doi:10.14132/j.cnki.1673-5439.2022.02.013

一种基于矩阵补全的智能电网数据隐私保护方案

王德庆,许 建

(南京邮电大学 计算机学院,江苏 南京 210023)

摘要:针对智能电网中不同信任域间的数据隐私保护问题,提出了一种基于矩阵补全的智能电网数据隐私保护方案。该方案从少量节点产生的模糊值子集中估计出原始数据相关子空间,通过矩阵补全来修复相关矩阵中的缺失元素。为保证数据隐私,通过添加一种与原始数据具有相同统计特性的噪声来增加随机扰动。仿真结果表明,该方案具有较好的隐私保护效果和数据可用性,同时能够有效降低智能电网中的通信开销。

关键词:智能电网;隐私保护;矩阵补全;数据扰动

中图分类号:TP309 **文献标志码**:A **文章编号**:1673-5439(2022)02-0103-08

A matrix completion based data privacy protection scheme in smart grid

WANG Deqing, XU Jian

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: To protect data privacy among different trust domains of the smart grid, a matrix completion based data privacy protection scheme is proposed for smart grids. The scheme estimates the original data correlation subspace through a subset of obfuscated values generated by a small number of nodes, and repairs the missing elements in the correlation matrix via matrix completion. Further, to protect data privacy, the specific noises that have the same statistical properties as the original data have been added to increase random perturbation. Simulation results demonstrate that the proposed scheme achieves a good performance from the perspective of data availability and privacy protection, and it can also reduce network communication overhead effectively.

Keywords: smart grid; privacy protection; matrix completion; data perturbation

智能电网是指电网的智能化,也被称为“电网2.0”,它建立在集成的、高速双向通信网络的基础之上。具体来说,智能电网将传统的电力系统和先进的智能通信系统、控制技术、采集与传感技术结合起来,配合全面完善的安全策略,实现电网用户与电能供应商之间的互动,以保证电网智能、可靠、安全、友好、高效地运行[1]。因此,精确、高效、开放的信息系统是未来电网的特性,也是智能电网与传统电网的本质区别。

近年来,随着可再生能源等分布式发电资源数量不断增加以及各类智能化家居等智能终端设备的大量接入,电网企业与电力用户之间、电气设备与控制中心之间会产生大量的数据流[2]。同时,智能电网的真正价值不在于物理互联设备本身,而在于它们所包含的大量数据集和粗糙、未经提炼的信息,以及如何高效、快速、有意义地处理这些信息[3]。智能电网产生空前数量的原始信息,以准确评估态势感知,提高多个工业系统的智能、效率和可持续性。然而,正是由于海量数据的采集、传输和处理,使得智能电网中的数据隐私问题变得越来越严重,尤其是敏感数据的获取会直接暴露用户的隐私信息。因此,需要通过隐私保护技术来对智能电网数据进行

收稿日期:2021-11-02;**修回日期**:2022-02-16 **本刊网址**:http://nyzr.njupt.edu.cn

基金项目:中国博士后科学基金(2019M651922)资助项目

作者简介:王德庆,男,硕士研究生;许建(通信作者),男,博士,副教授,xuj@njupt.edu.cn

引用本文:王德庆,许建.一种基于矩阵补全的智能电网数据隐私保护方案[J].南京邮电大学学报(自然科学版),2022,42(2):103-110.

处理,以保障智能电网中数据的隐私安全,促进智能电网应用的实际开展[4]。

在实际应用中,为了最大限度地实现智能电网数据信息的采集获取,不同机构会在同一个物理区域中部署各自的数据采集网络。这些网络覆盖范围相互重叠,但是在通信过程中属于不同的信任域,这也就意味着分属不同信任域的网络节点获取的数据需要对其他信任域保持隐私性。为实现不同信任域之间数据的隐私保护,一种可能的解决方案是进行同态加密,文献[5]提出了一种使用同态加密的隐私保护方案来保护明文数据不被攻击者访问。文献[6]在智能电网中数据聚合的情况下提出一种基于区块链和同态加密的数据聚合(Blockchain and Homomorphic Encryption-based Data Aggregation, BHDA)方案。文献[7]提出了一种基于原始对偶次梯度分布式优化技术和全同态加密的算法,来解决智能电网中的隐私问题。文献[8]提出了一种保护数据隐私并具有一定容错的安全聚合方案来防止虚假数据注入的攻击。但是这些方案因为需要可信的第三方而受到限制。另一种可行的解决方案是安全多方计算,比较有代表性的如文献[9]针对智能电网数据聚合提出的一种基于雾的安全多方计算(Fog-Enabled Secure Multiparty Computation, FESMPC)方案,该方案在加密、聚合、解密方面是有效的。但面临的问题是随着参与方数量的增加,问题变得复杂;同时,在传输错误的情况下,会因为数据缺失导致难以计算。另外,差分隐私作为隐私保护机制也被应用到智能电网中来,文献[10]提出了一种基于信任距离的个性化差分隐私方案来为智能电网中敏感数据提供隐私保护。文献[11]提出了基于稀疏编码的本地化差分隐私随机响应扰动机制,通过对输入数据进行随机响应与数据重构,实现本地化差分隐私保护。这一类解决方案往往对数据聚合后的结果进行隐私保护,直接对数据流进行隐私保护的研究很少。此外,为减少智能电网中的通信开销,可能会采用一些能量调度算法,如文献[12]提出了一种基于能量和密度的低功耗自适应集簇分层算法,但是这类方法还需要额外能量来实现算法优化。同时在真实环境中仍会有多种因素导致数据无法传输成功,因此需要对缺失的电网数据进行补全,文献[13]提出了一种基于稀疏表示的电力负荷缺失数据补全方法,但是需要提供先验知识让算法学习。文献[14]提出了另一种缺失数据补全方法,该方法利用一种基于奇异值阈值的算法可以有效对电网数据中缺失部分进行补全。

