文章编号:1674-2974(2020)02-0125-09 DOI:10.16339/j.cnki.hdxbzkb.2020.02.017

基于经验小波变换的基因关联隐私保护实验研究

陈红松[1†],孟彩霞[2],刘书雨[1]
(1. 北京科技大学 计算机与通信工程学院,北京 100083; 2. 铁道警察学院,河南 郑州 450053)

摘 要:为了解决某类风湿性关节炎与致病基因单核苷酸多态性(Single-Nucleotide Polymorphism,SNP)的相关度研究中,针对病人隐私保护强度与数据可用性的权衡问题,提出一种新型的基于经验小波变换(Empirical Wavelet Transform,EWT)的隐私保护方法. 该方法通过对差分隐私加噪机制产生的数据进行 EWT 变换和分解,然后计算各 EWT 分量的峭度值并筛选出可能的噪声分量,去除一定的噪声分量后对信号进行重构得到新数据,基于该数据进行致病基因相关度排序. 实验结果表明使用该方法能在保证差分隐私保护强度的情况下提高数据可用性,实现了隐私保护强度与数据可用性的合理权衡.

关键词:隐私保护;经验小波变换;差分隐私;相关度;数据可用性
中图分类号:TP309 **文献标志码**:A

Privacy Protection Experimental Research on Genes Association Ranking Based on Empirical Wavelet Transform

CHEN Hongsong[1†],MENG Caixia[2],LIU Shuyu[1]
(1. School of Computer and Communication Engineering,University of Science and Technology Beijing,Beijing 100083,China;
2. Railway Police College,Zhengzhou 450053,China)

Abstract:Due to privacy concerns in the genome-wide association studies of rheumatoid arthritis,there has been applying differential privacy to protect phenotype information (disease status) from being leaked while returning highly associated SNP(Single-Nucleotide Polymorphism). The trade-off between privacy protection intensity and data availability is a great problem. In order to solve the problem,a novel differential privacy protection method based on EWT (Empirical Wavelet Transform) was proposed. This method achieved the balance between privacy protection intensity and data availability by processing the noise introduced by differential privacy. Firstly,the data with differential privacy noise mechanism was processed by EWT approach; secondly,the kurtosis values of each EWT component were calculated,then some account of noise components was filtered out. At last,the data was reconstructed. After the above steps,the new data was obtained; it would be sorted according to the correlation degree of pathogenic genes. The experimental results show that the novel method can improve the data availability while ensuring the differential privacy protection intensity,and achieve a reasonable trade-off between the privacy protection intensity and the data availability.

Key words:privacy protection;Empirical Wavelet Transform (EWT);differential privacy;association degree;data availability

* **收稿日期**:2019-05-16
基金项目:国家社会科学基金资助项目(18BGJ071),National Social Science Foundation of China(18BGJ071)
作者简介:陈红松(1977—),男,山东济宁人,北京科技大学教授,博士
† 通讯联系人,E-mail:chenhs@ustb.edu.cn

致病基因关联分析是全基因组关联研究[1](Genome-wide Association Studies,GWAS)中的一项分析 DNA 序列集以发现疾病遗传基础的流行方法,这项研究主要检查特定患者群体的基因中数千个单核苷酸多态性位点(SNP)与疾病之间的相关度,对 SNP 进行评分,并根据这个评分对相关度较高的 SNP 排序. 但对于 GWAS 发布的数据而言,即使是只发布统计数据,患者的疾病状态也可以从每个 SNP 与疾病相关联的统计检验中推断出来,这使得患者的隐私面临着泄露的风险.

目前已有许多研究人员研究使用差分隐私技术来解决这一问题,差分隐私保护技术是当前数据发布中最主要的隐私保护方法,它通过向查询数据中添加噪声来干扰攻击者泄露原始数据的目的,从而达到隐私保护效果. 差分隐私保护技术的应用使得数据发布的效率得到了很大的提高,但为了满足差分隐私保护要求需要注入过高的噪声,影响数据的正确性和可用性,最终导致数据效用降低. 为了解决这一问题,本文提出了一种基于 EWT 变换的差分隐私保护方法,不仅依赖于注入噪声,还通过适当过滤部分噪声实现隐私保护与数据可用性的合理折中,由于只是针对噪声的注入、变换和过滤,所以不会还原出用户隐私信息. 主要研究目的是在致病基因相关度研究中,使用差分隐私保护患者隐私的同时,降低由于添加差分隐私噪声带来的误差.

1 相关技术

1.1 差分隐私

1.1.1 定 义

差分隐私的主要思想是给数据集中的每条记录都添加一个噪声,使在一个数据集上计算的给定统计量的结果类似于在另一个任意的数据集上计算的相同的统计量,以此来把数据泄露的概率控制在一定的范围内,从而达到隐私保护的目的. 满足以上两个数据集中最多只有一条记录不同,即如果一个数据库是另一个数据库的正确子集,那么较大的数据库只比另一个多包含一行数据.

本文专注于保护表型数据,因此对差分隐私的定义进行略微修改,如定义 1.

定义 1[1] 设 F 是一个随机函数,它接受一个 $n\times m$ 的基因型矩阵 $\boldsymbol{D}$ 和一个 n 维表型向量 $\boldsymbol{y}$,并输出结果 $F(D,y)$,Ω 表示随机函数 F 的输出范围,那么随机函数 F 对于任意的 $\varepsilon>0$,满足 ε-表型差分隐私. 对于任意的基因型矩阵 $\boldsymbol{D}$,任意的表型向量 $\boldsymbol{y}$,$y'\in\{0,1\}^n$(y 与 y' 仅有一个坐标不同)以及任意的输出集合 $S\subset\Omega$,我们规定了 ε-表型差分隐私:

$$P[F(D,y)\in S]\leqslant e^{\varepsilon}\times P[F(D,y')\in S] \tag{1}$$

ε 为隐私保护预算,由数据拥有者公开定制,ε 的值越接近 0 表示差分隐私保护级别越高,但同时这也意味着 F 的输出越不准确.

