\leq k)$条.租户T_1的数据中,包含隐私属性N_P个,已被分布在不同的数据块上.在N_P个隐私属性中,"姓名"属性只被分在一个数据块上,在这个数据块上,可能存在属性值相同的情况,即同名数据,假设有N_N条,则获取到正确姓名对应信息的概率为$P(NN)=1/N_N$.假设隐私属性PA_i在数据分块i中有N_V种属性取值,数据分块i中的总数据条数为R_i,则每种取值的概率$P(N_{V_j})=R_{N_{V_j}}/R_i(0<j\leq N_V)$,由全概率①定义可知隐私属性$PA_i$的取值有$N_V$种,分别为$EV_1, EV_2, \cdots, EV_{N_V}$,为一个完备事件组,每个事件都是独立且互不相容的,每个属性取值的概率为$P_{N_{V_j}}$,假设事件$B_i=\{PA_i$的泄露$\}$,那么事件B_i发生的概率为:

$$P(B_i)=\sum_{j=1}^{N_V}P(EV_j)P(B_i \mid EV_j) \tag{1}$$

由式(1)可知,在数据分块i中,隐私属性PA_i的块泄露概率为$P(B_i)$,所以PA_i的$PALP=P(B_i)\times P(N_N)$,即:

$$PALP_{Ai}=\frac{1}{N_N}\times\sum_{j=1}^{N_V}\left(\frac{R_{NV_j}}{R_i}\right)^2 \tag{2}$$

假定$PALP$的标准值为λ,且每个隐私属性的$PALP$值都不大于λ,则可根据式(2)计算出租户T_1的所有隐私属性的泄露概率,并记为隐私属性泄露概率向量$\boldsymbol{P}_1=[PALP_{PA1}, PALP_{PA2}, \cdots, PALP_{PAN_P}]$,如果$PALP_{Ai}>\lambda$,表示隐私属性$PA_i$发生了隐私泄露.

定义3. 隐私属性泄露比率(Privacy Attribute Leakage Ratio,$PALR$),假设数据存储模式(Data Storage Mode)DSM_1中有P_1个隐私属性,其中泄露隐私的属性有P_2个,则$PALR=P_2/P_1$.

在数据分布不均方面,本文采用隐私属性信息熵来衡量隐私数据的分布情况,信息熵用来衡量事物的稳定程度,而隐私属性信息熵能够衡量每个属性中数据取值是否均衡,从而衡量数据分布是否均衡,如定义4所示.

定义4. 隐私属性信息熵(Privacy Attribute Value Entropy,$PAVE$),假设数据分块i中的隐私属性$PA_i(1\leq i\leq n)$的取值有N_{PA_i}个,分别为$\{V_1, V_2, \cdots, V_{NPA_i}\}$,每种属性取值的概率为$VR_j(1\leq j\leq N_{PA_i})$,根据信息熵的计算公式,$PA_i$的隐私属性信息熵$PAVE_i$为:

$$PAVE_i=-\sum_{j=1}^{N_{PA_i}}VR_j\times\log_2(VR_j) \tag{3}$$

隐私属性信息熵用来衡量隐私属性取值是否均衡,其计算值越大,则证明隐私属性取值越分散,越不易被攻击者获取;反之,则证明隐私属性取值越集

① http://www.baike.com/wiki/全概率公式.

中，越容易被攻击者获取到.

数据块之间的关联关系采用块关联度来衡量.块关联度能够表示出数据分块在进行增删改操作之后数据块之间的关联程度，如定义 5 所示.

定义 5.　块关联度(Associate Rule Attributes Degree，*ARAD*)，在一个 DSM 中，对样本进行多次增加、删除、修改操作后，记录每次操作后每个数据分块中的元组变更次数.假设操作元组数据总数为 D_{all}，数据块 A 中元组的变化数目为 D_A，数据块 B 中元组的变化数目为 D_B，数据分块 A、B 同时变化的元组数目为 $D_{A\&B}$，则块 A 和块 B 的块关联度为：

$$ARAD=\frac{P(A\cap B)}{P(A)}/\frac{1}{P(B)}=\frac{\frac{D_{A\&B}}{D_{\text{all}}}}{\frac{D_A}{D_{\text{all}}}}/\frac{1}{\frac{D_B}{D_{\text{all}}}}=\frac{D_{A\&B}\times D_{\text{all}}}{D_A\times D_B}\tag{4}$$

由式(4)可以看出，当 $ARAD$ 的值越接近于 0 时，数据块 A、B 的关联性越大，数据分块内的隐私属性泄露的可能性就越大；$ARAD$ 越接近于 1，数据块 A、B 的关联性越小，数据分块内的隐私属性泄露的可能性越小，隐私保护效果越好.

4　基于概率统计的层次分析模型

3.2 节定义了一些对于分块混淆技术的隐私保护评价指标，本节将综合运用概率统计的知识，针对分块混淆隐私保护策略的数据存储特点，融合层次分析的思想，构建隐私保护层次分析模型，如图 2 所示.模型的终极目标是评价隐私保护后的数据的隐私保护程度如何，因此模型的顶层为隐私保护后的数据.经过隐私保护后的数据被分到不同的存储模式 DSM 中，见图 2 中第二层.在每个存储模式中，根据不同的隐私保护策略，存储不同的数据分块，见图 2 中第三层.第四层为隐私保护评价指标.第五层则是底层回逆过程，根据第四层的隐私保护评价指标计算出每个租户的每个数据块的对应评价指标值；然后根据每个数据块的评价指标值计算每个 DSM 的评价指标值；最后，根据每个 DSM 的评价指标值计算出每个租户的隐私保护总评价值，详细介绍见 4.1 节和 4.2 节.

