DOI:10.13193/j.issn.1673-7717.2025.04.004

基于安全多方计算和区块链技术的中医药临床数据共享方法研究:以频数分析和描述性统计为例

何畅[1],赵冉[1],张雯[2],张泽丹[3],温宵宵[3],古欣怡[4],庄荟瑛[5],王斌[1,2]
(1.中国中医科学院中医药信息研究所,北京 100700;2.中国中医科学院中医药科技合作中心,北京 100700;
3.中国中医科学院中医药数据中心,北京 100700;4.中国中医科学院医学实验中心,北京 100700;
5.内蒙古民族大学公共卫生学院,内蒙古 通辽 015099)

摘要:区块链和安全多方计算协议可以解决数据共享时的安全隐私问题。区块链确保计算过程的可追溯性,防止恶意节点篡改上传数据;安全多方计算协议将互不可信的用户作为参与方整合起来,建立安全可靠的合作机制。基于区块链和安全多方计算协议设计中医药临床数据分布式算法能整合来自多方的临床数据,扩大研究样本量,解决目前中医药临床数据共享存在的"不愿,不敢,不能"问题,促进中医药真实世界研究的发展。创造性地将频数分析与描述性统计融入安全多方计算协议,融合了区块链技术增强数据共享的安全性和透明度,为中医药临床数据设计了一套高效且保护隐私的分布式计算方案。分析了该方案相较于集中式计算在数据安全和隐私保护方面的优势,还设计实验证明了该方案在计算真实中医临床数据时的准确性和可行性。该方案为中医药领域的数据共享开创了新方向,既保障了患者隐私,又实现了临床数据的有效聚合,加速了中医药现代化和科学化进程。

关键词:中医药临床数据;数据共享;区块链;安全多方计算;分布式算法
中图分类号:R241 **文献标志码**:A **文章编号**:1673-7717(2025)04-0018-08

Research on Secure Sharing Methods for Clinical Data of Traditional Chinese Medicine Based on Secure Multi-Party Computation and Blockchain Technology: Taking Frequency Analysis and Descriptive Statistics As Examples

HE Chang[1],ZHAO Ran[1],ZHANG Wen[2],ZHANG Zedan[3],WEN Xiaoxiao[3],GU Xinyi[4],ZHUANG Huiying[5],WANG Bin[1,2]
(1. Institute of Information on Traditional Chinese Medicine,China Academy of Chinese Medical Sciences,Beijing 100700,China;
2. Chinese Medicine Science and Technology Cooperation Center,China Academy of Chinese Medicine Sciences,Beijing 100700,China;
3. National Data Center of Traditional Chinese Medicine,China Academy of Chinese Medicine Sciences,Beijing 100700,China;
4. Experimental Research Center,China Academy of Chinese Medical Sciences,Beijing 100700,China;
5. School of Public Health,Inner Mogolia Minzu University,Tongliao 015099,Inner Mogolia,China)

Abstract:Blockchain and secure multi-party computation(SMPC) are feasible ways to solve security and privacy issues of data sharing. Blockchain ensures the traceability of computing processes and prevents malicious computing party to modify its upload data. The secure multi-party computation protocols can integrate untrustworthy users as participants to establish safe and trusted cooperation mechanism. The design of traditional Chinese medicine(TCM) clinical data algorithm based on blockchain and secure multi-party computing protocols can integrate the multiparty clinical data,expand the research sample size,solve the problem of "being unwilling,afraid and unable" to share TCM clinical data and promote the development of TCM real-world study. This article creatively integrated frequency analysis and descriptive statistics into SMPC protocols,integrated blockchain technology to enhance the security and transparency of data sharing,so as to design a set of efficient and privacy-protecting distributed computation scheme for TCM clinical data. This article not only analyzed the advantages of this scheme in data security and privacy protection compared with centralized computation,but also designed experiments to prove the accuracy and feasibility of this scheme in calculating real TCM clinical data. The scheme creates a new direction for TCM data sharing,which not only guarantees patient privacy,but also realizes effective aggregation of clinical data,and accelerates the modernization and scientific process of TCM.

Keywords:TCM clinical data;data sharing;blockchain;secure multi-party computation;distributed algorithm

基金项目:国家自然科学基金面上项目(82274685);中国中医科学院科技创新工程重大攻关项目(CI2021A01009)
作者简介:何畅(2000-),男,浙江义乌人,硕士在读,研究方向:数据挖掘、中医药机器学习。
通讯作者:王斌(1976-),男,辽宁建昌人,研究员,博士研究生导师,博士,研究方向:中医药数据、计算机与信息工程。E-mail:wangbin135@139.com。

近年来,随着我国信息化的深入,新一代互联网、物联网、云计算技术的广泛应用,中医药行业的数据呈爆发式增长。中医药临床数据是中医临床诊疗过程的直观体现,包括患者个人信息、入院信息、住院或门诊病历、中药处方、理化检查等内容。共享中医药临床数据能提高数据利用率,减少低水平重复试验,为政府医疗决策提供支持。由于中医药临床数据通常保存在医院或中医馆的本地数据库中,内容涉及患者隐私,在共享时需要保障它们离开本地环境后的安全性。为此,安全多方计算和区块链技术提供了一种可行方案[1]。

安全多方计算和区块链都用于处理分布式计算环境中的安全和信任问题[2]。安全多方计算是隐私计算的三大方向之一,指参与方在不共享各自数据且无可信第三方的情况下安全地计算约定函数。姚期智最早提出安全多方计算的概念并设计了第一个安全多方计算的基础协议:安全两方比较协议,用于解决百万富翁问题[3]。随着研究的深入,更多基础协议,如点积协议、加到乘变换协议、安全多方求和协议、秘密比较协议、茫然传送协议等被设计出来[4]。安全多方计算协议是解决隐私保护问题的工具包,可以基于它设计算法。区块链是记录数据交易的去中心化账本,通过多方存证的方式为交易过程提供可信的执行环境[5]。GAO H M 等[6]提出基于商用应用系统区块链的安全多方计算方案,包含公开的可验证秘密共享和激励机制,具有公平性和鲁棒性。在原始数据存储于不同机构的前提下,研究者可以通过区块链平台将任务信息发往各数据库,数据库管理员分享各自的计算结果再由研究者汇总后计算得出最终结果,实现安全多方计算[7]。

