文章编号:1671-4598(2023)11-0221-07 DOI:10.16526/j.cnki.11—4762/tp.2023.11.033 中图分类号:TP309 文献标识码:A

基于属性分类的分布式大数据隐私保护加密控制模型设计

姜春峰
(西北大学现代学院,西安 710130)

摘要:在分布式大数据的存储和传输过程中,数据极易被恶意用户攻击,造成数据的泄露和丢失;为提高分布式大数据的存储和传输安全性,设计了基于属性分类的分布式大数据隐私保护加密控制模型;挖掘用户隐私数据,以分布式结构存储;根据分布式隐私数据特征,判断数据的属性类型;利用Logistic混沌映射,迭代生成数据隐私保护密钥,通过匿名化、混沌映射、同态加密等步骤,实现对隐私数据的加密处理;利用属性分类技术,控制隐私保护数据访问进程,在传输协议的约束下,实现分布式大数据隐私保护加密控制;实验结果表明,设计模型的明文和密文相似度较低,访问撤销控制准确率高达98.9%,在有、无攻击工况下,隐私数据损失量较少,具有较好的加密、控制性能和隐私保护效果,有效降低了隐私数据的泄露风险,提高了分布式大数据的存储和传输安全性。

关键词:属性分类;分布式大数据;隐私保护;加密控制模型;Logistic混沌映射

Design of Distributed Big Data Privacy Protection Encryption Control Model Based on Attribute Classification

JIANG Chunfeng
(Modern College of Northwest University, Xi'an 710130, China)

Abstract: In the storage and transmission process of distributed big data, data is highly susceptible to malicious user attacks, resulting in data leakage and loss. To improve the storage and transmission security of distributed big data, a distributed big data privacy protection encryption control model based on attribute classification was designed. Mine user privacy data, and store it in a distributed structure. Determine the attribute type of the distributed privacy data based on its characteristics. Using Logistic chaotic mapping, the data privacy protection key is generated iteratively, and the encryption of private data is realized through the anonymization, chaotic mapping, homomorphic encryption and other steps. Classification technology is used to control the access process of privacy protection data, and achieve the distributed big data privacy protection encryption control under the constraints of transmission protocols. The experimental results show that the similarity between plaintext and ciphertext in the designed model is low, and the accuracy of access revocation control reaches by 98.9%. Under the attack and no attack conditions, the loss of privacy data is relatively small, and it has good encryption, control performance, and privacy protection effects. It effectively reduces the risk of privacy data leakage and improves the storage and transmission security of distributed big data.

Keywords: attribute classification; distributed big data; privacy protection; encryption control model; logistic chaotic mapping

0 引言

目前大数据迅速发展,对于提升新一代信息技术和服务业态具有重要作用。大数据需要分布式的体系结构,而分布式的数据挖掘离不开云存储和虚拟化的支持[1]。分布式是指通过可扩展的体系架构,将海量数据分散到多个独立的机器上,利用多台存储服务器分担存储负荷的计算技术。分布式大数据已经应用到多个行业和领域,然而在分布式大数据的存储与传输过程中,数据极易受到恶意用户的攻击,导致数据出现泄露、丢失等现象,尤其是分布式隐私数据[2]。隐私保护是指对个人信息进行保护,防止被非法获取、利用、传播等,以确保个人信息的安全和私密性。数据加密是指将明文数据通过一定的方法和密钥进行处理,转换成特定的加密形式,从而达到保护数据机密性的目的。因此,对数据隐私保护加密控制具有重要意义。

文献[3]提出了基于同态加密的DBSCAN聚类隐私保护方案。根据数据特点,考虑数据精度和计算开销,选择数据预处理策略,并依据用户端与云服务器之间的协议,完成密文比较。该方法具有较低的时间开销。文献[4]提出了基于可搜索加密机制的数据库加密方案。构建密态数据库查询框架,提出了满足IND-CKA1安全的数据库加密

收稿日期:2023-04-23; **修回日期**:2023-06-01。
基金项目:国家自然科学基金项目(12345678)。
作者简介:姜春峰(1968-),女,硕士,副教授。
引用格式:姜春峰.基于属性分类的分布式大数据隐私保护加密控制模型设计[J].计算机测量与控制,2023,31(11):221-227.

方案。建立可搜索加密方案中的安全索引，获取密态数据库安全索引结构，对数据库中的数据进行加密。该方法具有一定的有效性。文献［5］研究了基于差分隐私的医疗大数据隐私保护模型。在目前医疗大数据隐私保护技术的基础上，阐述差分隐私保护技术的基本原理和研究方向，构建医疗大数据隐私保护模型，实现医疗大数据隐私保护。该模型具有一定的可行性。文献［6］提出了基于大数据的区块链数据隐私文本智能加密方法。依据文本加密需求，对区块链文本进行预处理，将格栅化大数据加密作为目标，在大数据技术下建立多层级文本加密模型，实现区块链数据文本加密。该方法具有较大的实际应用价值。文献［7］提出了基于数据消冗技术的隐私大数据属性加密方法。采用 Bloom 过滤技术，降维处理大数据，使用 hash 函数对消冗过程中误判率进行计算，依据映射位数组，对最优扩列函数进行确定，优化 ABE 加密算法，完成云数据的安全共享访问。该方法具有较高的实用性。但上述方法仍存在加密和控制效果较差的问题。