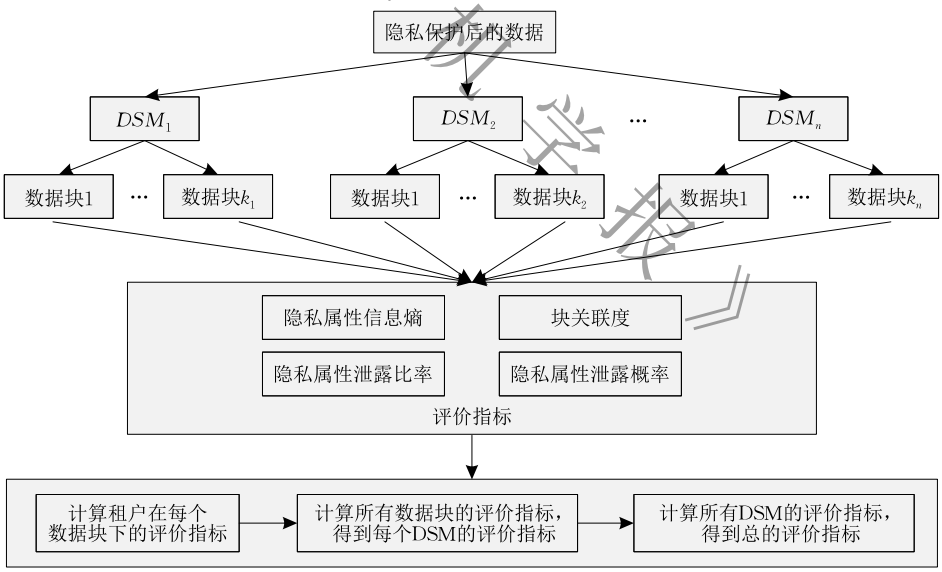


图 2　基于概率统计的层次分析模型

4.1　数据块层向 DSM 层回逆

假设 DSM_1 中有 N_1 个数据块，为了计算该数据存储模式下隐私保护评价指标，将数据块进行分组，每个分组被看成一个独立的样本，即将 N_1 个数据块分成 X_1 个组，每个组中数据块的数量为 N_2（$N_2\ll N_1$），则 DSM_1 可以看成由无数个分组样本 X 组成的总体 Y，因此可以通过研究单个样本的隐私保护评价指标值估算出样本总体的隐私保护评价指标值，计算过程如下所示.

(1) 隐私属性泄露概率

通过分析和研究样本 X 的 $PALP$，评估 DSM_1 的 $PALP$，步骤如下：

① 计算出所有 X 的隐私属性泄露概率向量，即 $\boldsymbol{XPP}_t=[PALP_{PA1},PALP_{PA2},\cdots,PALP_{PANp}]$，$0<t\leq X_1$.

计算 $\boldsymbol{XPP}$ 的均值

$$\overline{\boldsymbol{XPP}}=\frac{1}{X_1}\sum_{t=1}^{X_1}\boldsymbol{XPP}_t\tag{5}$$

$$\boldsymbol{XPP}_t=\left[\frac{1}{N_N}\times\sum_{j=1}^{N_{V_i}}\left(\frac{R_{N_{V_j}}}{R_i}\right)^2,\cdots,\frac{1}{N_N}\times\sum_{j=1}^{N_{V_{N_p}}}\left(\frac{R_{N_{V_j}}}{R_{N_p}}\right)^2\right] \tag{6}$$

② 计算 ***XPP*** 的方差

$$(\boldsymbol{XPP})^2=\frac{1}{X_1-1}\sum_{t=1}^{X_1}(\boldsymbol{XPP}_t-\overline{\boldsymbol{XPP}})^2 \tag{7}$$

(2) 隐私属性泄露比率

隐私属性泄露比率 $PALR$ 可通过 $XP_t(0<t\leq X_1)$ 计算得出，详细过程如下：

① For each $XP_t\in\boldsymbol{XPP}$

If $XP_t>\lambda$

$PALR_t$++ //隐私属性泄露比率

② 计算 $PALR$ 的均值

$$\overline{PALR}=\frac{1}{X_1}\sum_{t=1}^{X_1}PALR_t \tag{8}$$

③ 计算 $PALR$ 的方差

$$PALR^2=\frac{1}{X_1-1}\sum_{t=1}^{X_1}(PALR_t-\overline{PALR})^2 \tag{9}$$

(3) 隐私属性值熵

通过样本的 $PAVE$，评估 DSM_1 的 $PAVE$，评估过程如下所示：

① 计算出所有 X 的隐私属性值熵向量 ***XPE***，即 $\boldsymbol{XPE}_t=[PAVE_{PA_1},PAVE_{PA_i},\cdots,PAVE_{PA_{N_p}}]$，$0<t\leq X_1$

② 计算 ***XPE*** 的均值

$$\overline{\boldsymbol{XPE}}=\frac{1}{X_1}\sum_{t=1}^{X_1}\boldsymbol{XPE}_t \tag{10}$$

$$\boldsymbol{XPE}_t=[-\sum_{j=1}^{N_{PA_i}}R_j\times\log_2(R_j),\cdots,-\sum_{j=1}^{N_{PA_{N_p}}}R_j\times\log_2(R_j)] \tag{11}$$

③ 计算 ***XPE*** 的方差

$$(\boldsymbol{XPE})^2=\frac{1}{X_1-1}\sum_{t=1}^{X_1}(\boldsymbol{XPE}_t-\overline{\boldsymbol{XPE}})^2 \tag{12}$$

(4) 块关联度

① 对于 X_1 个分组中的每个数据块，分别进行增加、删除和修改操作各 N_u 次

② 计算每个数据块的元组变化数量

③ 根据式(5)计算该 X_1 中所有块两两之间的块关联度 $ARAD$

④ 计算 DSM_1 中 $ARAD$ 的均值

$$\overline{ARAD}=\frac{1}{N_{21}}\sum_{t=1}^{N_{21}}ARAD_t \tag{13}$$

⑤ 计算 DSM_1 中 $ARAD$ 的方差

$$ARAD^2=\frac{1}{N_{21}-1}\sum_{t=1}^{N_{21}}(ARAD_t-\overline{ARAD})^2 \tag{14}$$

4.2 DSM 层向顶层回逆

在 4.1 节中，通过样本估算的方式从数据块层向 DSM 层回逆，得到 DSM 层的评价指标值，本节将利用 DSM 层的评价指标值得到最终的隐私保护评价值.

(1) 隐私属性泄露概率

① 计算每个 DSM 的 $PALP$ 均值

For (int $i=0$; $i<n$; i++)

$$\overline{DSM_i\_PALP}=\frac{1}{N_{pi}}\sum_{j=0}^{N_{pi}}\overline{\boldsymbol{XPP}}[j] \tag{15}$$

② 计算 DSM_PALP 的均值

$$\overline{DSM\_PALP}=\frac{1}{n}\sum_{i=0}^{n}\overline{DSM_i\_PALP} \tag{16}$$

③ 计算 DSM_PALP 的方差

$$DSM\_PALP^2=\frac{1}{n-1}\sum_{i=0}^{n}(DSM_i\_PALP-\overline{DSM\_PALP})^2 \tag{17}$$

(2) 隐私属性泄露比率

① 计算 DSM_PALR 的均值

$$\overline{DSM\_PALR}=\frac{1}{n}\sum_{i=0}^{n}\overline{DSM_i\_PALR} \tag{18}$$