这与差分隐私的通常定义不同,因为在差分隐私中 $\boldsymbol{D}$ 通常不是固定的,而我们假设基因型矩阵 $\boldsymbol{D}$ 是固定的. 直观地,以上定义表明,当一个人患有疾病时,F 返回的结果在统计上与他们没有疾病时返回的结果没有区别.

1.1.2 差分隐私强度影响参数

1)隐私保护预算[2]. 隐私保护预算 ε 一般体现了 F 所能提供的隐私保护程度. 因为 ε 取值越小,隐私保护程度就越高,反之亦然,因此选取多大隐私保护预算,ε 是一项非常重要的参数,需要根据具体需求定义 ε 的取值范围.

2)敏感度[3]. 敏感度是一个衡量加入噪声量的参数信息,指的是对数据集中任意删除操作对结果所造成的最大改变力度.

定义 2 全局敏感度(Global Sensitivity). 设有函数 $f:D\to R^d$,输入为一数据集,输出为一 d 维实数向量. 对于任意的邻近数据集 D 和 D',若满足公式(2),

$$\mathrm{GS}_f=\max_{D,D'}\|f(D)-f(D')\|_1 \tag{2}$$

则 GS_f 称为函数 f 的全局敏感度,全局敏感度用于量化表示对原始数据集 D 增加或删除一条记录时,对于整个算法 f 输出结果的最大影响. 其中,$\|f(D)-f(D')\|_1$ 是 $f(D)$ 和 $f(D')$ 之间的 1 阶范数距离. 函数本身决定了全局敏感度的选取,函数不同,全局敏感度也不相同.

1.1.3 实现机制

面向致病基因相关度研究分析中,差分隐私的实现主要采用 3 种实现机制,包括拉普拉斯机制、指数机制以及高斯机制.

1)拉普拉斯机制.

定义 3 拉普拉斯机制(Laplace Mechanism)[4]. 给定数据集 D,设有函数 $f:D\to R^d$,其敏感度为 Δf,那么随机算法 $K(D)=f(D)+Y$ 提供 ε-差分隐私保护,其中 $Y\sim\mathrm{Lap}(\Delta f/\varepsilon)$ 为随机噪声,服从尺度参数为 $\Delta f/\varepsilon$ 的 Laplace 分布,其中 $\mathrm{Lap}(\Delta f/\varepsilon)$ 概率密度函数为:

$$p(\eta)=\frac{1}{2\lambda}e^{\frac{|\eta|}{\lambda}} \tag{3}$$

根据公式(3)可得 Laplace 分布的期望值为 0,方差为 $2\lambda^2$.

2)指数机制.

定义 4 指数机制(Exponential Mechanism)[5]. 设

随机算法 K 输入为数据集 D,输出为一实体对象 $r\in$ Range,$q(D,r)$为可用性估价函数,Δq 为函数 $q(D,r)$ 的敏感度. 若算法 K 满足输出为 r 的概率与 $\exp(\varepsilon q(T,r)/2S(q))$成比例关系,那么算法 K 满足服从指数机制的 ε–差分隐私.

3)高斯机制(Gauss Mechanism). 与拉普拉斯机制相类似,同样是通过向查询请求结果 $f(T)$中添加服从高斯分布的噪声 η,得到 $f(T)+\eta$ 来实现 ε–差分隐私保护,其概率密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

根据公式(4)可得高斯分布的期望值为 μ,方差为 $2\lambda^2$, 其中 λ 由全局敏感度和隐私预算 ε 决定,λ 体现了添加噪声的幅度大小以及隐私保护的强度大小,与隐私保护强度成正比.

1.1.4　质量评估指标

1)数据查询准确度. 一个具有敏感信息的数据集在经过隐私保护算法处理后, 除了要保证敏感信息不外泄, 还要保证处理过的数据集中的信息还能够用于研究分析,所以要充分保证数据的可用性. 因此, 数据查询准确度是衡量隐私保护方法的一个重要指标. 本文通过将隐私保护方法得到的数据表与原数据计算重合比,来检验数据查询准确度.

2)隐私保护强度表示在所设计的方法中满足差分隐私的同时,确保隐私信息不被泄露,通常采用差分隐私的定义方法来评价算法是否满足差分隐私的要求. 由于差分隐私算法的隐私保护强度目前没有一种定量的测量机制, 本文在对隐私保护强度进行评估的时候使用 ε 以及噪声的方差来近似表示.

3)时间复杂度. 本文使用时间复杂度这项指标对所设计的算法进行评价, 具体方法是通过计算各个实验的运行时间来进行比较分析.

1.2　经验小波变换

经验小波变换 (Empirical Wavelet Transform, EWT)[6]是 Gilles 提出的一种构建适合处理信号小波族的方法. 为清楚起见,只考虑实际信号(它们的频谱相对于频率对称,$\omega=0$), 但通过在正负频率中构建不同的滤波器, 可以很容易地将以下推理扩展到复杂信号.

把傅里叶频谱划分 N 份, 每个分割的区间定义为 $\Lambda_n=[\omega_{n-1},\omega_n]$,$n=1,2,\cdots,N$. 其中,围绕每个 ω_n 都定义一个过渡段 T_n(宽度是 $2n$),这样就需要 N+1 个边界,除去已知的 0 和 π 两个边界,还需要 N–1 个边界,如图 1 所示.

考虑到归一化的傅里叶轴具有 2π 周期性,为了遵守 Shannon 标准, 将信号的频谱变化范围限制在 $\omega\in[0,\pi]$,首先假设傅里叶支持$[0,\pi]$被分割成 N 个连续的段. 将 ω_n 表示为每个段之间的界限 (其中 $\omega_0=0$、$\omega_N=\pi$),参见图 1. 每个段表示为 $\Lambda_n=[\omega_{n-1},\omega_n]$, 则很容易看出 $U_{n=1}^{N}\Lambda_n=[0,\pi]$. 以每个 ω_n 为中心,在 ω_n 周围定义了一个灰色阴影过渡区域 T_n,宽度为 $2\tau_n$.