属性分类指的是划分分布式大数据属性类型的过程，通过对分布式大数据属性的分类，可以确定大数据的隐私保护等级，从而生成具有针对性的数据加密方案。为此，设计了基于属性分类的分布式大数据隐私保护加密控制模型，挖掘分布式隐私数据，划分分布式隐私数据属性类型，利用 Logistic 混沌映射，生成数据隐私保护密钥，加密处理隐私数据。采用属性分类技术，控制隐私保护数据访问进程，根据传输协议约束，实现分布式大数据隐私保护加密控制，以期能够提高分布式大数据的存储与传输安全。

1 分布式大数据隐私保护加密控制模型设计

设计的分布式大数据隐私保护加密控制模型，通过数据加密和访问控制两个步骤，实现保护大数据隐私的目的。在分布式大数据加密过程中，首先利用属性分类算法，确定数据的隐私保护等级，从而选择相应的加密强度和方式。在此基础上，根据分布式隐私数据属性的分类结果，设置用户的访问权限，实现数据的访问控制。设计模型的输入项为分布式大数据，输出项为隐私数据的保护加密结果以及访问控制指令，在模型运行过程中，设置属性类型以及控制协议作为约束条件。

1.1 挖掘分布式隐私数据

隐私数据主要包括个人生活安宁权、个人生活情报保密权、个人通信保密权等，以用户在网络环境中的通信信息为研究对象。挖掘的隐私数据以分布式结构进行存储，分布式存储结构由设备、虚拟、运营和业务 4 个层次组成。设备层为加密、防火墙等安全技术提供硬件支持。通过对大数据的分析，为用户提供从终端到网络的全方位的安全保障。虚拟层的作用是对移动终端进行虚拟化，以保证用户的身份认证，虚拟化应用程序的安全性，以及对移动终端的日志进行审核。操作层可以利用虚拟层的功能，来管理虚拟装置，并进行虚拟服务的配置，从而达到对用户和访问的控制。而运营层提供了基于运营层次的各类存储和计算等云计算服务。并承担数据加密，文件加密，病毒检测，用户身份验证，访问控制等功能。隐私大数据分布式存储结构如图 1 所示。

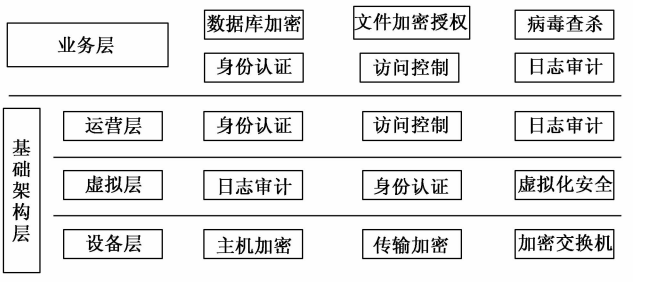

图 1 隐私大数据分布式存储结构图

以挖掘的隐私大数据为处理对象，首先需要对初始隐私数据进行归一化处理，处理结果如下：

$$x_g = \frac{x_0 - \mu}{\sigma} \tag{1}$$

其中：x_0 为初始挖掘的隐私数据，μ 和 σ 对应的是隐私数据集的均值和方差。在此基础上，对隐私数据进行分片处理。每一个分区都由至少一个服务器组成，利用内部包含的辅助节点实现异步主从式拷贝机制。其中，主节点以读为主，子节点以写为主。在这两种类型的节点之间，利用操作日志来确保数据的一致性。在这两种类型的节点之间，会将所有的操作数据和时间戳都写到操作日志中，而所有的副节点都会对操作日志进行监视，以便与主节点进行同步。当任意一个集群服务器发生故障时，分布式结构中的隐私大数据就会自动变成只读状态，这样可以有效地避免存储结构在不稳定时，由于误操作而造成初始数据信息被更改，也可以避免在配置服务器节点之间出现数据不一致。当某个配置服务器出现故障时，并不会影响到整体存储结构的运行，并可保证挖掘的数据能够成功写入到分布式存储结构中。

1.2 划分分布式隐私数据属性类型

根据分布式隐私数据特征，判断数据的属性类型。属性分类技术的基本运行流程如图 2 所示。

从图 2 中可以看出，属性分类可以分为 5 个环节，首先利用公式（2）计算属性的敏感系数。

$$\kappa_{\text{sensitive}} = \frac{\eta_{\text{change}}}{\eta_{\text{Uncertainties}}} \tag{2}$$

其中：η_{change} 和 $\eta_{\text{Uncertainties}}$ 分别表示的是隐私数据变化率及其内部不确定因素变化率[5]。根据公式（2）的计算结果，确定当前数据属性是否为敏感属性，若判定数据属性为敏感属性，则可以直接输出分类结果，否则需要提取数据信息熵、信息增益、最大信息系数等特征向量，其中，信息熵和信息增益特征的提取结果如下：

$$\begin{cases} \tau_{\text{Information entropy}} = \sum_{i=1}^{n_{\text{conceal}}} \left(P_i \lg \left(\frac{1}{P_i} \right) \right) \\ \tau_{\text{gain}} = W(X) - W(X|S) \end{cases} \tag{3}$$