② 计算 DSM_PALR 的方差

$$DSM\_PALR^2=\frac{1}{n-1}\sum_{i=0}^{n}(DSM_i\_PALR-\overline{DSM\_PALR})^2 \tag{19}$$

(3) 隐私属性值熵

① 计算每一个 DSM 的 $PAVE$ 均值

For (int $i=0$; $i<n$; i++)

$$\overline{DSM_i\_PAVE}=\frac{1}{N_{pi}}\sum_{j=0}^{N_{pi}}\overline{\boldsymbol{XPE}}[j] \tag{20}$$

② 计算 DSM_PAVE 的均值

$$\overline{DSM\_PAVE}=\frac{1}{n}\sum_{i=0}^{n}\overline{DSM_i\_PAVE} \tag{21}$$

③ 计算 DSM_PAVE 的方差

$$DSM\_PAVE^2=\frac{1}{n-1}\sum_{i=0}^{n}(DSM_i\_PAVE-\overline{DSM\_PAVE})^2 \tag{22}$$

(4) 块关联度

① 计算该用户所有 DSM 中 $ARAD$ 的均值

$$\overline{DSM\_ARAD}=\frac{1}{n}\sum_{i=1}^{n}\overline{ARAD_i} \tag{23}$$

② 计算该用户所有 DSM 中 $ARAD$ 的方差

$$DSM\_ARAD^2=\frac{1}{n-1}\sum_{i=1}^{n}\overline{(ARAD_i-\overline{DSM\_ARAD})^2} \tag{24}$$

本节主要讲述了从数据块层向顶层两阶段回逆的计算过程，通过计算可以得到隐私保护后数据总体分布的评价指标值，为隐私保护综合评价模型提供基础.

5　基于德尔熵值法的综合评价模型

第 4 节中，通过构建层次分析模型，可以得到隐私保护评价指标的计算方法，本节根据这些评价指标，将德尔菲法与熵值法相结合，计算得到评价指标的权重，并建立隐私保护评价模型，详细介绍如下.

① 确定评价对象的因素集合

根据第 4 节介绍可知，评价指标集合为 $ES=$｛隐私属性泄露概率，隐私属性泄露比率，隐私属性值熵值，块平均关联度｝，本节将针对分块混淆后的数据进行评价指标值的计算.

② 确定评语集合

本节共给出 4 个评价评语，表示为 $CS=$｛非常好，好，一般，差｝，各个评价指标的评价标准如表 1 所示，其中 $\lambda,\lambda_1,\lambda_2,\lambda_3$ 是云服务商与租户在 SLA 协议中签订的标准值.

表 1　评价标准

指标	等级			
	非常好	好	一般	差
$PALP$	$<0.5\lambda$	$0.5\lambda\sim0.75\lambda$	$0.75\lambda\sim\lambda$	$>\lambda$
$PALR$	$<0.5\lambda_1$	$0.5\lambda_1\sim0.75\lambda_1$	$0.75\lambda_1\sim\lambda_1$	$>\lambda_1$
$PAVE$	$=\lambda_2$	$0.95\lambda_2\sim\lambda_2$	$0.90\lambda_2\sim0.95\lambda_2$	$<0.9\lambda_2$
$ARAD$	$<0.5\lambda_3$	$0.5\lambda_3\sim0.75\lambda_3$	$0.75\lambda_3\sim\lambda_3$	$>\lambda_3$

③ 单因素评价

根据第 4 节的定义，针对每个评价指标计算评价指标值，从而得到每个租户的评价指标向量，详细计算过程如算法 1 所示，算法复杂度为 $O(n)$. 算法中对数据进行 X_{check} 次抽查检测，每次抽检标准一致.

算法 1.　单因素评价算法.

输入：抽查的数据，评价因素

输出：单因素评价向量

1. FOR (int $i=0$; $i<X_{\text{check}}$; $i++$)
2. FOR (int $j=0$; $j<4$; $j++$)
3. 根据式(1)到式(25)，计算所需的评价指标值
4. 计算评价指标值的波动范围
5. 对比每个评价指标值隶属于哪个等级
6. 将每个评价指标的等级写进评价向量
7. Return 所有评价指标的评价向量

通过算法 1 可以得到每个评价指标的值，进而得到评价向量，将所有数据分块的评价向量组合起来，构成评价矩阵 $\boldsymbol{EM}$.

④ 综合评价

通过步骤③可以得到评价矩阵 $\boldsymbol{EM}$，下面将研究如何计算得到模糊综合评价的另一个重要因素，评价指标的权重 $\boldsymbol{W}$. 目前针对权重 $\boldsymbol{W}$ 的确定方法主要分为两种，一种是主观法，一种是客观法. 主观法主要以德尔菲法[①]为代表，该方法虽然在一定程度上消除了权威人士对其他人的影响，但是专家的意见可能不切合实际，人为因素太强；客观法中，熵值法[②]比较突出，该方法通过样本数据统计分析得到权值，克服了人为因素的干扰，但是其受样本数据的影响较大. 上述两种方法各有优势，因此本文综合两种方法的优势，提出基于德尔熵值的权重确定方法.

德尔菲法是通过对一组评审专家进行多次问询，最终得到一个统一的结果，为了降低主观因素对评价的影响，我们将问询过程进行改进，从 n 个评审专家中进行 m 次独立抽取，抽取出的专家组成 m 个小组，每个小组中人数不同，对评价指标进行打分，即可得到决策矩阵[③]，结果如表 2 所示.

表 2　决策矩阵表

分组	指标			
	$PALP$	$PALR$	$PAVE$	$ARAD$
分组 1	X_{11}	X_{12}	X_{13}	X_{15}
分组 2	X_{21}	X_{22}	X_{23}	X_{25}
⋮	⋮	⋮	⋮	⋮
分组 m	X_{m1}	X_{m2}	X_{m3}	X_{m5}

针对上述决策矩阵，运用熵值法计算每个评价指标的权重，过程如算法 2 所示，算法复杂度为 $O(n)$.

算法 2.　评价指标权重分配算法.

输入：评价因素，决策矩阵

输出：评价指标权重向量 $\boldsymbol{W}$

1. FOR (int $j=0$; $j<4$; $j++$)
　//计算所有分组的指标 j 的得分，并求和 $Score$
2. FOR (int $i=0$; $i<m$; $i++$)
3. $Score=Score+X_{ij}$

① http://wiki.mbalib.com/wiki/德尔菲法.
② http://wiki.mbalib.com/wiki/熵值法.
③ http://wiki.mbalib.com/wiki/决策矩阵.