本文描述了区块链的构成,设计了适用于中医药临床数据频数分析和描述性统计的安全多方计算协议,满足了多方协同共享数据时保护隐私的需要。安全多方计算的链上计算由智能合约自动执行,采用 JAVA 语言编写,由于代码冗长、不易理解,本文仅提供计算的实现过程。

1 区块链实现原理

中医临床数据的安全共享依托于区块链互联网平台[8]实现。平台具有链下的前端界面,用户在此界面发送和接收任务,但数据交易和安全计算在链上完成。智能合约作为交易的接收方执行安全计算,发出包含计算结果的事件。事件会记录在交易收据的日志中供外部系统监听。交易信息经节点验证后被打包进区块连接成链[9]。

1.1 交易 用户调用智能合约的函数上传数据需要创建一笔交易。交易包含 8 个字段:发送地址、接收地址、账户交易数量、发送者签名、机构 id、任务 id、节点信息、上传数据。

其中发送地址由 ANSI X9.62 标准编码的非压缩格式公钥经 keccak-256 算法取后 20 个字节并在前面加“0x”生成,公钥由创建账户时系统随机生成的私钥经 secp256k1 算法得出。接收地址是合约地址。账户交易数量是从发送地址发出交易的序列号。发送者签名由私钥经椭圆曲线数字签名算法(Elliptic Curve Digital Signature Algorithm,ECDSA)生成,用于证明交易发起者的身份和交易内容的完整性。节点信息是多字段对象,包含节点的类型、状态、名称、节点区块高度和时间。上传数据由原始明文数据经 UTF-8 编码和 SM4 算法得出。原始明文数据包含表标题、表内容、元素总数、状态码、任务类型 5 个字段,示例如下。

“uploaddata”:[{“title”:{“no”:“序号”,“elementname”:“中医主证”,“gendercount”:“数量”},“result”:[{“no”:“1”,“sexname”:“脾虚不固证”,“gendercount”:3},{“no”:“2”,“sexname”:“脾失健运,湿浊流注证”,“gendercount”:1}],“total”:4,“code”:“200”,“tasktype”:“0”}]。

示例中的表标题和表内容包含序号、元素选项、数量 3 个字段,状态码为 200 表示上传数据成功,任务类型为 0 表示任务类型是频数分析。

交易创建后会发生以下情况。

①后端系统将交易的非签名字段编码后通过 MD5 加盐算法生成交易哈希,发送者使用私钥签署交易哈希生成数字签名,签名被附加到交易信息中。

②发送者将交易传播给其他节点确认。若确认有效,节点将其添加到各自的交易池中。若确认无效,节点会拒绝该交易,交易将不会执行,发送者需要纠正错误并重新创建交易。

③某个节点挑选这笔交易并尝试将其打包进新区块。他们在本地运行智能合约代码确保交易有效并计算合约的状态改变。合约执行完毕后,系统生成交易收据记录交易状态和合约执行结果。节点计算出正确的区块哈希表示新区块创建完成。

④创建新区块的节点把这个区块发送给其他节点验证,若验证通过,节点将其添加到各自区块链末尾并将其包含的交易从各自交易池中移除。若验证未通过,节点会拒绝该区块并播报无效通知,该区块包含的交易会因区块失效而被放回交易池,区块内的合约操作将不会执行。

1.2 区块 该平台区块链的一个区块仅容纳一笔交易,通过包含前一区块哈希与前一区块相连。每个区块包含 8 个字段:前一区块哈希、区块哈希、交易哈希、收据哈希、时间戳、区块高度、区块版本号、交易详情[10]。

其中区块哈希由前一区块哈希、交易哈希、收据哈希、时间戳、区块版本号编码后经 MD5 加盐算法得到,区块中的任何改变都会导致后续区块哈希值变动,从而被用户发现。收据哈希由交易收据的交易状态和日志编码后经 MD5 加盐算法得出。时间戳是以 Unix 时间戳形式表示的区块打包时间。区块高度指区块连接在区块链上的个数,由于不同节点传播和验证新区块存在时间差,同一时间观测不同节点的区块高度可能不同。交易收据存储在链下的状态数据库中,可根据其哈希值定位和获取。区块链示意图见图 1。

1.3 智能合约 智能合约是运行在区块链上的程序,能解析交易信息、存储上传数据、执行链上计算、发出包含计算结果的事件。它的执行过程透明、不可篡改、不受人为干预。智能合约包含以下元素:用于接收和解密上传数据、将必要参数存储在合约内部的公开函数 uploaddata;用于封装计算数据的 JAVA 对象;用于触发合约链上计算的公开函数 triggercomputation;由 triggercomputation 函数调用、能自动遍历合约内部数据并完成计算的内部函数,详见章节 2.3、3.3;包含计算结果的事件,在计算完成后发出。

部署合约需要发送一笔包含编译后合约代码的交易到区块链上,系统将账户地址和账户交易数量通过 keccak256 算法得出合约地址[11]。智能合约拥有合约账户,能成为交易的接收者。

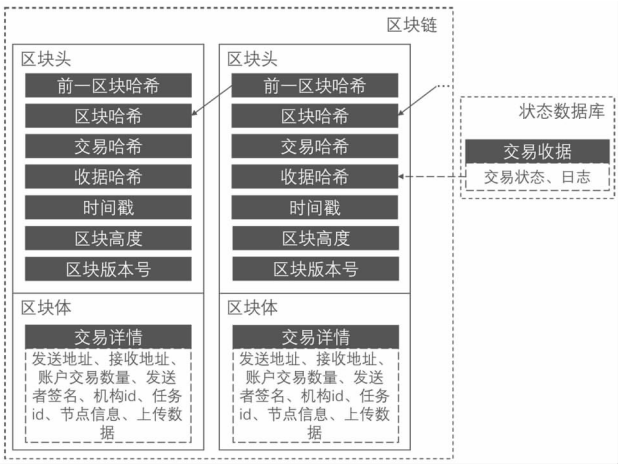


图 1 区块链示意图

1.4 安全多方计算流程 使用区块链平台进行安全多方计算的流程如下。

管理员用户在前端界面执行以下操作:新增算法模型,选择基础算法类型(描述性统计或频数分析),设定过滤指标和观察指标;新增分析任务,选择算法模型和计算节点;发布任务,服务器通过 websocket 通道将任务发送给节点用户。