综上所述,现有研究成果大多仅仅针对其中一个方面提出了相应的解决方案,并没有从根本上解决智能电网中的数据隐私保护问题。首先,智能电网的不同信任域在同一物理区域中采集数据并传输,但它们担心传输的原始数据可能会被不同信任域所窃听,因此要在传输前进行隐私保护处理。然而一般的数据扰动处理可能会遭受恶意重构导致原始数据泄露。其次,在智能电网的数据采集过程中,由于节点无法抵抗内在因素(例如节点的功耗、链接故障、数据包丢失)以及外部故障(例如恶意节点),会有大量的数据缺失,同时节点之间频繁的通信造成了较高的通信开销。

针对以上问题,本文在智能电网不同信任域应用场景下,提出了一种基于矩阵补全的智能电网数据隐私保护方案。该方案的主要创新点如下:

(1) 提出了一种直接对原始数据流进行模糊处理的方法,该方法对每个时刻的数据流,基于平均遍历定理来估计相关矩阵,再通过主成分分析得到统计特性类似于原始数据流的随机噪声。由于该噪声和原始数据流具有相同的统计特性,因此攻击者很难通过过滤方法获取原始数据。

(2) 由于数据传输过程中可能会丢失一些数据,以及智能电网中的数据采集节点为降低通信开销可能会与相邻节点间歇性通信,为实现数据矩阵中未知元素的重构,本文采用一种不精确的拉格朗日乘子法(Augmented Lagrange Multiplier, ALM)来补全每个节点上进化的全网络相关子空间。

1 系统模型和理论基础

1.1 系统模型

针对存在多个信任域的智能电网应用场景,即在同一个物理区域的智能电网中有多个属于不同信任域的信息采集传输网络,这些分属不同信任域的网络负责不同种类数据信息的采集和传输。由于是在同一物理区域和时间范围内,所以同一个信任域中数据采集节点收集的数据具有一定相关性。如图1所示,一个智能电网中有分属于3个不同信任域的信息采集网络,其中每个圆圈、三角、矩形分别代表不同信任域中的节点,同一种形状的节点属于同一个信任域。在通信过程中,每个数据采集节点只与属于同一信任域的相邻节点通信,如在信任域1中的节点 k 只与相邻的节点 k_1、k_2 进行通信。

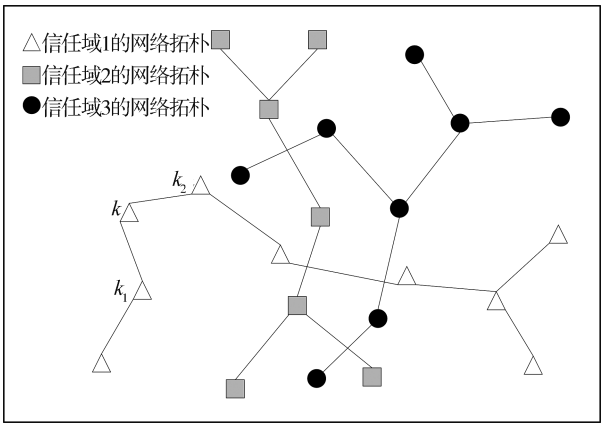

图 1　具有多个信任域的智能电网通信环境示意图

1.2　矩阵补全

矩阵补全(Matrix Completion,MC)技术指的是通过采样矩阵中的部分元素恢复出原始的完整矩阵,其标准形式如下[15]

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X})$$
$$\text{s.t. } P_{\Omega}(\boldsymbol{X}) = P_{\Omega}(\boldsymbol{M}) \tag{1}$$

式中,$\boldsymbol{X}$ 和 $\boldsymbol{M}$ 都是 $m * n$ 的矩阵,$\boldsymbol{X}$ 为待求的矩阵,$\boldsymbol{M}$ 为不完整的矩阵,$\operatorname{rank}(\boldsymbol{X})$ 为矩阵 $\boldsymbol{X}$ 的秩,采样符号 P_{Ω} 的定义为

$$P_{\Omega}(\boldsymbol{X}) = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & \text{其他} \end{cases} \tag{2}$$

然而式(1)是一个 NP-hard 问题,想要直接求解比较困难。可以通过将秩函数松弛为核范数,将秩函数最小化问题松弛为如下凸优化问题[16]

$$\min_{\boldsymbol{X}} \| \boldsymbol{X} \|_{*}$$
$$\text{s.t. } P_{\Omega}(\boldsymbol{X}) = P_{\Omega}(\boldsymbol{M}) \tag{3}$$

式中,$\| \boldsymbol{X} \|_{*}$ 为矩阵 $\boldsymbol{X}$ 的核范数,它等于矩阵 $\boldsymbol{X}$ 的奇异值之和(一个秩为 r 的矩阵 $\boldsymbol{X}$ 有 r 个非零的奇异值,它们的和等于该矩阵的核范数 $\| \boldsymbol{X} \|_{*}$),即 $\| \boldsymbol{X} \|_{*} = \sum_{k=1}^{r} \sigma_k(\boldsymbol{X})$。其中,$\sigma_k(\boldsymbol{X})$ 是 $\boldsymbol{X}$ 的第 k 大的奇异值。