//P_{ij} 为指标 j 下第 i 组评分的贡献值

4. FOR (int $i=0$; $i<m$; $i++$)

5. $P_{ij}=\frac{X_{ij}}{Score}$

//计算熵值 E_j

6. FOR (int $i=0$; $i<m$; $i++$)

7. $E_j=E_j+P_{ij}\ln(P_{ij})$

8. $E_j=-KE_j$

//计算第 j 个指标的重要程度 D_j

9. $D_j=1-E_j$

10. $D=D+D_j$

11. FOR (int $j=0$; $j<4$; $j++$)

//计算权重 $\boldsymbol{W}_j$

12. $\boldsymbol{W}_j=\frac{D_j}{D}$

13. $\boldsymbol{W}=(W_1W_2W_3W_4)$

14. Return $\boldsymbol{W}$

算法第 8 行,常数 $K=1/\ln(m)$,由此可保证 $0\leq E_j\leq 1$. 当某一评价指标的所有分组评分完全相等时,$E_j=1$,此时,$D_j=0$,则可以忽略掉该属性. 第 9 行中,D_j 表示指标 j 的权重,指标值的差异越大,熵值 E_j 越小,D_j 值越大,则指标 j 对评价影响越大.

模糊算子 $M(\cdot,\oplus)$ 具有较强的综合能力,且能充分体现权重在评价过程中的作用,因为本文使用该算子计算最终的隐私保护程度评分. 评价矩阵和评价指标权重由算法 2 可以求得,根据模糊算子,评价向量 $\boldsymbol{S}$ 的求解公式如式(25)所示.

$$\boldsymbol{S}=\boldsymbol{W}\cdot\boldsymbol{EM} \tag{25}$$

最终,隐私保护模糊综合评价算法如算法 3 所示,算法复杂度为 $O(1)$,通过算法 3 可以得到每个租户所采用的隐私保护策略的评价分数,分数越高,隐私保护效果越好.

算法 3. 隐私保护模糊综合评价算法.

输入:评价矩阵 $\boldsymbol{EM}$,评价指标权重 $\boldsymbol{W}$

输出:隐私保护程度评分 S

//根据式(26)计算评价向量

1. FOR (int $j=1$; $j\leq 4$; $j++$)

2. FOR (int $i=0$; $i\leq 4$; $i++$)

3. $S_j=S_j+W_i\times R_{ij}$

4. $S=(S_1S_2S_3S_4)$ //归一化处理

5. $S'=(S_1/(S_1+S_2+S_3+S_4)S_2/(S_1+S_2+S_3+S_4)$
$S_3/(S_1+S_2+S_3+S_4)S_4/(S_1+S_2+S_3+S_4))$
//加权平均获得最后的评价结果

6. $S=(100\times S'_1+75\times S'_2+50\times S'_3+25\times S'_4)/(S'_1+S'_2+S'_3+S'_4)$

7. Return S

6 实验评估

6.1 实验环境

为了验证本文提出的基于概率统计和德尔熵值法的综合评价机制的评价效果,对原始的分块混淆算法 PPCC[1]、隐私保护调整算法 SAG[22] 和数据关系隐藏算法 HMDR[23] 进行了测试. 3 种算法在以实验室社保项目内部数据为租户的平台上进行评分.

实验环境:5 台服务器作为数据节点,系统采用 Red Hat Enterprise Linux 6.2 版本;Apache Tomcat 作为应用服务器,配置为 4 核 CPU Inter(R) Xeon(R) 2.40 GHz,10 GB 内存,500 GB 硬盘;测试数据库采用 5.5.25 MySQL Community Server (GPL);编程环境选用 Eclipse-SDK-4.3-win64,编程语言为 Java 1.7. 测试数据集使用的是本实验室社保项目中的内部测试数据集,300 万条的社会医疗保险参保人员的基本信息,隐私属性选取了姓名(已模糊处理)、性别、年龄、单位、家庭住址、参保类型、信任等级、医疗账户等 20 个属性进行测试,其中不相容隐私约束为姓名、性别、年龄、参保类型、单位类型.

6.2 结果分析

6.2.1 评价指标

实验 1,验证本文提出的 4 种评价指标是否能够公平合理地对分块混淆的隐私保护技术进行评价,在 3 种隐私保护算法下,计算评价指标的值,看是否有差异. 实验中,在进行隐私保护策略之前,平台中有存储模式 15 个,每个存储模式下分配 20 万条左右的数据,每个存储模式中 1~20 个数据分块不等. 首先用 PPCC 算法对所有存储模式中的数据进行基础处理,然后随机进行增加、删除、修改操作各 50 次,过程中调整隐私保护需求 20 次,根据第 4 节定义的公式计算评价指标值;然后将数据恢复原样,使用 SAG 和 HMDR 算法先后对数据进行隐私保护处理,随机进行增加、删除、修改操作 50 次,计算评价指标值;最后将数据恢复原貌,先后用 PPCC、SAG、HMDR3 种算法处理上述数据,进行相同的增删改操作后,计算评价指标值,实验结果如图 3 到图 6 所示.

图 3 展示了 15 种数据存储模式下经过 3 种算法隐私保护处理后,计算得到隐私属性泄露概率的分布情况,其中 $PALP$ 的安全范围为 0~0.2.

从图 3 可知,3 种算法下,$PALP$ 的值各不相

同，在 PPCC 算法下，*PALP* 分布在 0.05～0.3 之间，其中有 66.7%的概率分布在 0.05～0.2 之间，另外 33.3%存在隐私泄露风险，这是因为增删改操作的过程中，数据关联关系可能导致隐私泄露. 在 HMDR 算法下，*PALP* 主要分布在 0.00～0.25 之间，0.05～0.20 之间的概率有 80%，其隐私保护程度要比 PPCC 算法略有提高，但是隐私属性泄露的概率仍然有 20%. 在 SAG 算法下，*PALP* 主要在 0.00～0.20 之间，隐私保护程度已经在标准范围内，但隐私属性泄露的概率也有 6.7%. 因此从 *PALP* 层面，隐私保护效果还有提升的空间.

