DOI:10.13878/j.cnki.jnuist.20220303001

刘小都[1]　赵慧奇[2]

基于混合密码体制的大数据隐匿性特征安全提取技术

摘要

传统大数据隐匿性特征安全提取技术忽略了大数据密文的公钥及密钥封装,且大数据隐匿性特征类别混乱,导致该技术的提取精度偏低、冗余度较高.为此,本文提出一种基于混合密码体制的大数据隐匿性特征安全提取方法.通过混合密码体制中的公钥封装以及密钥封装机制生成大数据密文;根据密文内容设计对称加密方法和非对称加密方法,基于此分类隐匿性特征,利用不同类的隐匿性特征构建大数据隐秘性特征相空间,计算大数据间的关联维值,实现大数据隐匿特征的安全提取.实验结果表明,与传统方法相比,所提出的大数据隐匿特征提取方法冗余度低,大数据隐匿特征平均分类正确率高达95%,且特征安全提取误差低,验证了所提方法具有更好的应用性能.

关键词

混合密码体制;大数据;隐匿性特征;安全提取;混合算法;关联维数

中图分类号 TP393

文献标志码 A

收稿日期 2022-03-03

资助项目 山东省重点研发计划重大科技创新工程(2020SO10103-00517)

作者简介

刘小都,男,工程师,主要研究方向为信息安全技术等.beiyao83185380@163.com

赵慧奇(通信作者),男,博士,副教授,主要研究方向为网络安全、数据隐私保护、工业信息安全等.huasha4887479476@163.com

1 中国科协信息中心,北京,100863

2 山东科技大学 智能装备学院,泰安,271019

0 引言

为了保证大数据传输安全,应对大数据实施加密处理[1-2].因数据量剧增,大数据隐匿性特征类别混乱,导致原有的加密技术无法达到大数据加密要求[3].目前,各种网络入侵行为加剧,如不法黑客收取、复制、发布信息等.为了规避大数据信息发生上述风险,必须采用大数据隐匿性特征安全提取技术保证大数据交互安全[4].而传统大数据隐匿性特征安全提取技术,已经无法满足大数据发展的要求.

传统大数据隐匿性特征安全提取方法具有局限性.王安琪[5]讨论了网络用户协议语言存在的专业术语堆砌、表达模糊、文本语句不规则、子语言信息隐藏陷阱等问题,并提出了加强网络用户协议监管的特征安全提取方法.在大数据隐匿性特征安全提取过程中,该方法主要解决的是用户协议语义中存在的数据监管及提取问题,计算过程非常复杂,忽略了大数据密文的公钥及密钥封装,导致大数据隐匿性特征安全提取效果不佳.蔡柳萍等[6]基于稀疏表示和特征加权的大数据挖掘方法,采用求解线性方程稀疏解的方法对大数据进行特征分类,在稀疏解的求解过程中利用向量的范数将此过程转化为最优化目标函数的求解.在完成特征分类后进行特征提取以降低数据维度,最后充分结合数据的分布情况进行有效加权来实现大数据挖掘.在大数据隐匿性特征安全提取过程中,该方法主要通过特征提取降低数据维度来实现大数据挖掘,但未考虑大数据中的冗余特征数据,使得大数据隐匿性特征安全提取效率低.同时,上述两种方法均忽略了大数据密文的公钥及密钥封装,提取精度较低、冗余度较高.混合密码体制是对称加密方法和非对称加密方法的综合技术.因此,本文基于混合密码体制的大数据隐匿性特征安全提取技术,通过混合算法提高大数据的加密速度,实时提取关联维数,并利用不同种类的大数据隐匿性特征构建大数据隐匿性特征相空间.密钥对称与公钥封装机制相融合,提升了大数据隐匿性,通过椭圆加密算法对数据摘要实施加密处理,提高了密钥传输的安全性.实验结果表明,本研究能够提升大数据隐匿性特征安全提取效率,满足大数据时代的要求.