各计算节点获取任务,根据任务信息准备数据、进行链下计算、调用智能合约的 uploaddata 函数上传计算结果。前端系统自动对上传的原始明文数据进行 utf-8 编码和 SM4 加密。

智能合约解析交易信息、使用 SM4 密钥解密上传数据、将必要参数存储在合约内部状态中。

所有计算节点上传完数据后,监管节点调用 triggercomputation 函数向合约发送交易,合约执行链上计算、发出事件。

外部系统监听交易收据中的事件,将合约执行结果呈现在前端界面。具体步骤见图 2。

2 频数分析

频数分析是对一组数据中不同类别样本的频数,或是样本落入指定区域内的频数进行统计,了解其分布情况的方式。频

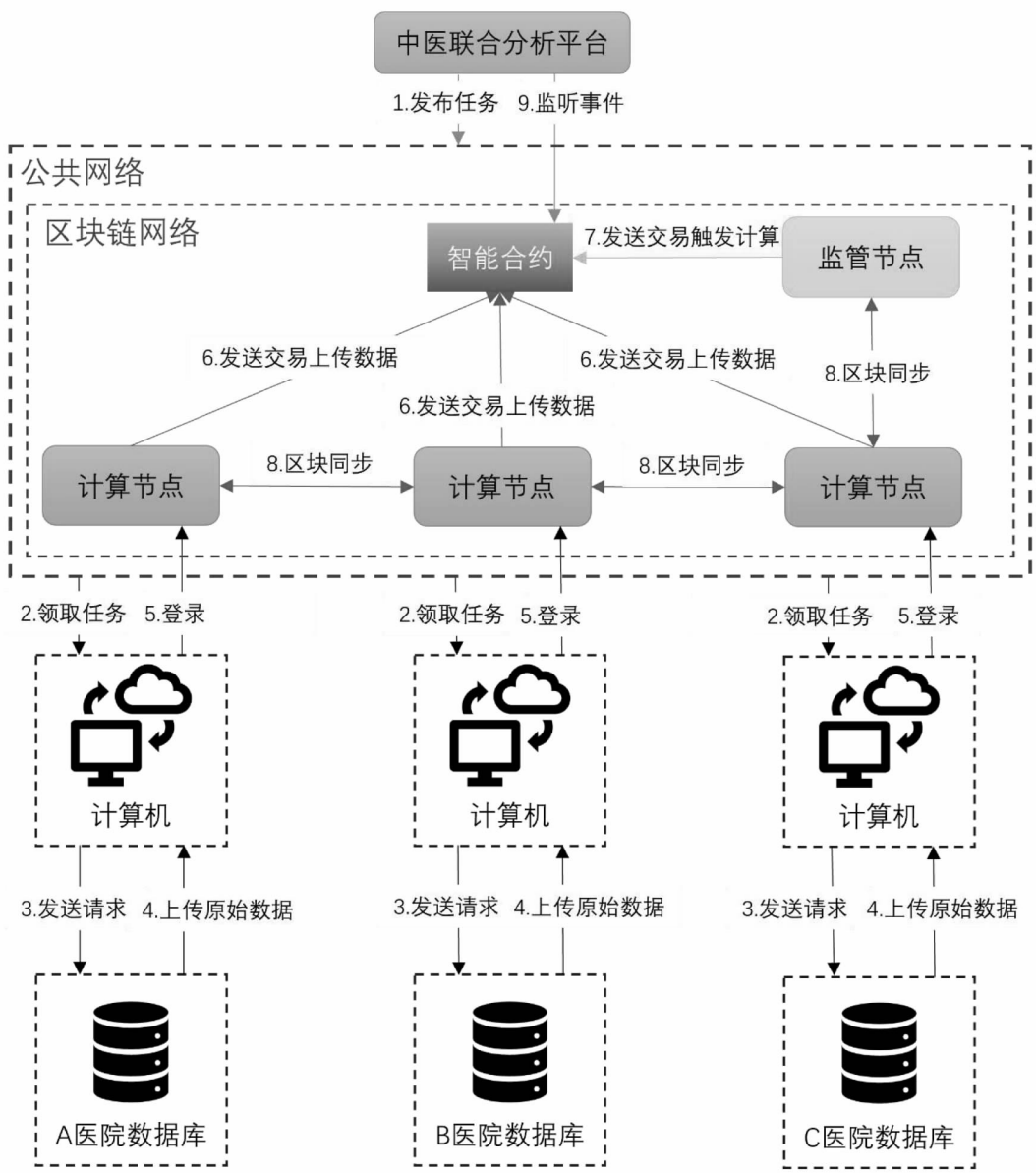


图 2 安全多方计算流程

数分析方法适用于年龄、性别、中医主证等数据的统计,用于发现临床数据的分布规律,反映样本是否具有总体代表性、抽样是否存在系统偏差。

2.1 频数分析集中式计算方法 频数分析的集中式算法由分节点直接向中心节点上传病历数据。中心节点合并上传数据,统计样本名称、频数、总频数、百分比、累计百分比,示例如下。

节点 A 获取病历数据,患者 id:{001,002,003,004,005},脉象:{脉弦,脉滑,脉数,脉沉,脉弦}。节点 B 获取病历数据,患者 id:{001,002,003,004,005,006},脉象:{脉弦,脉细,脉滑,脉数,脉沉,脉沉}。节点 A、B 将病历数据发送给中心节点,中心节点统计:样本名称:{脉弦,脉细,脉滑,脉数,脉沉},频数:{3,1,2,2,3}、总频数:11,百分比:{27.27%,9.09%,18.18%,18.18%,27.27%},累计百分比{27.27%,36.36%,54.54%,72.72%,99.99%}。

