基于差分隐私的海量数据发布方法研究

颜 飞 张 兴* 李 畅 李万杰 李 帅
(辽宁工业大学电子与信息工程学院 辽宁 锦州 121001)

摘 要 海量静态数据直方图发布过程中分组划分存在离群点,导致误差增大和离群点判定效率低的问题。对此提出一种适用于 Spark 框架的满足 ε-差分隐私保护的海量静态数据直方图发布方法。对 k-means 聚类算法进行避免距离重复计算的优化改进;利用改进后的 k-means 聚类算法进行直方图最优分组划分,实现快速聚合相似分组,形成最优分组融合;对分组结果添加噪声处理,并将经过差分隐私保护处理后的数据进行发布。利用实际数据进行仿真实验,结果表明,所提方法在海量静态数据集隐私保护处理中可提高发布效率和保证数据隐私安全性,同时保证发布数据具有较好的可用性。

关键词 差分隐私 分组融合 噪声干扰 数据发布

中图分类号 TP393.08 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.11.053

MASSIVE DATA PUBLISHING METHOD BASED ON DIFFERENTIAL PRIVACY

Yan Fei Zhang Xing* Li Chang Li Wanjie Li Shuai
(*School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, Liaoning, China*)

Abstract There are outliers in group partition in the process of massive static data histogram publishing, which may lead to increased errors and low efficiency of outlier decision. To solve this problem, we presented a histogram publishing method for massive static data satisfying differential privacy protection for Spark framework. k-means clustering algorithm was optimized to avoid distance duplication calculation. The improved k-means clustering algorithm was used to partition the histogram into the best grouping, and the similar grouping was quickly aggregated to form the optimal grouping fusion. We added noise to grouping results, and published data with differential privacy protection. The simulation experiment was carried out by real data. The results show that the proposed method can improve the publishing efficiency and ensure the data privacy security in the privacy protection processing of massive static data sets. It can also ensure the availability of publishing data.

Keywords Differential privacy Grouping fusion Noise interference Data publishing

0 引 言

随着信息技术,特别是大数据技术和人工智能领域研究的飞速发展,海量数据的收集、存储、发布和分析变得越来越容易。但从数据安全和个人隐私保护层面来看,大数据应用也带来了很大的数据安全隐患。而且数据的安全和隐私数据的泄露不仅会影响到个人利益,甚至会威胁到国家的网络空间安全。面对如此复杂的大数据背景,大数据面临着诸多安全问题,其中如何从大数据中分析挖掘出更多的价值而又很好地保护数据的隐私安全显得尤为重要[1]。

从个人隐私安全层面来看,一旦隐私信息被泄露,用户个人隐私无异于“裸奔”。对于企业来说,确保用户数据安全和隐私安全是必须面对和解决的问题,若数据安全和隐私保障存在问题,将会影响大数据和人工智能的进一步推广应用。在未来发展中,如果国家在数据安全控制方面失去了主动权,那么必将受制于

收稿日期:2018 - 03 - 01。辽宁省高等学校杰出青年学者成长计划项目(LJQ2014066);辽宁省自然科学基金项目(20170540434)。**颜飞**,硕士生,主研领域:大数据安全,隐私保护。**张兴**,教授。**李畅**,硕士生。**李万杰**,硕士生。**李帅**,硕士生。

他人。因此,确保大数据安全和隐私安全十分重要,针对大数据安全和隐私保护的关键技术的研究值得更进一步探索,而且大数据安全和隐私保护也将逐渐上升至国家战略层面。

为了有效地保护个人隐私安全,研究人员提出了许多隐私保护模型,例如基于匿名技术的 K-anonymity[2]、L-diversity[3]、M-invariance[4]、T-closeness[5] 等。由于以匿名为基础的隐私保护模型均需特殊的攻击假设和一定的背景知识,且未能对隐私保护强度进行量化分析,因此在实际应用中具有较大的局限性。尤其是在海量数据的背景下,用户的原始信息可能在经过数据挖掘分析和深度学习的某个过程中被非法者破坏、攻击和篡改,用户信息的隐私安全面临着严重的威胁[6]。因此,差分隐私[7]作为一种新型、轻量级的隐私保护算法,引起了研究人员的关注。它通过对发布数据进行随意扰动,使得在传统意义上无论攻击者具有何种背景知识都无法识别一条记录是否在原数据表中[8],可以解决数据发布所隐藏的潜在隐私威胁。但针对海量数据的隐私保护数据发布往往会出现数据敏感度增大、隐私预算枯竭和数据噪声过大等问题,对后期的数据分析造成较严重影响。并且在采用直方图发布方法存在离群点导致数据高敏感问题,更容易泄露隐私。因此,针对离群点的分组划分问题,文献[9]在等宽划分的基础上提出了一个差值集的概念来处理在分组时由于离群点存在可能会导致的划分误差。该方法在面对充满离群点的数据集时有着很好的表现,但在分组划分阶段需计算所有数据的差值集,对于大数据集来说差值集计算效率问题成为必须解决的问题。

