

大数据环境下网络数据隐私保护算法研究

潘明波

(云南工商学院 信息工程学院, 云南 昆明 651701)

摘 要: 针对传统的数据隐私保护算法存在执行时间较长, 隐匿率较高, 适应性较差等问题, 提出基于密度聚类的网络数据隐私保护算法, 在大数据环境下, 随机生成一个满足给定的大数据环境下网络隐私数据特征的变换函数, 然后利用这个变换函数对原始的网络隐私数据的数值进行变换, 并将变换后的数值作为随机化回答的结果; 然后根据网络的特征对网络节点进行密度聚类分析, 生成符合大小的任意形状的簇, 通过量化网络隐私数据信息丢失量和网络隐私数据结构信息丢失量对隐私数据进行有效性分析; 最后在成功生成的簇内插入真实网络节点, 通过增加边等技术, 完成对网络数据信息的隐私保护。仿真实验结果证明, 所提算法具有较好的隐私保护效果, 在对多种不同背景知识攻击的情况下, 具有较好的适应效果。

关键词: 大数据环境下; 基于密度聚类; 网络数据隐私保护算法

中图分类号: TN309

文献标识码: A

文章编号: 1000—7180(2017)07—0101—04

Big Data Environment Network Data Privacy Protection Algorithm Research

PAN Ming-bo

(Information Engineering Institute, Yunnan Technology and Bussiness University, Kunming 651701, China)

Abstract: There are many problems in the traditional data privacy protection algorithm, such as long execution time, high concealment and poor adaptability. The proposed network data density clustering algorithm based on privacy protection, in the big data environment, the transformation function generates a random given the big data environment of network privacy data, then use the numerical transform function of network privacy data are changed, and the numerical transform as the result of random response then according to the characteristics of the network; the network node density clustering analysis, generated in accordance with any size and shape of the clusters, through the data network privacy and network privacy quantitative information loss data structure information for effective analysis of data privacy loss amount; and finally into the real network node in the successful generation of cluster, through increasing etc. technology, complete privacy protection of network information. The simulation results show that the proposed algorithm has a better privacy protection effect, and it has a good adaptability to the attack of a variety of different background knowledge.

Key words: big data environment; based on density clustering ; network data privacy protection algorithm

1 引言

随着互联网技术的快速发展, 推进整个社会进入了大数据时代^[1]. 大数据环境中蕴藏着众多有价值的信息, 是政府机构以及大中型企业的宝贵财富. 同样的, 大数据时代的到来也意味着风险和挑战^[2],

个人数据的网络化与透明化已经成为不可阻挡的大趋势, 网络数据隐私保护就是其中表现尤为突出的一个问题^[3]. 隐私泄露在大数据时代防不胜防. 在网络中留下的数据信息很可能被恶意利用, 造成意外甚至是难以挽回的经济损失^[4-5]. 因此需要对大数据环境下的网络数据隐私保护算法进行研究. 现行的

大数据环境下基于可扩展加密的传感器网络数据隐私保护算法^[6],通过使用复数的加法法则进行可扩展网络数据聚合,通过分段密钥来实现大数据环境下的网络数据隐私保护.存在网络隐私数据丢失严重,算法执行时间过长等问题^[7].对于如何更好地对大数据环境下的网络数据进行隐私保护引起了社会各界人士的广泛关注.

文献[8]提出基于主成分分析的网络数据隐私保护算法,该方法在对原始网络数据训练样本集进行转换的过程中,灵活性较低,导致网络数据的隐私保护安全性低.文献[9]提出基于数据分割与分级的网络数据隐私保护算法,该算法对网络数据的隐私保护执行时间较长,适应性较差.

针对上述问题的产生,本文提出基于密度聚类的网络数据隐私保护算法.

2 大数据环境下网络数据隐私保护算法研究

2.1 基于转换随机化回答的网络隐私数据处理与重构

转换随机化回答的网络隐私数据处理思想:首先要随机生成一个满足给定的大数据环境下网络隐私数据特征的变换函数,然后利用这个变换函数对原始的网络隐私数据的数值进行变换^[10],并将变换后的数值作为随机化回答的结果,具体描述过程如下.

大数据环境下假设给定网络隐私数据集 $A = \{a_1, a_2, \dots, a_l\}$,则其中网络隐私数据元素的均值和方差分别用式(1)、(2)表示:

$$\bar{A} = \frac{1}{l} \sum_{i=1}^l a_i \quad (1)$$

$$\sigma_A^2 = \frac{1}{l-1} \sum_{i=1}^l (a_i - \bar{A})^2 \quad (2)$$

式中, l 表示网络隐私数据的属性值; i 表示网络隐私数据的特征向量; a_i 为网络隐私数据的信息损失.