1 混合密码体制的大数据隐匿性特征安全提取技术

建立混合密码体制,设计对称加密方法和非对称加密方法,选择大数据隐匿性特征,构建大数据隐匿性特征相空间,引入关联的隐匿

性特征安全提取.

1.1 混合密码体制研究

混合密码体制的建立融合了密钥对称与公钥封装机制,提升大数据隐匿性.混合密码体制建立原理如图 1 所示.

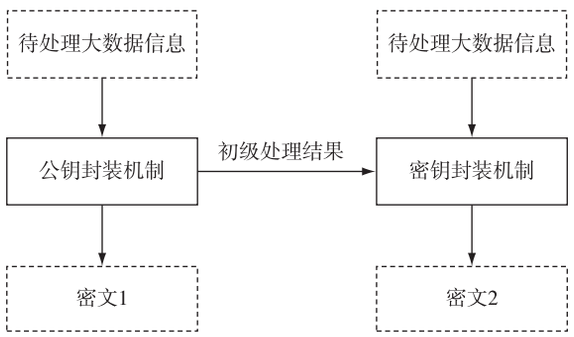

图 1 混合密码体制建立原理

Fig. 1 Schematic of hybrid cryptosystem establishment

由图 1 可知,为了生成安全性高的大数据混合密文,混合密码体制在接收待处理的大数据信息之后[7],实施如下操作:部分大数据密文的生成通过公钥封装机制实施大数据初级处理实现,其他大数据密文通过密钥封装机制实施深度处理实现.

1.2 混合算法

为了提高大数据的加密速度,采用混合密码体制中混合算法实现,流程如图 2 所示.

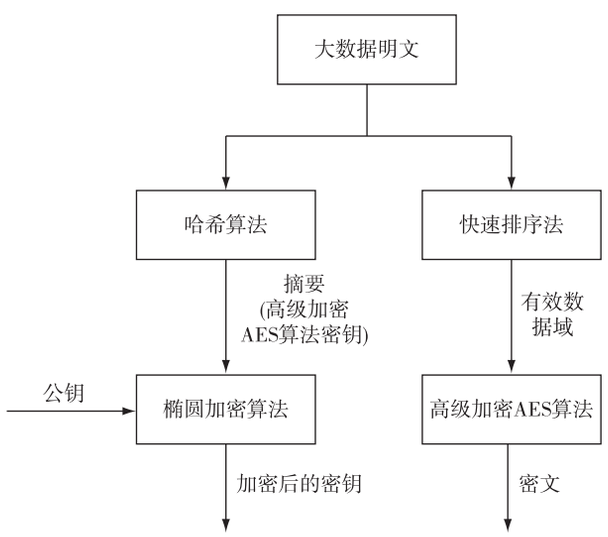

图 2 混合算法流程

Fig. 2 Hybrid algorithm flow

由图 2 可知,高级加密算法的密钥通过哈希算法将大数据明文生成一个数据摘要.为了增强密钥传输的安全性,采用椭圆加密算法对数据摘要实施加密处理.在搜寻有效数据域的基础上,采用高级加密 AES 算法生成大数据密文.经过加密后的密钥和密文[8-10],通过数据传输至指定对象.

根据上述加密后的密文内容,设计对称加密方法和非对称加密方法,两种加密方法分类构建,如图 3 所示.

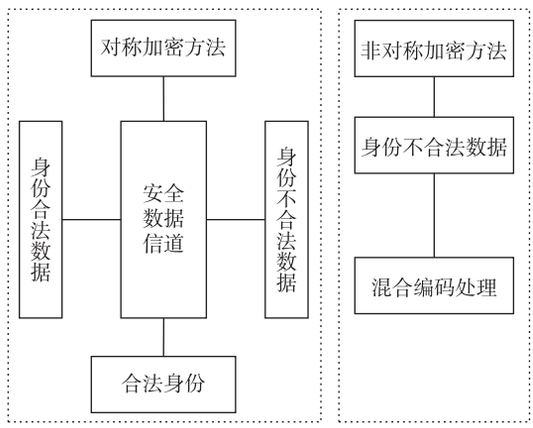

图 3 两种加密方法分类构建原理

Fig. 3 Classification and construction of two encryption methods

图 3 中,根据安全数据信道,采用对称加密方法将数据分为身份合法数据与身份不合法数据并呈对称形式,只传给合法的身份.在实施混合编码处理时,身份不合法的数据采用非对称加密方法,对两个数据请求里的内容实施判断[11],对大数据隐匿性安全特征信息实施分类.