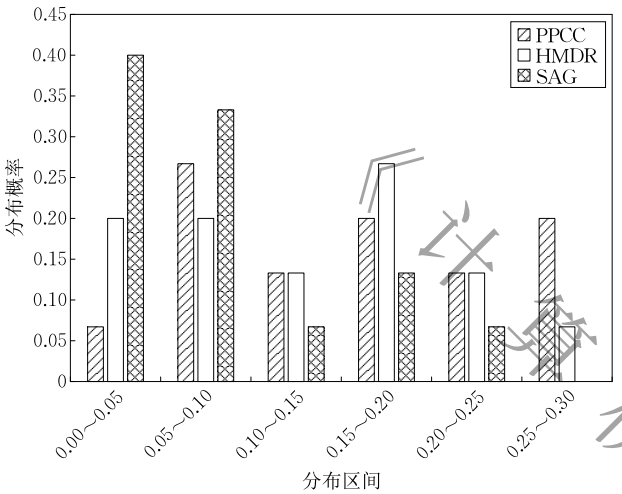


图 3　隐私属性泄露概率在 3 种算法下的分布情况(*PALP*)

图 4 展示了 15 种数据存储模式下经过 3 种算法隐私保护处理后，计算得到隐私属性泄露比率的分布情况，我们规定 *PALR* 不超过 0.3 为安全范围.

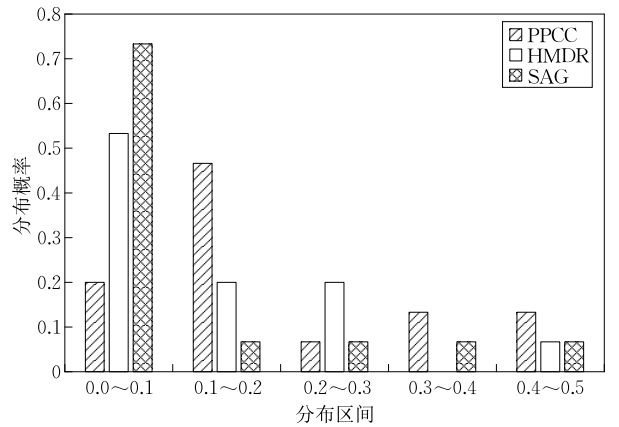


图 4　隐私属性泄露比率在 3 种算法下的分布情况(*PALR*)

从图 4 中可看出 PPCC 算法发生隐私泄露的概率为 26.6%，但是隐私保护效果在标准范围内的概率有 66.6%；HMDR 只有 6.7%的概率发生隐私泄露，比 PPCC 算法有了明显提高，而且隐私保护效果在标准范围内的概率有 73.3%，显著提升了保护效果；与前两个算法相比，SAG 发生 *PALR* 的概率有 13.3%，但有 80%的概率具有非常好的隐私保护效果.

图 5 中展现了隐私属性值熵 *PAVE* 在 3 种算法下的变化情况.

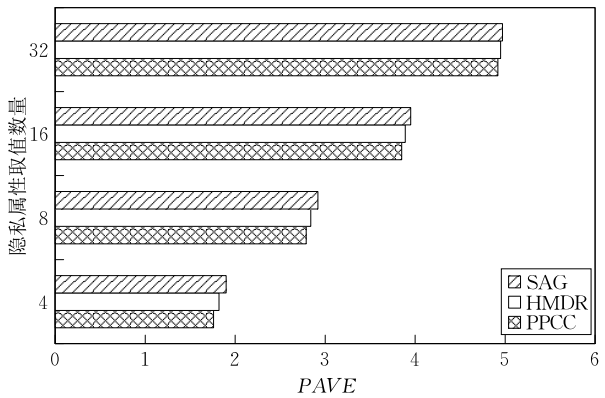


图 5　3 种算法下隐私属性值熵的变化情况

由图 5 可看出，隐私属性取值数量关系到熵值的大小，数量越多，熵值就越多，相对稳定，隐私保护效果就越好. 例如，当取值数量为 32 时，隐私属性熵值接近于理想状态下的熵值 5；而当属性个数取值为 4 时，3 种算法的熵值都不到 2，这是因为隐私属性取值个数越多，*PAVE* 值就越大，隐私保护属性取值就越分散，隐私保护程度就越高. 由上图可以看出，3 种算法的熵值降序排列为 SAG、HMDR 和 PPCC，与我们的预期结果相符.

图 6 展示了 15 种数据存储模式下经过 3 种算法隐私保护处理后，计算得到数据块间的关联度情况，我们规定，块间关联度大于 0.4 为安全范围.

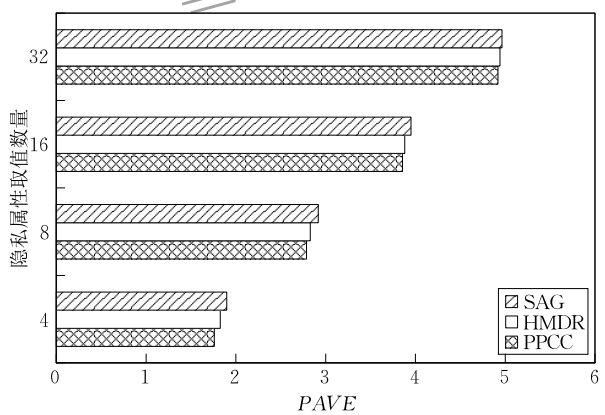


图 6　块关联度在 3 种算法下的分布情况

如图 6 所示，在 PPCC 算法下，有 40%的块间关联度小于 0.4，这说明，在该算法下，数据块之间的关联性比较大，这是因为在 PPCC 算法中，没有对数据块之间的关联性做处理；算法 SAG 和算法 HMDR 有 50%以上的比例落在 0.4～0.6 之间，

13%左右的概率落在了 0.8～1.0 之间，块间关联性不是很大，隐私保护效果较好，这是因为这两种算法对数据块之间的关联性都做了隐藏，所以数据块之间的关联性较小；从实验结果也可看出与 PPCC 相比，算法 SAG 和 HMDR 都增强了隐私保护效果，与预期结果相符.

6.2.2 综合对比评价

实验 1 验证了评价指标的正确性，实验 2 我们将对 3 种算法进行总体评价和等级划分，验证本文算法的有效性. 实验 2 中，在每个数据存储模式下每次抽取 70%数据块的数据，抽取 100 组，每组抽取 50 次. 假设 $\lambda=0.2$，$\lambda_1=0.3$，$\lambda_2=|\log_2(1/N_{PA_i})|$，$\lambda_3=0.4$，$i=1,2,\cdots,N_p$. 对抽取完的数据依次使用 3 种隐私保护算法进行处理，然后进行增加、删除、修改操作各 50 次，更改隐私需求 20 次，计算每次操作前后数据的评价指标值，实验结果如表 3 所示.