研究人员在提高大数据集的计算效率方面也做了许多研究工作。文献[10]采用高效的 MapReduce 并行计算模型实现了 k-means 聚类算法,有效提高了 k-means 算法的运行效率。针对云平台的开放性使攻击者拥有大量的攻击背景知识[11],攻击者可以通过关联背景知识和聚类结果来窃取数据隐私[12-13]。文献[14-15]将隐私保护机制融入 Hadoop 平台下的 MapReduce 分布式计算框架,实现了海量分布式数据的隐私保护算法。华为研究人员为满足数据挖掘需求,实施部署了满足差分隐私保护的大数据分析平台[16]。

虽然 MapReduce 分布式计算框架的使用可高效处理海量数据,但该框架在算法迭代过程需多次读写硬盘数据,消耗大量 I/O 通信资源,并且过多的噪声扰动也会增加隐私保护算法的复杂度开销[17]。针对以上分析,为了提高数据隐私保护程度和数据的可用性,解决差值集计算效率问题,本文以海量静态数据的发布需求为出发点,提出一种满足 ε-差分隐私保护的适用于 Spark 内存迭代的 SPDP-GS(Spark Differential privacy-Grouping Smothing)算法。该方法可提高离群点判断速度和差值集计算效率,并有效控制基于直方图的数据发布方法中的离群点对数据发布的敏感度的影响,具有一定的应用价值和理论研究意义。

1　差分隐私保护理论

差分隐私保护主要通过对发布数据进行随意扰动,使得攻击者使用传统方法攻击时,无论拥有何种背景知识均无法轻易识别出某条记录是否一定在原数据表中。

1.1　差分隐私定义

定义 1　对于给定的 2 个至多相差 1 条记录的数据集 D_1 以及 D_2,f 为随机算法,$range(f)$ 表示算法 f 的所有输出构成的集合,S 为 $range(f)$ 的子集。若算法 f 满足 $\Pr[f(D_1)\in S]\leqslant e^{\varepsilon}\times\Pr[f(D_2)\in S]$,则算法 f 具有 ε-差分隐私性。

其中,ε 为隐私保护预算,代表算法的隐私保护水平,ε 的取值越小,隐私保护水平越高。

1.2　Laplace 噪声机制

差分隐私保护的实现机制是采用数据扰动,数据扰动常用方法之一是采用 Laplace[19] 噪声机制来实现数据加噪,该机制使用拉普拉斯分布所产生噪声添加到真实输出值中来实现差分隐私保护。

定义 2　对于任意一个函数 $f:D\rightarrow R^d$,算法 Y 满足 $Y(D)=f(D)+<Lap_1(\Delta f/\varepsilon),Lap_2(\Delta f/\varepsilon),\cdots,Lap_d(\Delta f/\varepsilon)>$。

其中,函数 $Lap_i(\Delta f/\varepsilon)(1\leqslant i\leqslant d)$ 表示拉普拉斯密度函数;$\Delta f=\max_{D_1,D_1}|f(D_1)-f(D_2)|$ 为函数 $f(D)$ 的查询敏感度。D_1、D_2 为兄弟数据集;d 为查询维度。

1.3　差分隐私组合特性

在差分隐私保护研究中,为证明算法满足差分隐私,需满足如下差分隐私组合特性:序列组合性和并列组合性。

性质 1[19]　给定数据库 D 与 n 个随机算法 f_i,且 f_i 满足 ε_i-差分隐私,那么 $f_i(D)$ 序列组合满足 ε-差分隐私,且 $\varepsilon=\sum\varepsilon_i$。

性质 2[19]　设将给定数据库 D 划分成 n 个不相交的子集,$D=\{D_1,D_2,\cdots,D_n\}$,若任意算法 f_i 满足 ε-差分隐私,则序列 f_i 在 D 上的操作结果仍满足 ε-差分隐私。

2 Spark 框架下 SPDP-GS 直方图数据发布方法

为了提高对隐私数据的保护程度和挖掘结果的可用性,解决文献[9]中差值集计算效率问题,以海量数据的统计特征为出发点,提出一种适用于 Spark 框架的满足 ε-差分隐私保护的海量静态数据直方图发布方法。