2.2 频数分析安全多方计算协议 频数分析的安全多方计算协议要求计算节点统计后上传样本名称和频数。智能合约根据上传数据,统计样本名称、频数、总频数、百分比、累计百分比,示例如下。

节点 A 获取病历数据,统计:样本名称:$\{x_1,x_2,x_3,x_4\}$,频数:$\{f_1,f_2,f_3,f_4\}$。节点 B 获取病历数据,统计:样本名称:$\{x_1,x_2,x_3,x_4,x_5\}$,频数:$\{f_1,f_2,f_3,f_4,f_5\}$。节点 A、B 将统计结果发送给智能合约,合约计算如下。

样本名称:

$$A\cup B=\{x_1,x_2,x_3,x_4,x_5\}$$

样本频数:

$$A\cup B=\{f_1+f_5,f_2+f_6,\cdots,f_9\}$$

总频数:

$$S=f_1+f_2+f_3+f_4+f_5+f_6+f_7+f_8+f_9$$

定类数据 x_1 的百分比:

$$k_1=\frac{f_1+f_5}{s}*100\%$$

累计百分比:

$$\{\sum_{i=1}^{1}k_i,\sum_{i=1}^{2}k_i,\sum_{i=1}^{3}k_i,\sum_{i=1}^{4}k_i,\sum_{i=1}^{5}k_i\}$$

2.3 频数分析链上计算 智能合约执行频数分析链上计算的实现步骤如下。

①声明频数分析的 JAVA 对象 Frequencystatisticsvo 包含元素选项、数量、元素名称、元素类型、百分比、累计百分比 6 个字段。

②智能合约遍历内部存储的计算数据,以元素名称为键、相同元素名称对象的数组为值构建哈希表 map。

③获取 map 中一个键值对的值,将相同元素选项对象的字段“数量”累加赋值给变量 aLong,以元素选项为键,aLong 为值存储在树表 map2 中,若元素类型为定量变量且元素选项种类大于 8,则将元素选项的区间划分为等长的 4 个区域,以 4 个区域为键,落入区域对象的字段“数量”之和为值存储在树表 map2 中。

④将 map2 中键值对的值相加赋值给变量 all,以元素名称为键,all 为值存储在哈希表 map3 中。

⑤声明 Frequencystatisticsvo 类型变量 vo,将 map3 的键、map2 的键、map2 的值分别设为 vo 的字段“元素名称”“元素选项”和“数量”,将 aLong 除以 all 的结果设为 vo 的字段“百分比”,将百分比逐个累加设为 vo 的字段“累计百分比”。

⑥以元素名称为键,vo 组成的数组为值存储在 map1 中,将 map1 封装在合约事件中发出。

其中哈希表和树表是存储键值对的数据结构,哈希表中的键值对是无序的,树表中的键值对按照键的大小进行排序。键值对由唯一的键和与之配对的值组成,可以根据键快速访问对应值。模拟 2 个元素总数为 100 的数据集,数据结构与实现步骤见图 3。

3 描述性统计

描述性统计是分析一组定量数据集中、离散程度的方法,适用于年龄、身高、体质量、药物用量、实验室检查结果等定量数据的分析,能发现疾病发生和药物使用规律。描述性统计需要计算总样本量、最值、算数平均数、方差、标准差、变异系数、偏度系数、峰度系数的数值。

3.1 描述性统计集中式计算方法 描述性统计集中式算法由分节点向中心节点发送病历数据。中心节点合并病历数据并计算一系列数值,示例如下。

节点 A 获取病历数据,患者 id:{001,002,…,00a},样本:$\{x_1,x_2,\cdots,x_a\}$。节点 B 获取病历数据,患者 id:{001,002,…,00b},样本:$\{y_1,y_2,\cdots,y_b\}$。节点 A、B 将病历数据发送给中心节点,中心节点统计如下。节点样本量:a、b,最值:Max、Min,总样本量:

$$n=a+b$$

算术平均数:

$$\mu=\frac{\sum_{i=1}^{a}+x_i+\sum_{i=1}^{b}y_i}{a+b}$$

方差:

$$\sigma^2=\frac{\sum_{i=1}^{a}(x_i-\mu)^2+\sum_{i=1}^{b}(y_i-\mu)^2}{a+b}$$

标准差:

$$\sigma=\sqrt{\frac{\sum_{i=1}^{a}(x_i-\mu)^2+\sum_{i=1}^{b}(y_i-\mu)^2}{a+b}}$$

变异系数:

$$CV=\frac{\sigma}{\mu}*100\%$$

偏度系数:

$$S_k=\frac{\sqrt{n(n-1)}}{n-2}*\frac{\{\frac{1}{n}[\sum_{i=1}^{a}(x_i-\mu)^3+\sum_{i=1}^{b}(y_i-\mu)^3]\}}{\{\frac{1}{n}[\sum_{i=1}^{a}(x_i-\mu)^2+\sum_{i=1}^{b}(y_i-\mu)^2]\}^{\frac{3}{2}}}$$

峰度系数:

$$K_u=\frac{(n-1)n(n+1)}{(n-2)(n-3)}*\frac{[\sum_{i=1}^{a}(x_i-\mu)^4+\sum_{i=1}^{b}(y_i-\mu)^4]}{[\sum_{i=1}^{a}(x_i-\mu)^2+\sum_{i=1}^{b}(y_i-\mu)^2]^2}-\frac{3(n-1)^2}{(n-2)(n-3)}$$

3.2 描述性统计安全多方计算协议 描述性统计安全多方计算协议为避免上传原始数据,需要计算节点进行 2 次链下计算并上传 2 次计算结果,智能合约执行 2 次链上计算,示例如下。