表 3 隐私保护等级分配表

指标	等级				
	算法	非常好	好	一般	差
PALP	PPCC	8	12	55	25
	HMDR	15	40	37	8
	SAG	36	42	18	4
PALR	PPCC	10	41	39	10
	HMDR	35	38	25	2
	SAG	48	26	20	6
PAVE	PPCC	6	27	46	21
	HMDR	21	38	29	12
	SAG	23	41	27	9
ARAD	PPCC	15	37	21	27
	HMDR	27	39	25	9
	SAG	30	40	26	4

由表 3 计算 3 种算法的评价矩阵，表示如下：

$$\boldsymbol{EM}_{PPCC}=\begin{bmatrix}8/100 & 12/100 & 55/100 & 25/100\\ 10/100 & 41/100 & 39/100 & 10/100\\ 6/100 & 27/100 & 46/100 & 21/100\\ 15/100 & 37/100 & 21/100 & 27/100\end{bmatrix}=\begin{bmatrix}0.08 & 0.12 & 0.55 & 0.25\\ 0.10 & 0.41 & 0.39 & 0.10\\ 0.06 & 0.27 & 0.46 & 0.21\\ 0.15 & 0.37 & 0.21 & 0.27\end{bmatrix}$$

$$\boldsymbol{EM}_{HMDR}=\begin{bmatrix}15/100 & 40/100 & 37/100 & 8/100\\ 35/100 & 38/100 & 25/100 & 2/100\\ 21/100 & 38/100 & 29/100 & 12/100\\ 27/100 & 39/100 & 25/100 & 9/100\end{bmatrix}=\begin{bmatrix}0.15 & 0.40 & 0.37 & 0.08\\ 0.35 & 0.38 & 0.25 & 0.02\\ 0.21 & 0.38 & 0.29 & 0.12\\ 0.27 & 0.39 & 0.25 & 0.09\end{bmatrix}$$

$$\boldsymbol{EM}_{SAG}=\begin{bmatrix}36/100 & 42/100 & 18/100 & 4/100\\ 48/100 & 26/100 & 20/100 & 6/100\\ 23/100 & 41/100 & 27/100 & 9/100\\ 30/100 & 40/100 & 24/100 & 6/100\end{bmatrix}=\begin{bmatrix}0.36 & 0.42 & 0.18 & 0.04\\ 0.48 & 0.26 & 0.20 & 0.06\\ 0.23 & 0.41 & 0.27 & 0.09\\ 0.30 & 0.40 & 0.24 & 0.06\end{bmatrix}$$

在实验 2 中，我们选择 20 名的研究人员，随机抽取 15 名研究人员，抽取 8 次，得到 8 个研究小组，从每个研究小组中依次抽取 10 名研究人员，进行多轮问询，记录每轮每组研究人员的评价指标打分情况，结果如表 4 所示.

表 4 10 位专家对评价指标打分的决策矩阵表

分组	指标			
	PALP	*PALR*	*PAVE*	*CDD*
分组 1	7	4	6	6
分组 2	8	5	7	6
分组 3	9	8	4	5
分组 4	5	6	10	8
分组 5	9	8	5	4
分组 6	9	2	8	6
分组 7	6	5	10	3
分组 8	10	5	4	9

每个分组的贡献度有决策矩阵和算法 2 计算得出，结果如表 5 所示.

表 5 贡献度表

PALP	*PALR*	*PAVE*	*CDD*
0.111111	0.093023	0.111111	0.127660
0.126984	0.116279	0.129630	0.127660
0.142857	0.186047	0.074074	0.106383
0.079365	0.139535	0.185185	0.170213
0.142857	0.186047	0.092593	0.085106
0.142857	0.046512	0.148148	0.127660
0.095238	0.116279	0.185185	0.063830
0.158730	0.116279	0.074074	0.191489

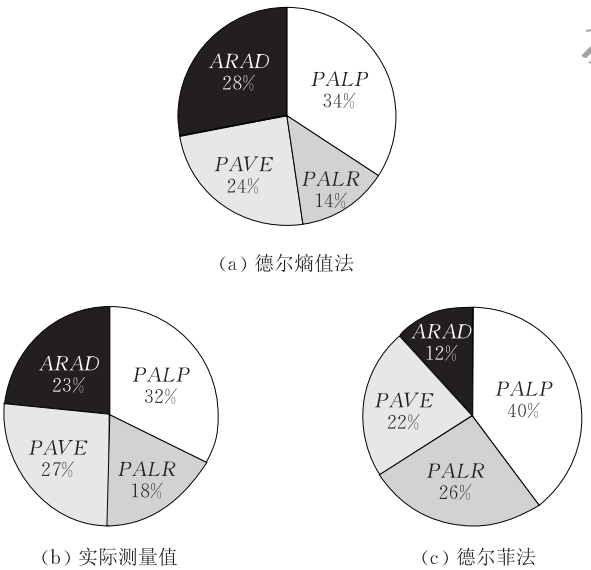
实验共分 8 组，因此算法 2 中的常数取值为 $K=1/\ln(8)=0.481$，由此计算出标价指标的熵值 $E=(-0.96343,-0.98427,-0.97425,-0.96949)$，评价指标的重要程度 $D=(0.036574,0.015729,0.025753,0.030505)$，因此最终求得评价指标的权重 $\boldsymbol{W}=(0.34,0.14,0.24,0.28)$. 根据式(25)，3 种算法归一化之后的评价向量 $\boldsymbol{S}_{\mathrm{PPCC}}=\boldsymbol{W}\cdot\boldsymbol{EM}_{PPCC}=(0.0976,0.2666,0.4108,0.2550)$，同理 $\boldsymbol{S}_{\mathrm{HMDR}}=(0.2260,0.3896,0.3004,0.084)$，$\boldsymbol{S}_{\mathrm{SAG}}=(0.3288,0.3896,0.2212,0.0604)$. 最后，加权平均后，3 种算法的得分分别为 56.67、68.94 和 74.67，由表 1 可

知，3 种隐私保护算法的保护程度为好级别，而且保护效果 SAG 优于 HMDR，HMDR 算法优于 PPCC，这与实际情况相符. 由此可知，从隐私保护等级上看，3 种算法具有较好的隐私保护效果，但是从分数上看，隐私保护程度还有待进一步的增强和优化.

6.2.3　指标权重分析

为了验证改进后德尔菲法和熵值法的效果，本文设计了实验 3：分别使用德尔熵值法和德尔菲法计算评价指标权重，然后再根据实际测量值与二者进行对比，实验条件和环境与实验 2 相同，结果如图 7 所示.