2.1 数据发布处理框架

本文提出一个满足差分隐私保护需求的非交互式计算框架系统,其结构示意如图 1 所示。该框架主要由 3 个部分组成:原始数据收集和存储,Spark 框架下的数据处理和存储,数据的隐私保护处理。

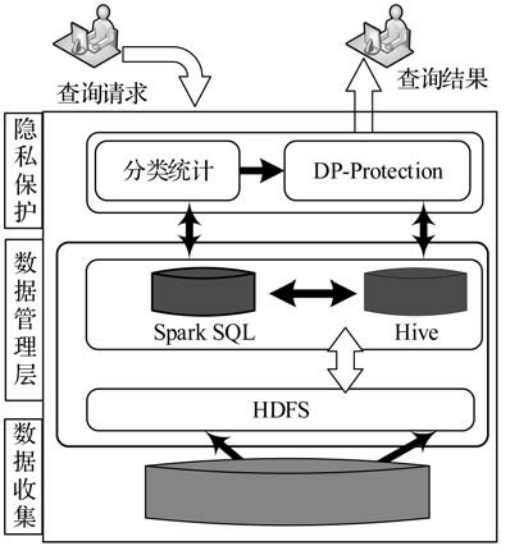


图 1 Spark 框架下差分隐私保护模型

数据管理层,首先将原始数据集导入 HDFS 进行数据的管理,然后数据从 HDFS 读取到 Spark 框架形成 RDD 数据集,并进行 map 操作、执行 join 操作和 Shuffle 过程,最后将 RDD 处理结果输出并保存到 HDFS。

隐私处理,对待发布数据的隐私保护处理主要是借助 Spark 并行计算框架对经过预处理的数据进行分类统计、特征提取和聚类分组等计算任务,并对分组进行添加 Laplace 噪声。

2.2 Spark 框架下 k-means 并行化改进

k-means 聚类算法是通过计算每个样本到每个中心的欧氏距离,并选择距离最近的中心,将样本进行归类。如此会造成很大的计算量,尤其是对于分布式计算而言,由于样本数据存放在不同的节点上,更是会带来大量的通信开销。所以本文引用了文献[20]的距离优化方法,采用将样本数据与其二范数进行关联的优化措施,避免了距离的重复计算,降低了 k-means 聚类过程的计算开销。就是将数据点的坐标(x,y)与其二范式进行关联,构成 $<(x,y),\|(x,y)\|^2>$ 的键值对形式,将二范数之差的平方值(*boundDistance*)与最近的中心点的距离(*bestDistance*)进行比较。若前者小于后者,则进行更新替代操作。k-means 主要步骤如下:

1)初始化 k 个初始聚类中心,形成样本聚类。

2)遍历数据样本,若 *boundDistance* < *bestDistance*,则进行真正的欧氏距离计算(*realDistance*),若 *realDistance* < *bestDistance*,将距离最小的归类到聚类中心形成 k 个聚类。

3)计算各聚类内数据均值,更新聚类中心。

4)循环 Step1 – Step3,直到达到指定迭代次数或聚类收敛聚类中心不再变化。

5)输出聚类处理结果。

2.3 SPDP-GS 算法设计

在大数据应用背景下,基于 Spark 框架的差分隐私保护直方图发布方法主要目的在于满足海量数据计算效率的要求下,提供有效的隐私保护方法。对于满足差分隐私保护的直方图发布方法,文献[21]通过对数据集进行排序、分组以及求各分组均值,再添加 Laplace 噪声。但是在可能会存在大量离群点数据集时,会导致隐私泄露,而且简单的等宽分组方法容易导致误差增大问题。所以,文献[9]在文献[21]的基础之上提出了采用插值集概念处理分组过程中由于离散点问题而导致的划分误差问题。但该方法对于海量数据集的处理来说差值集的计算量巨大。

因此,本文提出了借助 Spark 平台采用 k-means 改进算法对分组进行最优划分,对每个分组求均值,再在各分组的平均数上添加 Laplace 噪声,对隐私算法保护处理后的数据进行发布。基于 Spark 框架的 SPDP-GS 算法主要由:统计分类和 DP-protection 两部分,主要步骤描述如下:

1)采用 Hash_map 按属性进行分类统计。

2)对各类内部求均值:$avg = \dfrac{\sum_{i=1}^{n} x_i}{n}$。

3)采用 Laplace 机制添加噪声:$Y(D) = f(D) + lap(\Delta f/\varepsilon)$。

4)对待发布数据进行直方图发布。

2.4 SPDP-GS 算法描述

本节描述 SPDP-GS 算法包括:数据的类型统计、k-means 聚类分组、分组求均值和添加 Laplace 噪声,具体过程如下所述:

算法 1 k-means 聚类分组划分算法

输入:经 Hash_map 算法统计分类后数据集 $D\{x_1,x_2,\cdots,x_n\}$,聚类簇数 k。

输出:聚类分组 $C=\{c_1,c_2,\cdots,c_k\}$,组内均值 u_{c_j},组内数据数量 num_{c_j}。

```
1) KMeansCluster(hashmapResult)
2) { Kmeans.setMax(k);                    //设定聚类中心数 k
3)
   sourcedata = kmeans.loadData(hashmapresult);
                                          //读取统计分类结果数据集
4)  for i = 1 to n
5)  { for j = 1 to k
6)    {
7)        if (boundDistance < bestDistance & realDistance <
bestDistance)
8)     then bestDistance←realDistance;
9)      else realDistance_ij = ||x_i − u_j||_2;
                                          //计算 x_i 与各均值向量 u_j 距离
10)      c_j = arg min_i(realDistance_ij);
                                          //将对应值加入相应簇
11)      u_cj = 1/|c_j| Σ_{x∈c_j} x;      //更新均值向量
12) num_cj = num_cj + 1;                  //类内数据个数统计
13)          }
14)       }
15)     result:RDD[(int, C_k)]; //聚类结果存入 RDD
16)  输出:组内均值 u_cj,组内数据数量 num_cj
17)  }
```

算法 2　差分隐私直方图发布

输入:聚类分组 $C=\{c_1,c_2,\cdots,c_k\}$,查询任务 Q,查询敏感度 $\Delta f(Q)$。

输出:满足 ε-差分隐私的数据集 D_ε。

1) 依据查询任务 $f(Q)$ 返回分组 C 中对应查询记录;

2) 对查询结论添加 Laplace 噪声后得到 C'_k

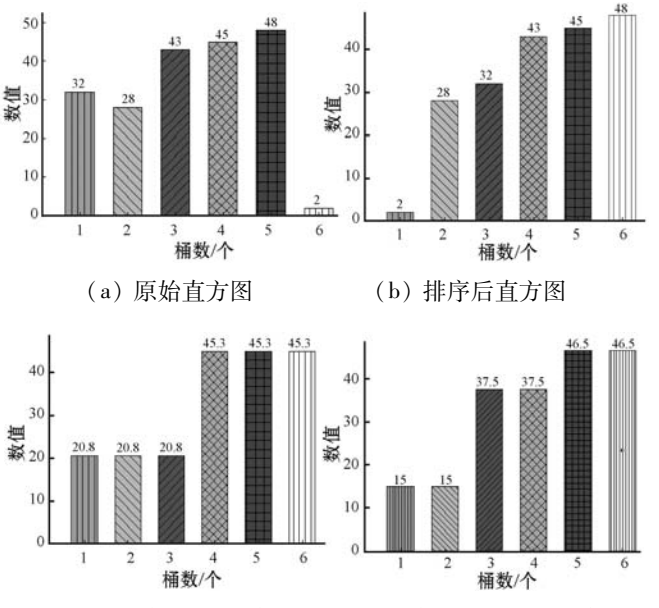
$C'_k=C_k+<Lap_1(\Delta f/\varepsilon),Lap_2(\Delta f/\varepsilon),\cdots,Lap_{d_k}(1/\varepsilon)>$;

3) 发布满足 ε-差分隐私的数据 $D_\varepsilon=\{c'_1c'_2,\cdots,c'_k\}$

2.5　SPDP-GS 数据发布方法

针对待发布数据集可能存在大量离群点导致隐私泄露风险增大和海量数据集的差值计算效率低的问题,满足差分隐私保护的海量数据发布成为本文研究的着眼点。通常,数据集中难以避免地存在一些离群点,离群点的存在可能诱发隐私泄露和误差增大问题。

例如,某疾病监控中心,需周期性更新某些疾病确诊患者,而所发布数据又不能泄露确诊患者年龄、住址等隐私信息。因此,可采用差分隐私保护方法对发布数据进行处理,数据隐私保护后再发布。实例具体说明如下:若将图 2(a)所示数据直接发布,拥有相关背景知识的人很容易推断离散点数据的隐私信息。采用文献[21]和本文所提出的方法将数据集 $D=\{32,28,43,45,48,2\}$ 进行分组划分,可有效解决隐私泄露问题。