在上述对称加密方法和非对称加密方法对大数据隐匿性安全特征信息实施分类的基础上,大数据隐匿性特征选择通过对特征集合实施评价实现.安全特征提取的实质是从最广泛描述数据的特征集合中,按照一定的规则和度量标准提取出最佳隐匿性特征子集,在有效降低数据冗余度的同时保持同初始特征集合相同的辨识能力.第 u 个样本第 i 个隐匿性特征、大数据第 u 个隐匿性特征分别用 F_{ui},$\boldsymbol{X}_u$ 表示,则第 u 个隐匿性特征用 $\boldsymbol{F}_u=[F_{u1},F_{u2},\cdots,F_{ui}]^{\mathrm{T}}$ 表示.大数据隐匿性特征选择步骤如下:

Step 1:建立大数据样本的部分结构用 $G_a=(\boldsymbol{V},\boldsymbol{E})$ 表示最近邻图,$\boldsymbol{E}$ 表示节点间连接边构成的集合,$\boldsymbol{V}$ 表示节点集合,同时 $\boldsymbol{V}=\boldsymbol{A}$,指定大数据样本 $\boldsymbol{A}\in\mathbf{R}^{n\times d}$,$\boldsymbol{E}\in\mathbf{R}^{n\times n}$ 表示 G_a 的权重矩阵.

$$W_{ij}=\begin{cases}\mathrm{e}^{-\frac{\|F_{ui}-\boldsymbol{X}_u\|^2}{G_a}}, & (F_{ui}-\boldsymbol{X}_u)\in\eta,\\ 0, & \text{其他.}\end{cases}\tag{1}$$

其中：W_{ij} 表示大数据样本部分结构特征；η 为常数.

Step 2：设 $\boldsymbol{B}=\mathrm{diag}(\boldsymbol{W}_1)$，$\boldsymbol{W}_1=[1,1,\cdots,1]^{\mathrm{T}}$，$\boldsymbol{X}=\boldsymbol{B}-\boldsymbol{W}$，其中 $\boldsymbol{X}$ 表示隐匿性样本特征，$\boldsymbol{W}$ 表示集合特征.

目标函数值为大数据第 u 个隐匿性特征 $\boldsymbol{X}_u$，目标函数值最小则可提取更好的大数据隐匿性特征. B_{ii} 表示隐匿性特征系数，I 表示隐匿性特征参量，λ_u 表示隐匿性目标特征函数，目标函数值的计算式如下：

$$\boldsymbol{X}_u=\frac{\sum_{ij}(F_{ui}-F_{uj})^2W_{ij}}{I(\boldsymbol{F}_u)},\tag{2}$$

$$\sum_{ij}(F_{ui}-F_{uj})^2W_{ij}=2\boldsymbol{F}_u{}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{F}_u-2\boldsymbol{F}_u{}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{F}_u=2\boldsymbol{F}_u{}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{F}_u,\tag{3}$$

$$I(\boldsymbol{F}_u)=\sum_i(F_{ui}-\lambda_u)^2B_{ii},\tag{4}$$

$$\lambda_u=\sum_i\left(F_{ui}\frac{B_{ii}}{\sum_i B_{ii}}\right)=\frac{\sum_i F_{ui}B_{ii}}{\sum_i B_{ii}}=\frac{\boldsymbol{F}_u^{\mathrm{T}}\boldsymbol{B}\boldsymbol{1}}{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{1}}.\tag{5}$$

