

基于改进关联规则的报送信息大数据特征 隐匿性加密算法

陈威宇^{1†}, 王浣¹, 何建锋¹, 苏怀振², 徐延文¹

(1. 国网甘肃省电力公司, 甘肃 兰州 730000; 2. 国网甘肃省电力公司定西供电公司, 甘肃 定西 743000)

摘 要:报送信息大数据的安全性直接关系到“数字化纪检云平台”数据挖掘的准确性,一旦泄露或者被篡改,纪检工作可靠性将大大降低,由此,提出一种基于改进关联规则的报送信息大数据特征隐匿性加密算法。手工获取报送信息大数据,通过连接转换器实时连通异构数据库,对不符合规范要求的数据进行自动修正,包括缺失数据填补、离群值处理和数据标准化。利用改进遗传算法改进 Apriori 算法,挖掘数据项或属性之间隐含的关联规则,通过关联规则识别报送信息大数据特征。以报送信息大数据特征为输入,利用神经网络层级之间的权重矩阵作为密钥,将明文数据转变为密文数据。结果表明:所研究加密算法应用下,“三公经费”开支数据加密后的密文和明文相关性均在 0~0.2 之间,说明所研究算法的加密效果较好,保证了报送信息大数据安全。

关键词:改进关联规则; Apriori 算法; 报送信息大数据; 特征识别; 神经网络; 隐匿性加密算法

中图分类号:TP333.6

文献标识码:A

Hidden Encryption Algorithm of Big Data Characteristics of Submitted Information Based on Improved Association Rules

CHEN Weiyu^{1†}, WANG Long¹, HE Jianfeng¹, SU Huaizhen², XU Yanwen¹

(1. State Grid Gansu Electric Power Company, Lanzhou, Gansu 730000, China;

2. Dingxi Power Supply Company, State Grid Gansu Electric Power Company, Dingxi, Gansu 743000, China)

Abstract: The security of big data of submitted information is directly related to the accuracy of data mining of "digital discipline inspection cloud platform". Once it is leaked or tampered, the reliability of discipline inspection work will be greatly reduced. Therefore, a feature hiding encryption algorithm of submitted information big data based on improved association rules is proposed. Manually obtain the big data of the submitted information, connect the heterogeneous database in real time through the connection converter, and automatically correct the data that does not meet the specification requirements, including missing data filling, outlier processing and data standardization. The improved genetic algorithm is used to improve the Apriori algorithm to mine the hidden association rules between data items or attributes, and identify the characteristics of big data of submitted information through association rules. Taking the big data characteristics of the submitted information as the input, and using the weight matrix between the neural network levels as the key, the plaintext data is transformed into ciphertext data. The results show that under the application of the encryption algorithm studied, the correlation between ciphertext and plaintext after the encryption of "three public funds" expenditure data is between 0 ~ 0.2, which shows that the encryption effect of the studied algorithm is good and ensures the security of big data of submitted information.

Key words: improved association rules; Apriori algorithm; submitted information big data; feature recognition; neural network; hidden encryption algorithm

一直以来,由于纪检监察工作缺乏系统的统一规范性,监督工作人员力量单薄,基层派驻巡察机构的工作目标不够明确,难以发挥应有的作用^[1,2]。面对这种情况,为适应集约化、信息化、数据化的管理需求,通过物联网、大数据、云计算、区块链、移动终端微应用等多种最前沿的技术,创新纪检方法手段,大力推进数字化纪检模式,开发“数字化纪检云平台”,深度挖掘基础数据,分析掌握监督对象“前景后场”总体情况,促进公司纪检业务全覆盖、全方位、全过程开展和管控,有效发挥内部“免疫系统”作用,保证报送信息大数据安全至关重要^[3]。

陈卉等^[4]运用互联网信息技术开发自然资源部门门户网站信息报送系统,主要讨论信息报送系统的技术架构、主体功能、关键技术和应用成效等,解决了门户网站信息报送渠道不统一、审批单保存与查询难度大、信息安全有风险等问题,实现了用户在电脑端、移动端报送和审核信息。曹守启等^[5]提出安全性更高的身份认证协议,提出了一种改进的远程用户身份认证协议,采用 BAN 逻辑进行验证,进行了性能比较与计算效率分析,其提出的协议在增加较低计算量的前提下具有更高的安全性。