由图 7 可知，图 7(a)中的 *ARAD* 的值为 28%，图 7(c)中的值为 12%，而实际的测量值为 23%，与图 7(a)更为接近；同理，其他指标在德尔熵值法下的权重值与德尔菲法下的值相比，与实际测量值的测距更小，效果更好. 德尔菲法主要依赖于人的感官判断，主观性太强，由图 7(c)可知，有些评价指标权重分配过多，有些权重分配又过少，而德尔熵值法虽然也部分依赖主观判断，但通过熵值法一定程度上降低了人为因素的干扰，使得权重的分配更为合理.



(a) 德尔熵值法

(b) 实际测量值　　(c) 德尔菲法

图 7　权重准确性对比分析

6.2.4　评价效果分析

为了验证评价机制的有效性以及持续性，我们设计了实验 4：在优化后的分块混淆方法保护下，对数据进行监控和评价，评价指标和权重不变；实验中，每隔一周对数据进行一次评价，共进行 4 次评价，评价过程与 6.2.2 节类似；在非评价期间，模拟日常的数据操作和业务对数据进行处理；实验结果如图 8 和图 9 所示.

由图 8 可知，随着时间的变化，评价的分数总体有上升的趋势，但波动幅度不太，其中第三周的时候，还略微下降；从隐私等级上来看，第一周、第二周和第四周属于非常好的状态，而第三周属于好的状态. 总体而言，评价效果比较稳定.

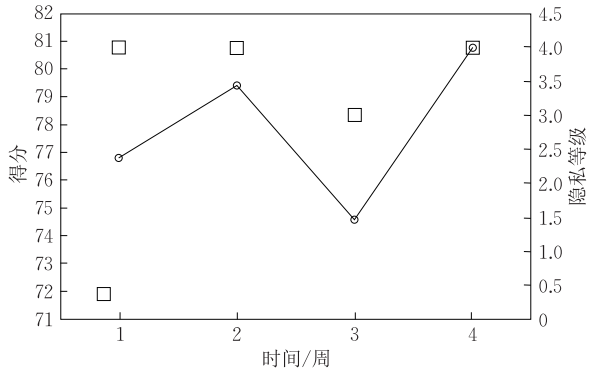


图 8　隐私评价效果与时间的关系

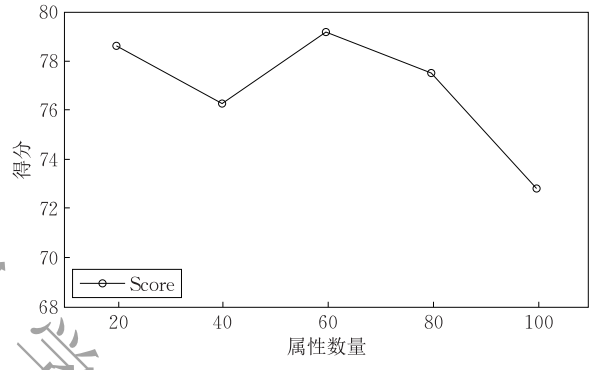


图 9　隐私评价效果与属性数量的关系

由图 9 可知，随着数据属性数量的增加，评价分数前期比较稳定，波动起伏不大，有时候高一点，有时候低一点，但当属性数量达到 100 个的时候，评价分数出现比较大的下降，这是因为随着数据属性数量的增多，数据分块数也越来越多，但数据节点的数量是有限的，因此存在部分不能放在一起的数据块放在了同一个节点的现象，从而增加了隐私泄露的概率，使得评价分数下降. 由实验可知，分块混淆隐私保护技术虽然具有较好的隐私保护效果，但还需要进一步完善.

7　总　结

在云计算平台中，基于分块混淆的隐私保护技术虽然能够保护租户的数据隐私性，然而对于租户而言，无法直观地感受该类隐私保护技术的隐私保护程度到底如何，基于此问题，本文综合概率统计和德尔熵值法，提出一种模糊综合评价模型，实现分块混淆的隐私保护技术的可视化评价. 首先，在分析了

影响隐私保护效果的因素后,定义了评价指标,并提出了一种基于概率统计的层次分析模型,采用底层回逆的方式,计算租户在隐私保护策略之后数据分布的评价指标值;然后综合德尔菲法和熵值法的优点,提出德尔熵值权重确定方法,建立隐私保护综合评价模型,该模型通过德尔熵值法计算出每个评价指标的权重,提高了权重分配的准确性和合理性,同时减小了人为主观因素的干扰程度,实现了分块混淆的隐私保护技术的可视化评价,并为隐私策略的调整提供了理论基础和方向;最后本文通过实验验证了本文提出的评价指标的合理性和算法的有效性.

参 考 文 献

[1] Zhang Kun, Li Qing-Zhong, Shi Yu-Liang. Research on data combination privacy preservation mechanism for SaaS. Chinese Journal of Computers, 2010, 33(11): 2044-2054(in Chinese)
(张坤, 李庆忠, 史玉良. 面向 SaaS 应用的数据组合隐私保护机制研究. 计算机学报, 2010, 33(11): 2044-2054)

[2] Shao Ya-Li, Shi Yu-Liang, Li Hui. A novel cloud data fragmentation cluster-based privacy preserving mechanism. International Journal of Grid & Distributed Computing, 2014, 7(4): 21-32

[3] Shi Yu-Liang, Jiang Zhen, Zhang Kun. Policy—Based customized privacy preserving mechanism for SaaS applications//Proceedings of the International Conference on Grid and Pervasive Computing. Berlin, Germany, 2013: 491-500

[4] Khan K M, Shaheen M. Data obfuscation for privacy and confidentiality in cloud computing//Proceedings of the IEEE International Conference on Software Quality, Reliability and Security-Companion. Vancouver, Canada, 2015: 195-196

[5] Wang Yu-Li, Tian Jiay-Jin, Yang Cheng, Zhu Ya-Ping. Research on anonymous protection technology for big data publishing//Proceedings of the 7th IEEE International Conference on Software Engineering and Service Science. Beijing, China, 2016: 438-441

[6] Shalin Eliabeth S, Sarju S. Bigdata anonymization using one dimensional and multidimensional map reduce framework on cloud. International Journal of Database Theory and Application, 2015, 8(6): 253-262

[7] Tran Q, Sato H. A solution for privacy protection in mapreduce//Proceedings of the 36th IEEE Annual International Computer Software and Applications Conference. Izmir, Turkey, 2012: 515-520

[8] Ádám Erdélyi, Winkler T, Rinner B. Privacy protection vs. utility in visual data: An objective evaluation framework. Multimedia Tools & Applications, 2018, 77(2): 2285-2312