通过式(3)对第 u 个隐匿性特征实施标准化处理，得出式(6)：

$$\overset{*}{\boldsymbol{F}}_u=\boldsymbol{F}_u-\frac{\boldsymbol{F}_u^{\mathrm{T}}\boldsymbol{B}\boldsymbol{1}}{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{1}}.\tag{6}$$

第 u 个隐匿性特征计算公式为

$$\boldsymbol{X}_u=\frac{\overset{*}{\boldsymbol{F}}_u^{\mathrm{T}}\boldsymbol{X}\overset{*}{\boldsymbol{F}}_u}{\overset{*}{\boldsymbol{F}}_u^{\mathrm{T}}\boldsymbol{B}\overset{*}{\boldsymbol{F}}_u}.\tag{7}$$

大数据隐匿性特征选择依据是求出的目标函数值，并提取前 m 个较小的大数据隐匿性特征.

大数据隐匿性特征安全提取的关键是相空间的构建，以保障数据隐匿性特征不变[12-14]. 根据大数据隐匿性特征选择方法，利用不同种类的大数据隐匿性特征构建大数据隐匿性特征相空间.

相空间构建方法：用 $\{q_1,q_2,\cdots,q_N\}$ 表示 1 维时间序列，构建相空间用式(8)描述：

$$\boldsymbol{Q}=[\boldsymbol{Q}_1,\boldsymbol{Q}_2,\cdots,\boldsymbol{Q}_M]=\begin{bmatrix}q_1 & q_2 & L & q_M\\ q_{1+t} & q_{2+t} & L & q_{M+t}\\ M & M & O & M\\ q_{1+(r-1)t} & q_{2+(r-1)t} & L & q_{M+(r-1)t}\end{bmatrix},\tag{8}$$

其中，$M=N-(r-1)\tau$，N 表示 1 维时间序列数，r 表示嵌入维数，τ 表示时延，O 表示嵌入序列，L 表示关联序列. 隐匿性特征是当 $r\geqslant 2d'+1$ 时完全打开数据的几何结构，其中 d' 表示混沌吸引子的维度.

大数据隐匿特征安全提取主要采用关联维数实时提取. 在大数据样本间的关联维程度，为数据在多维空间内疏密程度的表现[15]. 因此，采用离散智能优化算法计算大数据间的关联维值，降低特征之间的冗余度，使分类辨识度最大限度地接近原始特征集合，以实现大数据隐匿特征的安全提取.

一组空间矢量通过大数据实施相空间重构方式获取，二者的间距用 2 个矢量的最大分量差描述：

$$|\boldsymbol{Q}_i-\boldsymbol{Q}_j|=\max_{1\leqslant\delta\leqslant r}|\boldsymbol{Q}_{i\delta}-\boldsymbol{Q}_{j\delta}|.\tag{9}$$

关联积分是在相关矢量计算的基础上对全部 K^2 种组合内的比重，关联积分为

$$D_k(l)=\frac{1}{K^2}\sum_{i,j=1}^{K}H(l-|\boldsymbol{Q}_i-\boldsymbol{Q}_j|),\tag{10}$$

其中 K 表示构建相空间内点数量，l 为既定数，H 表示 Heaviside 函数：

$$H(q)=\begin{cases}0, & q\leqslant 0,\\ 1, & q>0.\end{cases}\tag{11}$$

通过相关分析显示，当关联积分 $D_k(l)$ 在 $l\to 0$ 时，与 l 间关联，如式(12)所示：

$$\lim_{l\to 0}D_k(l)\propto l^{C},\tag{12}$$

其中 C 表示关联维数. 用 C 描述混沌吸引子的自相似结构，需选取适合的直线 l，近似值为

$$C_p=\frac{\ln D_k(l)}{\ln l}.\tag{13}$$

通常分析双对数 $\ln D_k(l)\to\ln l$，选取拟合直线，此时斜率为 C.