为了系统提高报送信息大数据的安全性,本文提出一种基于改进关联规则的报送信息大数据特征隐匿性加密算法,以期提高报送信息的安全性,保证纪检工作的可靠性。

1 大数据特征隐匿性加密

本文加密算法的构建主要分为三部分:第一部分为报送信息大数据采集与预处理,包括缺失数据填补、离群值处理、数据标准化;第二部分为基于改进关联规则的报送信息大数据特征提取,通过改进遗传算法改进 Apriori 算法,挖掘数据项或属性之间隐含的关联规则;第三部分为报送信息大数据特征隐匿性加密实现,将神经网络层级之间的权重矩阵作为密钥,将明文数据转化为密文数据,可以同时加密多项特征数据。

1.1 报送信息大数据集成与预处理

前期为手工获取阶段,由现场实施小组将各业务系统数据集中,后期为自动化获取阶段,通过连接转换器实时连通异构数据库,来自动获取报送信息大数据^[6,7]。

1.1.1 手工获取阶段

加密处理之前需要做的第一件事情就是原始数据导入,利用“数字化纪检云平台”导入功能进行导入后,如果有重复,还需要允许用户进行挑选处理。系统提供完善的导入功能,支持字段超长报警、自动超长截取;支持 GB2312 和 UTF-8 字符集^[8,9];支持 CSV、TXT、ELX 等自定数据格式;支持表中导入数据清空等功能。同时系统支持 JSON 数据导入^[10]。数据导入后,提供数据审核、替换等功能保证数据的准确性。

1.1.2 自动连接阶段

当整个项目各个流程畅通后,初始数据改为自动连接,数据来源数据链接管理支持 MySQL 数据库^[11,12],同时增加 Excel 数据源(即保留手工获取功能)^[13]。为增加其适用性,需增加数据来源类型,提供其他常用数据库的数据接入,如 sg-rdb、pgsql、非关系型数据库等,通过 JDBC 连接串连通数据,并支持更多的不同的数据库版本,扩大数据来源。

集成的报送信息大数据还需要依据数据规范^[14],将不符合规范要求的数据自动修正,修正措施包括以下三种。

(1) 缺失数据填补

纪检报送信息大数据中极易出现部分数据缺失的现象,若该数据为关键数据,一旦缺失,该数据的可靠性就极大地降低,因此需要进行填补^[15]。方法过程如下:首先利用聚类算法获取到与存在缺失数据的同属性数据集合,然后计算每一个同类数据集的均值,记为 $\bar{x}_i, i = 1, 2, \dots, n$ 。其中, n 代表同类数据集的数量。然后再次计算均值,得到填补值,计算公式如下:

$$X = \frac{\sum_{i=1}^n \bar{x}_i}{n}, i = 1, 2, \dots, n \quad (1)$$

式中, X 代表填补值^[16]。

(2) 离群值处理

离群值,即错误数据,并不代表纪检报送数据集本身中的异常值。采用绝对值差中位数法进行处理,首先确定报送数据的中值,记为 x_{med} ,然后计算每个数据与 x_{med} 之间绝对偏差值,记为 $s_j, j = 1, 2, \dots, m$, m 为绝对偏差值的数量^[18]。然后确定 m 个绝对偏差值的中值,记为 s_{med} ,最后划分正常值范围,即 $[x_{med} - m \cdot s_{med}, x_{med} + m \cdot s_{med}]$ 。

遍历所有纪检报送数据,超过正常值范围的数据被认为是离群值,进行删除或者当作缺失值,重复缺失填补过程,重新插入即可^[18]。

(3)数据标准化

数据标准化,即将数据样本属性规范到某个指定的范围,统一量纲,规范公式如下:

$$x' = \frac{x - \min x}{\max x - \min x} \tag{2}$$

式中, x' 代表标准化后的报送信息数据; x 代表原始报送信息数据; $\min x$ 、 $\max x$ 代表数据集中最小值和最大值。

对整理转换完成的符合规范数据,导入新系统,提交并保存,用于后续研究。

1.2 报送信息大数据特征识别

利用改进后的 Apriori 算法对报送信息大数据进行挖掘,找出数据项或属性之间隐含的关联规则,通过关联规则识别报送信息大数据特征^[19,20]。Apriori 算法基本流程如图 1 所示。