在大数据环境下给定网络隐私数据集 $B = \{b_1, b_2, \dots, b_m\}$,则网络隐私数据元素均值和方差为

$$\bar{B} = \frac{1}{m} \sum_{i=1}^m b_i \quad (3)$$

$$\sigma_B^2 = \frac{1}{m-1} \sum_{i=1}^m (b_i - \bar{B})^2 \quad (4)$$

式中, m 表示网络隐私数据所占样本的比例; b_i 表示网络隐私数据的训练样本.对数值类的网络隐私数据 x ,网络隐私数据的随机化函数为

$$r(x) = ax + b \quad (5)$$

式中, $a \in A$,是从网络隐私数据集 A 中随机选择的一个元素,而 $b \in B$,是从网络隐私数据集 B 中随机选择的一个元素,则通过转换后的网络隐私数据可以通过以下公式计算得到:

$$y = r(x) \quad (6)$$

则发布网络隐私数据集 D' 中的 n 条记录所对应的网络隐私数据属性 At 的取值为

$$Y = \{y_1, y_2, \dots, y_n\} \quad (7)$$

可以转换为式(9)进行计算:

$$Y = R(X) \quad (8)$$

式中, $R(X)$ 表示网络隐私数据的相似数据样本,且 $y_i = r(x_i), i = 1, 2, \dots, n$.

按照上述转换随机化回答方法,将原始网络隐私数据 x 转换成如下公式计算:

$$y = r(x) = ax + b \quad (9)$$

当网络隐私数据元素的均值 $\bar{A} \neq 0$ 时,原始网络隐私数据 x 的估计量为

$$\hat{x} = \frac{y - \bar{B}}{\bar{A}} \quad (10)$$

式中,网络隐私数据的估计量 $\hat{x}$ 是原始网络隐私数据的无偏估计量.则原始网络隐私数据集 D 中的 n 条记录所对应的属性 At 的取值通过式(8)计算可得.

假设 X 符合高斯分布 $N(\mu_x, \sigma_x^2)$, σ_x^2 表示网络隐私数据的标准差; $\bar{x}$ 是 μ_x 的无偏估计量,表达式如下:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (11)$$

式中, x_i 表示网络隐私数据的随机化参数,是未知的,用 $\bar{x}_i$ 代替 x_i ,得到网络隐私数据估计量为

$$\bar{x}_i = \frac{1}{n} \sum_{i=1}^n \hat{x}_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i - \bar{B}}{\bar{A}} = \frac{\bar{y} - \bar{B}}{\bar{A}} \quad (12)$$

为了减小计算量,可设 $\bar{B} = 0$,得到原始网络隐私数据 x 均值的估计值为

$$\bar{x} = \frac{\bar{y}}{\bar{A}} = \frac{1}{n\bar{A}} \sum_{i=1}^n y_i \quad (13)$$

另一方面原始网络隐私数据均值方差计算公式为

$$\text{var}(x) + \sigma_A^2 \cdot \bar{x}^2 + \sigma_B^2 \quad (14)$$

式中, $\bar{A}_2$ 表示网络隐私数据的伪造样本均值; $\bar{x}^2$ 表示网络隐私数据的伪造样本无偏估量.并进行加权

融合,可以转换为以下公式进行计算:

$$\text{var}(x) = \frac{\text{var}(y) - \sigma_A^2 \cdot \bar{x}^2 - \sigma_B^2}{\sigma_A^2 + A^2} \quad (15)$$

式中, $\bar{x}$ 为未知量,用 $\bar{x} = \frac{\bar{y}}{A}$ 代替,可以得到原始网络隐私数据 x 的方差估计值:

$$\hat{S}^2 = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}) - \frac{\sigma_A^2 \cdot \bar{y}^2}{A^2} - \sigma_B^2}{\sigma_A^2 + A^2} \quad (16)$$

利用网络隐私数据集 A 中元素的均值 $\bar{A}$ 和方差 σ_A^2 ,以及网络隐私数据集 B 中元素的方差 σ_B^2 ,并通过发布网络隐私数据集 D' 中属性 A_t 的取值,根据式(13)和式(16),计算出 $\bar{x}$ 和 $\hat{S}^2$ 的取值.然后以此作为原始网络隐私数据集 D 中属性 A_t 的取值的分布特征 $N(\bar{x}, \hat{S}^2)$,进而实施网络隐私数据的分类处理.