1.3　主成分分析

主成分分析(Principal Component Analysis,PCA)是一种降维方法[17]。该方法通常用于将一大组变量转换为较少的变量,同时较少的变量中仍包含原数据集中的大部分信息,在这个过程中主要以准确性为代价。PCA 通过线性变换将原始数据转换为各维线性无关的表示,可用来提取数据的主要特征分量。

PCA 的具体思路如下:设一个 d 行 d 列的矩阵 $\boldsymbol{R}_{d\times d}$,通过特征分解计算协方差矩阵的特征值与特征向量

$$\boldsymbol{R}_{d\times d} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} \tag{4}$$

式中,$\boldsymbol{\Sigma}$ 为主对角元素均为特征值的对角矩阵;$\boldsymbol{U}$ 为由特征向量组成的正交矩阵,特征向量也称为主成分。特征分解后,特征值越大,说明该方向上的变化越大,也就越重要。

按特征值从大到小重新进行排列有 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$,给定一个特征值贡献值 a,$0 \leqslant a \leqslant 1$。计算要保留的主成分个数 k,使其满足 $b \geqslant a$,其中

$$b = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{5}$$

选择其中最大的 k 个,然后将其对应的 k 个特征向量分别作为列向量组成特征向量矩阵。

2　基于矩阵补全的智能电网数据隐私保护方案

2.1　方案设计

为了解决智能电网环境中不同信任域间的数据隐私保护问题,本文提出了一种基于矩阵补全的智能电网数据隐私保护方案,该方案的基本流程如图 2 所示。

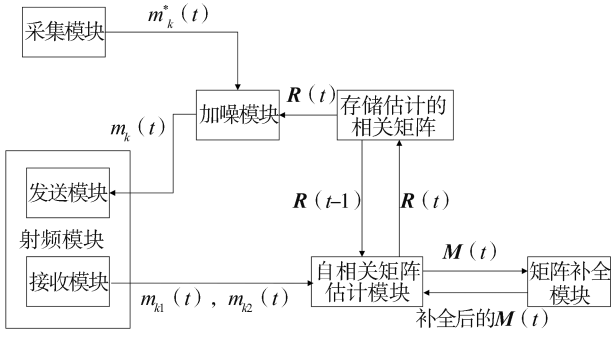

图 2　本文方案流程示意图

图 2 所示的方案可以被智能电网中的每个数据采集节点所采用。首先节点的采集模块采集到原始数据,将其发送到加噪模块准备进行加噪处理;然后接收模块将收集到的其他节点数据发送给自相关矩阵估计模块;接着在自相关矩阵估计模块中构建整个网络相关矩阵,但由于采集的数据不全,所以需要先把当前时刻的矩阵 $\boldsymbol{M}(t)$ 发送到矩阵补全模块;把矩阵中缺失的元素补全后,再返回给自相关矩阵模块中;在自相关矩阵模块中把网络相关矩阵计算出来,并更新估计的相关矩阵;在加噪模块将零均值

噪声映射到数据协方差矩阵的主成分上,生成与原始数据具有相同统计特性的噪声,向原始数据投射相关噪声;最后把加噪后的数据发送出去。

2.2 噪声生成及添加

下面介绍在每个时刻估计整个相关矩阵 $\boldsymbol{R}(t)$,并由 $\boldsymbol{R}(t)$ 生成该方案所需的噪声的过程。

假设信任域 1 由 K 个节点组成,并且第 k 个节点在时刻 t 记录的原始数据为 $m_k^*(t)$,它可以被建模为离散时间广义平稳随机过程。原始数据若想发送给其他节点,则需要进行加噪处理为 $m_k(t) = m_k^*(t) + n_k(t)$。而 $\boldsymbol{n}(t) = [n_1(t),\cdots,n_k(t)]^{\mathrm{T}}$ 是为每个节点添加噪声的向量,其中 $\boldsymbol{n}(t) \sim N(0,\boldsymbol{C}_n)$,$\boldsymbol{C}_n \in \mathbb{R}^{K\times K}$ 是噪声协方差矩阵。因为采集的数据具有高度相关性,所以原始数据协方差矩阵 $\boldsymbol{C} \in \mathbb{R}^{K\times K}$ 是低秩的。

把所有节点在时刻 t 收集到的有噪声数据表示为向量形式,如 $\boldsymbol{m}(t) = [m_1(t),\cdots,m_K(t)]^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$。过程 $\boldsymbol{m}(t)$ 的相关矩阵定义为 $\boldsymbol{C} = E\{\boldsymbol{M}(t)\} \in \mathbb{R}^{K\times K}$,其中 $\boldsymbol{M}(t) = \boldsymbol{m}(t)\boldsymbol{m}(t)^{\mathrm{T}}$。基于平均遍历定理,可以使用指数衰减窗口来估计相关矩阵 $\boldsymbol{C}$

$$\boldsymbol{R}(t) = \frac{t-1}{t}\boldsymbol{R}(t-1) + \frac{1}{t}\boldsymbol{M}(t) = \frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{M}(\tau) \tag{6}$$

其中 $\lim\limits_{t\to\infty}\boldsymbol{R}(t) = \boldsymbol{C}$。

以往在自相关矩阵估计模块中构建相关矩阵,每个节点需要在每个时刻 t 接收 K 个测量值。但由于节点并不拥有所有的测量数据,因此在每个时刻 t,每个节点都需要基于不完整的测量来补全 $\boldsymbol{M}(t)$,具体的补全过程见 2.3 节。

