ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 659－674
DOI:10. 16337/j. 1004-9037. 2025. 03. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于压缩的本地差分隐私的序列数据收集方法

金　严[1]，朱友文[1]，吴启晖[2]

(1. 南京航空航天大学计算机科学与技术学院,南京 211106;2. 南京航空航天大学电子信息工程学院，南京 211106)

摘　要：压缩的本地差分隐私是本地差分隐私的一种基于度量的放松形式,它具有比本地差分隐私更好的效用性和灵活性。但是,现有方案在序列模式捕捉和效用性方面存在不足。为了克服这些局限性,提出了一种新颖的基于压缩的本地差分隐私的序列数据收集方法SCM-CLDP。该方法在收集过程中充分考虑了序列数据的长度、转移等重要信息,通过这些信息,数据收集者能够合成接近原始数据集的隐私保护的数据集。根据扰动对象的不同,本文提出了两种收集方法,分别是基于值扰动的SCM-VP方法和基于转移扰动的SCM-TP方法。理论证明了SCM-VP和SCM-TP满足序列级别的压缩的本地差分隐私,并基于两个真实数据集,在Markov链模型准确性、合成数据集效用性及频繁序列模式挖掘准确性上,与现有方案进行了对比实验。结果表明,SCM-CLDP表现出显著的优势,其中SCM-VP的性能在大多数情况下都要优于SCM-TP。并且在最优的情况下,相较于现有方法,SCM-CLDP在Markov链模型及合成数据集分布误差方面至少降低了一个数量级。同时,SCM-CLDP在合成数据集中各项频率排序的准确性以及频繁序列模式挖掘的准确性方面,相较于现有方法提升了近30%。

关键词：压缩的本地差分隐私;序列数据;Markov链模型;数据收集;隐私保护

中图分类号：TP309　　**文献标志码**：A

Sequential Data Collection Method with Condensed Local Differential Privacy

JIN Yan[1], ZHU Youwen[1], WU Qihui[2]

(1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; 2. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Condensed local differential privacy is a metric-based relaxation of local differential privacy with better utility and flexibility than local differential privacy. However, existing solutions are deficient in terms of sequence pattern capture and utility. To address these limitations, this paper proposes SCM-CLDP, a novel sequential data collection method based on condensed local differential privacy. SCM-CLDP fully takes into account important information such as the length and transitions of sequential data during the collection process, through which the data collector is able to synthesize privacy-preserving dataset close to the original dataset. Specifically, according to different perturbation objects, we propose two collection methods, SCM-VP based on value perturbation and SCM-TP based on transition perturbation,

基金项目:江苏省重点研发计划(产业前瞻与关键核心技术)(BE2022068,BE2022068-1);国家自然科学基金(62172216);中央高校基本科研业务费项目(NP2024117);稳定支持国防特色学科基础研究项目(ILF240061A24)。
收稿日期:2024-06-19;**修订日期**:2024-10-22

respectively. We theoretically prove that SCM-VP and SCM-TP satisfy sequence-level condensed local differential privacy, and comparative experiments are conducted with existing solutions based on two real datasets in terms of Markov chain model accuracy, synthetic dataset utility, and frequent sequence pattern mining accuracy. The results show that SCM-CLDP performs significantly better than the existing solutions, with SCM-VP outperforming SCM-TP in most cases. In the optimal situation, SCM-CLDP reduces the error of the Markov chain model and the distribution of the synthetic dataset by at least one order of magnitude compared to the existing method. Meanwhile, SCM-CLDP improves the accuracy of item frequency ranking of the synthetic dataset and the accuracy of frequent sequence pattern mining by nearly 30% compared to existing solutions.

Key words: condensed local differential privacy; sequential data; Markov chain model; data collection; privacy protection

引　言

序列数据是具有时间相关性或逻辑相关性的一系列数据点，这些数据点按照它们出现的时间顺序或逻辑顺序排列，如网页点击数据、移动轨迹等。在实际生产生活中，序列数据的应用极其广泛，比如，从网页点击数据中分析用户行为模式、从个人轨迹数据中做旅游路线推荐[1-2]以及从文本序列中构建语言模型[3]等。隐私已成为当前大数据时代的重要关注点，随着个人隐私保护意识的提高，各国已逐步建立和完善了相关的个人隐私保护法律法规，如中国在2021年相继实施了《中华人民共和国数据安全法》[4]和《中华人民共和国个人信息保护法》[5]。序列数据中包含有大量的用户隐私信息，如从网页点击数据中可推断出个人偏好、从个人轨迹数据中可以推断出家庭住址甚至健康状况[6]。因此，在序列数据的收集中应避免直接收集用户真实的敏感序列数据，而是通过隐私保护的方法来收集。

近年来，本地差分隐私(Local differential privacy, LDP)[7]是一种被广受关注的隐私保护模型，它是传统的依赖于可信第三方的差分隐私(Differential privacy, DP)[8]的本地化版本，它给出了严格的可证明的隐私保护，且无需可信的第三方。LDP要求用户在本地扰动自身的隐私数据且不同用户有一定概率扰动输出相同的值，使得攻击者无法从截获到的个人数据中准确推断出隐私信息，但是数据收集者可从扰动的数据中聚合得到统计信息。目前，LDP已在工业界得到广泛应用，例如，Google公司在Chrome浏览器中嵌入LDP模块来收集浏览器默认主页设置信息[9]、Apple公司在终端输入法中利用LDP来收集emoji表情使用信息以预测用户输入[10]。

然而，尽管LDP在处理大规模数据集时能够利用数据量来抵消扰动带来的影响，从而显示出较高的效用性，但在小规模数据集上，其性能表现并不尽如人意。除此之外，LDP平等地保护所有数据，即真实值扰动到其他所有非真实值的概率相等，但在某些场景希望不同数据对间的隐私保护程度与数据对的距离相关，为此，有学者提出压缩的本地差分隐私(Condensed LDP, CLDP)[11]，它是一种基于度量的LDP放松形式。相比LDP，CLDP在数据规模较小的场景下表现更优。本文专注于序列数据收集问题，LDP下的序列数据收集任务已经有了广泛而深入的研究。然而，自CLDP提出以来，针对满足CLDP的序列数据收集方法研究鲜少。现存的满足CLDP的序列数据收集方法单一地逐位扰动序列的数据项，忽视了序列模式。简单来说，序列模式是数据集中经常出现的子序列，它在序列数据分析中发挥着重要作用[12]，它蕴含着数据间的内在关系和规律。对序列数据进行分析和建模的一种主流方法是Markov链，它是描述一系列可能事件的随机模型，其中每个事件都依赖于前面一个或多个事件，利用Markov链模型能很好地模拟数据间的转移。现已有不少利用Markov链模型来实现隐私保护的序列数据收集方法，如满足DP/LDP的频繁序列模式挖掘[13]、轨迹合成等[6,14-15]。

针对上述问题,本文提出了一种新颖的满足压缩的本地差分隐私的序列数据收集方法(Sequential data collection method with CLDP,SCM-CLDP)。不同于现有方法中用户直接独立地扰动序列中的各项,且数据收集者直接以收集到的扰动集作为最终得到的隐私保护数据集,SCM-CLDP中用户扰动序列并上传给数据收集者,数据收集者在扰动集的基础上先学习Markov链模型,再利用Markov链模型合成数据集。具体来说,SCM-CLDP将序列数据的收集转为数据集的生成问题。首先,由于序列长度分布不均,为后续扰动聚合方便,根据序列长度的估计分布确定了统一的长度,这一分布在后续的序列合成中起到了控制序列结束的作用。其次,针对用户端的扰动,一种直观的做法是直接单独逐一扰动序列中的每个单值,但是希望尽可能保留原始序列模式。因此还考虑了以多个连续值为一组的转移,并将以这两种扰动对象为基础的收集方法分别命名为SCM-VP和SCM-TP。为了增强序列数据的真实性,在SCM-TP中单独处理了序列的起初数据和终止数据。然后,数据收集者在扰动集上学习Markov链模型,SCM-VP中根据序列中每一位的数据分布来进一步构建Markov链模型,SCM-TP中直接估计所有转移的分布并映射到Markov链模型。最后,结合序列长度的估计分布和Markov链模型来合成隐私保护的序列数据集。理论分析表明,SCM-VP和SCM-TP具有良好的隐私保护性。为验证其性能,在相同的隐私保护程度下,在多个评价指标下进行了大量的实验,旨在确保合成的隐私保护序列数据集尽可能地接近原始数据集。实验结果充分证明了SCM-CLDP相比于现有方法的显著优越性,且在大多数情况下SCM-VP优于SCM-TP。