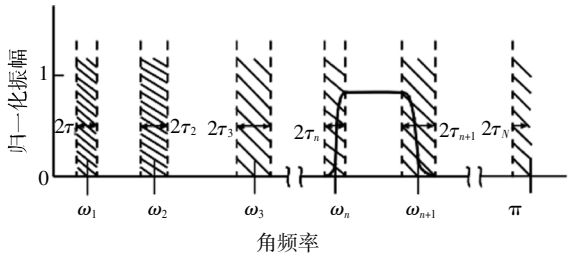


图 1　EWT 频谱的分割示例

Fig.1　Example of EWT spectrum segmentation

EWT 算法自适应性的好坏,很大程度上取决于信号频谱中的有用信息能否被包含在相应的过渡区间内. 因此, 分段数 N 及其边界点 ω_n 的选取至关重要. 分段数 N 选取的具体流程如图 2 所示,其中 α 值取 0.3~0.4. 确定分段数 N 后, 取 M 个极大值点中前 N 个最大值点,即$\{M_i\}_{i=1}^{N}$,找出它们在频谱中的具体位置,取相邻两极大值点所对应频率的中值,记为边界点 ω ($n=1,2,\cdots,N-1$). 经验小波就被定义为每个 Λ_n 上的带通滤波器. 为此, 利用 Littlewood–Paley 和 Meyer 的小波构造中使用的思想,对于 $\forall n>0$,分别通过方程(5)和(6)定义经验尺度函数和经验小波.

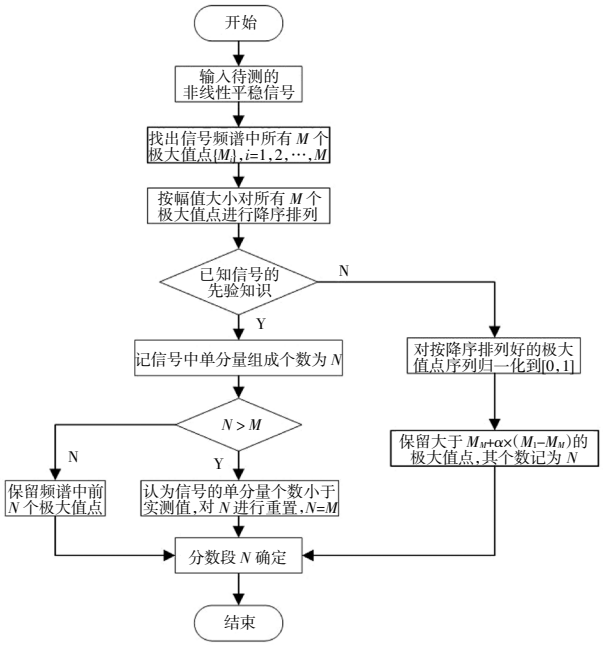


图 2　EWT 算法中分段数 N 的计算方法

Fig.2　Calculation method of segment number N in EWT algorithm

$$\hat{\phi}_n(\omega)=\begin{cases}1, |\omega| \leqslant \omega_n-\tau_n \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega|-\omega_n+\tau_n)\right)\right], \\ \qquad \omega_n-\tau_n \leqslant |\omega| \leqslant \omega_n+\tau_n \\ 0, \text{其他}\end{cases} \tag{5}$$

$$\hat{\psi}_n(\omega)=\begin{cases}1, \omega_n+\tau_n \leqslant |\omega| \leqslant \omega_{n+1}-\tau_{n+1} \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_{n+1}}(|\omega|-\omega_{n+1}+\tau_{n+1})\right)\right], \\ \qquad \omega_{n+1}-\tau_{n+1} \leqslant |\omega| \leqslant \omega_{n+1}+\tau_{n+1} \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega|-\omega_n+\tau_n)\right)\right], \\ \qquad \omega_n-\tau_n \leqslant |\omega| \leqslant \omega_n+\tau_n) \\ 0, \text{其他}\end{cases} \tag{6}$$

其中:函数 $\beta(x)$是任意的 $C^k([0,1])$函数,许多函数满足这些属性,比如式(5).

$$\beta(x)=x^4(35-84x+70x^2-20x^3) \tag{7}$$

关于 τ_n 的选择,有几种选择是可能的. 最简单的是选择与 τ_n 成比例的 $\omega_n:\tau_n=\gamma\omega_n$,其中 $0<\gamma<1$.

EWT 的构建. 根据经典小波变换理论构建经验小波,细节系数 $\omega_f^{\varepsilon}(n,t)$和逼近系数 $\omega_f^{\varepsilon}(0,t)$由待测信号与经验小波函数和经验尺度函数分别做内积得到,如式(8)和式(9)所示:

$$\omega_f^{\varepsilon}(n,t)=<f,\psi_n>= \int f(\tau)\overline{\psi_n(\tau-t)}\mathrm{d}\tau=\left(\hat{f}(\omega)\overline{\hat{\psi}_n(\omega)}\right)^{\vee} \tag{8}$$

$$\omega_f^{\varepsilon}(0,t)=<f,\phi_1>= \int f(\tau)\overline{\phi_1(\tau-t)}\mathrm{d}\tau=\left(\hat{f}(\omega)\overline{\hat{\phi}_1(\omega)}\right)^{\vee} \tag{9}$$

故信号重构公式为:

$$f(t)=\omega_f^{\varepsilon}(0,t)*\phi_1(t)+\sum_{n=1}^{N}\omega_f^{\varepsilon}(n,t)*\psi_t \tag{10}$$

公式(10)中:* 表示卷积. 根据重构公式,信号 $f(t)$可以由公式(11)得到:

$$\begin{cases}f_0=\omega_f^{\varepsilon}(k,t)*\phi_1(t) \\ f_k(t)=\omega_f^{\varepsilon}(k,t)*\psi_k(t), k=1,2,\cdots,N-1\end{cases} \tag{11}$$

通过经验小波变换将信号分解,获取一系列的调频调幅分量,然后对这些分量处理获取瞬时频率和瞬时幅值.

1.3 致病基因关联分析概述

致病基因相关度分析是本文的研究基础,该技术来源于全基因组关联研究(Genome-wide Association Studies,GWAS),目的是确定群体中哪些常见的单核苷酸多态性(SNP)与给定疾病相关. 这是通过采集大量个体,在常见的 SNP 上对它们进行基因分型,并且对于每个 SNP,进行统计测试来检查该 SNP 是否与所述疾病相关,然后计算每个 SNP 的相关度并根据相关度进行排序来完成[7].