①节点 A 获取病历数据,统计:样本量:a,最大值:x_{max},最

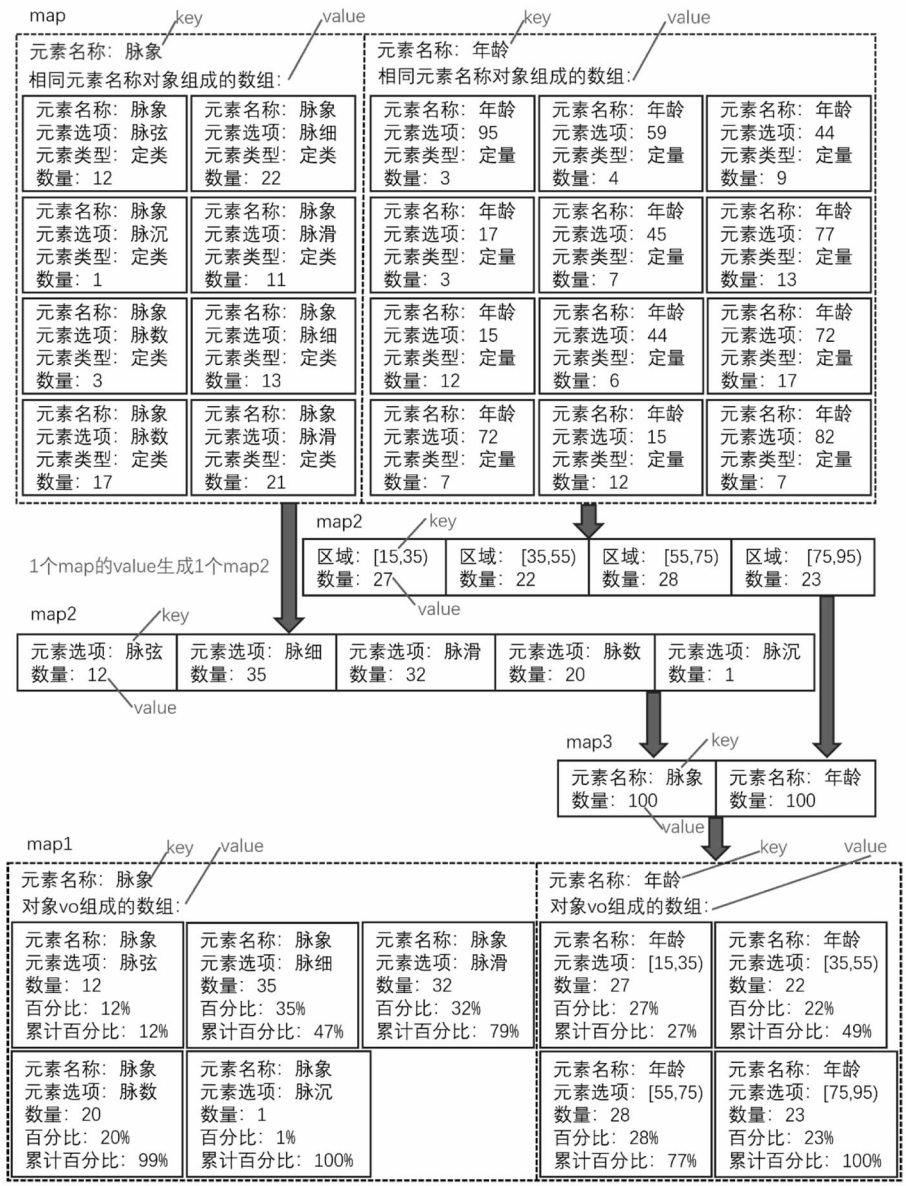


图 3 频数分析的数据结构和实现过程

小值：x_{min}，算术平均数：$\bar{x}$，将统计结果发送给智能合约。

②节点 B 获取病历数据，统计：样本量：b，最大值：y_{max}，最小值：y_{min}，算术平均数：$\bar{y}$，将统计结果发送给智能合约。

③智能合约获取第 1 次上传的数据，分别比较 x_{max}、y_{max} 和 x_{min}、得到最大值 Max，最小值 Min，计算如下，并将结果通过事件发出。

总样本量：

$$n = a + b$$

算术平均数：

$$\mu = \frac{a\bar{x} + b\bar{y}}{n}$$

④节点 A 获取 μ，计算如下，并将结果发送给智能合约。

离均差平方和：

$$T_1 = \sum_{i=1}^{a} (x_i - \mu)^2$$

离均差立方和：

$$U_1 = \sum_{i=1}^{a} (x_i - \mu)^3$$

离均差四次方和：

$$V_1 = \sum_{i=1}^{a} (x_i - \mu)^4$$

⑤节点 B 获取 μ，计算如下，并将结果发送给智能合约。

离均差平方和：

$$T_2 = \sum_{i=1}^{b} (y_i - \mu)^2$$

离均差立方和：

$$U_2 = \sum_{i=1}^{b} (y_i - \mu)^3$$

离均差四次方和：

$$V_2 = \sum_{i=1}^{b} (y_i - \mu)^4$$

⑥智能合约获取第 2 次上传的数据并计算。

方差：

$$\sigma^2 = \frac{T_1 + T_2}{n}$$

标准差：

$$\sigma = \sqrt{\frac{T_1 + T_2}{n}}$$

变异系数：

$$CV = \frac{\sigma}{\mu} * 100\%$$

偏度系数：

$$S_k = \frac{\sqrt{n(n-1)}}{n-2} * \frac{\frac{1}{n}(U_1 + U_2)}{\sigma^3}$$

峰度系数：

$$K_u = \frac{(n-1)n(n+1)}{(n-2)(n-3)} * \frac{(V_1 + V_2)}{(T_1 + T_2)^2} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

3.3 描述性统计链上计算 智能合约执行描述性统计链上计算的实现步骤如下。

①声明描述性统计的 JAVA 对象 DescriptiveStatisticsVo，包含 14 个字段：描述信息、最大值、最小值、平均值、样本数、样本总数、方差、标准差、峰度、偏度、变异系数、离均差平方和、离均差立方和、离均差四次方和。

②从合约内部获取第 1 次上传数据转化为 DescriptiveStatisticsVo 类对象组成的数组，以描述信息为键、相同描述信息对象组成的数组为值构建哈希表 map。

③获取 map 中 1 个键值对，获取值中每个对象的字段："样本数""最大值""最小值""平均值"，比较得出最大值 max1，最小值 min1，将样本数相加得到样本总数，将样本数乘以平均值的结果相加赋值给变量 all，将 all 除以样本总数的结果保留 1 位小数赋值给变量 avg1，将描述信息、avg1 封装进合约事件发出。