1　相关工作

现有的序列数据的隐私保护工作多集中在传统的差分隐私和本地差分隐私,针对问题主要是频繁序列模式挖掘和数据合成,其中,数据合成方法不局限于某一具体数据分析任务,其适用场景更加广泛。Xu等[16]提出了满足差分隐私频繁序列模式挖掘方法(Differentially private frequent sequence mining via sampling-based candidate pruning,PFS^2),设计了基于采样的候选剪枝技术,该技术是实现PFS^{2}的基础。Xia等[17]提出了满足LDP的频繁序列模式挖掘方法(Sequential patterns mining with LDP,LDPSPM),该方法中,各用户随机报告一组隐私保护的序列模式。Wang等[18]利用前缀树实现了满足LDP的频繁序列挖掘方法PrivTrie。黄硕等[19]基于现有的满足LDP的频繁项集挖掘算法[20]提出PrivSPM(Private sequential pattern mining),该方法旨在识别top-k频繁序列模式并估计其频率。现有的序列数据合成方法大多针对轨迹数据。Gursoy等[6]提出了一个可扩展的位置轨迹合成方法AdaTrace,该方法具有可证明的隐私保护性、对特定攻击的抵抗力和优越的效用性,在轨迹合成中利用Markov模型来模拟实际的轨迹移动。Wang等[14]提出基于自适应Markov模型的满足DP的轨迹合成方法PrivTrace。Du等[15]第一个提出了LDP下的轨迹合成解决方案LDPTrace,该方法中用户本地扰动轨迹移动模式,服务端同样利用Markov模型来模拟轨迹运动。Guner等[21]分析了现有方法的局限性,提出了在LDP下,从序列数据中学习Markov链模型的方法PRIMA,该方法中各用户在本地单独扰动序列中的单值,然后在服务端学习Markov链模型。

CLDP下现存一种序列数据保护的工作,Gursoy等[11]提出了CLDP下的序列数据保护方法Sequence-CLDP,同时保护序列的长度和内容。序列扰动时,利用停止扰动概率halt和生成概率gen两个参数来控制扰动序列的长度。对序列中的每一项,以halt的概率终止,以1－halt的概率扰动;若序列未达到预先设定的序列最大长度,则以gen的概率从$\mathcal{X}$中随机选择一项填充序列,以1－gen的概率终止。Sequence-CLDP逐位扰动序列中的每个值,为了保护序列长度,利用halt和gen来控制序列的结束和数据项的生成,但是这增加了额外的噪声,且数据收集者不做任何后处理,这些都降低了数据的效用性。除此之外,序列模式是序列数据的重要信息,Sequence-CLDP在扰动过程中,直接单独扰动值且在

服务端未做任何后处理,这难以反映序列数据中的重要模式。

2 基础知识与问题描述

本文所提方法基于压缩的本地差分隐私,本节给出压缩的本地差分隐私的相关概念和本文针对的问题模型。

2.1 压缩的本地差分隐私

本节介绍压缩的本地差分隐私的基本概念、常用的扰动机制以及序列级别的定义。压缩的本地差分隐私是本地差分隐私的一种放松形式,在介绍压缩的本地差分隐私之前先介绍本地差分隐私。

定义 1 ϵ-本地差分隐私[7]:扰动算法 K 满足 ϵ-本地差分隐私,当且仅当对任意两真实数据 v_1 和 v_2 有

$$\frac{\Pr[K(v_1)=t]}{\Pr[K(v_2)=t]}\leqslant \mathrm{e}^{\epsilon}\quad \forall t\in \mathrm{Range}(K) \tag{1}$$

式中:ϵ 表示隐私预算;Range(K)表示算法 K 的输出域;t 表示输出域中任意值。

LDP中,用户在本地扰动隐私数据,然后将扰动结果发送给服务器方,使得攻击者无法从传输数据中推断出用户的真实数据。LDP提供了严格的可证明的隐私保证,其中隐私预算 ϵ 控制着隐私保护强度。但是,LDP提供所有数据以相同程度的隐私保护,且大都适用于大规模数据,在小规模数据下表现不如意,针对上述问题,压缩的本地差分隐私应运而生。

定义 2 α-压缩的本地差分隐私[11]:扰动算法 K 满足 α-压缩的本地差分隐私,当且仅当对任意两真实数据 v_1 和 v_2 有

$$\frac{\Pr[K(v_1)=t]}{\Pr[K(v_2)=t]}\leqslant \mathrm{e}^{\alpha\cdot \mathrm{dist}(v_1,v_2)}\quad \forall t\in \mathrm{Range}(K) \tag{2}$$

式中:α 为隐私预算;dist(v_1, v_2)表示 v_1、v_2 间距离。

CLDP是LDP的一种基于度量(距离函数)的放松形式,距离的引入使得CLDP能更灵活调节隐私性和效用性间的平衡。隐私预算不变,数据对间距离越小,这一数据对扰动输出相同值的概率就越接近,这一数据对间的不可区分性就越强;相反,数据对间距离越大,这一数据对扰动输出相同值的概率就越疏远,这一数据对间的可区分性就越强。在相同程度的隐私保护下,由于距离的影响,CLDP的隐私参数 α 要远小于LDP的隐私参数 ϵ[11]。特别地,当任意数据对间距离固定时,CLDP就退化成LDP。

指数机制是满足CLDP的扰动机制,它是DP下指数机制的变体,其基本思想是:用户根据数据域中各值和真实值间的距离对各值打分,距离越小的分数越高,以较高的概率输出分数较高的值,输出概率与分数成正比。给定数据域 $\mathcal{X}$,对于任意真实值 $x\in\mathcal{X}$,扰动输出 $y\in\mathcal{X}$ 的概率为

$$\Pr[K_{\mathrm{ExpM}}(x)=y]=\frac{\mathrm{e}^{-\frac{\alpha}{2}\cdot \mathrm{dist}(x,y)}}{\sum_{z\in\mathcal{X}}\mathrm{e}^{-\frac{\alpha}{2}\cdot \mathrm{dist}(x,z)}} \tag{3}$$

序列数据下的隐私保护可分为事件级别的和序列级别的,序列级别的隐私保护旨在保护整个序列而非单个事件,因此采用序列级别的CLDP定义。

定义 3 序列级别的 α-压缩的本地差分隐私[11]:扰动算法 K 满足序列级别的 α-压缩的本地差分隐私,当且仅当对任意两真实序列数据 S_1 和 S_2 有

$$\frac{\Pr[K(S_1)=t]}{\Pr[K(S_2)=t]}\leqslant \mathrm{e}^{\alpha\cdot \mathrm{dist}_{\mathrm{seq}}(S_1,S_2)}\quad \forall t\in \mathrm{Range}(K) \tag{4}$$

式中$\mathrm{dist}_{\mathrm{seq}}(S_1,S_2)$表示$S_1$、$S_2$间距离。

从上述定义中可以看出,相较于α-CLDP,序列级别的只是扰动对象不同,原始的CLDP限制了单个数据值扰动到其他数据值的概率,而序列级别的CLDP限制了整个序列扰动到其他序列的概率。特别地,本文中将序列间的距离函数定义为$\mathrm{dist}_{\mathrm{seq}}(X,Y)=\sum_{i=1}^{|X|}\mathrm{dist}(X[i],Y[i])$,其中$X[i]$和$Y[i]$分别表示序列$X$和序列$Y$的第$i$项。