(a) 原始直方图　　(b) 排序后直方图

(c) 采用 GS 划分　　(d) 采用 SPDP-GS 划分

图 2　不同方法所得直方图划分图

若采用文献[21]方法将直方图分为两组进行等宽划分:首先需对直方图进行递增排序得:$D'=\{2,28,32,43,45,48\}$,如图 2(b)所示;然后,将直方图分组划分得 $GS_1=\{2,28,32\}$、$GS_2=\{43,45,48\}$。图 2(c)为分组划分后的直方图,在分组合并过程中,容易引起误差增大的问题,导致数据可用性降低;若采用文献[9]所述 S-GS 方法对分组划分采用差值计算。利用差值计算所形成的差值集,将离散点与相邻分组合并,可进一步控制误差增大问题,但对于海量数据集的发布存在差值集计算效率低和分组划分不合理的问题。若在实际应用中,需要进一步改进。因此,本文借助 Spark 平台采用聚类方法将相似分组进行聚类划分,并采用就近原则将离群点数据与近似分组融合。如图 2(d)所示,采用本文所述方法首先对原分组进行分组划分:$D'_{dp1}=\{2,28\}$,$D'_{dp2}=\{32,43\}$,$D'_{dp3}=\{45,48\}$。对于海量数据来说,可实现快速聚合相似分组,达到最优分组融合。然后,对分组数据进行差分隐私保护处理,提高发布数据的隐私性,同时保证发布数据具有较好的可用性。

若采用和方差 $\{SSE(x)=\sum_{i=0}^{n}(x_i-x'_i)^2$,其中 x_i 为递增排序后数据;x'_i 为分组划分后失真数据$\}$对分组误差进行度量,由图 2 可得:$SSE(b)\approx 543.4$、$SSE(c)=403$、$SSE(d)=342.5$,若仅对此例进行和方差计算可得:本文所述方法可在一定程度上降低了发布数据的误差。

2.6 数据隐私安全性分析

2.6.1 隐私验证

本文所述算法的隐私性主要从算法满足 ε-差分隐私的定义和性质角度加以论证。由于噪声添加在分组划分后的各分组之中，所以主要证明直方图发布算法是否满足 ε-差分隐私。

定理 1 算法 SPDP-GS 满足 ε-差分隐私。

证明：由算法 1 中分组策略和噪声添加方法可知，每次滑动窗口的经过将会产生 d 个分组，而每分组所分得的隐私预算为 d/ε。差分隐私方法中敏感度 Q 设置为 1，即 $\Delta(Q)=1$。假设数据集 D_1 和 D_2 最多相差一条记录，即 $|D_1-D_2|\leqslant 1$，$|D_2-D_1|\leqslant 1$。由定义 1 可知，$\Pr[f(D_1=D')]\leqslant e^{\varepsilon}\times\Pr[f(D_2=D')]$。由性质 2 可知该直方图发布算法满足 ε-差分隐私。

2.6.2 数据可用性分析

本节所述算法采用聚类方法对数据进行聚类分组，将相似数组（相似分组指的是直方图数值相近的若干个桶）划分在一个分组内，并对同一分组内的数据以平均值表示。因此，发布数据会产生两种误差：一是由各分组均值产生的近似误差 SSE（Sum Squared Error）；二是因添加拉普拉斯噪声而产生的误差。

定理 2 给定一个分组 $D_{\varepsilon}=\{C'_1,C'_2,\cdots,C'_k\}$，使用 SPDP-GS 方法对动态数据统计直方图发布实现 ε-差分隐私，则该分组所形成的总体误差 $Err(C'_i)$ 如下：

$$Err(C'_i)=SSE(C'_i)+Lap(d/\varepsilon) \tag{1}$$

式中：$SSE(C'_i)=\sum_{l=i}^{j}(x_l)^2-\frac{\left(\sum_{l=i}^{j}x_l\right)^2}{j-i+1}$，$\overline{C'_i}=\frac{\sum_{l=i}^{j}x_j}{j-i+1}$，$Lap(d/\varepsilon)$ 为每分组所添加的 Laplace 噪音。

证明：由和方差度量定义可知，$SSE(C'_i)=\sum_{x_j}(x_j-\overline{C'_i})^2$，其中 $\overline{C'_i}=\sum_{x_j}\frac{x_j}{|C_j|}$。由统计学样本方差公式可计算得：$SSE(C'_i)=\sum_{l=i}^{j}(x_l)^2-\frac{\left(\sum_{l=i}^{j}x_l\right)^2}{j-i+1}$，其中 x_l 表示第 i 个桶所对应的数值，i 表示第 i 个分组。