④声明 DescriptiveStatisticsVo 类型变量 vo1，将描述信息、max1、min1、avg1、样本总数设为 vo1 字段："描述信息""最大值""最小值""平均值""样本总数"，将 vo1 存储在合约内部。

⑤从合约内部获取第 2 次上传的数据转化为 DescriptiveStatisticsVo 类型对象组成的数组，以描述信息为键，相同描述信息对象组成的数组为值存储在哈希表 map2 中。

⑥分别累加相同描述信息对象的字段"离均差平方和""离均差立方和""离均差四次方"，与 vo1 字段"样本总数"一起代入方差、标准差、偏度系数、峰度系数的公式进行计算，将结果四舍五入保留 3 位小数后分别设为 vo1 的字段"方差""标准差""峰度""偏度"，将 vo1 的字段"标准差"除以 vo1 的字段"平均值"的结果设为 vo1 的字段"变异系数"。

⑦以描述信息为键，相同描述信息的 vo1 为值存储在哈希表 map1 中，将 map1 封装进合约事件发出。以章节 4 中数据为例，数据结构和计算过程见图 4。

4 集中式和分布式计算的结果比较

验证数据来源于广安门医院脾胃科 2013—2018 年门诊病历。纳排标准：患者年龄 18～60 周岁，诊断包含慢性胃炎，排除重复就诊和包含异常值、缺失值的记录。经过数据提取得到 1812 例患者数据，选择"年龄"作为分析指标进行描述性统计和频数分析，将样本随机分为 3 个样本量为 695、122、995 的数据集传入 3 个计算节点模拟分布式环境。集中式计算在 excel 中完成，分布式计算在区块链平台中完成，计算节点上传数据保留 2 位小数，比较集中式和分布式的计算结果。见表 1、表 2。

5 讨论

5.1 研究背景和意义 由于计算耗时、内存有限、无法保障隐私安全、通信成本高等问题，集中式计算已不适用于海量数据的处理和分析。分布式计算因其快速、高效、安全的特点，成为近年来大数据统计研究的热点[12]。

分布式计算指由多个节点执行计算，体现出"分而治之"的思想。以计算均值为例，分节点计算各自样本的均值，再把均值和样本量发送给中心节点，中心节点接收后通过加权平均计算全样本均值。在此过程中，"计算全样本均值"任务被划分成了 K 个可同时进行的"计算子样本均值"任务，将计算时间缩短为原来的 1/K。此外，在需要保护数据隐私的真实应用场景中，分布式计算即使遭遇中心节点受损的情况也能有效限制信息泄露的程度，确保敏感数据的细节不被揭露，仅涉及局部汇总信息（如分节点统计指标）的泄露风险[13]。但分节点之间是物理隔离的，会产生额外通信成本和计算时间，计算步骤的增加也可能导致统计精度下降[14]。因此在设计面向医疗等高度敏感领域的分布式算法时，不仅要提升统计精确性、优化通信成本和缩短计算耗时，还要强化算法的保密性，全方位满足应用需求[15]。对于频数分析和描述性统计这类基础的统计任务要优先保障计算精度和保密性，其次考虑降低计算时间和通信成本。

安全多方计算协议是分布式算法的基础，规定节点进行指定计算。与集中式计算方法相比，安全多方计算避免分节点直接上传病历数据，防止敏感信息泄露。此外，因为链上计算在链下计算的结果上进行，减少了数据传输量，减轻了数据存储的负担。频数分析和描述性统计的分布式算法都基于安全多方计算协议设计，两者具有相似性，都构建了 JAVA 对象和哈希表，但描述性统计涉及更多数值的计算，过程更复杂。

区块链技术为安全计算提供了安全可靠的执行环境，这一特性主要体现在两方面：交易数据在区块链上由多方验证并存证，确保信息不可篡改；所有交易都包含发送者签名，能追溯来源，提升透明度[16]。智能合约的引入进一步增强了执行环境的安全性，它能自动解析交易指令并执行相应计算[17]。这一机制使分布式算法摆脱了对中心节点的依赖，规避了中心节点存在的单点故障和信任风险。

5.2 应用前景和价值 安全多方计算为中医药临床数据共享平台的构建提供算法支撑。建设中医药临床数据共享平台满足了开展大规模真实世界研究、流行病学调查和数据挖掘的需要[18]。真实世界研究指系统收集真实医疗环境中的患者数据，分析得出干预的价值及潜在的获益或风险，是符合中医学的内在规律的研究方法[19]。基于中医药临床数据的真实世界研究可以总结规律、形成证据[20]。由于中医药临床数据的来源广泛、结构性差、标准不统一、共享成本高、敏感性和隐私性强，难以直接收集、合并和共享，导致很多真实世界研究只能基于单一数据来源开展[21]，无法保证样本的代表性。如何安全高效地进行数据共享和分析，是中医药真实世界研究发展过程中必须解决的问题[22]。

利用区块链技术和安全多方计算协议构建高效安全的互联网平台，进行中医药临床数据的采集和分析，可以实现数据共享的去中心化，交易记录的防篡改和可追溯性，减轻数据存储和传输的负担，为开展中医药临床研究提供新思路。目前，区块链和安全多方计算协议在中医药领域的应用还处于起步阶段，但其应用前景广阔，具有颠覆中医药行业传统模式的能力[23]。