2.2　问题描述

假设有n个用户和1个数据收集者,用户u拥有1敏感的序列数据$S_u=x_i\rightarrow x_j\rightarrow x_k\rightarrow\cdots$,其中$x_i\in\boldsymbol{X}=\{x_1,x_2,\cdots,x_d\}$,$d=|x|$,$S_u[i]$表示用户$u$的序列数据的第$i$项。数据收集者收集所有用户的序列数据,并在此数据集上做统计分析任务,如收集个人轨迹做道路流量预测、收集个人网页浏览数据做网页推荐等。然而,这些数据中往往包含有隐私信息,比如从个人轨迹中可分析出家庭住址、工作场所,甚至健康状况。用户并不希望提交自己真实的敏感序列数据给数据收集者,因此,本文关注隐私保护的序列数据收集问题。

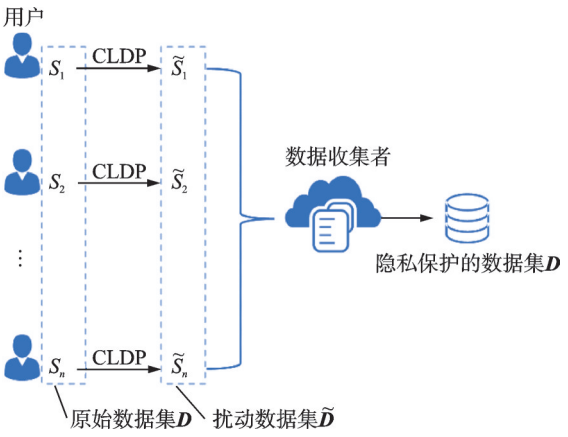

图1　问题模型图

Fig.1　Problem model diagram

如图1所示,为保护隐私,用户u在本地使用扰动机制对S_u扰动,得到$\tilde{S}_u$,攻击者即使知道$\tilde{S}_u$也无法准确推断出S_u。收集者对所有扰动数据$\tilde{\boldsymbol{D}}$做聚合,得到隐私保护的序列数据集$\hat{\boldsymbol{D}}$。本文的目标是设计一个满足序列级别CLDP的序列数据收集方法,在保护隐私数据的同时,提供与原始序列数据集$\boldsymbol{D}$高度相似的隐私保护的序列数据集$\hat{\boldsymbol{D}}$,其中,序列数据集的相似性从Markov模型的相似性、$\boldsymbol{X}$上数据分布的相似性和频繁序列模型的相似性来综合衡量。

3　序列数据收集方法

3.1　方法概述

序列数据中各项按一定顺序排列,如轨迹序列数据中各位置点按照时间先后顺序排列,网页浏览数据中各网页按照用户浏览的先后顺序排列,文本数据中各字符、单词按照特定的顺序排序。Markov链模型是分析研究序列数据的主要方法,它可以很好地表示各状态间的转移关系,本文提出的方法SCM-CLDP利用一阶Markov链模型来生成与原始序列数据集高度相似的隐私保护的序列数据集。

首先,收集序列长度。用户在本地扰动真实序列长度并提交给数据收集者,数据收集者估计长度分布,并选择一个合适的值作为后续序列扰动的最大长度。其次,收集序列数据,生成Markov链模型。用户在本地扰动序列数据,并将扰动结果发送给数据收集者。数据收集者根据所有用户的扰动结果学习Markov链模型。进一步地,依据不同的扰动对象设计了不同的序列数据收集方法,分别是基于值扰动的SCM-VP和基于转移扰动的SCM-TP。最后,合成隐私保护的序列数据集。已知序列长度分布和该用户群体组成的序列数据集上的Markov链模型,数据收集者可据此合成隐私保护的序列数据集,且每条序列数据的长度由长度分布和Markov链模型同时控制。

下面首先介绍序列长度的收集,然后分别介绍基于值扰动的序列数据收集方法SCM-VP和基于转移扰动的序列数据收集方法SCM-TP,并在SCM-VP的介绍中给出序列数据集的生成算法。

3.2 序列长度收集

序列长度是序列数据集中的一个关键信息,利用长度分布可以更好地合成序列数据。假设序列长度所属的数据域公开已知,每个用户利用指数机制扰动其真实序列长度,其中,定义距离函数为绝对距离,数据收集者收集所有用户的扰动结果,然后利用期望最大化算法估计长度分布$\boldsymbol{L}$。现已有不少工作利用期望最大化算法估计频率[22-23],由于文章篇幅问题,本文不再详细介绍利用期望最大化算法做分布估计的具体步骤和相关特性。

由于每个用户提交的序列数据长短不一,为了简单起见,假设每个用户的序列数据长度为l,l是一个向用户和数据收集者公开的参数。如果用户的序列数据长度大于l,则截断数据只扰动前l项,否则填充虚假项至序列长度为l,在后续的聚合估计过程中会忽略这些虚假项。l的取值影响着数据的效用性和算法的时间空间代价,随着l的增大,参与扰动的数据项增加,数据的效用性提高,但算法的时间空间代价也会随之增加。本文遵循了现有的LDP下集值数据[15]、轨迹数据收集的设置[24-25],将l设置为序列长度估计分布的90%分位数。

3.3 基于值扰动的序列数据收集方法

SCM-VP中,用户的扰动对象是序列中的单项,已知序列中数据项所属数据域$\mathcal{X}$,则序列S的扰动输出$\tilde{S}$的概率为

$$\Pr[\tilde{S}[i]|S[i]]=\frac{\mathrm{e}^{-\alpha\cdot\mathrm{dist}(S[i],\tilde{S}[i])/2}}{\sum_{z\in\mathcal{X}}\mathrm{e}^{-\alpha\cdot\mathrm{dist}(S[i],z)/2}}\qquad i=1,2,\cdots,l \tag{4}$$

式(1)确保了序列中每个数据项都能以较高的概率扰动到离它本身较近的数据项,使得最终得到的扰动序列尽可能接近原始真实序列。数据收集者收集所有用户的扰动序列,首先采用期望最大化算法逐位估计$\mathcal{X}$上各值的频率,再根据每位上的数据分布学习Markov链模型,最终基于Markov链模型和序列长度分布合成序列数据集。

令$\boldsymbol{P}$表示估计频率矩阵,P_{ij}表示序列第i位上值为x_j的估计频率,已知$\boldsymbol{P}$可学习到Markov链模型。令$\boldsymbol{M}$表示Markov链模型,状态集为$\boldsymbol{X}$,M_{ij}表示从状态$x_i\in\boldsymbol{X}$到状态$x_j\in\boldsymbol{X}$的转移,则M_{ij}转移概率为序列中当前项是x_i且下一项是x_j的概率和,即

$$\Pr(M_{ij})=\sum_{k=1}^{l-1}P_{ki}\cdot P_{(k+1)j}\Big/\sum_{k=1}^{l-1}\sum_{t=1}^{d}P_{ki}\cdot P_{(k+1)t} \tag{5}$$

Markov链模型中,初始状态为后续状态提供了初始条件,在诸如天气预测、股票价格预测和交通流量预测等应用中,初始状态的选择可能影响模型预测的准确性。初始状态和终止状态还有助于后续的数据合成。然而,式(5)仅反应了序列的内部转移,忽略了对序列开始和结束的建模,为此,单独考虑了初始转移和终止转移的收集。具体来说,增加一虚拟项x_0用来表示开始和结束,初始转移表示为$\{M_{0i}: x_0\rightarrow x_i,\ i=1,2,\cdots,d\}$,类似地,终止转移表示为$\{M_{i0}: x_i\rightarrow x_0,\ i=1,2,\cdots,d\}$,那么,$M_{0i}$转移概率为序列第一位是$x_i$的概率,$M_{i0}$转移概率为序列最后一位是$x_i$的概率,即

$$\Pr[M_{0i}]=P_{1i},\ \Pr[M_{i0}]=P_{li} \tag{6}$$

算法1给出了基于值扰动的Markov链模型学习算法,该算法中,输入为数据收集者收集到的所有用户的基于值扰动的序列扰动结果,输出为从扰动序列集中学习到的Markov链模型。已知Markov链模型和序列长度分布,可通过算法2合成序列数据集。首先采样序列长度得到l',然后根据$\boldsymbol{M}$来采样序列中的每个值,当采样的值为x_0或当前序列中已有l'个值时,则终止当前序列的合成,开始下一条序列