本文基于差分隐私的 GWAS 主要集中在对致病基因相关度排序并返回高度相关的 SNP 这一任务,为了保护私人表型信息(疾病状态)在做致病基因相关度排序以及返回高度相关的 SNP 算法研究时不被泄露,以隐私保护的方式选择相关度较高的 SNP. 首先需要使用基于噪声的差分隐私方法进行隐私保护,然后对 SNP 的疾病相关性进行一系列的计算并评分,最终保证具有隐私保护的同时返回 m 个相关度较高的 SNP,其中 m 是用户定义的可变参数.

1.4 基于小波变换的差分隐私

为了在满足差分隐私的条件下,提高数据可用性,目前已有实现基于小波变换的差分隐私保护方法[8]. 该方法需要将数据以及参数 λ 作为输入,其中参数 λ 是由不同的噪声机制来确定的[9]. 图 3 为基于小波变换的差分隐私实现步骤,其中最主要的有 3 个步骤. 首先要对原数据进行小波变换处理,一般来说,小波变换是一个可逆线性函数,即它将数据集 M 映射到另一个矩阵 $\boldsymbol{C}$,这样 $\boldsymbol{C}$ 中的每个数据项都是 M 中数据项的线性组合,且 M 也可以从 $\boldsymbol{C}$ 中无损地重建. $\boldsymbol{C}$ 中数据项是由小波变换得到的小波系数,小波系数包含细节系数和近似系数(即高频系数和低频系数). 为了达到更好的降噪效果,经多次实验得出,将低频系数添加差分隐私噪声会得到更好的效果,算法的准确度会比较高.本节中 C_1 为小波的低频系数,C_2 为小波的高频系数.

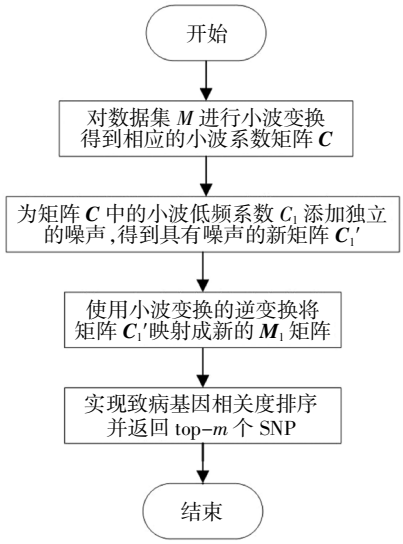


图 3 基于小波变换的差分隐私实现步骤

Fig.3 Differential privacy implementation steps based on wavelet transform

其次,在小波变换之后,为了保证实现差分隐私保护,需要为 C_1 中的小波低频系数添加独立的噪声(拉普拉斯噪声、指数分布噪声或者是高斯噪声),这一步将得到具有噪声系数的新矩阵 $\boldsymbol{C}_1'$.

最后,将矩阵 $\boldsymbol{C}_1'$使用小波变换的逆变换映射成具有差分隐私保护效用的矩阵 $\boldsymbol{M}_1$,并将该矩阵作为经过基于小波变换的差分隐私算法处理过的结果输出返回.

1.5 基因关联分析中的其他隐私保护技术

文献[10]采用同态加密和 Intel 软件保护扩展技术实现全基因组关联分析中的隐私保护. 文献[11]采用博弈论的方法实现大规模基因数据有效、定量的隐私保护. 文献[12]提出一种分析全基因组上位性的新方法,该方法采用二阶段框架的上位性分析方法,它包含特征过滤阶段以及上位性组合优化阶段,在上位性组合优化阶段采用贪婪算法启发式地搜索组合空间.

2 基于 EWT 的差分隐私

2.1 基于 EWT 的差分隐私的实现

本文是在满足差分隐私保护的前提下,对数据添加的噪声量进行一个降噪处理,最终得到较高可用性的数据集. 为了实现以上方法,本文提出将 EWT 变换应用到差分隐私的噪声处理中. 该方法的具体步骤如图 4 所示,首先使用差分隐私对数据进行处理,得到已满足差分隐私的数据集,然后对该数据集进行相关 EWT 的处理. 对于 EWT 的处理,首先,对数据集进行 EWT 分解,得到 N 个分段,并根据 N 个分段中信号的峭度值进行筛选并重构信号,得到最终降噪后的数据集. 实验最后使用该数据集来实现致病基因相关度排序,并对实验的隐私保护强度以及算法性能进行评估.

图 4 中根据峭度值筛选算法的具体步骤如下:

1)计算信号 $x(t)$经 EWT 分解后各分量的峭度值μ_n:

$$\mu_n = \frac{1}{N}\sum_{k=1}^{N} c_{nk}^4 \tag{12}$$

式中:N 为采样点数;c_{nk} 为 EWT 分解之后的分量.

2)根据 μ_n 得到各分量峭度值的集合 μ:

$$\mu = \{\mu_n\}, n = 1,2,\cdots,N \tag{13}$$

3)定义信号的调频调幅分量的峭度因子 Z_n:

$$Z_n = \frac{\mu_n - \min(\mu)}{\max(\mu) - \min(\mu)}, n = 1,2,\cdots,N \tag{14}$$

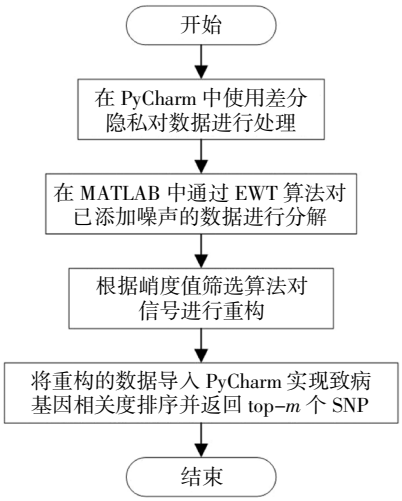


图 4　基于 EWT 变换的差分隐私实现步骤

Fig.4　Differential privacy implementation steps based on EWT transform

4)根据峭度因子选择峭度分量. 按照峭度因子从大到小的顺序将所有调频调幅分量进行重新排序,得到新的序列$\{F_n'\}, n = 1,2,\cdots,N; Z_1' \geqslant Z_2', \cdots, Z_n', \cdots, Z_{N-1}' \geqslant Z_N'$,$Z_n'$为排序后的峭度因子.