[9] Shi Yu-Liang, Chen Yu, Zhou Zhong-Min, Cui Li-Zhen. A privacy protection evaluation mechanism for dynamic data based on chunk-confusion. Journal of Signal Processing Systems, 2017, 89(1): 27-39

[10] Tahir N, James F. An annotation-free method for evaluating privacy protection techniques in videos//Proceedings of the IEEE Computer Society 2015. ISBN, Karlsruhe, Germany, 2015: 1-6

[11] Hugues D L S-G, David D, Guillaume R. Range reduction based on Pythagorean triples for trigonometric function evaluation// Proceedings of the 26th IEEE International Conference on Application-Specific Systems, Architectures and Processors. Toronto, Canada, 2015: 74-81

[12] Wang Yu-Biao, Wen Jun-Hao, Wang Xi-Bin, Zhou Wei. Cloud service evaluation model based on trust and privacy-aware. Optik-International Journal for Light and Electron Optics, 2017, 134: 269-279

[13] Zhu Jian-Ming, Srinivasan R. Evaluation model of information security technologies based on game theoretic. Chinese Journal of Computers, 2009, 32(4): 828-834(in Chinese)
(朱建明, Srinivasan R. 基于博弈论的信息安全技术评价模型. 计算机学报, 2009, 32(4): 828-834)

[14] Jiao Yu-Min, Wang Qiang, Su Fan-Tun, et al. Research on dynamic evaluation method for process-oriented virtual training. Acta Armamentarii, 2012, 33(7): 875-880

[15] Lund J R, Israel M. Information theory and multi-objective evaluation//Proceedings of the 1992 National Conference on Water Resources Planning and Management-Water Forum'92. Baltimore, USA, 1992: 486-491

[16] Hamed A, Ben Ayed H K. Privacy risk assessment for Web tracking: A user-oriented approach toward privacy risk assessment for Web tracking//Proceedings of the 2016 IEEE Canadian Conference on Electrical and Computer Engineering. Vancouver, Canada, 2016: 1-6

[17] Ren Dan-Dan, Du Su-Guo, Zhu Hao-Jin. A novel attack tree based risk assessment approach for location privacy preservation in the VANETs//Proceedings of the 2011 IEEE International Conference on Communications. Kyoto, Japan, 2011: 1-5

[18] Chen Yan-Bo, Liu Yang, Zhang Ji, et al. An evaluation method of design scheme for new generation smart substation based on improved SEC. Power System Technology, 2017, 41(4): 1308-1314

[19] Xue Zhe-Bin, Zeng Xian-Yi, Koehl L. Development of a method based on fuzzy comprehensive evaluation and genetic algorithm to study relations between tactile properties and total preference of textile products. Journal of the Textile Institute, 2016, 108(7): 1085-1094

[20] Wang Shao-Na, Dong Rui, Xie Hui, et al. The application of Delphi method and its construction index system. Journal of Bengbu Medical College, 2016, 41(5): 695-698(in Chinese)
(王少娜, 董瑞, 谢晖等. 德尔菲法及其构建指标体系的应用进展. 蚌埠医学院学报, 2016, 41(5): 695-698)

[21] Wang Sheng-Chang, Fu Di, Chen Juan-Juan, et al. A method to determine subjective evaluation index of vehicle dynamic performance based on entropy method. Journal of Highway and Transportation Research and Development, 2015, 32(7): 153-158(in Chinese)
(王生昌，付迪，陈娟娟等. 基于熵值法的汽车动力性能主观评价指标权重确定方法. 公路交通科技，2015，32(7)：153-158)

[22] Zhang Kun, Abraham A, Shi Yu-Liang. Data combination privacy preservation adjusting mechanism for software as a service//Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Manchester, UK, 2013: 2007-2012

[23] Chen Yu, Shi Yu-Liang, Cui Li-Zhen. Model for hiding data relationships based on Chunk-Confusion in cloud computing//Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA. Tianjin, China, 2016: 837-844

SHI Yu-Liang, born in 1978, Ph. D., professor. His research interests include cloud computing, database and privacy preserving.

ZHOU Wei, born in 1978, M. S., associate research librarian. Her research interests include database and privacy preserving.

ZANG Shu-Juan, born in 1992, M. S. candidate. Her research interests include cloud computing and privacy preserving.

CHEN Yu, born in 1990, M. S. candidate. His research interests include cloud computing and privacy preserving.

Background

The topic this paper researched belongs to problem of privacy protection in the cloud computing environment. The popular international solution to the kind of problem is data encryption. Besides, there are also some other privacy protection methods, such as data confusion technology, anonymous privacy protection technology and differential privacy protection. However, all the methods above have only consider the privacy protection, and do not take into account the issue of privacy protection adjustment. Therefore, in our earlier works, a privacy preserving technology based on chunk-confusion is proposed. First, according to the privacy constraints (PC) customized by tenants, the data of tenants are partitioned into a lot data chunks to ensure that the attributes in PC are in different data chunks. Next, we confused the relationships among attributes and saved them on the trusted third party. Finally, data chunks are assigned to multiple data nodes according to their load capacity and computing capability. However, due to the privacy needs of tenants and the data demands of tenants are variable, the underlying data chunks structure and storage location in the cloud will change, which makes there is a risk of leakage of privacy, so users can not apperceive the effect of this privacy protection method. To this end, this paper proposes a privacy preservation comprehensive evaluation model aimed at block confusing privacy protection technology based on probability and Del entropy method. The mechanism can not only objectively show the effects of privacy protection evaluation, but also proves the validity of the method of blocking and confusing privacy protection. The previous studies of our team in the direction of privacy preserving under the cloud computing environment.

Early work includes: A New Privacy-Preserving Scheme DOSPA for SaaS, Data Combination Privacy Preservation Mechanism for SaaS, Policy-Based Customized Privacy Preserving Mechanism for SaaS Applications, A Novel Cloud Data Fragmentation Cluster-based Privacy Preserving Mechanism, Document-oriented Database-based Privacy Data Protection Architecture, Game-Theoretic Strategy for Personalized Privacy Protection, A Sub Chunk-confusion Based Privacy Protection Mechanism For Association Rules, Model for Hiding Data Relationships Based on Chunk-Confusion in Cloud Computing, etc.

As a part of our project, the research work in this paper is supported by the National Key Research and Development Plan of China under Grant No. 2018YFC0114709, the TaiShan Industrial Experts Programme of Shandong Province under Grant No. TSCY20150305, the Primary Research & Development Plan of Shandong Province under Grant Nos. 2016GGX101008, 2016ZDJS01A09, and the Major Basic Research Project of Natural Science Foundation of Shandong Province under Grant No. ZR2017ZB0419.