数据的合成。

算法1　基于值扰动的Markov链模型学习算法

输入:所有用户的扰动结果。

输出:Markov链模型$\boldsymbol{M}$。

(1)　//期望最大化算法估计每一位上的频率分布

(2)　初始化分布矩阵$\boldsymbol{P}$

(3)　for $t=1$ to l do

(4)　　统计扰动序列中第t位上值为x_i的频率$\tilde{\theta}_i$

(5)　　while not converge do

(6)　　E-step: $Q_i=\sum_{j=1}^{d}\tilde{\theta}_j\cdot\frac{\hat{\theta}_i\Pr[x_j|x_i]}{\sum_{k=1}^{d}\hat{\theta}_k\Pr[x_j|x_k]}$, $i=1,2,\cdots,d$ //$\hat{\theta}_i$表示估计频率

(7)　　M-step: $\hat{\theta}_i=\frac{Q_i}{\sum_{j=1}^{d}Q_j}$

(8)　　end while

(9)　　$P_{ti}=\hat{\theta}_i$

(10)　end for

(11)　//学习Markov链模型$\boldsymbol{M}$

(12)　for $i=1$ to d do

(13)　　$\Pr[M_{0i}]=P_{1i}$, $\Pr[M_{i0}]=P_{li}$

(14)　end for

(15)　for $i=1$ to d do

(16)　　$\Pr[M_{ij}]=\sum_{k=1}^{l-1}P_{ki}\cdot P_{k+1j}$, $j=1,2,\cdots,d$

(17)　end for

(18)　归一化频率的每一行

(19)　return $\boldsymbol{M}$

算法2　序列数据集合成算法

输入:序列长度估计分布$\boldsymbol{L}$,Markov链模型$\boldsymbol{M}$。

输出:序列数据集$\hat{\boldsymbol{D}}$。

(1)　for $u=1$ to n do

(2)　　采样符合分布$\boldsymbol{L}$的序列长度l'

(3)　　$k=0$　//当前值为x_k

(4)　　for $i=1$ to l' do

(5)　　　以概率$\Pr[M_{kj}]$采样

(6)　　　if $j=0$ then　//采样到虚拟项则结束

(7)　　　　break

(8) end if

(9) $S'_u[i] \leftarrow x_j$ //$S'_u[i]$表示用户u的合成序列

(10) $k=j$

(11) end for

(12) $\hat{\boldsymbol{D}}$中添加用户u的合成序列S'_u

(13) end for

(14) return $\boldsymbol{D}$

3.4 基于转移扰动的序列数据收集方法

在序列分析中序列的模式至关重要,因此希望尽可能保留原始的序列模式。为此,在SCM-TP中,不同于SCM-VP的值扰动,而是扰动序列中的转移模式。为了后续的描述方便,首先介绍转移集$\boldsymbol{T}$。已知状态集大小为d,那么不考虑初始转移和终止转移共有内部转移d^2种转移,考虑初始转移$\{M_{0i}: x_0 \rightarrow x_i,\ i=1,2,\cdots,d\}$和终止转移$\{M_{i0}: x_i \rightarrow x_0,\ i=1,2,\cdots,d\}$共有$d^2+2d$种转移。进一步地,将转移域定义为$\boldsymbol{T}=\left\{t_1,t_2,\cdots,t_{|T|}\right\}$,且$|\boldsymbol{T}|=d^2+2d$,$t_{id+j}$表示转移$x_i \rightarrow x_j$,$\boldsymbol{T}$上的距离函数定义为$\mathrm{dist}_T(t_{id+j},t_{md+n})=\mathrm{dist}(x_i,x_m)+\mathrm{dist}(x_j,x_n)$。由此,可进一步得到初始转移域$\boldsymbol{T}_{\mathrm{start}}=\{t_1,t_2,\cdots,t_d\}$、终止转移域$\boldsymbol{T}_{\mathrm{end}}=\{t_{d^2+d+1},t_{d^2+d+2},\cdots,t_{d^2+2d}\}$和内部转移域$\boldsymbol{T}_{\mathrm{intra}}=\{t_{d+1},t_{d+2},\cdots,t_{d^2+d}\}$。

用户扰动时,首先将长为l的序列映射到转移集$\boldsymbol{T}$,其中$|\boldsymbol{T}|=l+1$且$T[i]\in\boldsymbol{T}$;然后采用指数机制分别扰动初始转移、内部转移和终止转移;最终提交扰动结果$\tilde{T}$给数据收集者。下面,将分别介绍初始/终止转移和内部转移的扰动。

希望扰动不改变转移所属的子域,即初始转移的扰动域为$\boldsymbol{T}_{\mathrm{start}}$,终止转移的扰动域为$\boldsymbol{T}_{\mathrm{end}}$,内部转移的扰动域为$\boldsymbol{T}_{\mathrm{intra}}$。确定扰动域后,可得到转移的扰动概率。初始转移的扰动概率为

$$\Pr[\tilde{T}[1]|T[1]]=\frac{\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[1],\tilde{T}[1])/2}}{\sum_{t\in\boldsymbol{T}_{\mathrm{start}}}\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[1],t)/2}} \tag{7}$$

类似地,终止转移的扰动概率为

$$\Pr[\tilde{T}[l]|T[l]]=\frac{\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[l],\tilde{T}[l])/2}}{\sum_{t\in\boldsymbol{T}_{\mathrm{end}}}\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[l],t)/2}} \tag{8}$$

$\boldsymbol{T}$中内部转移有$T[2],T[3],\cdots,T[l-1]$,它们的扰动概率为

$$\Pr[\tilde{T}[i]|T[i]]=\frac{\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[i],\tilde{T}[i])/2}}{\sum_{t\in\boldsymbol{T}_{\mathrm{intra}}}\mathrm{e}^{-\alpha\cdot\mathrm{dist}_T(T[i],t)/2}}\qquad i=2,3,\cdots,l-1 \tag{9}$$

数据收集者收集到所有用户的扰动结果,分别利用期望最大化算法估计初始转移、内部转移和终止转移的频率,然后将转移分布映射到Markov链模型,如算法3所示。已知Markov链模型和长度分布,利用算法2可合成序列数据集。

算法3 基于转移扰动的Markov链模型学习算法

输入:所有用户的扰动结果。

输出:Markov链模型$\boldsymbol{M}$。

(1) //期望最大化算法估计初始/终止转移、内部转移分布

(2) 统计任意$t_i\in\boldsymbol{T}$的频率$\tilde{\theta}_i$

(3)　//初始转移频率估计

(4)　while not converge do

(5)　　E-step: $Q_i=\sum_{j=1}^{d}\tilde{\theta}_j\cdot\frac{\hat{\theta}_i\Pr[t_j|t_i]}{\sum_{k=1}^{d}\hat{\theta}_k\Pr[t_j|t_k]}$, $i=1,2,\cdots,d$

(6)　　M-step: $\hat{\theta}_i=\frac{Q_i}{\sum_{j=1}^{d}Q_j}$, $j=1,2,\cdots,d$

(7)　end while

(8)　//内部转移频率估计

(9)　while not converge do

(10)　　E-step: $Q_i=\sum_{j=d+1}^{d^2+d}\tilde{\theta}_j\cdot\frac{\hat{\theta}_i\Pr[t_j|t_i]}{\sum_{k=d+1}^{d^2+d}\hat{\theta}_k\Pr[t_j|t_k]}$, $i=d+1,d+2,\cdots,d^2+d$

(11)　　M-step: $\hat{\theta}_i=\frac{Q_i}{\sum_{j=d+1}^{d^2+d}Q_j}$, $i=d+1,d+2,\cdots,d^2+d$

(12)　end while

(13)　//终止转移频率估计

(14)　while not converge do

(15)　　E-step: $Q_i=\sum_{j=d^2+d+1}^{d^2+2d}\tilde{\theta}_j\cdot\frac{\hat{\theta}_i\Pr[t_j|t_i]}{\sum_{k=d^2+d+1}^{d^2+2d}\hat{\theta}_k\Pr[t_j|t_k]}$, $i=d^2+d+1,\cdots,d^2+2d$