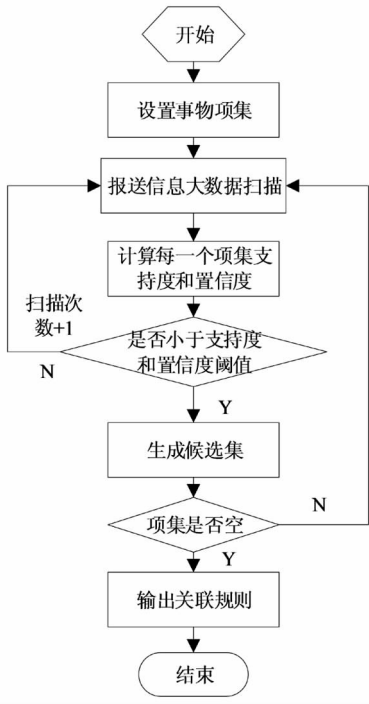


图 1 Apriori 算法基本流程

这种算法在工作时,需要重复扫描同样的报送信息大数据,效率并不高,因此引入遗传算法,通过遗传算法搜索频繁项集,改进关联规则生成速度^[21,22]。改进关联规则构建过程如下:

步骤 1:设置相关参数;

步骤 2:用 0 和 1 对报送信息大数据进行二进制编码,生成初始化种群;

步骤 3:计算每个个体的适用度。计算公式如下:

$$f(r_i) = \frac{S|\chi \cap \lambda|}{C(\chi|\lambda)} \tag{3}$$

式中,

$$S(\chi \cap \lambda) = \frac{N|\chi \cap \lambda|}{N|\gamma|} \tag{4}$$

$$C(\chi|\lambda) = \frac{N|\chi \cap \lambda|}{N|\chi|} \tag{5}$$

式中, $f(r_i)$ 代表第 i 个个体的适用度; $S|\chi \cap \lambda|$ 代表支持度; $N|\chi \cap \lambda|$ 代表报送信息大数据中项集 χ 和 λ 同时出现的次数; $N|\gamma|$ 代表报送信息大数据中事务数量。 $C(\chi|\lambda)$ 代表置信度; $N|\chi|$ 代表包含 A 的事务数量。

步骤 4:对种群进行三步遗传操作,更新粒子群;

步骤 5:重复上述步骤,直至满足迭代终止条件,得到频繁项集。

步骤 6:执行图 1 Apriori 算法基本流程,建立报送信息大数据集中数据项或属性之间隐含的关联规则。

基于得到的改进关联规则,识别报送信息大数据特征:按照置信度、支持度对挖掘到的关联规则进行从大到小的排序;从第一条规则开始逐一进行遍历,判断规则中是否包含一个属性,则加入特征库,否则同时删除所有属性属于规则体的数据;重复上述过程,完成报送信息大数据的约简,剩下的则是关键的、具有代表性的特征数据。

1.3 特征隐匿性加密算法

基于神经网络算法进行加密的原理是:利用神经网络层级之间的权重矩阵作为密钥^[23-25],将明文数据转化为密文数据,可以同时加密多项特征数据。具体过程如下:

步骤 1:初始化神经网络算法相关参数;

步骤 2:利用神经网络层级之间的权重随机生成一密钥 P 。该密钥是一个为 7×6 的矩阵,如下:

$$P = \begin{bmatrix} 0, 1, 0, 0, -1, 0 \\ 0, -1, 1, 0, 1, 1 \\ 1, 1, 1, -1, -1, 0 \\ 0, 1, 1, 0, 1, -1 \\ 0, 0, 0, 0, 0, 0 \\ 1, 1, 0, 0, 0, -1 \\ 0, 1, 0, 1, 0, -1 \end{bmatrix} \tag{6}$$

步骤 3:将待加密的报送信息大数据特征转换成与密钥矩阵同等大小的矩阵形式,记为 Q ;

步骤 4:以 Q 中每一个报送信息数据为输入,经过神经网络三层处理,将文本嵌入密钥矩阵中,嵌入公式如下:

$$\psi = \sum_{i=1}^n P \cdot q_{ij}$$

(7)