2.2 基于密度聚类的网络数据隐私保护算法

根据大数据环境下网络的特征,生成符合大小的任意形状的簇,最后对成功生成的簇,计算每个簇中应生成网络隐私数据真实节点的数目,通过增加边等技术手段,完成对网络数据信息隐私保护.具体描述过程如下.

定义 1 网络隐私数据量化信息丢失量.假设 Cl_t 是网络中的一个簇,准标识符表示为

$$Q = (n_1, n_2, \dots, n_s, c_1, c_2, \dots, c_r) \quad (17)$$

则泛化准标识符产生网络隐私数据信息丢失量为

$$QL(Cl_t) = |Cl_t| * \left[\sum_{i=1}^s \left(\frac{\text{Range}(n_i)}{\text{sub}(c)} \right) \right] \quad (18)$$

式中, $|Cl_t|$ 表示网络簇中节点的数量; n_i 表示网络隐私数据中簇的节点度.设共生成簇的数量为 m 个,则网络隐私数据总信息丢失量为

$$TQL(G) = \sum_{i=1}^m (QL(Cl_k)) \quad (19)$$

定义 2 网络隐私数据结构信息丢失量.假设大数据环境下网络 $G = (V, E)$, $V = \{v_1, v_2, \dots, v_n\}$,网络中某一个簇 Cl_k 节点集合为 $V = \{v_1, v_2, \dots, v_n\}$,则该网络簇中节点 v_i 的度为 $k(v_i)$,与网络节点 v_i 相连接的其他网络节点可以形成的边的最大数目表达式为

$$T(v_i) = k(v_i) * (k(v_i) - 1) / 2 \quad (20)$$

式中, k 表示网络数据的隐私参数.在大数据环境中

与网络节点 v_i 相连接的其他网络节点形成边的实际数目为 $E(v_i)$,则网络节点 v_i 的网络聚集系数计算为

$$C(v_i) = E(v_i) / T(v_i) \quad (21)$$

故大数据环境下网络 G 的聚集系数之和为

$$FQL = \sum_{i=1}^n C(v_i) \quad (22)$$

根据式(22)计算网络簇 Cl_t 聚集系数和为

$$EQL = \sum_{i=1}^m \left(\sum_{i=1}^{|CL|} C(v_i) \right) \quad (23)$$

式中, $|CL|$ 表示网络隐私数据的限定性阈值.求取网络隐私数据总结构信息丢失量计算公式如下

$$NTQL = FQT - EQL \quad (24)$$

假设生成网络真实节点的总数目为 N ,共同生成簇数量为 m ,则每个簇中生成真实网络用户数量为 N_i 为

$$N_i = \left(\frac{n_i}{n_1 + n_2 + \dots + n_m} \right) * N \quad (25)$$

式中, n_i 代表网络第 i 个簇中的节点数目,且 $1 \leq i \leq m$.最后,在不改变其他网络节点连接的前提下,选择度数最少的网络节点,分别与新生成的网络真实节点相连接.从而实现大数据环境下网络数据隐私保护.

3 仿真实验结果与分析

本实验环境使用 CPU 为主频 3.2 GHz Core i5,内存为 8 GB,操作系统使用虚拟机下的 Windows7 系统,开发语音使用 VisualC++ 6.0 和 MATLAB8.5.实验数据来自英国 BP 公司的有邮件数据集,其中包含 200 名用户、80 万封以上的邮件,放在 4 000 个左右的文件夹中.选取其中 2010.1.3,2011.4.8,2013.6.7 和 2014.8.10 的数据,删除其中重复数据,将不同的发件人作为大数据环境中的网络节点,邮件往来作为节点间的边,与客户之间的通信次数作为边权重,统计结果得到网络中共有 82 000 多个节点,320 000 多条边.仿真实验将从算法执行时间,防御不同类型攻击时隐匿率的变化,以及算法适应效果等方面对本文所提算法进行分析.

在大数据环境下对网络节点更改数量变化时,比较 3 种更新方式(添加节点、修改节点、删除节点)执行时间的差异.如图 1 所示的更改网络节点数目分别为 100,200,300,400,500,600,700,800.