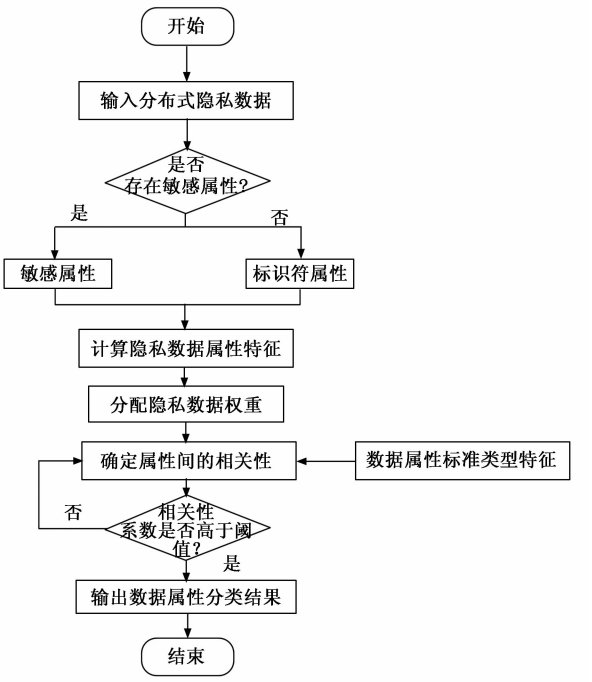

图 2　属性分类技术流程图

其中：n_{conceal}为挖掘的隐私数据量，P_i 为信息属于任意属性的概率，$W(X)$ 和 $W(X \mid S)$ 分别对应的是隐私数据样本集合量以及存在 S 属性的样本集合量[6]。同理可以得出隐私数据中其他特征的提取结果。属性对于分类的重要性，可以根据信息增益值来排序，各个属性权重的计算公式：

$$\omega_i = e^{\tau_{\text{gain}}(i)} \tag{4}$$

将公式（3）计算得出的信息增益特征提取结果代入到公式（4）中，即可得出属性权重的计算结果。根据各个属性权重的计算结果，对提取的隐私数据特征进行融合处理，处理过程如下：

$$\tau = \sum \omega_i \cdot \tau_i \tag{5}$$

式中，τ_i 为提取的特征分量。在此基础上，利用公式（6）计算隐私数据特征与标准属性类型特征之间的相关性系数：

$$\lambda = \frac{\tau \cdot \tau_{\text{standard}}}{\|\tau\| \cdot \|\tau_{\text{standard}}\|} \tag{6}$$

其中：τ_{standard}为设置的数据属性标准特征，τ 为提取的隐私数据融合特征，可通过公式（5）计算得出。若公式（6）的计算结果高于阈值 λ_0，则将对应隐私数据与 τ_{standard}对应类型划分成一类，否则需要更换标准属性类型特征，并重新计算相关性系数，直至得出满足阈值条件的属性类型为止[7]。按照上述流程，对挖掘的所有分布式隐私数据进行属性分类，完成分布式隐私数据属性类型的划分工作。

1.3　生成数据隐私保护密钥

分布式数据隐私保护密钥由公钥和私钥两部分组成，利用 Logistic 混沌映射[8-10]迭代产生的二值序列经过 y_{Sign}（）阈值函数转换为二进制序列，输出结果即为数据的隐私保护公钥，其中 y_{Sign}（）阈值函数表达式如下：

$$y_{\text{Sign}}(x_n) = \begin{cases} 0 & 0 \leqslant x_n < 0.5 \\ 1 & 0.5 \leqslant x_n < 1 \end{cases} \tag{7}$$

其中：x_n 为待加密的分布式隐私数据。那么生成的隐私保护公钥可以量化表示为：

$$\begin{cases} k_{\text{pub}} = (n, m) \\ m = y_{\text{Sign}}(x_n) \end{cases} \tag{8}$$

其中：n 为加密数据编号，那么在已知加密数据的情况下，利用公钥生成对应私钥，即：

$$k_{\text{pri}} = m^{x_n} \beta^n \bmod n^2 \tag{9}$$

式中，β 为随机数。将生成的数据隐私保护公钥分配给多个通过身份认证的用户，而私钥主要通过定点传输的方式分配给指定用户[11]。

1.4　分布式数据隐私保护加密处理

以挖掘的分布式隐私大数据为处理对象，考虑隐私数据属性的分类结果，通过匿名化、混沌映射以及同态加密等步骤，实现对隐私数据的加密处理，进而实现对隐私数据的保护。

1.4.1　隐私数据匿名化

匿名算法的主要工作是遍历由属性泛化层构成的泛化格，按照预定义的属性泛化层次以及节点中各属性所选择的泛化层次等级，对其进行遍历，并检验每个节点是否符合对应的匿名原则，再以当前节点的匿名状态为依据，对整个搜索空间上进行修剪，在优化设计的加密控制模型中，使用 k-匿名原则对节点进行验证[12]。在隐私数据匿名化处理过程中，首先需要对隐私数据进行泛化切片处理，处理结果如下：