数据样本标准差与关联维度相关，当数据样本标准差较高时，关联维低. 大数据分布关系为

$$Y_i=\beta\frac{\sigma_i}{C},\tag{14}$$

其中，σ_i 表示第 i 层分解的近似系数标准差，β 表示倍频因子. 由式(14)可知，数据样本标准差与关联性成反比. 通过 Y_i 可对数据隐匿特征实施安全提取.

2 实验分析

选取某公司的大量财务数据作为实验数据集，选用 Matlab 软件为实验平台，硬件配置为 3.20 GHz CPU、4.00 GB 内存，软件配置为 Windows7 SP1 的电脑，运行环境为 Visual Studio 2010. 在 Matlab 平台搭建实验环境，数据参数如表 1 所示. 实验对比方法为文献[5]加强网络用户协议监管的特征安全提取方法和文献[6]基于稀疏表示和特征加权的大数据挖掘方法.

表 1　数据参数

Table 1　Data parameters

大数据隐匿性特征代码	财务数据数量/个
1	175
2	732
3	508
4	905
5	993
6	652
7	622
8	517
9	932
10	946
11	761
12	832
13	745
14	931
15	1 022

2.1　冗余度测试

采用本文方法和文献[5]加强网络用户协议监管的特征安全提取方法、文献[6]基于稀疏表示和特征加权的大数据挖掘方法提取实验数据集中的大数据隐匿特征冗余度,对比结果如图 4 所示.

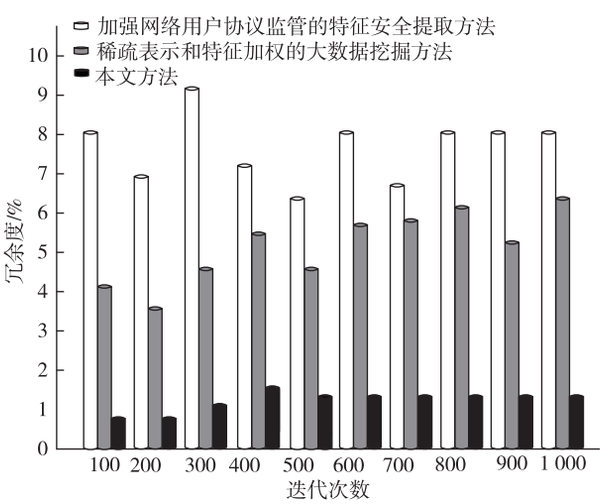

图 4　冗余度对比

Fig. 4　Redundancy comparison

由图 4 可知,本文方法提取的大数据隐匿特征冗余度平均值仅为 1.5%,相比其他两种特征安全提取方法,本文方法特征提取的冗余度较低,表明本文方法可有效去除大数据隐匿特征内的冗余特征量,提取出更加有效的大数据隐匿特征.这是因为本文方法采用关联维数实时提取,利用不同种类的大数据隐匿性特征构建大数据隐秘性特征相空间.

2.2　鲁棒性测试

为了进一步验证大数据隐匿特征安全提取性能,测试了 3 种方法的大数据隐匿特征安全提取的鲁棒性.鲁棒性越高,表明提取过程越稳定,具体对比结果如图 5 所示.