(16)　　M-step: $\hat{\theta}_i=\frac{Q_i}{\sum_{j=d^2+d+1}^{d^2+2d}Q_j}$, $i=d^2+d+1,\cdots,d^2+2d$

(17)　end while

(18)　//学习Markov链模型$\boldsymbol{M}$

(19)　for $i=1$ to d do

(20)　　$\Pr[M_{0i}]=\hat{\theta}_i$, $\Pr[M_{i0}]=\hat{\theta}_{d^2+d+i}$

(21)　end for

(22)　for $i=1$ to d do

(23)　　$\Pr[M_{ij}]=\hat{\theta}_{id+j}$, $j=1,2,\cdots,d$

(24)　end for

(25)　归一化$\boldsymbol{M}$的每一行

(26)　return $\boldsymbol{M}$

3.5　隐私性分析

SCM-VP中用户单独扰动序列中的每一项。SCM-TP中用户以序列中相邻两项组成的转移为单

位做扰动，考虑初始转移和终止转移，SCM-TP相当于做了两次值扰动。令SCM-VP的隐私预算为α_{VP}，SCM-TP的隐私预算为α_{TP}，下面将分别证明SCM-VP和SCM-TP满足序列级别的α-CLDP，并给出在相同隐私保护下α_{VP}和α_{TP}间的关系。

定理1 SCM-VP算法满足序列级别的$\boldsymbol{\alpha_{\mathrm{VP}}}$-CLDP

证明：若要证明SCM-VP算法满足序列级别的α_{VP}-CLDP，只需证明输入任意两长为l的序列S_1和S_2，输出长为l的序列S的概率之比满足

$$\frac{\Pr[S|S_1]}{\Pr[S|S_2]}\leqslant \mathrm{e}^{\alpha_{\mathrm{VP}}\cdot \mathrm{dist}_{\mathrm{seq}}(S_1,S_2)} \tag{10}$$

证明

$$\frac{\Pr[S|S_1]}{\Pr[S|S_2]}=\frac{\Pr[S[1]|S_1[1]]}{\Pr[S[1]|S_2[1]]}\cdot\frac{\Pr[S[2]|S_1[2]]}{\Pr[S[2]|S_2[2]]}\cdots\frac{\Pr[S[l]|S_1[l]]}{\Pr[S[l]|S_2[l]]}\leqslant$$

$$\mathrm{e}^{\alpha_{\mathrm{VP}}\cdot \mathrm{dist}(S_1[1],S_2[1])}\cdot\mathrm{e}^{\alpha_{\mathrm{VP}}\cdot \mathrm{dist}(S_1[2],S_2[2])}\cdots\mathrm{e}^{\alpha_{\mathrm{VP}}\cdot \mathrm{dist}(S_1[l],S_2[l])}=\mathrm{e}^{\alpha_{\mathrm{VP}}\cdot\sum_{i=1}^{l}\mathrm{dist}(S_1[i],S_2[i])}=\mathrm{e}^{\alpha_{\mathrm{VP}}\cdot \mathrm{dist}_{\mathrm{seq}}(S_1,S_2)} \tag{11}$$

得证。SCM-VP算法满足序列级别的α_{VP}-CLDP。

定理1给出了SCM-VP所满足的隐私性证明，由于SCM-TP每次扰动的对象是相邻两值组成的转移，因此给定隐私参数，不同于SCM-VP中实际隐私预算等于隐私参数，在SCM-TP中实际隐私预算是隐私参数的2倍。

定理2 SCM-TP算法满足序列级别的$\boldsymbol{2\alpha_{\mathrm{TP}}}$-CLDP

证明：若要证明SCM-TP算法满足序列级别的α_{TP}-CLDP，只需证明输入任意两长为l的序列S_1和S_2，输出长为l的序列S的概率之比满足

$$\frac{\Pr[S|S_1]}{\Pr[S|S_2]}\leqslant \mathrm{e}^{2\alpha_{\mathrm{TP}}\cdot \mathrm{dist}_{\mathrm{seq}}(S_1,S_2)} \tag{12}$$

SCM-TP中将长为l的序列转为长为$l+1$的转移集上，因此有

$$\frac{\Pr[S|S_1]}{\Pr[S|S_2]}=\frac{\Pr[T[1]|T_1[1]]}{\Pr[T[1]|T_2[1]]}\cdot\frac{\Pr[T[2]|T_1[2]]}{\Pr[T[2]|T_2[2]]}\cdots\frac{\Pr[T[l+1]|T_1[l+1]]}{\Pr[T[l+1]|T_2[l+1]]}\leqslant \mathrm{e}^{\alpha_{\mathrm{TP}}\cdot \mathrm{dist}_T(T_1[1],T_2[1])}\times$$

$$\mathrm{e}^{\alpha_{\mathrm{TP}}\cdot \mathrm{dist}_T(T_1[2],T_2[2])},\cdots,\mathrm{e}^{\alpha_{\mathrm{TP}}\cdot \mathrm{dist}_T(T_1[l+1],T_2[l+1])}=\mathrm{e}^{\alpha_{\mathrm{TP}}\cdot \mathrm{dist}(S_1[1],S_2[1])}\times$$

$$\mathrm{e}^{\alpha_{\mathrm{TP}}\cdot(\mathrm{dist}(S_1[1],S_2[1])+\mathrm{dist}(S_1[2],S_2[2]))},\cdots,\mathrm{e}^{\alpha_{\mathrm{TP}}\cdot \mathrm{dist}(S_1[l],S_2[l])}=\mathrm{e}^{2\alpha_{\mathrm{TP}}\cdot\sum_{i=1}^{l}\mathrm{dist}(S_1[i],S_2[i])}=\mathrm{e}^{2\alpha_{\mathrm{TP}}\cdot \mathrm{dist}_{\mathrm{seq}}(S_1,S_2)} \tag{13}$$

得证。SCM-TP算法满足序列级别的$2\alpha_{\mathrm{TP}}$-CLDP。

通过定理1和理论2，可以进一步得到，当SCM-VP和SCM-TP均满足序列级别的α-CLDP时，有$\alpha_{\mathrm{VP}}=2\alpha_{\mathrm{TP}}=\alpha$。实验中始终将$\alpha_{\mathrm{TP}}$设置为$\alpha_{\mathrm{VP}}/2$。

3.6 复杂度分析

在SCM-CLDP方法中，每个用户首先对其序列长度进行扰动，并将扰动结果发送给数据收集者。这里令$|L|$表示序列长度区间大小。在本文中，所有扰动均采用指数机制。因此，序列长度扰动的时间复杂度为$O(|L|)$。数据收集者在收集到所有用户的长度信息后，将利用期望最大化算法来估计长度分布。分布估计的时间复杂度主要取决于期望最大化算法的迭代次数和数据域的大小。用r来表示期望最大化算法的迭代次数，因此，长度分布估计的时间复杂度为$O(|L|r)$。

数据收集者将长度分度的90%分位数l发送给各用户用于序列扰动。在序列扰动阶段，SCM-VP方法中，用户对序列中的l个项进行独立地扰动，每项的数据域大小为d，因此SCM-VP方法的扰动时

间复杂度为$O(ld)$。而在SCM-TP方法中,用户对$l-1$个内部转移以及起始和终止转移进行扰动,其中内部转移的数据域大小为d^2,起始和终止转移的数据域大小为d,因此SCM-TP方法的总扰动时间复杂度为$O(ld^2)$。

数据收集者在收集到所有用户的扰动序列后,将进行序列分布估计。在SCM-VP方法中,数据收集者逐位估计序列中每一位的数据分布,每位都需要调用一次期望最大化算法,因此总的分布估计时间复杂度为$O(ldr)$。而在SCM-TP方法中,数据收集者直接估计转移分布,转移域大小为d^2+2d,总的分布估计时间复杂度为$O(d^2r)$。数据收集者将基于估计的序列分布生成Markov链模型,该过程的时间复杂度为$O(d^2)$。最后利用Markov链模型和长度分布合成数据集,这一过程的时间复杂度为$O(nl)$。