$$f(x):\{c\} \to s \tag{10}$$

其中：$f(x)$ 为隐私数据 x 对应的泛化切片结果，c 表示属性分类树中拥有相同父节点的泛化树兄弟节点的最大集合，s 为隐私数据对应的属性值。对泛化后的数据记录进行处理，看当前泛化的记录是否符合匿名化阈值，若符合则该泛化切片为可行的泛化结果。当匿名化阈值不符合时，那么这个泛化就不能被采用[13]。针对通过匿名验证的隐私数据节点，执行匿名化转换操作。匿名化转换操作包括 3 个部分，路径寻找算法负责建立一条由未被标记的节点构成的路径，路径审核算法负责审查建立路径上的节点是否满足匿名原则，从而找到一个符合条件的最优解决方案，最后一部分是算法整体的外部循环，负责遍历整个泛化格结构，找出既符合相对匿名原则，又信息损失较小的节点[14]。第 0 个泛化等级开始遍历表格中的所有泛化层次，并对各个层级的节点进行列举，若节点未被标注，就对其进行路径搜索，将其添加到该路径中，然后进行下一步的检查；在对发现的路径执行完毕路径检查算法后，对于检查路径中不满足 k-匿名的节点，从底层节点出发，依次对后继节点进行处理，为未标记的后继节点建立新的路径，并对新的路径进行检查。最终得出的隐私数据匿名化处理结果如下：

$$x_{\text{anonymous}} = \kappa_{\text{intensity}} \cdot A_{\text{anonymous}} \cdot f(x) \tag{11}$$

其中：$\kappa_{\text{intensity}}$为匿名强度，$A_{\text{anonymous}}$为匿名转换矩阵。重复上述操作，得出分布式隐私大数据的匿名化处理结果。

1.4.2 混沌映射

混沌映射是分布式隐私数据的一种加密处理方式，利用 Logistic 混沌映射原理，得出的分布式隐私数据的处理结果为：

$$x_{\text{chaos}} = x_i(\alpha - \delta x_i) \tag{12}$$

式中，α 为分布式隐私数据的增加率，δ 为考虑外界因素的映射关系饱和度[15]。将挖掘的分布式隐私数据代入公式（12）中，即可得出混沌映射处理结果。

1.4.3 同态加密

同态加密将分布式隐私数据从可读模式转换成编码模式，经过同态加密的隐私数据智能在解密后进行读取或处理。数据的同态加密原理如图 3 所示。

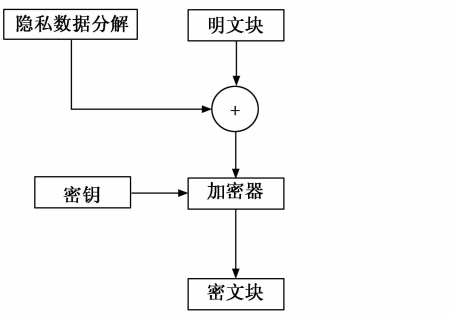

图 3 分布式隐私数据同态加密原理图

按照图 3 表示加密原理，利用生成的数据隐私保护密钥执行隐私数据的加密操作。在开始加密之前，首先对数据进行分解处理，首先确定其上的所有极值点，用 3 次样条曲线连接所有极大值点形成下包络线，将其标记为 b_1，则隐私数据的一级分解结果为：

$$x_{\text{resolve}}(1) = x - b_1 \tag{13}$$

重复上述步骤，直到满足同态加密操作对数据的分解精度条件。在此基础上，利用生成的密钥通过正向或反向得出隐私数据的加密结果，可以量化表示为：

$$\begin{cases} M_{\text{straight}} = x_{n-1} \oplus \kappa_{\text{encryption}} \oplus k(x_n) \bmod 127 \\ M_{\text{counter}} = x_{n+1} \oplus \kappa_{\text{encryption}} \oplus k(x_n) \bmod 127 \end{cases} \tag{14}$$

其中：$\kappa_{\text{encryption}}$为隐私数据的加密强度，$k(x_n)$ 为隐私数据密钥，融合了公钥和私钥两个部分，x_{n-1} 和 x_{n+1} 为隐私保护数据的前后两个相邻数据[16]。在实际的隐私数据同态加密过程中，根据数据属性的分类结果确定参数 $\kappa_{\text{encryption}}$ 的具体取值。由此完成分布式数据的隐私保护加密处理。