补全完成后,估计出相关矩阵 $\boldsymbol{R}(t)$。通过将生成的向量 $\boldsymbol{n}(t)$ 投影到 $\boldsymbol{R}(t)$ 的估计主成分上,可以在加噪模块处估计相关噪声,即通过 $\boldsymbol{R}(t)$ 的特征值分解:$\boldsymbol{R}(t) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$,其中 $\boldsymbol{U} = [\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_K] \in \mathbb{R}^{K\times K}$ 是由 $\boldsymbol{R}(t)$ 的 K 个特征向量组成,$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1,\cdots,\sigma_K)$ 是特征值按降序排列的对角矩阵。那么所提出的投影可以表示为矩阵形式:$\boldsymbol{n}(t) = \boldsymbol{P}_{\boldsymbol{u}_1}\boldsymbol{n}(t)$,其中 $\boldsymbol{P}_{\boldsymbol{u}_1} = \boldsymbol{u}_1\boldsymbol{u}_1^{\mathrm{T}}$,$\boldsymbol{u}_1$ 是主要特征向量,它对应于最大特征值为 σ_1。

噪声 $\boldsymbol{n}(t)$ 被添加到原始数据中,以生成混淆后的数据。加噪处理具体为

$$m_k(t) = m_k^*(t) + u_{1,k}\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{n}(t) \tag{7}$$

加噪后的 $\boldsymbol{m}(t)$ 的相关矩阵主成分很好地对应于原始数据 $\boldsymbol{m}^*(t)$ 的数据自相关矩阵主成分的估计结果。更重要的是,这种噪声不能被轻易过滤掉,因为它反映了时间序列的主要趋势。

2.3 缺失数据补全

为了从不完整的测量中补全 $\boldsymbol{M}(t)$,本方案采用了一种不精确 ALM 算法[18]。该方法是对 ALM 算法的一种改进算法,它的收敛速度几乎与 ALM 算法一样快,但所需的部分奇异值分解的数目要少得多。

不精确 ALM 算法将矩阵补全问题视为 Robust PCA 问题的特例,将矩阵补全问题建模为如下形式

$$\begin{aligned}&\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_*\\ &\text{s.t. } \boldsymbol{A} + \boldsymbol{E} = \boldsymbol{D}, P_{\Omega}(\boldsymbol{E}) = 0\end{aligned} \tag{8}$$

式中 $\boldsymbol{D},\boldsymbol{A},\boldsymbol{E} \in \mathbb{R}^{m\times n}$,$\boldsymbol{D}$ 为观测矩阵。

最优化问题(8)的部分增广拉格朗日函数为

$$L(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y},u) = \|\boldsymbol{A}\|_* + \langle \boldsymbol{Y},\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}\rangle + \frac{u}{2}\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}\|_{\mathrm{F}}^2 \tag{9}$$

式中正则化参数 $u > 0$。

当 $\boldsymbol{Y} = \boldsymbol{Y}_k$,$u = u_k$ 时,ALM 算法使用交替方法求解优化问题:$\min\limits_{\boldsymbol{A},\boldsymbol{E}} L(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y}_k,u_k)$。然而不精确 ALM 算法改善了 ALM 算法,它不需要求 $\min\limits_{\boldsymbol{A},\boldsymbol{E}} L(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y}_k,u_k)$ 的精确解,即矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{E}$ 的迭代更新公式为

$$\boldsymbol{A}_{k+1} = \arg\min_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{E}_{k+1},\boldsymbol{Y}_k,u_k) = \boldsymbol{D}_{\frac{1}{u_k}}\left(\boldsymbol{D} - \boldsymbol{E}_{k+1} + \frac{\boldsymbol{Y}_k}{u_k}\right) \tag{10}$$

$$\boldsymbol{E}_{k+1} = \arg\min_{\boldsymbol{E}} L(\boldsymbol{A}_{k+1},\boldsymbol{E},\boldsymbol{Y}_k,u_k) = \boldsymbol{S}_{\frac{\lambda}{u_k}}\left(\boldsymbol{D} - \boldsymbol{A}_{k+1} + \frac{\boldsymbol{Y}_k}{u_k}\right) \tag{11}$$

不精确 ALM 算法具体流程为:首先输入不完整的矩阵 $\boldsymbol{D}_{i,j}$,$i,j \in \Omega$,它是 $\boldsymbol{D} \in \mathbb{R}^{m\times n}$ 的部分数据。设置初始值 $\boldsymbol{Y}_0 = 0$,$\boldsymbol{E}_0 = 0$,$u_0 > 0$,$\rho > 1$,$k = 0$。

当不收敛时,循环计算 $\boldsymbol{A}$ 和 $\boldsymbol{E}$ 的迭代更新公式:为了解决 $\boldsymbol{A}_{k+1} = \arg\min\limits_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{E}_{k+1},\boldsymbol{Y}_k,u_k)$ 问题,通过计算 $(\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}) = svd(\boldsymbol{D} - \boldsymbol{E}_k + u_k^{-1}\boldsymbol{Y}_k)$,$\boldsymbol{A}_{k+1} = \boldsymbol{U}\boldsymbol{S}_{u_k^{-1}}[\boldsymbol{S}]\boldsymbol{V}^{\mathrm{T}}$。为了解决 $\boldsymbol{E}_{k+1} = \arg\min\limits_{\pi_{\Omega}(\boldsymbol{E})=0} L(\boldsymbol{A}_{k+1},\boldsymbol{E},\boldsymbol{Y}_k,u_k)$ 问题,通过计算 $\boldsymbol{E}_{k+1} = \pi_{\bar{\Omega}}(\boldsymbol{D} - \boldsymbol{A}_{k+1} + u_k^{-1}\boldsymbol{Y}_k)$。然后计算 $\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k + u_k(\boldsymbol{D} - \boldsymbol{A}_{k+1} - \boldsymbol{E}_{k+1})$。