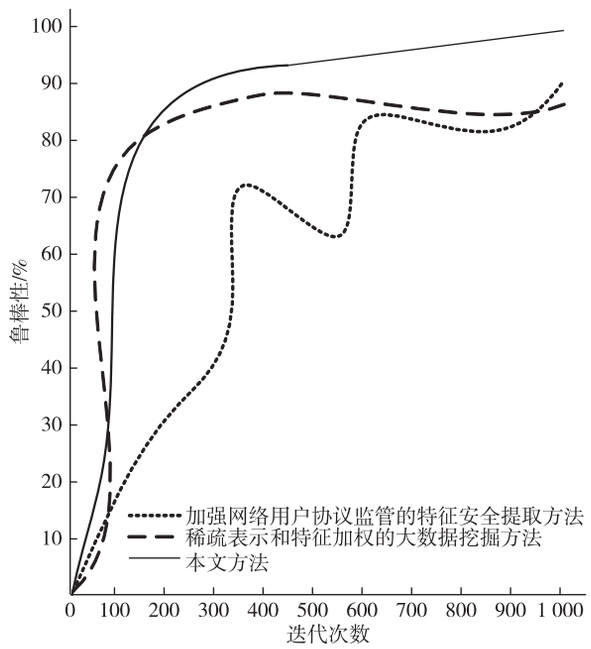

图 5　鲁棒性测试结果

Fig. 5　Robustness test results

由图 5 可知:本文方法大数据隐匿性特征安全提取效果较好,特征提取鲁棒性接近 100%,而其他两种方法仅达到 85%,表明本文方法大数据隐匿性特征安全提取过程更加稳定.这是因为本文方法采用混合密码体制中混合算法实现,并采用椭圆加密算法对数据摘要实施加密处理,增强了密钥传输的安全性.

2.3　精度测试

为验证本文方法的有效性,对比分析不同迭代次数情况下,3 种方法进行大数据隐匿特征安全提取的误差对比结果(表 2).分类正确率实验结果如图 6 所示.提取误差越小,分类正确率越高,表明安全提取精度越高.

由表 2 可知,本文方法的平均误差为 2.69%,分别比其他 2 种方法的平均标准差低 11.09 个百分点和 5.51 个百分点,表明本文方法具有更好的应用性能.这是因为本文方法中混合密码体制的建立融合了密钥对称与公钥封装机制,提升了大数据隐匿性,

降低了特征安全提取误差.

表 2　提取误差对比结果

Table 2　Comparison of extraction error　%

迭代次数	文献[5]方法	文献[6]方法	本文方法
100	13.41	7.83	3.20
200	13.70	8.12	2.32
300	12.91	7.33	2.65
400	13.91	8.33	2.75
500	14.02	8.44	2.66
600	13.53	7.95	2.58
700	14.24	8.66	2.68
800	15.44	9.86	2.68
900	13.84	8.26	2.68
1 000	12.81	7.23	2.68
平均值	13.78	8.20	2.69

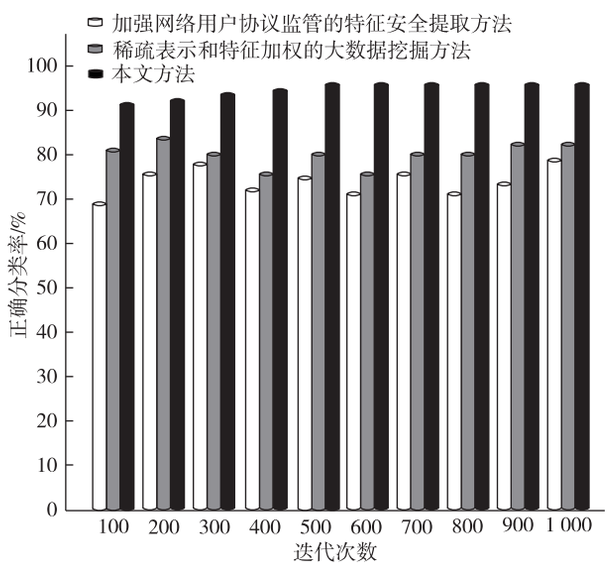

图 6　分类正确率对比

Fig. 6　Comparison of classification accuracy

由图 6 可知,其他 2 种方法平均分类正确率分别为 73%、80%,而本文方法平均分类正确率为 95%,表明本文方法提取精度高且能迅速地达到高收敛状态.这是因为本文方法在搜寻有效数据域的基础上,采用高级加密 AES 算法生成大数据密文.经过加密后的密钥和密文,通过数据传输至指定对象.

2.4　运行时间测试

采用 3 种方法对实验数据实施大数据隐匿特征安全提取,测试 3 种方法运行时间.运行时间越短,表明大数据隐匿特征安全提取效率越快,具体对比结果如图 7 所示.