式中, ψ 代表报送信息大数据特征密文数据; q_{ij} 代表 Q 中第 i 行第 j 列报送信息数据。

步骤 5:判断报送信息大数据特征密文数据的隐匿性是否满足要求,若满足设定的标准^[26],输出加密结果,否则回到步骤 4,在现有密文的基础上再次嵌入到密钥矩阵中,直至满足要求。

通过上述过程,完成报送信息大数据特征隐匿性加密。

2 算例分析

依托“数字化纪检云平台”,进行加密算法应用效果测试,如图 2 所示。

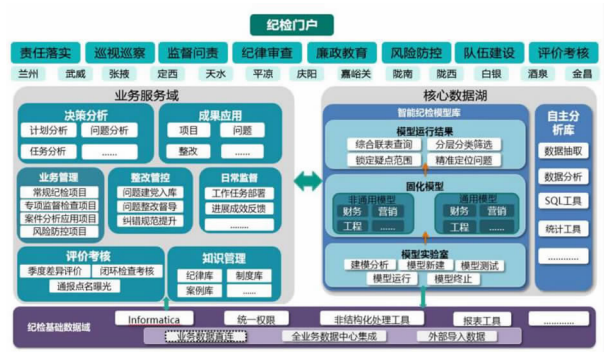


图 2 数字化纪检云平台

2.1 报送信息数据样本

以数据中台和全业务统一数据中心数据为基础,以纪检业务库数据为补充,在供电公司全业务统一数据中心建设纪检基础数据平台,对接 SG-186、SG-ERP、车辆管控平台、招标平台等业务系统,采用流式数据处理思路,业务按需采集全业务统一数据中心业务数据和 6 大纪检业务库数据。

而费用督查一直是纪检部门关注的重点,因此以监控机关本部和各基层单位“三公经费”开支数据为样本,具体包括因公出国(境)经费、公务用车费、公务接待费等,即应用 SQL 工具、Excel 高级应用工具等,直接对纪检数据进行查询和分析,在报送信息数据统一上传后,平台自动将报表进行整

合归纳,生成统一数据,并提供敏感、违规字段检测功能。

2.2 初始化参数设置

上文研究中涉及的遗传算法和神经网络算法的初始化参数如表 1 所示。

表 1 遗传算法和神经网络算法的初始化参数		
算法	名称	参数
遗传算法	种群规模/个	36
	编码长度/个	1076
	交叉概率/%	0.50
	变异概率/%	1.30
	最大迭代次数/次	200
神经网络算法	输入层节点数/个	25
	隐含层节点数/个	40
	输出层节点数/个	10
	初始权值	0
	输出层阈值	0.10
	训练目标	0.0001
	学习速率	0.02
	激励函数	Sigmoid 型
	最大迭代次数/次	500

2.3 特征识别测试结果

基于 1.2 节,挖掘关联规则,并识别报送信息大数据特征,结果如表 2 所示。

表 2 基于关联规则的特征识别测试结果			
项目	规则序号	支持度	置信度
关联规则	1	1.2565	2.3654
	2	1.2368	1.5783
	3	1.4262	0.8852
	4	1.8572	1.4224
	5	1.4265	2.6968
	6	1.3320	6.3255
	7	1.0215	4.2212
数据类型		特征维度	特征属性
报送信息 大数据特征	因公出国(境)经费	2	10
	公务用车费	6	12
	公务接待费	3	18

2.4 加密性能测试

以加密前后明文和密文的相关性来评价加密算法性能。相关性计算公式如下:

$$\text{Sim}(\phi,\varphi)=\frac{|F(\phi,\varphi)|}{\sqrt{D_x}\times\sqrt{D_y}}\tag{8}$$

式中, $\text{Sim}(\phi,\varphi)$ 代表明文 ϕ 和密文 φ 之间的相关性,该值取值 $0\sim 1$,划分 5 个等级,其中 $0\sim 0.2$ 之间代表加密效果最好,信息最安全; $|F(\phi,\varphi)|$ 代表明文 ϕ 和密文 φ 之间的协方差绝对值; D_x 、 D_y 分别代表明文 ϕ 和密文 φ 之间的标准差。