每次循环还要将 u_k 和 k 分别更新为 u_{k+1} 和 $k+1$。当算法经过足够次迭代之后结果收敛,最后输

出 $\boldsymbol{A}_k$,$\boldsymbol{E}_k$。

3　实验及结果分析

本次实验分析了本文方案在未授权重构的情况下保护电力数据隐私的效果。通过与现有的方案作对比,分析本文方案在引入噪声后的数据可用性,即噪声是否会对数据产生较大影响。最后分析本文方案的通信开销。

3.1　实验设置

实验采用模拟数据集对本文隐私保护方案的效果进行评估。模拟数据集模拟智能电网中节点采集到的电力数据,该电力数据集由矩阵 $\boldsymbol{D} \in \mathbb{R}^{K\times t}$ 来表示,它是由 $\boldsymbol{m}(t) = [m_1(t),\cdots,m_K(t)]^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$ 构成的。尽管测量的数据值受到多种因素(例如节点的功耗、数据包的丢失、噪声的存在、恶意节点)影响而产生随机偏差,但是大量的真实数据仍然会满足一定的分布规律。所以规定该电力数据集中的数据具有高度相关性,这也是本次实验的前提。

由于本方案所生成的相关噪声并不固定,所以每组实验均进行 100 次,记录其平均值。

3.2　隐私保护效果分析

3.2.1　隐私保护效果度量

假如不同信任域中的节点想要在未经授权的情况下获取其他信任域节点记录的原始数据,它们可以使用线性过滤操作。重构原始数据 $\boldsymbol{M}^*(t)$ 的操作可以表示为:$\widetilde{\boldsymbol{M}}^*(t) = \boldsymbol{F}\boldsymbol{M}(t)$,即使用低通滤波器来去除噪声。

有一种可实现去噪的方法是将数据投影到信号主成分的子空间上,以便在保留原始数据的同时消除大部分噪声,即使用 PCA 技术来进行数据重构。假设 $\boldsymbol{M}^* = [\boldsymbol{m}^*(1),\cdots,\boldsymbol{m}^*(t)]^{\mathrm{T}} \in \mathbb{R}^{K\times K}$ 是原始数据,$\boldsymbol{M} = [\boldsymbol{m}(1),\cdots,\boldsymbol{m}(t)]^{\mathrm{T}} \in \mathbb{R}^{K\times K}$ 是加噪后的数据。不同信任域通过使用 PCA 方法来去除噪声,来对 $\boldsymbol{M}$ 进行未授权的重构处理,得到重构数据 $\boldsymbol{M}_h$:$\boldsymbol{M}_h = \boldsymbol{U}_k(t)\boldsymbol{U}_k^{\mathrm{T}}(t)\boldsymbol{M}$。这里的 $\boldsymbol{U}_k(t)$ 对应于一个具有自相关矩阵 $\boldsymbol{R}(t)$ 的 K 个主要特征向量的 $K*K$ 矩阵。

为了能准确分析该方案的效果,通过使用 Frobenius 范数(即 F 范数)来比较原始数据与重构数据之间的隐私差异。$\|\cdot\|_{\mathrm{F}}$ 是矩阵数据的 F 范数,它是对矩阵元素的平方和再开平方。设 $\boldsymbol{A}$ 是一个 $m*n$ 矩阵,则 $\boldsymbol{A}$ 的 F 范数为:$\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$。

为了不失一般性,参考文献[19-20],本文实验采用一种衡量隐私保护效果的指标 PD:$\mathrm{PD} = \|\boldsymbol{M}^* - \boldsymbol{M}_h\|_{\mathrm{F}}^2 / \|\boldsymbol{M}^* - \boldsymbol{M}\|_{\mathrm{F}}^2$。其中 $\|\cdot\|_{\mathrm{F}}^2$ 是 F 范数的平方。若整个 PD 越接近 0,说明恶意重构数据越接近原始数据,方案的隐私保护效果越不好。反之,PD 越大,方案的隐私保护效果越好。

3.2.2　节点的连通数对隐私性的影响

本组实验研究节点连通数对隐私性的影响,采用固定节点数为 50 的模拟数据集进行实验,通过改变节点连通其他节点的个数,来判断节点间连通数对本方案隐私性的影响。

图 3 中横坐标轴代表单个节点的连通数,纵坐标轴代表 PD 值。本组实验分别计算本方案的隐私保护 PD 值和添加了加性噪声的隐私保护 PD 值。从图中可知,采用本文方案的 PD 值随着节点的连通个数的增加有轻微的上升趋势,说明节点连通数增加会提高隐私保护效果。本文方案的 PD 值都在 0.8 以上,而添加了加性噪声的 PD 值很小,一直在 0.2 附近。这是因为加性噪声机制本身的限制,噪声很容易被一些线性滤波器抵消,导致隐私保护效果下降。而本文方案添加的是与原始数据具有相同统计特性的噪声,因此很难被滤波攻击消除。这说明该方案的隐私保护效果远好于添加加性噪声。