1.5 利用属性分类技术控制隐私保护数据访问进程

分布式大数据隐私保护架构涉及主体包括隐私数据主体、服务运营商以及数据访问用户 3 种角色，其中隐私数据主体也就是分布式大数据的产生终端，该角色能够生成隐私保护数据，服务运营商能够为隐私数据的保护提供技术支持，实现访问用户的隐私保护[17]。而数据访问用户也就是申请访问、查询、调用隐私数据的用户终端，可通过对用户身份的验证，对数据访问用户赋予相应的访问权限。在分布式大数据环境下，考虑隐私保护数据属性的分类结果，根据用户角色类型对其访问权限进行初始化，按照上述加密方式生成公钥和私钥。在访问过程中，用户向分布式大数据端发送访问申请，通过身份认证的用户可以直接获得公钥，而满足隐私数据访问属性类型条件的用户能够获得私钥，用户利用公钥和私钥直接实现对隐私数据的访问[18-20]。在用户访问过程中，若检测出异常操作，则执行属性访问撤销操作，定义用户访问行为标识符为 I_{u_i}，则属性访问撤销指令可以表示为：

$$c = < I_{u_1}, k(u_1), s >, \cdots, < I_{u_h}, k(u_h), s > \tag{15}$$

其中：h 为用户访问过程中产生的标识符数量，$k(u_i)$ 代表的是用户在执行第 i 个访问环节使用的密钥，s 为属性分类结果[21-22]。在已知当前访问操作信息的情况下，可以推算出初始操作信息，从而恢复出 I_{u_0}，分布式大数据管理终端利用恢复的 I_{u_0} 更新密钥和用户隐私保护数据的加密，而撤销属性权限的用户因无法获取新密钥不能继续访问接下来的信息，从而在功能上实现对该用户访问权限的撤销。

1.6 实现分布式大数据隐私保护加密控制

上述分布式大数据隐私保护加密控制模型的运行需要加密策略安全传输协议作为约束条件，具体的协议运行时序如图 4 所示。

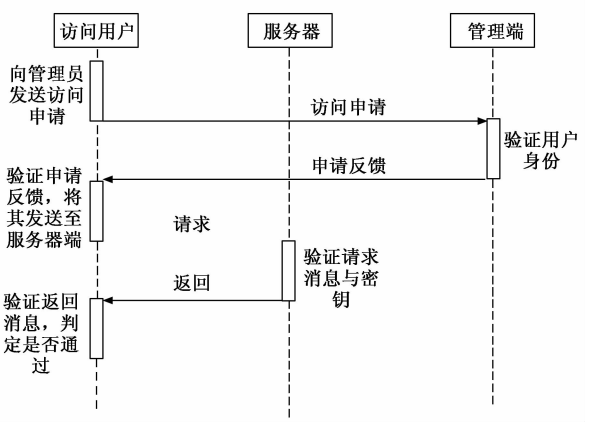

图 4 隐私保护加密传输协议运行时序图

在图 4 表示协议的约束下，合法用户可以在分布式大数据环境中进行隐私数据的缓存，同一用户无需重复执行身份验证程序，当用户退出系统终止缓存程序[23]。在传输协议的约束下，对实时产生的隐私数据进行加密与访问控制，实现模型的隐私保护功能。

2 模型验证实验分析

为了验证设计的基于属性分类的分布式大数据隐私保护加密控制模型的有效性，进行模型验证实验分析。此次

实验分为 3 个部分，分别为加密效果测试、控制效果测试以及隐私保护效果测试。其中，加密效果测试原理是验证密文与明文之间的相似度，相似度越高证明加密效果越差，无法达到隐私隐藏与保护效果。控制效果测试的目的是验证模型是否能够实现对非法用户、无授权用户的访问拦截，从而获取用户访问控制的细粒度，而隐私保护效果测试则是在有、无攻击环境下，对存储的分布式隐私大数据的泄露风险以及信息损失进行度量，最终通过与基于同态加密的数据隐私保护加密控制模型（文献［3］方法）和基于可搜索加密机制的数据隐私保护加密控制（文献［4］方法）进行对比，验证设计模型的效果。

2.1　配置加密模型运行环境

为了实现基于属性分类的分布式大数据隐私保护加密控制模型的开发，并为模型的运行提供硬件支持，需要对其运行环境进行配置。需要配置的硬件设备包括主测计算机、应用服务器和数据服务器。主测计算机设备选择 Y9000P2022 型号计算机，以异步双核设计方式配置 CPU 元件，内部集成两个 Scorpion 核心，工作频率最高达到 2.33 GHz，计算机中内置的 RAM 和 ROM 分别为 800 MB 和 2 GB[24]。应用服务器采用 HP 微机服务器，内存为 4 GB，硬盘存储空间为 300 GB，应用服务器为分布式大数据的处理提供支持。而数据库存储器用来存储分布式隐私大数据，数据库系统采用 Oracle9i。从软件分布来看，分布式大数据隐私保护加密控制模型均在 Windows7 的 64 位操作系统下运行，编程语言为 MATLAB。

2.2　准备隐私保护数据样本

选择 UCI 机器学习数据库中的 Adult 数据集作为实验数据来源，数据库中的数据内容包括：姓名、性别、年龄、教育程度、家庭住址、职业、婚姻状况等，初始准备的隐私保护数据共 58 450 条，输入到模型中的数据量为 45.6 GB。

2.3　编写分布式大数据隐私攻击程序

为了模拟分布式大数据的实际运行环境，模拟对分布式隐私数据的明文攻击过程，并编写相应的攻击程序。在已知隐私数据中部分明文或密文的情况下，反向推导出分布式隐私数据的加密方式，由此生成相应密钥，用户破解隐私数据。实验中编写的攻击程序为动态程序，即可以根据加密隐私数据的状态调整攻击方式与参数，但攻击程序只能同时对一个数据节点发起攻击[25]。按照上述原理，针对分布式存储的隐私数据，利用 thalheim 编码工具，实现分布式大数据隐私攻击程序的编写。在编写的攻击程序的首尾两端添加强制控制指令，最大程度地保证攻击程序运行状态的可控性。