通过图 1 可以发现,在对网络节点进行删除时,

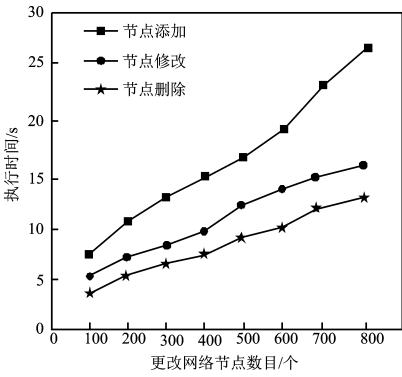


图 1 更改网络节点数目对算法执行时间的影响

执行时间变化幅度最大,当网络节点添加时,执行时间变化幅度最小,以上 3 种网络隐私数据变动的执行时间均随着网络节点更改数目的增加而增加。

大数据环境下网络数据隐匿率为网络数据隐匿网络节点数/网络节点边数占网络中总节点数/边数比例。从图 2 中可以看出,根据 2.2 节中式(20)可知,网络数据隐私参数 k 越大,大数据环境下的网络隐私数据的隐匿率越小,网络数据隐私泄露程度也越小。

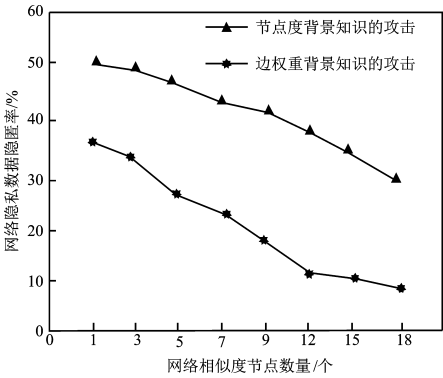


图 2 不同背景知识的攻击下对网络隐私隐匿率影响

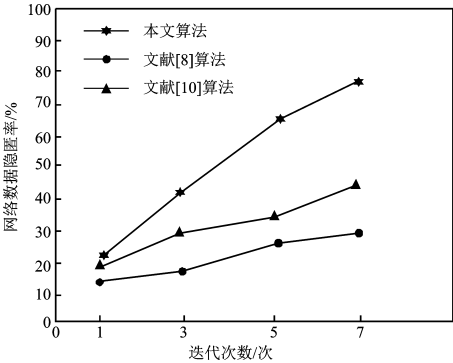


图 3 不同算法中隐匿率迭代次数变化趋势对比

通过图 3 可以看到,在经过多次迭代后,文献[8]的算法与文献[10]的算法中,网络数据隐匿率呈几何增长形式,在迭代 5 次之后已经无法满足网络

数据隐私保护对隐匿率(小于 0.6)要求。本文所提算法在网络数据隐匿率上有所升高,但效果并不明显,且网络数据隐匿率始终在要求范围之内,说明本文所提算法在多次迭代更新情况下对网络数据隐私保护的效果较强,适应性较好。

4 结束语

针对传统方法存在的问题,提出基于密度聚类的网络数据隐私保护算法,仿真实验结果证明,本文所提算法具有较好的网络数据隐私保护效果且网络数据隐匿率在要求范围内(隐匿率小于 0.6)。所提算法在对多种不同背景知识攻击的情况下,具有较好的适应效果。

参考文献:

[1] 潘晓,吴雷,胡朝君. 道路网络上基于网络 Voronoi 图的隐私保护算法[J]. 计算机研究与发展, 2015, 52(12): 2750-2763.

[2] 赵耘华,白光伟,沈航. 基于消息复用的位置服务隐私保护算法[J]. 微电子学与计算机, 2015, 32(2): 148-152.

[3] 王超,杨静,张健沛. 基于稀疏化最小生成树聚类的个性化轨迹隐私保护算法[J]. 电子学报, 2015, 43(11): 2338-2344.

[4] 孙丹丹,罗永龙,范国婷,等. 基于轨迹形状多样性的隐私保护算法[J]. 计算机应用, 2016, 36(6): 1544-1551.

[5] 胡德敏,郑霞. 基于 SpaceTwist 的 k-匿名增量近邻查询位置隐私保护算法[J]. 计算机应用研究, 2016, 33(8): 2402-2404.

[6] 朱庆华,常莹,赵凯. 基于加性汇聚函数的数据隐私保护算法[J]. 计算机工程与设计, 2016, 37(7): 1741-1745.

[7] 张晓琳,张臣,张文超,等. D-VSSP:分布式社会网络隐私保护算法[J]. 计算机科学, 2017, 44(2): 93-97.

[8] 鄂旭,杨明婧,励建荣,等. 能量均衡的数据融合隐私保护算法研究[J]. 计算机工程, 2016, 42(4): 126-130.

[9] 吴伟民,黄焕坤. 基于差分隐私保护的 DP-DBScan 聚类算法研究[J]. 计算机工程与科学, 2015, 37(4): 830-834.

[10] 智淑敏,刘静静. 一种用于网络位置隐私保护的动态路由算法[J]. 电子设计工程, 2016, 24(15): 28-31.

作者简介:

潘明波 男, (1984-), 讲师. 研究方向为计算机科学技术、计算机应用. E-mail: panmingbo6431@163. com.