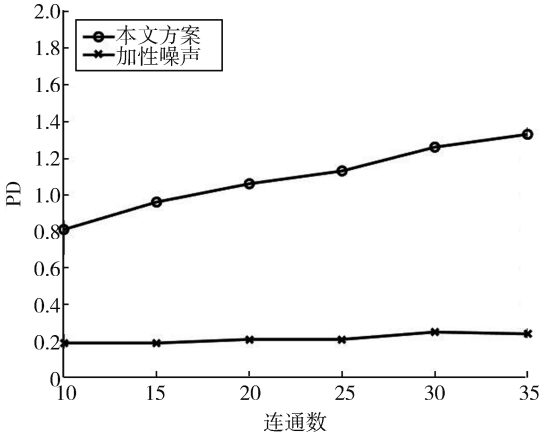

图 3　节点连通其他节点的个数对隐私性的影响

3.2.3　节点个数对隐私性的影响

本组实验研究节点个数对隐私性的影响,采用固定连通率(即每个节点连通的节点数与总节点数的比值)为 25% 的模拟数据集进行实验,通过改变节点个数,来判断节点个数对本方案隐私性的影响。

图 4 中横坐标轴代表一个信任域中的节点数,纵坐标轴代表 PD 值。本组实验分别计算本方案的隐私保护 PD 值和添加了加性噪声的隐私保护 PD

值。从图中可知,随着节点数的增加,本方案的 PD 值仍保持一个稳定的数值,说明节点数不同不会影响该方案的隐私保护效果。这意味着可以将此方案应用于大规模的智能电网场景下,同时基本没有影响隐私保护的效果。而添加了加性噪声的数据在隐私保护方面的效果仍不如本方案,这与预期结果一致,说明了此隐私保护方案是有效的。

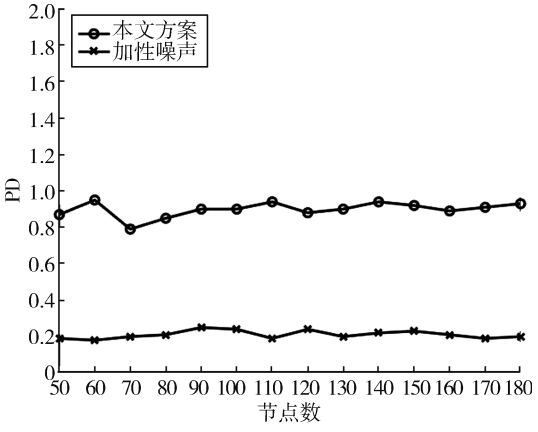

图 4 节点个数对隐私性的影响

3.3 数据可用性分析

3.3.1 数据可用性度量

为保护数据的隐私,为原始数据添加了噪声,但同时也引入了误差。本节将详细讨论引入误差的大小。文献[21]中使用的一种常用的测量误差的方法是 $\text{Error} = \| \boldsymbol{M} - \boldsymbol{M}^* \|_F$,其中 $\boldsymbol{M}$ 和 $\boldsymbol{M}^*$ 分别是扰动后的数据和原始数据,该值代表扰动后数据和原始数据之间的误差,Error 越小,数据可用性越高。

3.3.2 不同方案的数据可用性比较

为了验证本文所提出方案的数据可用性,本文分别与基于输出扰动的 PCA 隐私保护算法(PCA-Privacy-Preserving based on Output Perturbation,PCA-PP-based-OP)[21]和基于 Kalman 过滤的隐私保护算法[22]进行对比。

本组的 3 个实验分别使用的是节点数为 50、80、100 的模拟数据集,节点连通率均为 25%,即连通数分别为 13、20、25。实验通过比较 3 种隐私保护方案的近似误差大小,判断数据可用性的优劣。实验结果分别如图 5、图 6 及图 7 所示。

通过分析每一组实验的结果,可以发现随着时间增大,3 种方案的数据误差都呈上升趋势,但是在节点数相同的情况下,使用基于 Kalman 过滤的隐私保护算法和 PCA-PP-based-OP 算法的数据误差都远远大于本文方案,这说明本文方案所添加的噪声要少于其他 2 种方案添加的噪声。这主要是因为其他 2 种方案是向原始数据中添加随机 Laplace 噪声,引入额外误差的可能性更大,而本文方案添加的噪声与原始数据流具有相同的统计特性,不会造成过大的误差。因此本文所提出方案的数据可用性明显高于基于 Kalman 过滤的隐私保护算法和 PCA-PP-based-OP 算法。

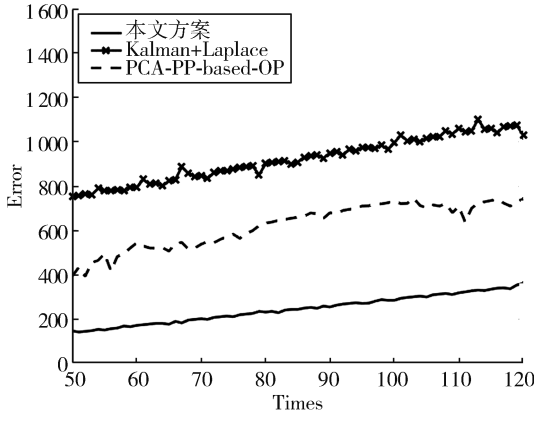

图 5 本文方案与其他算法的数据可用性对比(节点数为 50)