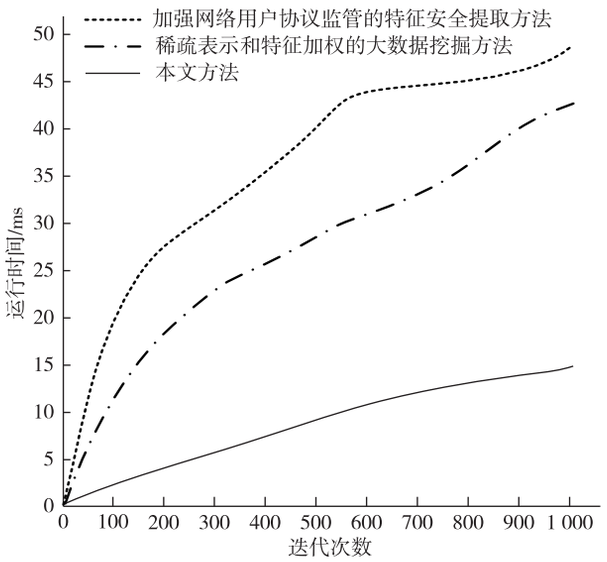

图 7　运行时间对比

Fig. 7　Comparison of running time

由图 7 可知,本文方法的运行时间低于 15 ms,其他 2 种方法的运行时间超过 40 ms,表明本文方法提取效率更高.这是因为本文方法在对称加密方法和非对称加密方法对大数据隐匿性安全特征信息实施分类的基础上,大数据隐匿性特征选择通过对特征集合实施评价实现.

3　结论

为解决传统方法大数据隐匿性特征类别混乱、提取精度偏低以及冗余度较高的问题,本文提出基于混合密码体制的大数据隐匿性特征安全提取方法.实验结果表明,本文方法大数据隐匿性特征安全提取误差低,数据提取效果较好、冗余度较低,提取精度高,且运行时间低于 15 ms,提取效率高.

在进行大数据隐匿性特征安全提取后,运用更先进的技术精细化处理大数据以及算法安全性,是下一步主要的研究方向.

参考文献

References

[1] 杨国强,丁杭超,邹静,等.基于高性能密码实现的大数据安全方案[J].计算机研究与发展,2019,56(10):2207-2215
YANG Guoqiang,DING Hangchao,ZOU Jing,et al.A big data security scheme based on high-performance cryptography implementation[J].Journal of Computer Research and Development,2019,56(10):2207-2215

[2] 徐超,陈勇,葛红美,等.基于大数据的审计技术研究[J].电子学报,2020,48(5):1003-1017

XU Chao,CHEN Yong,GE Hongmei,et al.Audit technology research based on big data[J].Acta Electronica Sinica,2020,48(5):1003-1017

[3] 王永坤,罗萱,金耀辉.基于私有云和物理机的混合型大数据平台设计及实现[J].计算机工程与科学,2018,40(2):191-199
WANG Yongkun,LUO Xuan,JIN Yaohui.A hybrid big data platform based on private cloud VMs and bare metals[J].Computer Engineering & Science,2018,40(2):191-199

[4] 杨丽丽.船用物联网大数据加密的混合密码体制[J].舰船科学技术,2020,42(4):196-198
YANG Lili.Mixed cryptography system encrypted by big data on marine internet of things[J].Ship Science and Technology,2020,42(4):196-198

[5] 王安琪.大数据战略下网络用户协议语言问题与监管建议[J].辽东学院学报(社会科学版),2019,21(3):69-73
WANG Anqi.Network user agreement language under big data strategy:problems and suggestions[J].Journal of Eastern Liaoning University (Social Sciences),2019,21(3):69-73