5)求出相邻两个调频调幅分量的峭度因子之差,之后找出最大差值 d_n.

$$d_n = Z_n' - Z_{n+1}' \tag{15}$$

利用峭度因子 Z_n'找出其对应的原序列 $F_{:,n}$,那么分解后从第 F_{n+1} 个分量开始以后都为噪声分量,前 n 个分量包含主要的峭度成分.

2.2 EWT 降噪实例

根据 2.1 节描述的基于 EWT 的差分隐私实现方式,对差分隐私的 3 种噪声机制进行实验对比,得出结果如图 5 所示. 由图 5 可以看出,基于 Gauss 噪声机制的差分隐私准确度较高,因此本文后续的差分隐私保护实验均以 Gauss 噪声机制为例.

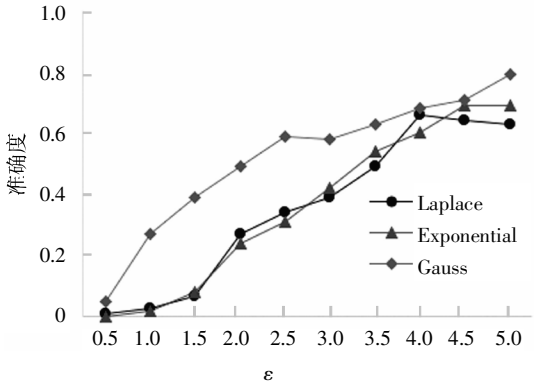


图 5　基于不同噪声机制的差分隐私算法准确度

Fig.5　Accuracy of differential privacy algorithm based on different noise mechanisms

在本文致病基因相关度研究实验中，使用差分隐私添加 Gauss 噪声进行 EWT 降噪. 首先添加 Gauss 噪声,然后采用 EWT 算法进行滤波仿真,最后将重构的数据提取后进行致病基因相关度排序实验并返回相关度最高的 m 个 SNP.

在使用 EWT 算法进行滤波仿真时,首先对包含 Gauss 噪声的数据做傅里叶变换,提取傅里叶分段的边界,根据对含噪声的信号进行有效估计,确定滤波器组的边界频率. 所得频谱分割结果如图 6 所示.

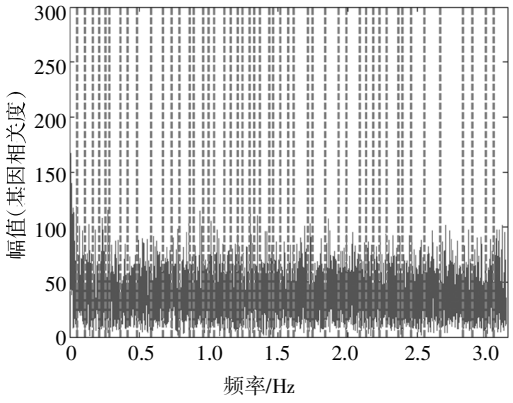


图 6 EWT 频谱分割

Fig.6 EWT spectrum segmentation

提取边界之后，通过镜像来扩展信号以处理边界,并建立相应的滤波器库,通过过滤信号来提取每个子带. 划分的频带共有 50 组，因此仿真信号的经验小波变换分解(EWT)的信号共有 50 组,用 F_1~F_{50} 表示. 图 7 为分解的 50 组 EWT 分量中的 7 组分量信号,从上到下依次表示为 F_1~F_6.

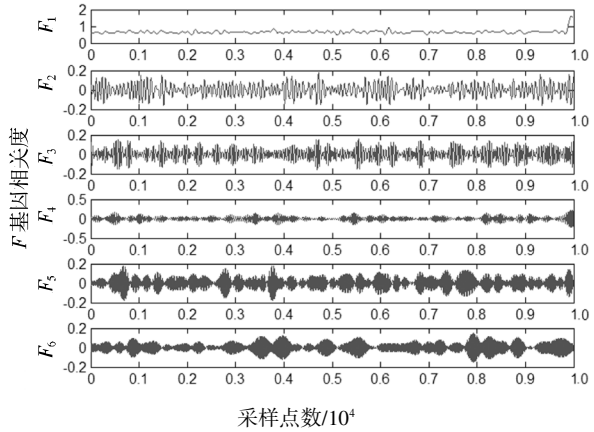
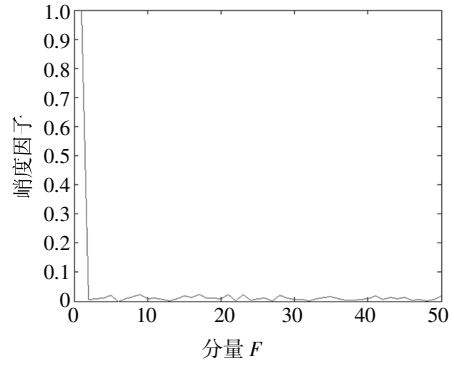


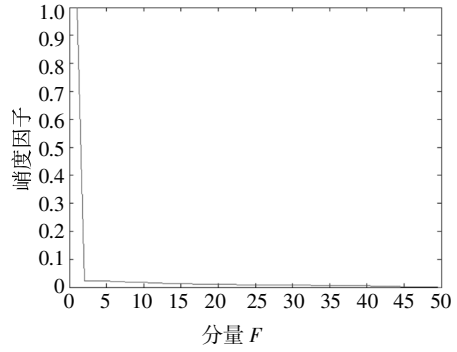
图 7 经 EWT 算法分解后的分量示例

Fig.7 Example of component decomposed by EWT algorithm

对于 EWT 分解得到的分量,根据 2.1 节中的筛选算法计算出峭度因子,并对其进行排序,排序前后的对比如图 8 所示. 根据进一步计算可得,排序后两个峭度因子之差的最大值位于 F_1 和 F_{23} 之间，最大差值为 0.976 6，依据 3.1 节中根据峭度值筛选重构信号的算法可以得出,F_1 以及 F_{24} ~ F_{50} 为噪声成分,而 F_2 和 F_{23} 之间包含主要的峭度成分，作为信号 $x(t)$的有效特征分量. 接下来,对 F_2 ~ F_{23} 个分量进行重构,形成重构信号. 图 9 为原信号、添加 Gauss 噪声后的信号以及重构信号的结果对比，纵轴相关度参数为致病基因相关度排序算法中计算疾病与基因相关度的实验数据. 从图 9 的对比结果可以看出,添加 Gauss 噪声后的信号分布较为杂乱，而重构之后的信号相关度系数均匀分布在–0.15~0.15 内，没有较大或者较小的信号.