3 实验设计与分析

3.1 实验环境

本文实验对算法的运行效率以及隐私保护数据的结果可用性进行考虑。实验选取 3 台主机搭建 Spark 平台，每台机器均为双核 IntelCorei3 处理器，4 GB 内存，操作系统选用 Ubuntu，hadoop-2.7.2 和 Spark 2.2.0；JDK 的版本是 1.8.0_121，Scala-2.12.3。

本实验所用数据集来自“Kaggle：The Home of Data Science”网站所以提供的 Transactions 商场交易数据，包括商品类型、品牌、交易日期、采购量和交易金额；另一个数据集为 US Census 1990 raw data，该数据集包含了来自 1990 年美国人口普查数据（PUMS）1% 的样本，详细信息见表 1。

表 1 实验数据集

名称	大小/GB	记录数/个	属性类型数/个
Transactions	19.6	349 655 790	836
US-Census	823	2 458 285	68

3.2 数据预处理

本实验选取交易数据集的 category（商品类型）和 US Census 的 age 属性作为数据处理对象。对数据集中 category 和 age 字段的各种商品类型进行统计。但 category 字段的值不应该为 0 值，因此需在数据统计过程中对取值为 0 的记录予以清除，不纳入统计。

3.3 数据可用性度量

本文选取选取交易数据集的 category 属性和 US Census 的 age 属性作为敏感属性进行数据组划分。主要采用和方差[23]和绝对误差（AE）两种评估标准度量算法的可用性。表达式如下：

$$AE=|C'_i-true| \tag{2}$$

式中：C'_i 表示添加 Laplace 噪声后的发布数据，$true$ 表示分组的真实值。

首先，本实验选取 Transactions 数据集，对隐私预算 ε 所产生的数据可用性的影响展开研究。实验过程中分别取隐私预算参数 ε 为 0.5、0.75、1 和 1.5。

图 3 给出了隐私预算 ε 变化下绝对误差的变化趋势。结果表明，算法绝对误差随着隐私预算 ε 的增大而减少。而且，本文所述方法的隐私保护效果和数据可用性上相较于 GS 方法和 S-GS 方法更优。

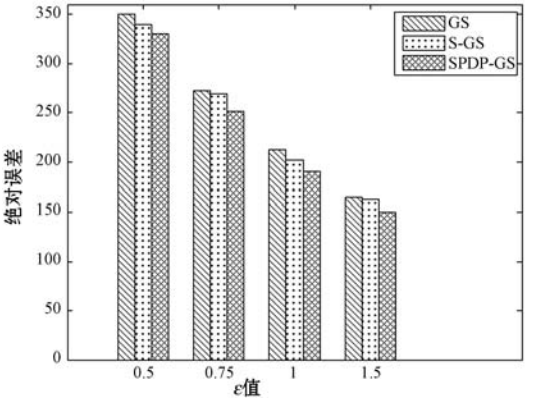


图 3 不同 ε 值下的绝对误差

接下来分别在 Transactions 和 US-Census 数据集上,通过改变离群点数量 *num* 来对 *AE* 结果进行研究,从而判断本文所述方法与文献[9,21]所述方法在数据发布结果可用性上的优劣情况。

实验过程中 ε 取值设置为 1.5,实验结果如图 4 和图 5 所示。在 Transactions 和 US-Census 数据集上,存在随着离群点的个数的增加,导致发布结果的绝对误差增大的现象。由绝对误差计算公式可得,离群点的数量的增多会导致分组划分的误差增大。本文方法比文献[9,21]所述方法表现更好的主要原因为采用固定的分组划分不可避免的出现离群点分组划分不合理的问题,从而导致分组过程中误差增大的问题。