[6] 蔡柳萍,解辉,张福泉,等.基于稀疏表示和特征加权的大数据挖掘方法的研究[J].计算机科学,2018,45(11):256-260
CAI Liuping,XIE Hui,ZHANG Fuquan,et al.Study on big data mining method based on sparse representation and feature weighting[J].Computer Science,2018,45(11):256-260

[7] Zhang C,Liu X J.Feature extraction of ancient Chinese characters based on deep convolution neural network and big data analysis[J].Computational Intelligence and Neuroscience,2021,2021:2491116

[8] 张启星,付敬奇.基于信道特征提取的物理层安全密钥生成方法[J].电子测量与仪器学报,2019,33(1):16-22
ZHANG Qixing,FU Jingqi.Physical layer security key generation method based on channel feature extraction[J].Journal of Electronic Measurement and Instrumentation,2019,33(1):16-22

[9] 王妍,李俊,曾辉,等.一种基于互信息的实时特征提取算法[J].小型微型计算机系统,2019,40(6):1242-1247
WANG Yan,LI Jun,ZENG Hui,et al.Real-time feature extraction algorithm based on mutual information[J].Journal of Chinese Computer Systems,2019,40(6):1242-1247

[10] 段大高,赵振东,梁少虎,等.基于条件变分自编码的密码攻击算法[J].计算机应用研究,2020,37(3):821-823,837
DUAN Dagao,ZHAO Zhendong,LIANG Shaohu,et al.Password cracking algorithm using conditional variational auto-encoders[J].Application Research of Computers,2020,37(3):821-823,837

[11] 吴颖,李晓玲,唐晶磊.Hadoop 平台下粒子滤波结合改进 ABC 算法的 IoT 大数据特征选择方法[J].计算机应用研究,2019,36(11):3297-3301
WU Ying,LI Xiaoling,TANG Jinglei.Internet of things big data feature selection method based on particle filter and improved ABC algorithm on Hadoop platform[J].Application Research of Computers,2019,36(11):3297-3301

[12] Cole J M.A design-to-device pipeline for data-driven materials discovery[J].Accounts of Chemical Research,2020,53(3):599-610

[13] 刘波涛,彭长根,吴睿雪,等.基于 MILP 方法的 LED 密码安全性分析[J].计算机应用研究,2020,37(2):505-509,517
LIU Botao,PENG Changgen,WU Ruixue,et al.Based on MILP method for security analysis of LED[J].Application Research of Computers,2020,37(2):505-509,517

[14] Zhang F,Yang Y H.Feature vector extraction algorithm based on big data in engineering quality[J].E3S Web of Conferences,2021,257:02029

[15] Elaggoune Z,Maamri R,Boussebough I.A fuzzy agent approach for smart data extraction in big data environments[J].Journal of King Saud University:Computer and Information Sciences,2020,32(4):465-478

Secure extraction of hidden big data features based on hybrid cryptosystem

LIU Xiaodu[1] ZHAO Huiqi[2]

1 Information Center of China Association for Science and Technology,Beijing 100863

2 College of Intelligent Equipment,Shandong University of Science and Technology,Tai'an 271019

Abstract The chaotic categories of hidden big data features,combined with the ignorance of the public key and key encapsulation of big data ciphertext,result in low extraction accuracy and high redundancy of traditional hidden big data feature extraction methods.Here,a secure extraction approach of hidden features of big data is proposed based on hybrid cryptosystem.First,the big data ciphertext is generated through public key encapsulation and cryptographic key encapsulation mechanisms in hybrid cryptosystem.Second,the hidden big data characteristics are cate-

gorized based on symmetric encryption and asymmetric encryption designed according to the content of big data ciphertext,which are then used to construct the phase space of big data hidden features and calculate the correlation dimension between big data,thus realize the secure extraction of hidden big data features.The experimental results show that,compared with traditional methods,the proposed approach has low redundancy,high accuracy of classification rate for big data hidden features up to 95%,and low error of feature extraction,verifying the feasibility and application prospect of the proposed approach.

Key words mixed cipher system;big data;occult characteristics;secure extraction;hybrid algorithm;correlation dimension