2.4　描述模型运行与测试过程

在配置的加密模型运行环境中，利用相应的编程语言实现设计的基于属性分类的分布式大数据隐私保护加密控制模型的开发。将准备的隐私保护数据样本导入其中，在不运行攻击程序的前提下，完成对数据样本的加密，加密结果如图 5 所示。

图 5　分布式大数据隐私保护加密结果

在此基础上，生成多个模拟用户节点，并设置用户身份与角色，分配用户权限，模拟用户的访问过程，在模型作用下得出分布式隐私保护加密数据的访问控制结果，如图 6 所示。

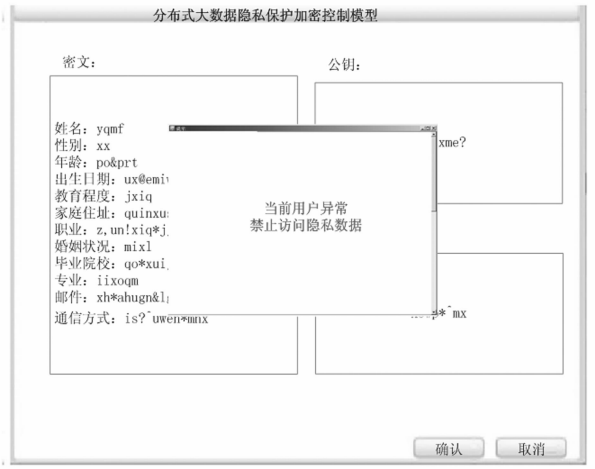

图 6　分布式隐私保护加密数据访问控制界面

由此完成基于属性分类的分布式大数据隐私保护加密控制模型的开发与运行任务。根据分布式隐私数据的动态变化情况，对加密结果以及访问控制指令进行实时更新。经过一段时间周期后，启动编写的明文攻击程序，获取攻击工况下隐私数据的变化情况，以此作为测试结果数据。为了验证设计模型在性能与效果方面的优势，选择对比模型，在相同的开发环境下对同一批隐私数据进行加密与控制处理。

2.5　设置模型量化测试指标

从加密性能、控制性能、隐私保护效果 3 个方面设置模型的量化测试指标，将明文/密文相似度作为模型加密性能的量化测试指标，其数值结果如下：

$$\mu = \sqrt{\sum (X_{\mathrm{mw}} - X_{\mathrm{ciphertext}})^2} \tag{16}$$

其中：X_{mw} 和 $X_{\mathrm{ciphertext}}$ 分别表示的是隐私数据的明文和密文。模型访问性能的测试指标设置为访问撤销控制准确

率，该指标的测试结果为：

$$\eta_{\text{visit}} = \frac{N_{\text{cancel}}}{N_{\text{abnormal}}} \times 100\% \tag{17}$$

式中，N_{cancel}和 N_{abnormal}分别为访问撤销的用户数量以及设置的异常用户数量。另外模型隐私保护效果的测试指标为隐私泄露风险指数和隐私数据损失量，其中隐私泄露风险指数的测试结果如下：

$$\vartheta = \frac{\varphi_{\text{loss}}}{N_{\text{mw}}} \tag{18}$$

其中：φ_{loss}为隐私数据的损失量，N_{mw}表示隐私数据明文总量，而指标隐私数据损失量的数值结果为：

$$\varphi_{\text{loss}} = N_{\text{mw}} - N_{\text{jm}} \tag{19}$$

式中，N_{jm}为解密后的隐私数据，将公式（19）的计算结果代入到公式（18）中，即可得出隐私泄露风险指数的测试结果。最终计算得出明文/密文相似度 μ 越低，说明模型加密性能越好，访问撤销控制准确率 η_{visit}越高，证明模型控制性能越好，隐私泄露风险指数 ϑ 越低、损失数据量 φ_{loss}越少，说明模型的隐私保护效果越好。