4　实验评估与验证

4.1　实验设置

本节将介绍仿真实验的设置,包括数据集设置和评价指标设置,实验环境是Intel i5-10500 6核CPU、16 GB内存、Windows11系统,代码采用Python实现。

4.1.1　数据集

在两个公开数据集上进行实验,分别是MSNBC和T-drive。MSNBC数据集[26]描述了1999年9月28日访问msnbc.com网站的用户页面访问数据。数据集中的每条序列对应一个用户在24 h内按时间顺序访问的页面。其中页面被分为17类,且被编号为1～17。实验中,随机选取其中100 000条记录,序列平均长度为4.68。T-drive数据集[27]包含了2008年2月2日至2月8日期间在北京市内的10 357辆出租车的GPS数据。实验中,只考虑经度范围在116.20°～116.60°,且纬度度范围在39.84°～40.04°的区域,并将该区域划分成2×4的网格,然后将出租车GPS数据映射到该网格上,网格单元编号为1～8,最终得到的数据集中轨迹平均长度为85.72。

4.1.2　评价指标

本文从3个方面来评估算法性能,分别是Markov链模型的效用性、合成数据集的效用性和频繁序列模式的效用性。下面将从这3个方面来分别介绍评价指标。

(1)Markov链模型的效用性。Markov链模型对合成数据集的生成至关重要,这里通过初始分布的误差和(内部)转移概率的误差来衡量Markov链模型的效用性。令$\boldsymbol{M}$表示真实Markov链模型,$\hat{\boldsymbol{M}}$表示隐私保护的Markov链模型,初始分布误差和转移概率误差定义如下。

(a) 初始分布误差(Initial distribution error, IDE)。衡量隐私保护Markov链模型与真实模型在初始分布上的差异,计算方式为

$$\mathrm{IDE}=\frac{1}{d}\sum_{i=1}^{d}\left(\Pr[M_{0i}]-\Pr\left[\hat{M}_{0i}\right]\right)^2 \tag{14}$$

(b) 转移概率误差(Transition probability error, TPE)。衡量隐私保护Markov链模型与真实模型在转移概率上的差异,计算方式为

$$\mathrm{TPE}=\frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\Pr[M_{ij}]-\Pr\left[\hat{M}_{ij}\right]\right)^2 \tag{15}$$

(2)合成数据集的效用性。希望合成数据集尽可能接近真实数据集,本文从数据集分布的误差和数据项排名的准确性来衡量合成数据集的效用性,其中数据项排名的准确性采用肯德尔τ系数来衡量。令θ表示真实分布且θ_i表示x_i的频率,$\hat{\theta}$表示合成数据集的分布且$\hat{\theta}_i$表示合成数据集中x_i的频率,数据分布误差和肯德尔τ系数定义如下。

(a) 数据集分布误差(Dataset distribution error, DDE)。衡量合成数据集与真实数据集中数据分布

的差异,计算方式为

$$\mathrm{DDE}=\frac{1}{d}\sum_{i=1}^{d}(\theta_i-\hat{\theta}_i)^2 \tag{16}$$

(b) 肯德尔τ系数。衡量合成数据集中数据项排名的准确性。当且仅当$\theta_i>\theta_j$且$\hat{\theta}_i>\hat{\theta}_j$时,$(x_i,x_j)$才被认为是和谐的数据对,否则是不和谐的。令$N_{\mathrm{c}}$表示和谐的数据对数,$N_{\mathrm{d}}$表示不和谐的数据对数,那么有

$$\tau=\frac{N_{\mathrm{c}}-N_{\mathrm{d}}}{d(d-1)/2} \tag{17}$$

(3)频繁序列模式的效用性。序列模式往往能反映数据之间的内在关系和规律,这里考虑的序列模式是连续的有序的子序列,并选择top-k频繁模式做进一步的度量。令真实数据集上top-k的频繁模式为$\boldsymbol{FP}$,对应地,合成数据集上的为$\boldsymbol{FP}_{\mathrm{syn}}$,利用频繁模式的$F_1$分数和频数误差来反映频繁序列模式的效用性。

(a) F_1分数:用于衡量二分类模型精确度的一种指标,这里用于评估合成数据集中top-k频繁模式挖掘的准确性,有

$$F_1=2\times\frac{\mathrm{Precision}(\boldsymbol{FP},\boldsymbol{FP}_{\mathrm{syn}})\times\mathrm{Recall}(\boldsymbol{FP},\boldsymbol{FP}_{\mathrm{syn}})}{\mathrm{Precision}(\boldsymbol{FP},\boldsymbol{FP}_{\mathrm{syn}})+\mathrm{Recall}(\boldsymbol{FP},\boldsymbol{FP}_{\mathrm{syn}})} \tag{18}$$

(b) 模式频数误差(Pattern frequency error, PFE):衡量隐私保护的合成数据集中模式出现频数的差异。这里只考虑top-k的频繁模式,统计真实的top-k频繁模式在真实数据集和合成数据集中频数的差异,有

$$\mathrm{PFE}=\frac{1}{|\boldsymbol{FP}|}\sum\nolimits_{fp\in FP}\frac{|c(fp)-c_{\mathrm{syn}}(fp)|}{c(fp)} \tag{19}$$

式中:$c(fp)$和$c_{\mathrm{syn}}(fp)$分别表示在真实数据集和合成数据集中频繁模式fp的出现次数。特别地,若$fp\notin\boldsymbol{FP}_{\mathrm{syn}}$,则$c_{\mathrm{syn}}(fp)=0$。

如无特殊说明,后文展示的所有结果为实验重复运行20次取平均值的结果。

4.1.3 对比方案

本文专注于满足压缩的本地差分隐私的序列数据收集问题,而Sequence-CLDP是目前最先进的且唯一满足CLDP的序列数据收集方案,因此将本文所提方法与Sequence-CLDP[11]进行了广泛且深入的对比,以评估和比较本文提出的方法的性能。为了公平比较,Sequence-CLDP同样只扰动序列的前l项,且隐私预算设置与SCM-VP的相同。

4.2 实验结果与分析

4.2.1 Markov链模型效用性分析

本节在不同数据集、不同隐私预算下对比了不同方案的Markov链模型效用性,由于距离函数的影响,CLDP中的隐私参数要远小于LDP中的隐私参数,因此这里设置的隐私参数范围$\alpha\in(0,1)$要远小于LDP研究中常见的设置。实验中和现有的Sequence-CLDP方案进行了对比,公平起见,将Sequence-CLDP方案中的最大序列长度设置为l,若无特殊说明,后续的实验均为这个设置。

MSNBC上的实验结果如图2所示。从图2中可以清楚地看出在绝大多数情况下,所有方案的误差都呈下降趋势。基于初始分布误差的实验结果,可以发现SCM-VP的IDE最低,SCM-TP次之,Sequence-CLDP最差,SCM-TP的IDE高于SCM-VP是因为SCM-TP中初始状态的实际隐私预算为SCM-VP的1/2。此外,随着α的增大,SCM-TP的效果逐渐逼近SCM-VP。在评估转移概率的误差时,观察到当$\alpha>0.1$时,SCM-TP的TPE显著低于SCM-VP和Sequence-CLDP,且随着α的增大,与其他方案的差异明显增大。除此之外,注意到SCM-VP的TPE受α的影响较小,甚至随着α的增大而轻

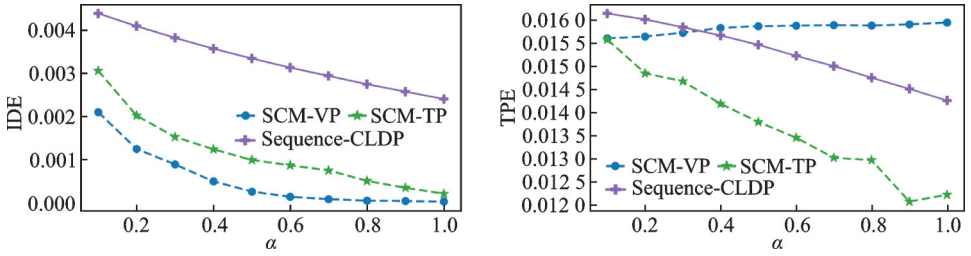

图2　MSNBC数据集上初始分布误差和转移概率误差在不同α下的变化

Fig.2　Variation of initial distribution error and transition probability error with α on the MSNBC dataset