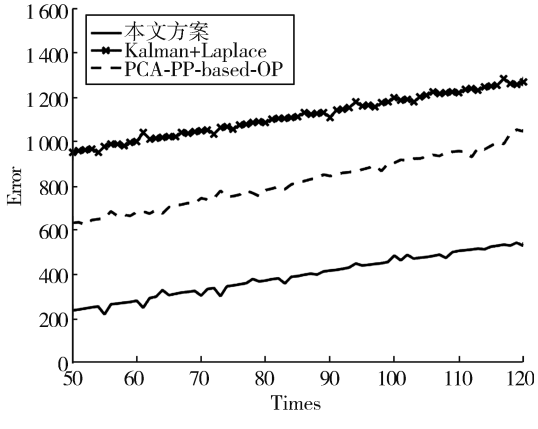

图 6 本文方案与其他算法的数据可用性对比(节点数为 80)

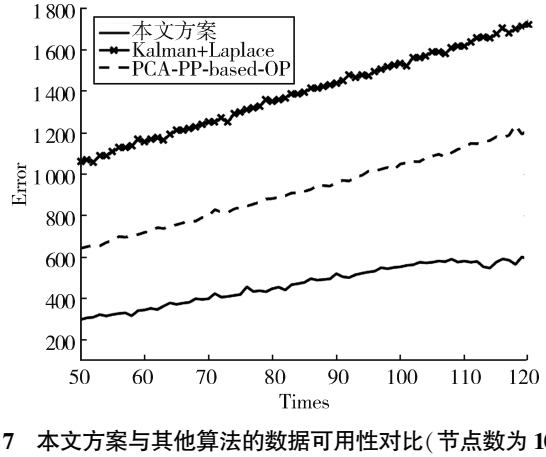

图 7 本文方案与其他算法的数据可用性对比(节点数为 100)

通过这 3 组节点数不同的实验对比可知,随着网络中的节点数越多,这 3 种方案的数据近似误差

都在增加。在节点数为 50 和 80 的实验中,这 3 种方案的近似误差增长趋势基本一致。但是在节点数为 100 的实验中,用于对比的 2 种方案的数据误差增长趋势要高于本文方案,这说明在较大规模的智能电网中这 2 种方案所使用的随机噪声会引入更大的误差,而本文方案添加的相关噪声所造成的误差仍接近于在较小规模网络中的误差。根据上述分析,本文方案在不同规模的智能电网中都有着不错的数据可用性。

3.4　通信开销分析

本文方案中若不使用矩阵补全,则需要节点之间频繁的通信。节点的通信开销主要由发送开销和接收开销构成。当一个节点的任意一个邻居节点发送数据包时,该节点都会监听到并接收该数据包,所以每个节点接收数据包消耗的能量的上界是一个与该节点邻居节点数目相关的常量。每个节点的发送开销一般是由所发送的数据包个数决定。

假设在智能电网环境中的数据采集节点采用较为常用的 CC2530 通信芯片,节点间的连通率为 25%,并且由于各种因素导致的丢包率会随着网络规模变大而变大。根据文献[23]测量所得的数据,一次发送的能耗为 0.008 59 mJ,一次接收的能耗为 0.037 316 mJ。本方案与不使用矩阵补全的方案的通信开销对比如图 8 所示。

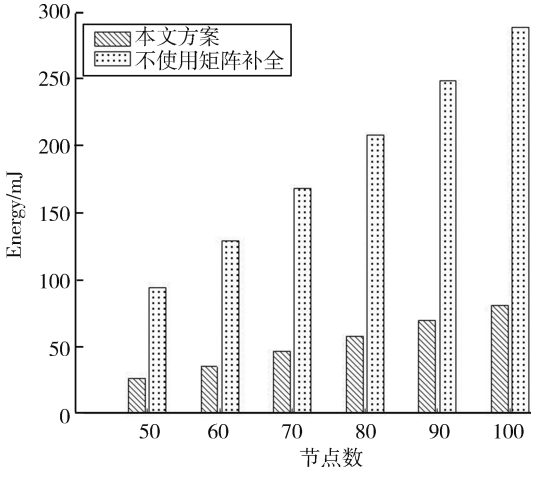

图 8　本文方案与不使用矩阵补全方案的通信开销对比

由此可见,本文方案的网络通信开销远低于不使用矩阵补全的方案。随着网络规模变大,不使用矩阵补全的方案消耗的能量越多,而本文方案可以大大减少通信产生的开销,并降低由于特殊情况导致的丢包问题所带来的影响。

4　结束语

本文提出了一种基于矩阵补全的智能电网数据隐私保护方案,用于解决智能电网中不同信任域之间的数据隐私保护问题。该方案通过给原始数据添加难以去除的噪声来避免数据被恶意重构。同时为了补全因特殊情况丢失的数据,本方案利用不精确的拉格朗日乘子算法进行相关矩阵的补全。最后的仿真实验进一步验证了本方案具有不错的隐私保护效果和数据可用性,并且通信开销较低。

参考文献:

[1] 张晶, 代攀, 吴天京, 等. 新一代智能电网技术标准体系架构设计及需求分析[J]. 电力系统自动化, 2020, 44(9): 12-20.
ZHANG Jing, DAI Pan, WU Tianjing, et al. Architecture design and demand analysis of new generation technical standard system for smart grid[J]. Automation of Electric Power Systems, 2020, 44(9): 12-20.(in Chinese)

[2] DILEEP G. A survey on smart grid technologies and applications[J]. Renewable Energy, 2020, 146: 2589-2625.