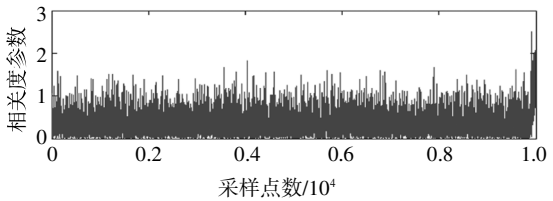
(a)排序前的峭度因子



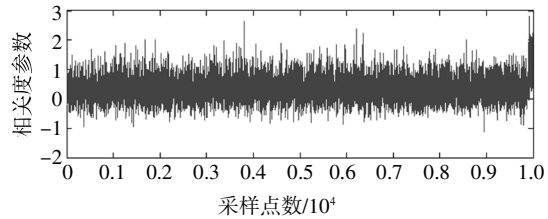
(b)排序后的峭度因子

图 8 排序前后的峭度因子

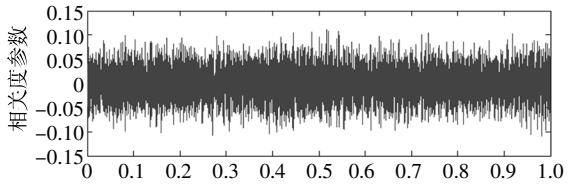
Fig.8 Kurtosis factor before and after sorting



(a)原始信号



(b)添加 Gauss 噪声后的信号



(c)重构信号(滤波结果)

图 9　采用 EWT 算法对 Gauss 噪声的降噪结果对比

Fig.9　Comparison of noise reduction results of Gauss noise using EWT algorithm

3　实验与结果分析

本节对 ε-差分隐私、基于 EWT 变换以及基于小波变换的 3 种差分隐私的保护效果进行实验对比分析. 具体实验中的差分隐私分别使用拉普拉斯机制、指数机制以及高斯机制 3 种实现机制,并设置不同的隐私参数 ε，以及返回前 m 个与疾病具有高相关度的 SNP 来对实验结果的影响进行测试.

3.1　数据集及实验环境

实验平台为 Intel(R) Core(TM) i7-6700HQ 2.60 GHz 处理器,8 G 内存,操作系统为 Windows10,编程环境为 MATLAB R2014a 以及 Pycharm 2016.3(64).

主要的测试数据来自某一类风湿性关节炎(RA)数据集 NARAC-1，这组数据由 Plink 工具生成,该数据集及其生成代码(基于 Plink 工具)可在线获取. 它包含 2 个种群,对于每一组,首先使用 plink 为 10 000 个 SNP 选择 MAF (Minor Allele Frequency),在某些条件下,最小等位基因频率可以使用统计方法来准确和稳健地解析在已知只有 MAF 的 DNA 样本混合物中存在已知基因型的个体，每个 SNP 从[0.05,0.5]中随机均匀选取. 然后,从每个人群中生成了 5 000 人，有 2 500 个病例和 2 500 个对照病例. 结果为每个样本有 10 000 个 SNP,9 900 个无效,100 个引起疾病(奇数比率 1.1). 然后将这 2 个群体组合起来生成模拟数据集,该模拟数据集是一个 1 × 10 000 的矩阵.

3.2　实验步骤

本文中实验的主要步骤如图 10 所示. 分别使用基本的 ε-差分隐私(DP)、基于小波变换(WT-DP)以及基于 EWT 变换(EWT-DP)这 3 种差分隐私算法来进行实验，并对这 3 种算法的实验结果进行对比分析. 在这 3 种算法的对比实验中,选择使用表现较好的 Gauss 机制来实现基本的差分隐私. 另外,对基于 EWT 变换的差分隐私噪声实现机制进行评估实验，该实验中分别使用 Laplace、Exponential 以及 Gauss 机制来实现基本的差分隐私.

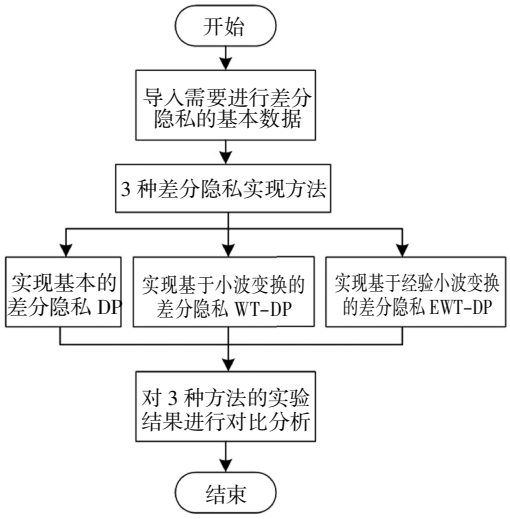


图 10　实验步骤

Fig.10　Experimental procedure

3.3　隐私保护强度评估

本节对各种差分隐私算法的保护强度进行比较评估. 由于差分隐私算法的隐私保护强度目前没有一种定量的测量机制，但是噪声参数 λ 体现了添加噪声的幅度大小以及隐私保护强度的大小. 因此,本文在对隐私保护强度进行评估的时候，通过计算噪声参数 λ 以及噪声分布的方差来近似表示.