微增大,并且当$\alpha>0.3$时,SCM-VP的效果要劣于Sequence-CLDP。

T-drive上的实验结果如图3所示。从图3中,可以发现在所有情况下,所有方案的误差都呈下降趋势,且SCM-VP始终表现最优。此外,综合考虑IDE和TPE,在绝大多数情况下,SCM-VP表现最优,其次是SCM-TP,而Sequence-CLDP表现最差。特别地,当$\alpha\leqslant0.4$时,Sequence-CLDP的IDE要略微低于SCM-TP,这是因为T-drive的数据集规模较小,且实际分配给SCM-TP的隐私预算是其他两方案的1/2,在隐私预算较小时无法有效弥补扰动带给SCM-TP的消极影响。与图2相比,发现MSNBC数据集上的实验结果要整体好于T-drive,且在TPE指标上,SVM-TP在MSNBC数据集上的效果明显更优,甚至远优于SVM-VP和Sequence-CLDP,猜测这是因为数据规模对于SVM-TP的影响更大,随着数据规模的增大,SVM-TP中转移分布估计的准确性随之提高,而转移分布直接且密切影响着TPE,进而TPE随之降低。

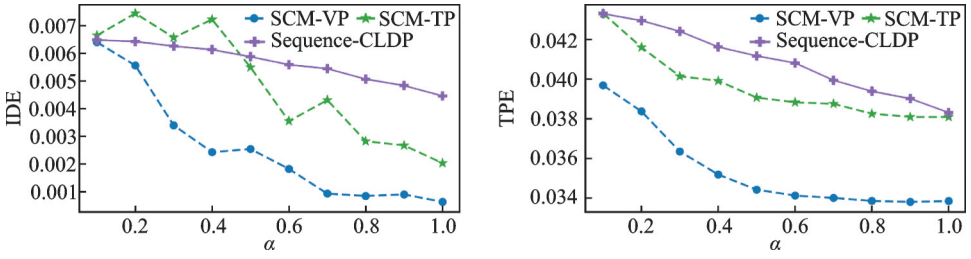

图3　T-drive数据集上初始分布误差和转移概率误差在不同α下的变化

Fig.3　Variation of initial distribution error and transition probability error with α on the T-drive dataset

4.2.2　合成序列数据集效用性分析

下面分析利用Markov链模型合成的隐私保护数据集的效用性,图4和图5分别描述了在MSNBC数据集和T-drive数据集上的实验结果。由图4、5可以发现:在所有数据集上,随着α的增大,所有方案的DDE均呈下降趋势、所有方案的τ均呈上升趋势,并且在所有情况下SCM-VP表现最优,SCM-TP次之,Sequence-CLDP表现最差。特别地,当$\alpha=0.1$时,三者间的差距较小,在T-drive数据集上,三者的DDE和τ几乎重合,随着α的增大,SCM-CLDP方法的优势显著显现出来。结合图2,SVM-TP的转移概率误差最小,但是这一优势在合成序列数据集上并没有显现出来,这是因为初始状态定义了过程的起点,为后续状态提供了初始条件,合成序列数据是初始分布和内部转移概率共同作用的结果。

4.2.3　频繁序列模式效用性分析

这里分析了隐私保护的数据集上的频繁序列模式的效用性,实验中选取top-25的频繁模式进行评估。从图6中发现在MSNBC数据集上,SVM-TP的F_1分数要远高于SCM-VP和Sequence-CLDP,但是,它的频繁模式频数误差要远高于SCM-VP,这是因为虽然SCM-VP的top-25序列模式的命中率较

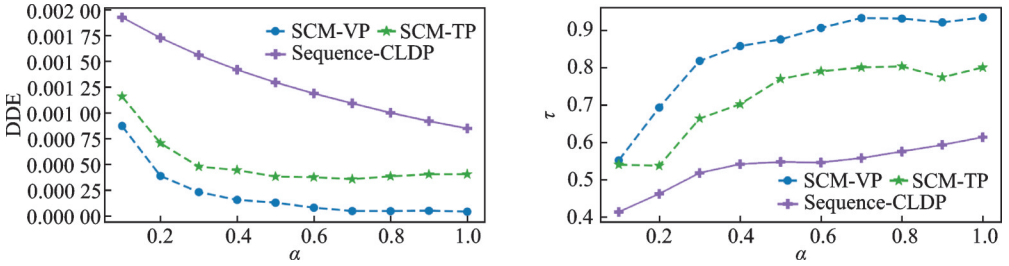

图4 MSNBC数据集上数据集分布误差和肯德尔τ系数在不同α下的变化

Fig.4 Variation of dataset distribution error and Kendall's τ coefficient with α on the MSNBC dataset

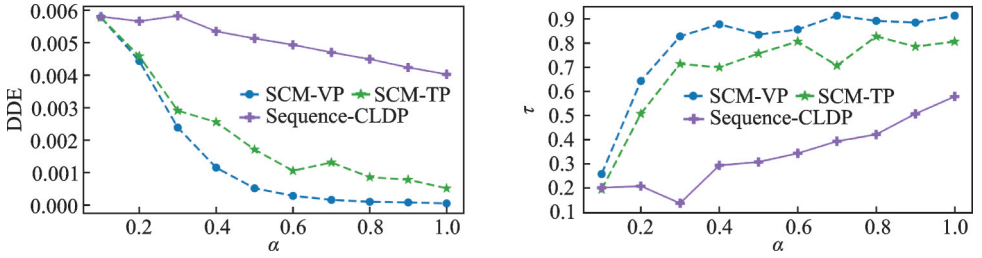

图5 T-drive数据集上数据集分布误差和肯德尔τ系数在不同α下的变化

Fig.5 Variation of dataset distribution error and Kendall's τ coefficient with α on the T-drive dataset

低,但是它有着更准确的较频繁的若干序列模式频数。另外,还发现F_1分数上3种方案的表现和图2中TPE的一致:SCM-TP的表现始终最优,SCM-VP的F_1分数和TPE都随着α的增大而变差,当α较大时,SCM-VP开始劣于Sequence-CLDP。这表明F_1分数和内部转移概率有着密切关联。

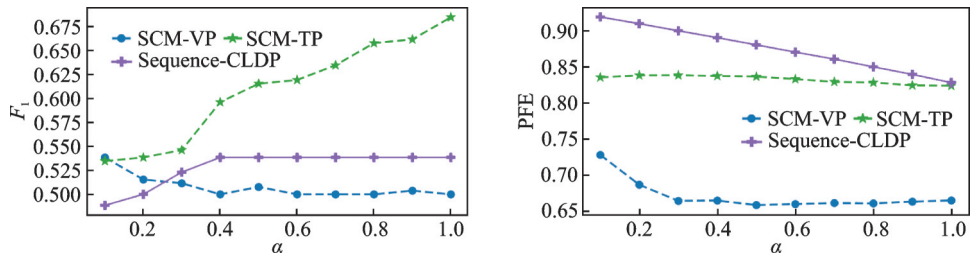

图6 MSNBC数据集上频繁模式的F_1分数和频数误差在不同α下的变化

Fig.6 Variation of F_1 score and frequence error of frequent pattern with α on the MSNBC dataset

图7展示了T-drive上的实验结果,可以发现这3个方案的F_1分数的变化趋势与前文中的实验保持一致,即随着α的增大,效果逐渐变好,且SCM-VP表现最好,Sequence-CLDP表现最差。SCM-TP的