2.6 模型性能测试结果与分析

2.6.1 加密性能测试

收集初始隐私明文数据以及不同模型下输出的加密结果，将相关数据代入到公式（16）中，得出模型加密性能的测试对比结果，如图 7 所示。

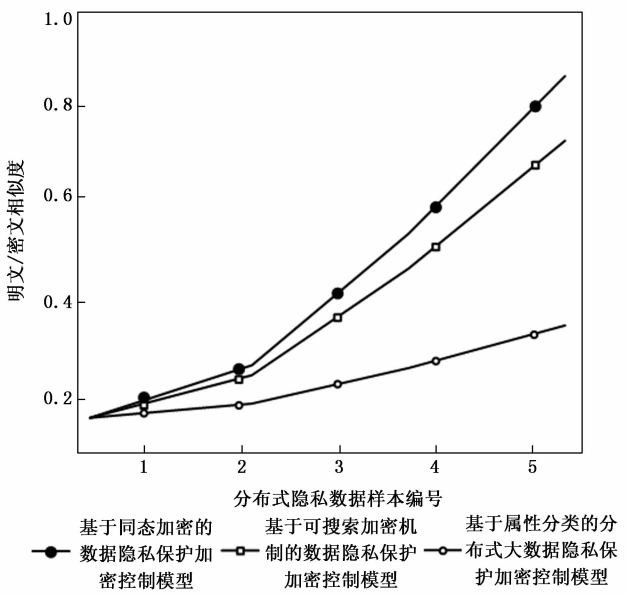

图 7 分布式大数据隐私保护加密性能测试对比曲线

从图 7 中可以直观地看出，与两种对比模型相比，设计模型输出的加密密文与初始明文之间的相似度更低，由此说明设计模型的加密性能较好。

2.6.2 控制性能测试

统计访问分布式隐私数据的用户数量，通过对异常数据的设置，确定变量 N_{abnormal}的具体取值，在模型作用下实时监测用户的访问进程数据，得出反映模型控制性能的测试结果，如表 1 所示。

表 1 隐私保护加密控制模型控制性能测试数据表

实验次数	异常用户设置数量/个	基于同态加密的数据隐私保护加密控制模型撤销的访问用户量/个	基于可搜索加密机制的数据隐私保护加密控制模型撤销的访问用户量/个	基于属性分类的分布式大数据隐私保护加密控制模型撤销的访问用户量/个
1	32	28	30	32
2	46	42	45	46
3	37	31	35	36
4	44	39	41	44
5	39	34	36	38

将表 1 中的数据代入到公式（17）中得出，两种对比模型的平均访问撤销控制准确率分别为 87.7% 和 94.3%，而在设计模型的访问撤销控制准确率的平均值为 98.9%。由此可知，设计模型的控制性能较好。

2.6.3 隐私保护效果测试

在有、无攻击条件下，对隐私数据明文与解密结果进行统计，得出隐私数据损失量指标的测试结果，如表 2 所示。

表 2 隐私数据损失量测试数据表

实验次数	明文数据量/GB	基于同态加密的数据隐私保护加密控制模型的解密数据量/GB		基于可搜索加密机制的数据隐私保护加密控制模型的解密数据量/GB		基于属性分类的分布式大数据隐私保护加密控制模型的解密数据量/GB	
		无攻击工况	有攻击工况	无攻击工况	有攻击工况	无攻击工况	有攻击工况
1	45.6	45.2	44.6	45.4	45.0	45.6	45.5
2	45.6	45.0	44.3	45.5	44.9	45.5	45.4
3	45.6	45.1	44.4	45.5	44.7	45.6	45.4
4	45.6	45.0	44.5	45.5	44.8	45.6	45.5
5	45.6	45.1	44.3	45.3	44.7	45.5	45.3

将表 2 中的数据代入到公式（19）中计算得出，在无攻击工况下，3 个模型的平均隐私数据损失量分别为 0.52 GB、0.16 GB 和 0.04 GB，而在有攻击工况下，3 个模型隐私数据损失量的平均值分别为 1.18 GB、0.78 GB 和 0.18 GB。由此可以看出，在有、无攻击工况下，设计模型的隐私数据损失量均较少，设计模型的隐私保护性能较好。另外通过公式（18）的计算，得出不同模型下隐私泄露风险的测试对比结果，如图 8 所示。

从图 8 中可以看出，在设计模型作用下，隐私数据的泄露风险更低，综合隐私数据损失量的测试结果，表明设计模型在隐私保护效果方面具有明显优势，能够有效提高分布式大数据的存储和传输安全性。

3 结束语

针对用户数据的隐私保护问题，设计了基于属性分类的分布式大数据隐私保护加密控制模型。该模型应用属性分类技术，确定加密等级与强度，实现对分布式大数据隐

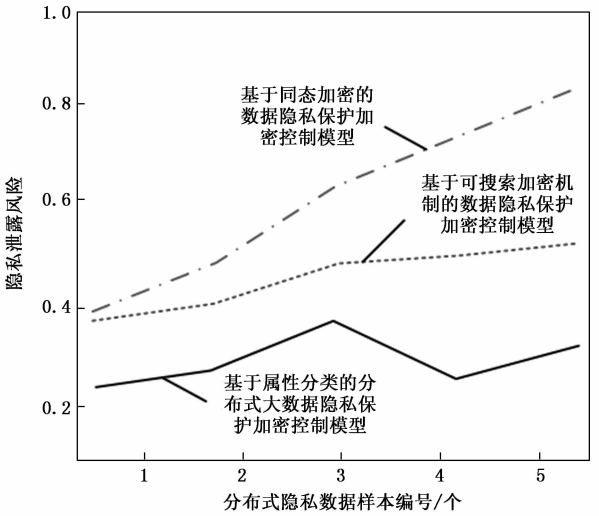

图 8　不同模型作用下隐私泄露风险测试对比结果

私保护的最大化。实验结果表明，设计模型在加密、控制以及隐私保护方面具有明显优势，能够有效提高分布式大数据的存储和传输安全性。

参考文献：

[1] 王加梁. 基于属性分类建模的入侵检测方法 [J]. 计算机工程与设计，2022，43 (4)：907－913.