[3] LIU Y N, GUO W, FAN C N, et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid[J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1767-1774.

[4] 陈思光, 杨熠, 黄黎明, 等. 基于雾计算的智能电网安全与隐私保护数据聚合研究[J]. 南京邮电大学学报(自然科学版), 2019, 39(6): 62-72.
CHEN Siguang, YANG Yi, HUANG Liming, et al. Fog computing based secure and privacy-aware data aggregation in smart grid[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019, 39(6): 62-72.(in Chinese)

[5] SALIM M M, KIM I, DONIYOR U, et al. Homomorphic encryption based privacy-preservation for IoMT[J]. Applied Sciences, 2021, 11(18): 8757.

[6] SINGH P, MASUD M, HOSSAIN M S, et al. Blockchain and homomorphic encryption-based privacy-preserving data aggregation model in smart grid[J]. Computers & Electrical Engineering, 2021, 93: 107209.

[7] CHENG Z Y, YE F, CAO X H, et al. A homomorphic encryption-based private collaborative distributed energy management system[J]. IEEE Transactions on Smart Grid, 2021, 12(6): 5233-5243.

[8] LU Y, TIAN Y, ZHOU T S, et al. Multicenter privacy-preserving cox analysis based on homomorphic encryption[J]. IEEE Journal of Biomedical and Health Informatics,

2021, 25(9): 3310-3320.

[9] KHAN H M, KHAN A, JABEEN F, et al. Fog-enabled secure multiparty computation based aggregation scheme in smart grid[J]. Computers & Electrical Engineering, 2021, 94: 107358.

[10] BHATTACHARJEE A, BADSHA S, SENGUPTA S. Personalized privacy preservation for smart grid[C]// IEEE International Smart Cities Conference.2021.

[11] 曹辉. 智能电网本地化差分隐私保护研究[D].武汉:武汉大学,2020.
CAO Hui. Research on local differential privacy protection of smart grid[D]. Wuhan: Wuhan University, 2020. (in Chinese)

[12] 王创. 智能电网中 WSN 能耗优化机制与覆盖算法研究[D]. 淮南: 安徽理工大学, 2021.
WANG Chuang. Research on WSN energy consumption optimization mechanism and coverage algorithm in smart grid[D]. Huainan: Anhui University of Science & Technology, 2021. (in Chinese)

[13] 李培冠, 於志勇, 黄昉菀. 基于稀疏表示的电力负荷数据补全[J]. 计算机科学, 2021, 48(2): 128-133.
LI Peiguan, YU Zhiyong, HUANG Fangwan. Power load data completion based on sparse representation[J]. Computer Science, 2021, 48(2): 128-133.(in Chinese)

[14] GENES C, ESNAOLA I, PERLAZA S M, et al. Recovering missing data via matrix completion in electricity distribution systems[C]// IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications. 2016: 1-6.

[15] 陈蕾, 陈松灿. 矩阵补全模型及其算法研究综述[J]. 软件学报, 2017, 28(6): 1547-1564.
CHEN Lei, CHEN Songcan. Survey on matrix completion models and algorithms[J]. Journal of Software, 2017, 28(6): 1547-1564.(in Chinese)

[16] ZHU Z H, LI Q W, TANG G G, et al. Global optimality in low-rank matrix optimization[J]. IEEE Transactions on Signal Processing, 2018, 66(13): 3614-3628.

[17] SHLENS J. A tutorial on principal component analysis [EB/OL]. [2021-09-20]. https: // arxiv. org/abs/1404.1100.

[18] LIN Z C, CHEN M M, MA Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices[EB/OL]. [2021-09-20].https: // arxiv.org/abs/1009.5055.

[19] LI F F, SUN J M, PAPADIMITRIOU S, et al. Hiding in the crowd: privacy preservation on evolving streams through correlation tracking[C]// IEEE 23rd International Conference on Data Engineering. 2007: 686-695.

[20] 李国瑞, 王颖, 王聪. 一种基于矩阵补全的无线传感网数据收集方案[J]. 电子学报, 2018, 46(12): 2950-2956.
LI Guorui, WANG Ying, WANG Cong. A matrix completion based data collection scheme in wireless sensor networks[J]. Acta Electronica Sinica, 2018, 46(12): 2950-2956.(in Chinese)

[21] 徐亚红. 面向主成分分析的差分隐私数据发布算法[D]. 南京: 南京邮电大学, 2020.
XU Yahong. Research on principal component analysis algorithm under differential privacy[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[22] 涂子璇, 刘树波, 熊星星, 等. 可穿戴设备的数值型流数据差分隐私均值发布[J]. 计算机应用, 2020, 40(6): 1692-1697.
TU Zixuan, LIU Shubo, XIONG Xingxing, et al. Differential private average publishing of numerical stream data for wearable devices [J]. Journal of Computer Applications, 2020, 40(6): 1692-1697.(in Chinese)

[23] 杜永文, 练云翔, 冯珂. 基于 TinyOS 的传感器节点能耗仿真研究[J]. 自动化仪表, 2018, 39(1): 96-98, 102.
DU Yongwen, LIAN Yunxiang, FENG Ke. Research on the energy consumption simulation of sensor nodes based on TinyOS [J]. Process Automation Instrumentation, 2018, 39(1): 96-98, 102.(in Chinese)

(责任编辑:李小溪)