统计“三公经费”开支数据加密后的密文和明文相关性 $\text{Sim}(\phi,\varphi)$,结果如图 3 所示。

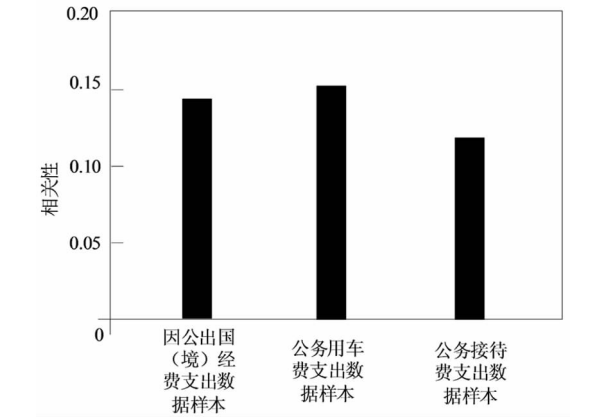


图 3 “三公经费”开支数据加密后明文和密文的相关性

从图 3 中可以看出,所研究加密算法应用下,“三公经费”开支数据加密后的密文和明文相关性均在 $0\sim 0.2$ 之间,说明所研究算法的加密效果较好,更能保证报送信息大数据的安全。

选取多个报送信息大数据特征,利用三种方法即本文方法、文献[5]方法和文献[6]方法对每组报送信息大数据特征进行转换,获取密钥矩阵同等大小的矩阵,已知在最理想的转换条件下,特征与元素可一比一转换,但是将文本嵌入密钥矩阵的过程中,数次迭代后容易出现损失值,对此,分析不同方法的转换效果,实验结果如图 4 所示。

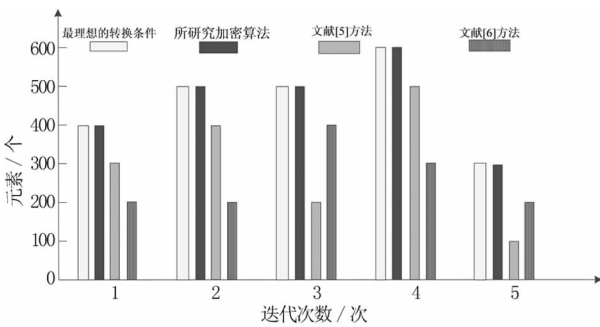


图 4 三种方法的转换效果

从图 4 中可以看出,当迭代次数为 1 时,最理想条件的元素个数为 400,所研究算法与理想条件一致,文献[5]方法元素个数为 400,文献[6]方法元素个数为 200;当迭代次数为 2 时,最理想条件的元素个数为 500,所研究算法元素个数也为 500,文献[5]方法元素个数为 400,文献[6]方法元素个数为 200;当迭代次数为 3 时,最理想条件的元素个数为 500,所研究算法与最理想条件一致,文献[5]方法元素个数为 200,文献[6]方法元素个数为 400;当迭代次数为 4 时,最理想条件的元素个数为 600,所研究算法元素个数也为 600,文献[5]方法元素个数为 500,文献[6]方法元素个数为 300;当迭代次数为 5 时,最理想条件与所研究算法的元素个数均为 300,文献[5]方法元素个数为 100,文献[6]方法元素个数为 200。由此可知,所研究算法的转换条件与理想转换条件基本一致,在转换过程中基本无损失值,因此转换效果较好。而文献[5]方法与文献[6]方法的转换条件与理想转换条件偏差较大,容易出现损失值,因此转换效果较差。因此,应用所研究算法不会影响最终的加密结果,可以保证实际应用效果。

3 结 论

“数字化纪检云平台”通过集成监督对象的个人学习、成长经历、工作经历、籍贯、审批项目的中标厂商等基本报送信息大数据。报送信息大数据直接关系到纪检平台功能的发挥,面对这种情况,提出一种基于改进关联规则的报送信息大数据特征隐匿性加密算法,研究结论如下。

- (1)利用改进遗传算法改进 Apriori 算法,挖掘数据项或属性之间隐含的关联规则,精确刻画项目与人、事与人的关联关系,发现规律,挖掘异常情况。
- (2)以关联规则为连接,找到报送信息大数据填补值,保证密文转换过程的完整性和逻辑性,及时在纪检平台进行比对、分析、甄别,精准发现问题,加密后密文和明文相关性均在 $0\sim 0.2$ 之间。

参考文献

[1] 李嘉. 国有上市公司纪检监察与内部审计合署办公实践探析[J]. 财会通讯,2022(11):128—133.