[2] 潘莹丽，刘　展，朱千慧子. 大数据背景下基于 Huber 回归模型的分布式优化方法研究 [J]. 数理统计与管理，2022，41 (4)：633－646.

[3] 贾春福，李瑞琪，王雅飞. 基于同态加密的 DBSCAN 聚类隐私保护方案 [J]. 通信学报，2021，42 (2)：1－11.

[4] 孙僖泽，周福才，李宇溪，等. 基于可搜索加密机制的数据库加密方案 [J]. 计算机学报，2021，44 (4)：806－819.

[5] 侯梦薇，卫　荣，兰　欣，等. 基于差分隐私的医疗大数据隐私保护模型应用研究 [J]. 中国数字医学，2019，14 (12)：86－88.

[6] 张　婵. 基于大数据的区块链数据隐私文本智能加密方法 [J]. 网络安全技术与应用，2023 (3)：26－28.

[7] 陈小娟，贺红艳，张慧萍. 基于数据消冗技术的隐私大数据属性加密仿真 [J]. 计算机仿真，2022，39 (11)：422－426.

[5] 蔡玉涵，王静宇. 使用模糊关键字可搜索同态加密的区块链隐私保护方案 [J]. 小型微型计算机系统，2022，43 (11)：2406－2413.

[6] 陈子秋，冯瑞珏，郑扬富，等. 非侵入式负荷监测系统数据隐私保护方法研究 [J]. 电子技术应用，2021，47 (12)：116－119.

[7] 杨佳辉，陈兰香，穆　怡，等. 结构化加密的 PSI 协议 [J]. 计算机学报，2022，45 (12)：2652－2666.

[8] HARSHITHA M，RUPA C，SAI K P，et al. Secure medical data using symmetric cipher based chaotic logistic mapping [C] //2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE，2021 (1)：476－481.

[9] LIN J，ZHAO K，CAI X，et al. An image encryption method based on logistic chaotic mapping and DNA coding [C] //MIPPR 2019：Remote Sensing Image Processing，Geographic Information Systems，and Other Applications. SPIE，2020，11432：363－369.

[10] ZHAO F，LI C，LIU C，et al. Analysis of the effects of scrambling and diffusion of logistic chaotic map on image encryption [C] //Eleventh International Conference on Digital Image Processing (ICDIP 2019). SPIE，2019，11179：98－107.

[11] 牛淑芬，杨平平，谢亚亚，等. 区块链上基于云辅助的密文策略属性基数据共享加密方案 [J]. 电子与信息学报，2021，43 (7)：1864－1871.

[12] 刁一晴，叶阿勇，张娇美，等. 基于群签名和同态加密的联盟链双重隐私保护方法 [J]. 计算机研究与发展，2022，59 (1)：172－181.

[13] 张人上，邱久睿. 基于混沌系统的扩频通信多源异构数据加密算法 [J]. 火力与指挥控制，2021，46 (8)：162－166.

[14] 杨尧林，和红杰，陈　帆，等. 基于预测误差自适应编码的图像加密可逆数据隐藏 [J]. 计算机研究与发展，2021，58 (6)：1340－1350.

[15] 张瑞瑞，牛宏侠. 安全性增强的无证书可搜索公钥加密方案 [J]. 微电子学与计算机，2022，39 (6)：89－98.

[16] 王　勇，王宏志. 基于 K 近邻的隐式位置访问隐私保护方法 [J]. 计算机仿真，2022，39 (6)：412－416.

[17] 曹素珍，丁宾宾，丁晓晖，等. 基于身份的具有否认认证的关键字可搜索加密方案 [J]. 电子与信息学报，2022，44 (3)：1086－1092.

[18] 郑振青，毋小省，王　辉，等. 移动社交网络中的轨迹隐私 PTPM 保护方法 [J]. 小型微型计算机系统，2021，42 (10)：2153－2160.

[19] PAVANI K，SRIRAMYA P. Enhancing public key cryptography using RSA，RSA-CRT and N-Prime RSA with multiple keys [C] //2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). IEEE，2021：1－6.

[20] SU J，SUN H，WANG H，et al. Topological public-key cryptography based on graph image-labellings for information security [C] //2020 IEEE International Conference on Information Technology，Big Data and Artificial Intelligence (ICIBA). IEEE，2020 (1)：366－370.

[21] 殷凤梅，陈　鸿. Geohash 编码的 k 匿名位置隐私保护方案 [J]. 武汉大学学报 (理学版)，2022，68 (1)：73－82.

[22] 姜海洋，曾剑秋，韩　可，等. 5G 环境下移动用户位置隐私保护方法研究 [J]. 北京理工大学学报，2021，41 (1)：84－92.

[23] 李幸昌，王　斌，王　超，等. 基于加密分割的位置隐私保护方法 [J]. 计算机应用研究，2021，38 (10)：3153－3156.

[24] 邓　桦，宋甫元，付　玲，等. 云计算环境下数据安全与隐私保护研究综述 [J]. 湖南大学学报 (自然科学版)，2022，49 (4)：1－10.

[25] 杜　刚，张　磊，马春光，等. 基于属性基隐私信息检索的位置隐私保护方法 [J]. 哈尔滨工程大学学报，2021，42 (5)：680－686.