图 11 的实验结果是对 ε-差分隐私(DP)、基于小波变换(WT-DP)以及基于 EWT 变换(EWT-DP)的差分隐私算法的保护强度的比较，这 3 种方法中使用到的差分隐私均使用 Gauss 噪声机制来实现. 本文中隐私保护强度是根据噪声的方差来进行计算的. 图 11 结果表明,3 种方法中 ε-差分隐私算法的隐私保护强度相对较高,基于 EWT 变换的差分隐私算法的隐私保护强度相对较低,但差距并不大,这表明使用 EWT 变换可以保证一定量的隐私保护效用.

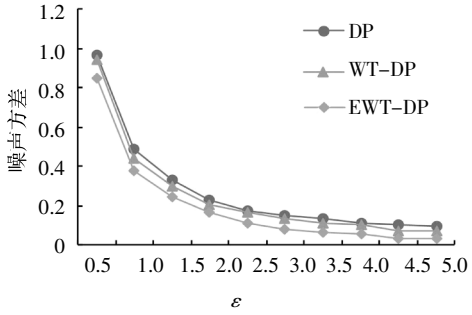


图 11　3 种差分隐私算法的隐私保护强度对比

Fig.11　Comparison of privacy protection strength of three differential privacy algorithms

图 12 的实验结果是基于 EWT 变换的 3 种差分隐私噪声实现机制的隐私保护强度的比较，实验中分别使用 Laplace、Exponential 以及 Gauss 机制来实现基本的差分隐私，横坐标为 ε 的值. 实验结果表明，差分隐私的保护强度与 ε 负相关，ε 的值越大，差分隐私所添加的噪声越小，噪声方差也越小，因此差分隐私的保护强度也越小.

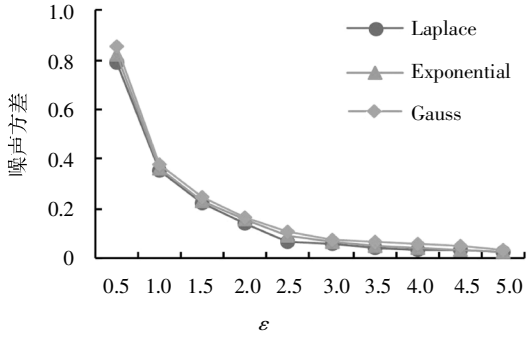


图 12 基于 EWT 变换的差分隐私实现机制的隐私保护强度对比

Fig.12 Comparison of privacy protection strength of differential privacy implementation mechanism based on EWT transform

3.4 算法时间复杂度

基于 EWT 变换的差分隐私算法时间复杂度主要由以下几步决定：1）对数据表进行差分隐私加噪处理，将数据表映射成便于计算的序列 M；2）对序列 M 进行 EWT 变换分解得到一系列的 EWT 分量；3）计算峭度值并根据峭度值来筛选重构信号；4）对信号进行重构，并提取降噪数据；5）使用降噪数据进行致病基因相关度排序算法，返回 m 个与疾病高度相关的 SNP. 总的时间复杂度近似为以上 5 个主要步骤的时间复杂度相加.

由于本文提出的基于 EWT 变换的差分隐私保护算法包括以上所列 5 个步骤，而 ε-差分隐私(DP)理论上只包括步骤 1)、步骤 5)；而基于小波变换(WT-DP)的差分隐私算法理论上虽然包括步骤 1)~步骤 5)，但是与本文所提的变换方法和降噪方法不同；所以，在运行时间上存在一定差异.

依据 3.2 节中的实验步骤进行实验并计算算法时间复杂度，对 3 种算法的运行时间计算 20 次并取平均值作为实验结果.

图 13 中的 3 种差分隐私算法均使用 Gauss 机制实现，可以看出基于 EWT 变换的差分隐私相较于其他 2 种差分隐私算法来说所花费的运行时间较长，性能比较低，这是因为 EWT 变换的过程相对于其他 2 种方法所花费的时间比较长.

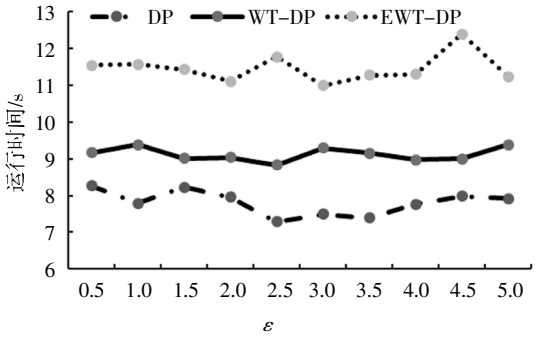


图 13 3 种差分隐私算法的运行时间对比

Fig.13 Comparison of run time of three differential privacy algorithms

图 14 的结果是基于 EWT 变换的 3 种差分隐私噪声实现机制的运行时间对比，可以看出基于 Gauss 机制的运行时间最短，性能最佳.

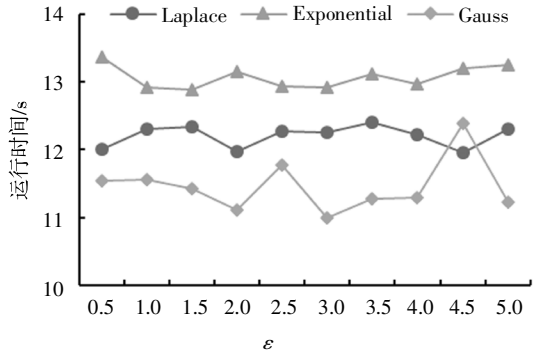


图 14 基于 EWT 变换的差分隐私实现机制的运行时间对比

Fig.14 Comparison of run time of differential privacy implementation mechanism based on EWT transform

3.5 致病基因相关度排序准确度评估

本文通过计算致病基因相关度排序的准确度来对经差分隐私处理过的数据的数据质量进行评估，具体做法是计算每次实验返回结果与真实结果重合的百分比，作为算法准确度的度量.

依据 3.2 节中的实验步骤进行实验并计算算法时间复杂度，得出结果分别如图 15、图 16 所示，这些结果均在 20 次迭代中取平均值.