目前关于区块链和安全多方计算应用于医疗信息平台构

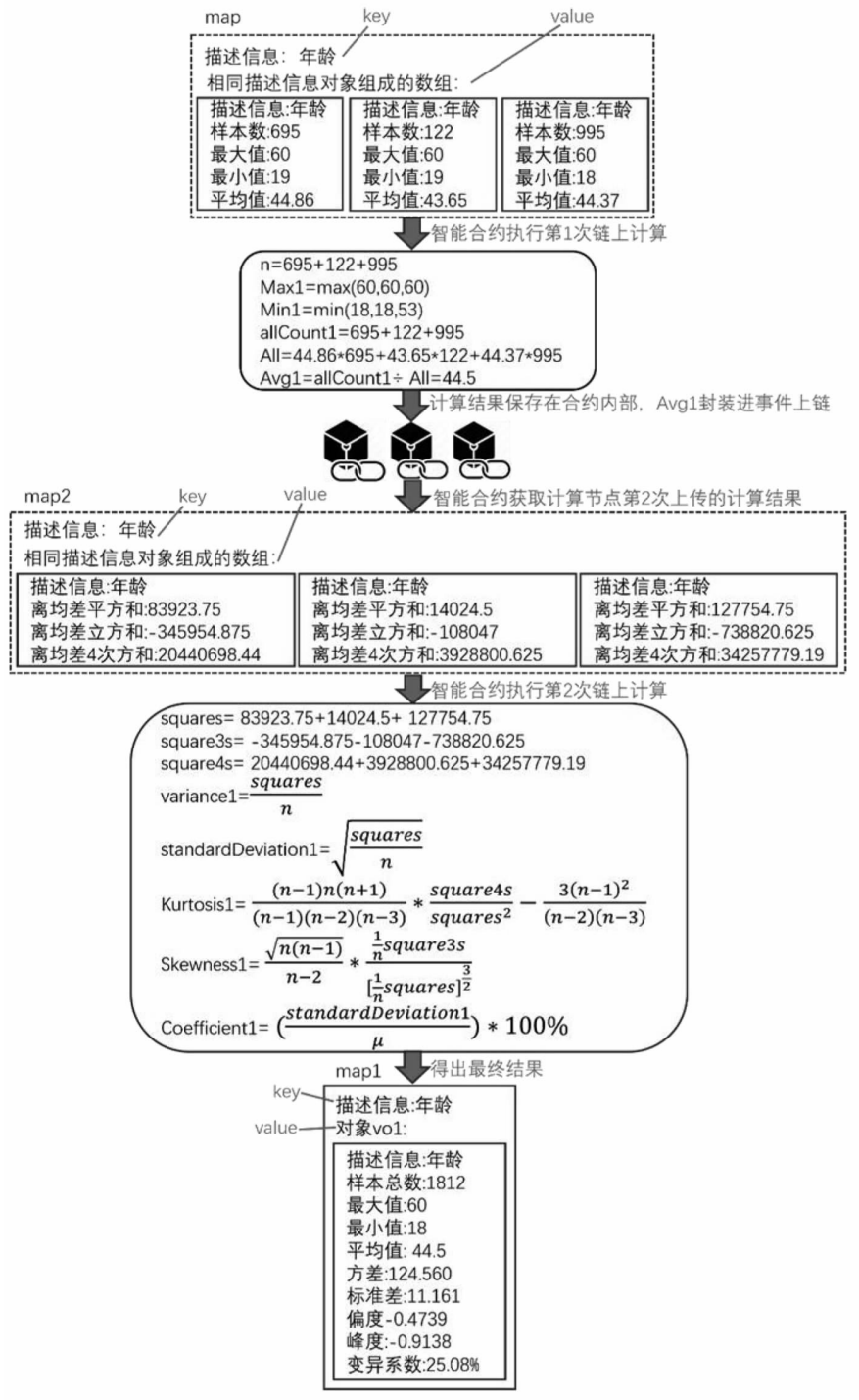


图 4 描述性统计的数据结构和实现过程

表 1 年龄描述性统计结果验证

算法	最大值（最小值）	均值	方差	标准差	偏度	峰度	变异系数/(%)
集中式	60(18)	44.5	124.560	11.161	−0.476 6	−0.912 1	25.07
分布式	60(18)	44.5	124.560	11.161	−0.473 9	−0.913 8	25.08

表 2 年龄频数分析结果验证

算法	频数/百分比(%)			
	[18,28.5)	[28.5,39)	[39,49.5)	[49.5,60)
集中式	205/11.31	378/20.86	479/26.43	750/41.39
分布式	205/11.31	378/20.86	479/26.43	750/41.39

建的研究有：王甜宇等[24]提出基于区块链的中医药大数据云存储共享平台，可以实现中医药健康教育、评估、咨询等功能。余健等[25]使用基于加法的同态加密和基于范围的零知识证明对医疗区块链的中医药临床数据进行隐私保护，能使多个医疗平台共享访问患者数据并保障数据安全。常源恒等设计了基于弱中心化区块链的中医药数据共享算法，使节点不暴露身份信息进行数据共享[26]。WETTSTEIN R 等[27]提出了基于流程模型 BPMN 2.0、数据模型 HL7 FHIR R4 和安全多方计算秘密共享技术的医疗数据共享平台，无须中央组件即可实现多方数据共享。

5.3 优势、局限和展望 本文将安全多方计算协议与集中式计算方法进行对比,论述了分布式算法的安全性,通过实验验证了分布式算法的准确性。分布式描述性统计的误差主要来自除法运算的舍入损失,为确保精度可将上传数据保留 2 位或更多位小数。

区块链使交易信息由多方存证、可追溯且无法篡改,智能合约使数据共享摆脱了对可信中介的依赖。但也存在交易处理速度慢、智能合约安全漏洞等问题。区块链系统使用 MD5 加盐算法计算区块哈希和交易哈希,用于验证区块和交易的完整性,防止篡改。MD5 加盐算法的优点是使用方便、扩展性强、安全系数比 MD5 算法高。链下系统使用 SM4 算法加密原始明文数据,能确保数据在传输和存储时的安全性,但对于数据敏感性和保密要求较高的计算方,SM4 算法可能存在密钥长度固定、容易破解、兼容性差等问题,可以使用非对称加密算法如 RSA 算法、椭圆加密算法加密上传数据。

本文重点关注分布式算法实现过程,没有比较分布式和集中式计算在时间和通信成本上的差异,需要设计实验进一步证明。在已发表的研究中,赵冉等使用区块链平台对多家医院的慢性胃炎患者数据进行分布式频数分析和描述性统计,验证了 2 种分布式算法具有良好的准确性、稳定性和效率[28]。