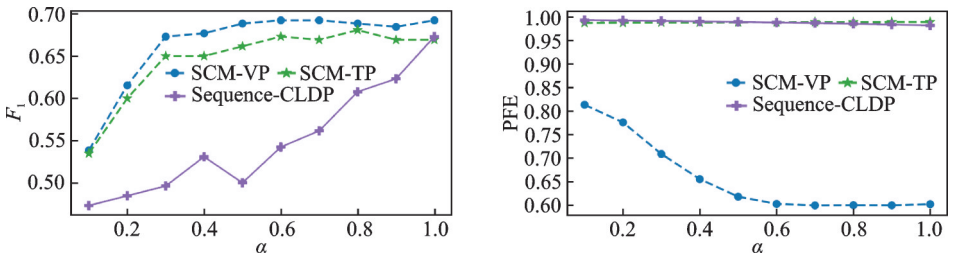

图7 T-drive数据集上频繁模式的F_1分数和频数误差在不同α下的变化

Fig.7 Variation of F_1 score and frequence error of frequent pattern with α on the T-drive dataset

PFE表现与在MSNBC数据集上的一致,即虽然能挖掘出较多的准确序列模式,但是这些序列模式的频数要远远低于真实频数,在数据集规模较小时表现得尤为明显。

5　结束语

针对目前基于压缩的本地差分隐私下的序列数据收集方法忽略序列模式,且效用性仍有提升空间的问题,本文提出了SCM-CLDP方法。根据扰动对象的不同,该方法又细分为SCM-VP和SCM-TP。其中,SCM-VP是针对值扰动的序列数据收集方法,SCM-TP是针对转移扰动的序列数据收集方法,这两种方法中都需要先估计数据值或转移分布再学习Markov模型,最终由Markov模型合成数据集。首先理论证明了SCM-VP和SCM-TP满足序列级别的CLDP,然后在两个真实数据集、三方面的评价指标上和现有方案进行了广泛且深入的实验对比,实验表明SCM-CLDP在各个方面都有着显著的优越性,且在大多数情况下,SCM-VP要优于SCM-TP。本文利用一阶Markov链模型来提取序列模式,未来希望能够通过更高阶的Markov链模型以及除序列长度外的其他重要信息来提取更丰富的序列模式,以此提升合成序列集的准确性,并探究其在各种真实场景下的应用。

参考文献:

[1] 吉根林,赵斌.时空轨迹大数据模式挖掘研究进展[J].数据采集与处理, 2015, 30(1): 47-58.
JI Genlin, ZHAO Bin. Research progress in pattern mining for big spatio-temporal trajectories[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 47-58.

[2] 张逸凡, 赵斌, 孙鸿艳, 等.基于时空轨迹的移动对象汇聚模式挖掘算法[J].数据采集与处理, 2018, 33(3): 487-495.
ZHANG Yifan, ZHAO Bin, SUN Hongyan, et al. Algorithm for mining converging patterns of moving objects from spatiotemporal trajectories[J]. Journal of Data Acquisition and Processing, 2018, 33(3): 487-495.

[3] 夏润泽,李丕绩.ChatGPT大模型技术发展与应用[J].数据采集与处理, 2023, 38(5): 1017-1034.
XIA Runze, LI Piji. Large language model ChatGPT: Evolution and application[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1017-1034.

[4] 全国人民代表大会常务委员会.中华人民共和国数据安全法[EB/OL]. (2021-09-01). http://www.gov.cn/xinwen/2021-06/11/content_5616919.htm.

[5] 全国人民代表大会常务委员会.中华人民共和国个人信息保护法[EB/OL]. (2021-11-01). http://www.gov.cn/xinwen/2021-08/20/content_5632486.htm.

[6] GURSOY M E, LIU L, TRUEX S, et al. Utility-aware synthesis of differentially private and attack-resilient location traces[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2018: 196-211.

[7] DUCHI J C, JORDAN M I, WAINWRIGHT M J. Local privacy and statistical minimax rates[C]//Proceedings of 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. [S.l.]: IEEE, 2013: 429-438.

[8] DWORK C. Differential privacy[C]//Proceedings of International Colloquium on Automata, Languages, and Programming. Berlin, Heidelberg: Springer, 2006: 1-12.

[9] ERLINGSSON Ú, PIHUR V, KOROLOVA A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2014: 1054-1067.

[10] ZDZIARSKI J. Apple's differential privacy is about collecting your data—but not your data[EB/OL]. (2016-06-14). https://www.oreilly.com/radar/apples-differential-privacy-collecting-data/.

[11] GURSOY M E, TAMERSOY A, TRUEX S, et al. Secure and utility-aware data collection with condensed local differential privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2019, 18(5): 2365-2378.

[12] MOONEY C H, RODDICK J F. Sequential pattern mining-approaches and algorithms[J]. ACM Computing Surveys (CSUR), 2013, 45(2): 1-39.

[13] CHEN R, ACS G, CASTELLUCCIA C. Differentially private sequential data publication via variable-length N-grams[C]//Proceedings of the 2012 ACM Conference on Computer and Communications Security. [S.l.]: ACM, 2012: 638-649.

[14] WANG H, ZHANG Z, WANG T, et al. PrivTrace: Differentially private trajectory synthesis by adaptive Markov models [C]//Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). [S.l.]: USENIX Association, 2023: 1649-1666.

[15] DU Y, HU Y, ZHANG Z, et al. LDPTrace: Locally differentially private trajectory synthesis[J]. Proceedings of the VLDB Endowment, 2023, 16(8): 1897-1909.

[16] XU S, CHENG X, SU S, et al. Differentially private frequent sequence mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 2910-2926.

[17] XIA H, HUANG W, XIONG Y, et al. Mining frequent sequential patterns with local differential privacy[J]. International Journal of Network Security, 2021, 23(5): 817-829.

[18] WANG N, XIAO X, YANG Y, et al. PrivTrie: Effective frequent term discovery under local differential privacy[C]//Proceedings of 2018 IEEE 34th International Conference on Data Engineering (ICDE). [S.l.]: IEEE, 2018: 821-832.

[19] 黄硕, 李艳辉, 曹建秋. 本地化差分隐私下的频繁序列模式挖掘算法PrivSPM[J]. 计算机应用, 2023, 43(7): 2057-2064.
HUANG Shuo, LI Yanhui, CAO Jianqiu. PrivSPM: Frequent sequential pattern mining algorithm under local differential privacy[J]. Journal of Computer Applications, 2023, 43(7): 2057-2064.

[20] WANG S, HUANG L, NIE Y, et al. PrivSet: Set-valued data analyses with locale differential privacy[C]//Proceedings of IEEE INFOCOM 2018 IEEE Conference on Computer Communications. [S.l.]: IEEE, 2018: 1088-1096.

[21] GUNER E, GURSOY M E. Learning Markov chain models from sequential data under local differential privacy[C]//Proceedings of European Symposium on Research in Computer Security. Cham: Springer Nature Switzerland, 2023: 359-379.

[22] LI Z, WANG T, LOPUHAÄ-ZWAKENBERG M, et al. Estimating numerical distributions under local differential privacy [C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. [S.l.]: ACM, 2020: 621-635.

[23] FANTI G, PIHUR V, ERLINGSSON Ú. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries[J]. Proceedings on Privacy Enhancing Technologies (PoPETS), 2016(3): 41-61.

[24] WANG T, LI N, JHA S. Locally differentially private frequent itemset mining[C]//Proceedings of 2018 IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2018: 127-143.

[25] QIN Z, YANG Y, YU T, et al. Heavy hitter estimation over set-valued data with local differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2016: 192-203.

[26] HECKERMAN D. MSNBC.com anonymous web data, UCI machine learning repository[EB/OL]. (1999-03-09). https://doi.org/10.24432/C5390X.

[27] YUAN J, ZHENG Y, ZHANG C, et al. T-drive: driving directions based on taxi trajectories[C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. [S.l.]: Association for Computing Machinery, 2010: 99-108.

作者简介:

金 严(2000-),女,硕士研究生,研究方向:隐私保护和数据安全,E-mail:jinyan1618@nuaa.edu.cn。

朱友文(1986-),**通信作者**,男,博士,教授,博士生导师,研究方向:应用密码学、人工智能安全、隐私计算,E-mail: zhuyw@nuaa.edu.cn。

吴启晖(1970-),男,博士,教授,博士生导师,研究方向:认知信息论、电磁空间频谱智能管控、天地一体化信息网络、无人机集群智能通信和安全等。

(编辑:刘彦东)