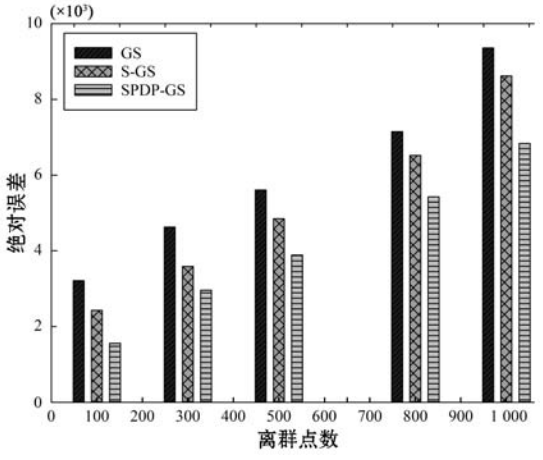


图 4　数据集 Transactions 下不同离群点的绝对误差

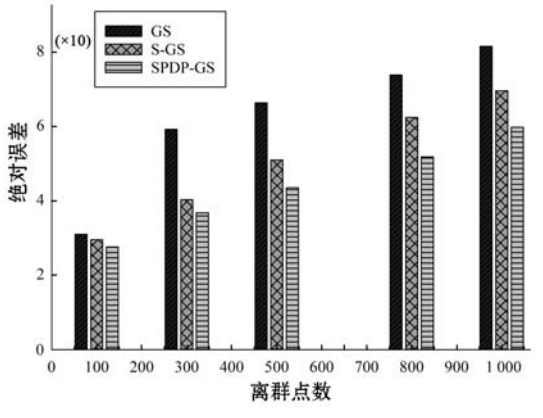


图 5　数据集 US-Census 下不同离群点的绝对误差

3.4　算法时间效率分析

本文采用了对原始数据集进行 Hash_map 算法进行分类统计,将统计结果进行数据对外发布。而数据发布之前,对其中离散点的处理采用 k-means 聚类方法进行合理组聚类,对同一类内数据进行求均值,从而减少 S-GS 方法在大数据背景下的差值集计算量巨大的问题。

每组实验分别做了 5 次测试,取 5 次平均时间作为最终结果,如图 6 所示。

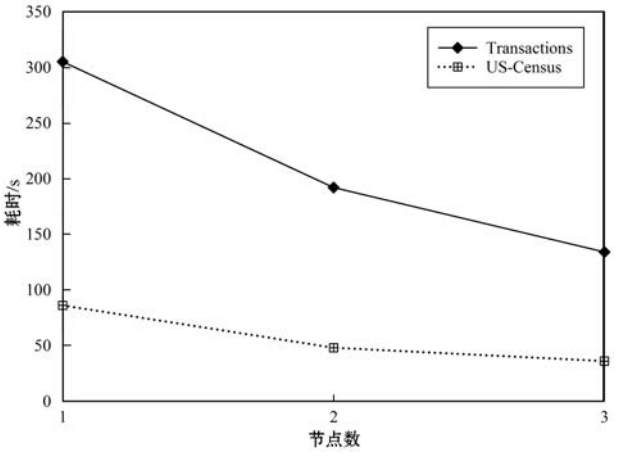


图 6　算法时间效率

由图 6 可知,Spark 平台参与运算的子节点数量越多,算法执行时间显著减少,说明 Spark 平台可以较好地解决大数据的运行效率问题。采用 Spark 平台进行差分隐私数据发布可有效保证发布数据的隐私安全及运算效率。

4　结　语

为了提高隐私数据的保护程度和保证数据挖掘结果的可用性,解决海量静态数据直方图发布过程中差值集计算效率低、存在隐私泄露安全隐患问题,研究了大数据背景下的差分隐私保护数据发布方法,提出一种 Spark 框架下的满足差分隐私保护的直方图数据发布方法。本文借助 Spark 计算平台实现对海量数据的分类统计、聚类分析和分析结果的差分隐私保护。文中给出了数据处理的计算框架,并对各部分做了简要阐述。通过对实验结果中的总体误差和隐私预算 ε 进行分析,相较于 GS 方法和 S-GS 方法数据可用性上更佳,而且解决了 S-GS 方法在海量数据计算中的差值集计算问题,满足数据隐私安全性需求,同时保证发布数据具有较好的可用性,具有一定的应用价值。

参　考　文　献

[1] 尹浩, 乔波. 大数据驱动的网络信息平面[J]. 计算机学报, 2016, 39(1): 126－139.

[2] Sweeney L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557－570.

[3] Machanavajjhala A, Gehrke J, Kifer D, et al. L-diversity: privacy beyond k-anonymity[C]//International Conference on Data Engineering. IEEE, 2006:24－24.

[4] Xiao X, Tao Y. M-invariance:towards privacy preserving republication of dynamic datasets[C]//ACM SIGMOD Inter-