开展高质量、大规模的多方数据共享是时代发展的必然趋势,这离不开共享技术的进步。目前关于中医药临床数据分布式算法的研究较少,在本研究的基础上还能进行分布式参数检验、分布式非参数检验、联邦学习的探索和研究。此外,为提升合并分析结果的准确性和可靠性,共享数据需满足同质性假设。

6 结论

本文提出了频数分析和描述性统计的中医药临床数据分布式算法以及中医消化领域可信联合循证分析区块链的构建。分布式算法可分为链上计算和链下计算两部分:从医院获取数据、生成预处理数据在链下完成,合并计算在链上完成。区块链提供了可信的计算环境,保障数据共享去中心化、可追溯、防篡改。安全多方计算确保数据的安全和隐私,减轻了数据的存储和传输压力。该方案的提出旨在高效完成分布式环境中的统计分析,促进临床数据高质量共享,为构建中医药临床数据共享平台提供算法支持。实验证明该方案具有可行性,能得到较为准确的统计结果。

参考文献

[1] 韩伟力,宋鲁杉,阮雯强,等. 安全多方学习:从安全计算到安全学习[J]. 计算机学报,2023,46(7):1494 - 1512.
[2] ZHOU J,FENG Y,WANG Z,et al. Using secure multi - party computation to protect privacy on a permissioned blockchain[J]. Sensors (Basel),2021,21(4):1540.
[3] YAO A C. Protocols for secure computations[C]//23rd Annual Symposium on Foundations of Computer Science (sfcs 1982). Chicago:IEEE,1982:3 - 5.
[4] 朱友文. 分布式环境下的隐私保护技术及其应用研究[D]. 合肥:中国科学技术大学,2012.
[5] 尹紫荆. 基于区块链的安全多方计算隐私保护技术研究[D]. 成都:电子科技大学,2023.
[6] GAO H M,MA Z F,LUO S S,et al. BFR - MPC:a blockchain - based fair and robust multi - party computation scheme[J]. IEEE Access,2019,7:110439 - 110450.
[7] 刘炜,唐琮轲,马杰,等. 区块链在隐私计算中的应用研究进展[J]. 郑州大学学报(理学版),2022,54(6):12 - 22.
[8] 唐旭东,陈钟,王斌,等. 基于区块链的中医消化领域可信联合循证分析平台及方法[P]. 中国:CN113921126A,2021 - 10 - 29.
[9] 张泽丹,王斌,陈婷,等. 基于区块链的中医消化领域临床科研一体化数据共享平台的研究与设计[J]. 中华中医药学刊,2024,42(6):29 - 33,265.
[10] AKCORA C G,GEL Y R,KANTARCIOGLU M. Blockchain networks:data structures of Bitcoin,monero,zcash,ethereum,ripple,and iota[J]. Wiley Interdiscip Rev Data Min Knowl Discov,2022,12(1):e1436.
[11] 高一琛,赵斌,张召. 面向以太坊的智能合约自动生成方法研究与实现[J]. 华东师范大学学报(自然科学版),2020(5):21 - 32.
[12] 杨鑫,吴密霞. 多源异质大数据线性模型的分布式统计推断[J]. 数学学报(中文版),2022,66(2):263 - 276.
[13] 郁淼淼,李子洋,周勇. 分布式计算下基于差分隐私保护的参数估计[J]. 应用数学学报,2023,46(2):145 - 165.
[14] YUAN G,LI W D,WEN H S,et al. A review of distributed statistical inference[J]. Statistical Theory and Related Fields,2023,6(2):154 - 159.
[15] KIRIENKO M,SOLLINI M,NINATTI G,et al. Distributed learning:a reliable privacy - preserving strategy to change multicenter collaborations using AI[J]. Eur J Nucl Med Mol Imaging,2021,48(12):3791 - 3804.
[16] DIAO Y Q,YE A Y,ZHANG Y X,et al. P - Chain:Towards privacy - aware smart contract using SMPC[J]. J Inf Secur Appl,2024,86:103872.
[17] HUANG Y X,WANG B,WANG Y G. Research and application of smart contract based on ethereum blockchain[J]. J Phys:Conf Ser,2021,1748(4):42016.
[18] 刘保延. 真实世界的中医临床科研范式[J]. 中医杂志,2019,54(6):451 - 455.
[19] 杨亦含,胡元会,陈婷,等. 基于真实世界的中医临床科研大数据平台研发现状、不足及展望[J]. 中国中医基础医学杂志,2022,28(11):1882 - 1886.
[20] 张泽丹,王斌,陈婷,等. 我国中医药临床数据存储与共享存在的问题及建议[J]. 中国中医药图书情报杂志,2022,47(4):1 - 5.
[21] 赵天易,雒琳,何丽云,等. 多源异构数据整合在中医药真实世界临床研究中的应用及展望[J]. 世界中医药,2023,17(5):614 - 619.
[22] 王晓丽,万生芳,魏昭晖,等. 真实世界大数据下中医临床研究特点探析[J]. 世界最新医学信息文摘,2019,19(6):244 - 246.
[23] 王海隆. 区块链技术在中医药领域中的应用展望[J]. 贵阳中医学院学报,2017,39(3):1 - 4.
[24] 王甜宇,张柯欣,孙艳秋,等. 基于区块链的中医药大数据云存储共享方案研究[J]. 中华中医药学刊,2022,40(2):132 - 135.
[25] 余健,胡孔法,丁有伟. 一种面向中医药临床数据的区块链安全与隐私保护方案[J]. 世界科学技术 - 中医药现代化,2021,23(10):3688 - 3695.
[26] 常源恒. 基于区块链的中医药数据安全共享方法研究[D]. 南京:南京中医药大学,2022.
[27] WETTSTEIN R,KUSSEL T,HUND H,et al. Secure multi - party computation based distributed feasibility queries - A HiGHmed use case[J]. Stud Health Technol Inform,2022,296:41 - 49.
[28] 赵冉,张雯,唐旭东,等. 基于中医药真实世界数据验证中医可信联合分析平台中的安全多方计算:以描述性统计和频数分析为例[J]. 中华中医药学刊,2024,42(12):20 - 24.