[2] 吴金群,曾智洪,游晨. 廉政治理的技术应用及超越:系统集成的智慧纪检监察模式研究[J]. 党政研究,2020(6):15—25.

[3] 田开友. 电子商务平台经营者涉税信息报送义务履行的规范化研究——基于我国《电子商务法》第 28 条规定的阐释[J]. 政治与法律,2021(3):137—148.

[4] 陈卉,叶兴茂,侯一俊,等. 自然资源部门户网站信息报送系统研发与应用[J]. 国土资源信息化,2020(4):28—32.

[5] 曹守启,何鑫,刘婉荣. 一种改进的远程用户身份认证方案[J]. 计算机工程与科学,2021,43(11):1959—1965.

[6] 张峻,阚银环. 基于量化行为的异构数据库强制访问控制方法[J]. 计算机仿真,2021,38(8):487—491.

[7] 李珺,刘鹤,朱良宽. 基于改进的 K-means 算法的关联规则数据挖掘研究[J]. 小型微型计算机系统,2021,42(1):15—19.

[8] 任克强,王传强. 基于 STC8F1K08S2 的串口驱动 OLED 显示系统设计[J]. 电子器件,2022,45(1):40—44.

[9] 葛光富. 面向信息系统跨平台轻量应用的汉字编码转换程序[J]. 计算机系统应用,2020,29(7):251—255.

[10] 李成严,辛雪,赵帅,等. Sp-IEclat:一种大数据并行关联规则挖掘算法[J]. 哈尔滨理工大学学报,2021,26(4):109—118.

[11] 蒋睿. MySQL 数据库安全研究[J]. 电脑知识与技术,2020,16(9):3—4+21.

[12] 孙妍. PHP 语言访问 MySQL 数据库的实现[J]. 数码设计(上),2020,9(10):23.

[13] 邱均平,崔腾腾,陈仕吉. 基于聚类 and 关联规则的 Altmetric TOP 榜文献特征分析[J]. 现代情报,2021,41(9):12—21+63.

[14] 李勇男. 基于多层次关联规则挖掘的反恐情报跨层特征关联分析[J]. 情报科学,2021,39(11):127—132.

[15] 赵毅强,宋凯悦,何家骥,等. 基于安全授权密钥的增强型逻辑加密算法[J]. 华中科技大学学报(自然科学版),2020,48(5):1—6.

[16] 张久岭,黄道超,沈时军. 加密文档排序中保序加密算法的最优化选取[J]. 北京航空航天大学学报,2022,48(2):266—272.

[17] 周辉,谢红薇,张昊,等. 混沌系统和 DNA 编码的并行遥感图像加密算法[J]. 中国图象图形学报,2021,26(5):1081—1094.

[18] 伍书重,莫思阳,张友良,等. 数值流形法局部网格加密算法[J]. 科学技术与工程,2021,21(7):2850—2856.

[19] 张雷,董万富,阚欢迎,等. 基于改进 Apriori 算法的客户需求数据分析方法[J]. 机械设计与制造,2020(5):185—188.

[20] 陈小宇,叶佳栋. 基于 CHIP ID 的 FPGA 加密算法设计与实现[J]. 电子技术应用,2020,46(11):100—103.

[21] 毛伊敏,邓千虎,陈志刚. 基于信息熵与遗传算法的并行关联规则增量挖掘算法[J]. 通信学报,2021,42(5):122—136.

[22] 乔东平,柏文通,文笑雨,等. 基于关联规则的作业车间调度问题改进遗传算法研究[J]. 河南理工大学学报(自然科学版),2022,41(2):138—148.

[23] 王剑锋,林永峰. 基于神经网络的电力调度信息加密传输方法研究[J]. 微型电脑应用,2022,38(1):134—138.

[24] 柯亚文,蔡挺,夏晓峰,等. 面向嵌入式系统的加密算法性能检测方法[J]. 重庆大学学报,2020,43(11):1—10.

[25] 谭静仪,周旭毓. 基于信息熵抑制的大数据特征隐匿性加密仿真[J]. 计算机仿真,2020,37(4):187—191.

[26] 蔡中民. 多层实时网络加密数据流频繁项集挖掘方法[J]. 沈阳工业大学学报,2021,43(3):301—306.