图 15 中的 3 种差分隐私算法均使用 Gauss 机制实现，可以看出，经 EWT 变换后的差分隐私所得的致病基因相关度排序的准确度相对于其他 2 种方法比较高，也就是说使用该种方法处理数据的数据质量比较好.

图 16 中基于 EWT 变换的 3 种差分隐私噪声实现机制中 Gauss 机制的准确度较高.

用户可以根据其实际需求，采用本文所提方法设置相应的噪声注入量和过滤量，实现合理的隐私保护效果.

本文所提方法基于经验小波变换，假设需要隐私保护的基因数据规模为 N，经验小波变换的时间复杂度是 $O(N\log(N))$[6],整数全同态加密算法时间复杂度是 $O(N^3)$[13],本文所提方法与同态加密隐私保护技术相比,具有较低的计算时间复杂度.

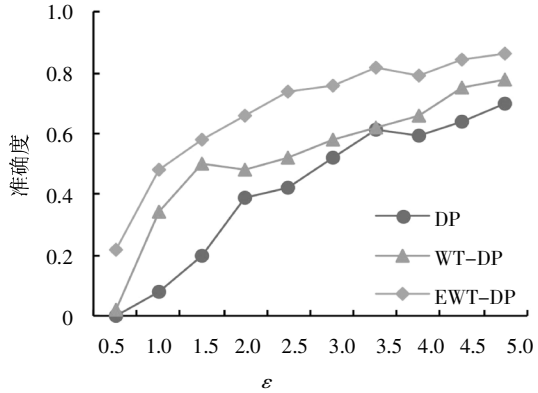


图 15　3 种差分隐私算法的准确度对比

Fig.15　Comparison of accuracy of three differential privacy algorithms

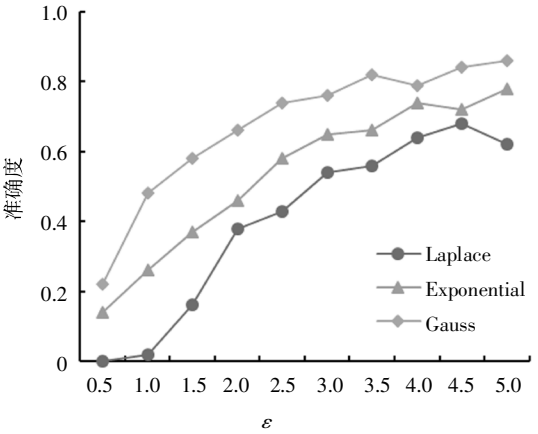


图 16　基于 EWT 变换的差分隐私实现机制的准确度对比

Fig.16　Comparison of accuracy of differential privacy implementation mechanism based on EWT transform

4 结　论

针对在致病基因相关度排序实验中数据因添加差分隐私噪声而导致的数据可用性较低这一问题，本文提出了一种基于 EWT 变换的差分隐私保护方法，设计了实现步骤并通过实验验证了该方法的可行性和正确性. 实验结果表明,该方法在保证了差分隐私保护强度的条件下，能够较为显著地提高致病基因相关度排序数据的可用性和准确度，实现了数据隐私保护强度与可用性的有效权衡. 下一步将继续研究如何在保证算法准确度的情况下降低隐私保护算法的时间复杂度.

参考文献

[1] JOHNSON A,SHMATIKOV V. Privacy-preserving data exploration in genome-wide association studies [C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago: ACM,2013:1079—1087.

[2] 何贤芒,王晓阳,陈华辉,等. 差分隐私保护参数 ε 的选取研究[J]. 通信学报,2015,36(12):124—130.
HE X M,WANG X Y,CHEN H H,*et al*. Study on choosing the parameter ε in differential privacy [J]. Journal on Communications,2015,36(12):124—130. (In Chinese)

[3] HAN Z,LIU H,WU Z. A differential privacy preserving framework with nash equilibrium in genome-wide association studies [C]// 2018 International Conference on Networking and Network Applications (NaNA). Xi'an:IEEE,2018:91—96.

[4] 熊平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报,2014,37(1):101—122.
XIONG P,ZHU T Q,WANG X F. A survey on differential privacy and applications [J].Chinese Journal of Computers,2014,37(1):101—122.(In Chinese)

[5] 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.
ZHANG X J,MENG X F. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers,2014,37(4):927—949.(In Chinese)

[6] GILLES J. Empirical wavelet transform [J]. IEEE Transactions on Signal Processing,2013,61(16):3999—4010.

[7] SIMMONS S,SAHINALP C,BERGER B. Enabling privacy-preserving GWASs in heterogeneous human populations [J]. Cell Systems,2016,3(1):54—61.

[8] DISHABI M R E,AZGOMI M A. Differential privacy preserving clustering using Daubechies-2 wavelet transform [J]. International Journal of Wavelets,Multiresolution and Information Processing,2015,13(4):1550028.

[9] 刘春,谢皓,肖奕霖,等. EWT 算法在 ECG 信号滤波中的研究[J]. 电子测量与仪器学报,2017,31(11):1835—1842.
LIU C,XIE H,XIAO Y L,*et al*. Research on empirical wavelet transform algorithm in ECG signal filtering[J]. Journal of Electronic Measurement and Instrument,2017,31(11):1835—1842.(In Chinese)

[10] SADAT M N,AZIZ A,MOMIN M,*et al*. SAFETY:Secure GWAS in federated environment through a hybrid solution [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2019,16(1):93—102.

[11] WAN Z,VOROBEYCHIK Y,XIA W,*et al*. Expanding access to large-scale genomic data while promoting privacy:A game theoretic approach [J]. The American Journal of Human Genetics,2017,100(2):316—322.

[12] 李泽军,陈敏,曾利军. 一种分析全基因组上位性的新方法[J]. 湖南大学学报(自然科学版),2016,43(10):160—165
LI Z J,CHEN M,ZENG L J. A genome-wide epistasis analysis method based on multiple criteria fusion[J]. Journal of Hunan University(Natural Sciences),2016,43(10):160—165.(In Chinese)

[13] ZHANG P,SUN X,WANG T,*et al*. An accelerated fully homomorphic encryption scheme over the integers[C]//2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS). Beijing:IEEE,2016:419—423.