national Conference on Management of Data, Beijing, China, June. DBLP, 2007: 689 - 700.
[5] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity[C]//2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007: 106 - 115.
[6] 孟小峰, 张啸剑. 大数据隐私管理[J]. 计算机研究与发展, 2015, 52(2): 265 - 281.
[7] Dwork C. Differential Privacy[C]//International Colloquium on Automata, Languages, and Programming. Springer, Berlin, Heidelberg, 2006:1 - 12.
[8] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014,37(4): 927 - 949.
[9] 邵荣敏, 张琳, 王汝传. 一种 D-ProPer 保护框架下的差分隐私数据发布方法[J]. 南京邮电大学学报(自然科学版), 2016, 36(5): 96 - 104.
[10] 李洪成, 吴晓平, 陈燕. MapReduce 框架下支持差分隐私保护的 k-means 聚类方法[J]. 通信学报, 2016, 37(2): 124 - 130.
[11] 肖人毅. 云计算中数据隐私保护研究进展[J]. 通信学报, 2014, 35(12): 168 - 177.
[12] Hua J, Gao Y, Zhong S. Differentially private publication of general time-serial trajectory data[C]//2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015: 549 - 557.
[13] Maruseac M, Ghinita G. Privacy-Preserving Mining of Sequential Association Rules from Provenance Workflows [C]//ACM Conference on Data and Application Security and Privacy. ACM, 2016: 127 - 129.
[14] Kenekar T V, Dani A R. An efficient private FIM on hadoop MapReduce [C]//International Conference on Automatic Control and Dynamic Optimization Techniques. IEEE,2017.
[15] Yin C, Zhang S, Xi J, et al. An improved anonymity model for big data security based on clustering algorithm[J]. Concurrency & Computation Practice & Experience, 2016, 29 (7).
[16] Hu X, Yuan M, Yao J, et al. Differential Privacy in Telco Big Data Platform[J]. Proceedings of the VLDB Endowment, 2015, 8(12): 1692 - 1703.
[17] 高志强,李庆鹏,胡人远. 基于 Spark 的支持隐私保护的聚类算法[J]. 网络与信息安全学报,2016,2(11):47 - 51.
[18] Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Springer, Berlin, Heidelberg, 2006:265 - 284.
[19] Mcsherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[J]. Communications of the Acm, 2010, 53(9):89 - 97.
[20] 朱光辉, 黄圣彬, 袁春风,等. SCoS:基于 Spark 的并行谱聚类算法设计与实现[J]. 计算机学报, 2018,41(4): 868 - 885.
[21] Ebadi H, Sands D, Schneider G. Differential Privacy: Now it's Getting Personal[J]. Acm Sigplan Notices, 2015, 50 (1):69 - 81.
[22] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication[J]. The VLDB Journal, 2013, 22(6): 797 - 822.
[23] 张啸剑, 孟小峰. 基于差分隐私的流式直方图发布方法[J]. 软件学报, 2016, 27(2):381 - 393.

(上接第 300 页)

[2] Lin Y, Yang Y. Stock markets forecasting based on fuzzy time series model[C]//IEEE International Conference on Intelligent Computing and Intelligent Systems. IEEE, 2009: 782 - 786.
[3] Sharma A, Mansotra V, Shastri S. Forecasting Public Healthcare Services in Jammu & Kashmir Using Time Series Data Mining[J]. International Journal of Computer Science & Software Engineering, 2015, 5(12):569 - 575.
[4] Wah W, Das S, Earnest A, et al. Time series analysis of demographic and temporal trends of tuberculosis in Singapore [J]. BMC Public Health, 2014, 14:1121.
[5] Medina D C, Findley S E, Guindo B, et al. Forecasting Non-Stationary Diarrhea, Acute Respiratory Infection, and Malaria Time-Series in Niono, Mali[J]. Plos One, 2007, 2 (11):e1181.
[6] 张美英,何杰. 时间序列预测模型研究综述[J]. 数学的实践与认识,2011,41(18):189 - 195.
[7] Box G E P, Jenkins G. Time Series Analysis Forecasting and Control [M]//Time Series Analysis, Forecasting and Control. Holden-Day, 1976.
[8] 朱东妹. 时间序列数据预测方法的应用研究[J]. 兰台世界,2011(10):73 - 74.
[9] 周奎. ARIMA 模型在我国 GDP 预测中的应用[J]. 广西职业技术学院学报,2016,9(1):19 - 23.
[10] Rahman A, Hasan M M. Modeling and Forecasting of Carbon Dioxide Emissions in Bangladesh Using Autoregressive Integrated Moving Average (ARIMA) Models[J]. Open Journal of Statistics, 2017, 7(4): 560 - 566.
[11] 张楠. 时间序列预测法简介[J]. 经济与管理研究,1981 (4):41 - 44,59.
[12] 周雄鹏. 自适应过滤法的原理、程序和应用[J]. 预测, 1985(2):36 - 41.
[13] 王坚强. 自适应过滤法在经济预测中的应用[J]. 工业技术经济,1996,15(4):88 - 91.
[14] 陶庭叶,高飞,吴兆福. 自适应过滤法及其在大坝监测中的应用[J]. 测绘科学,2009,34(5):181 - 182.