基于信用系数的动态改进的 PBFT 算法

王陈培

（杭州外国语学校，浙江杭州，310000）

摘要：随着以比特币为代表的数字货币的兴起，区块链作为其底层的技术受到越来越多的关注。区块链本身是一种点对点的分布式系统，共识算法是解决各节点达成共识的机制，以POW、POS为代表的公有链共识算法有算法效率低下，耗能严重，以Paxos、Raft为代表的传统分布式一支算法未考虑到拜占庭容错。因此，本文在对FBFT算法分析的基础上，提出了基于信用系数的动态改进算法，既考虑到了拜占庭容错、又增加了算法了灵活性，提高了算法的吞吐量、时延等性能。

关键词：区块链；共识算法；拜占庭容错；PBFT；信用系数

DOI:10.16589/j.cnki.cn11-3571/tn.2019.08.017

0 引言

区块链作为比特币的底层技术，随着比特币的兴起而进入公众视野，它的本质是一种去中心数据库，及点对点传输、分布式数据存储、共识机制、加密算法等于一体的新型计算机技术应用模式，它具有去中心化、时序数据、集体维护、信息不可篡改性等特征，其作为种新兴的计算机技术应用模式，目前已经引起政府部门、金融机构、科技企业和资本市场的高度重视。它起源于 2008 年由化名为中本聪的学者在密码学邮件组在奠基性论文《比特币：一种点对点的电子现金系统》[1]，该论文解释了区块链的内部机制，区块链是将数据区块按照时间顺序组成链条式的数据结构，这种数据结构组成了去中心化电子总账，通过密码学技术保证了总账的不可篡改和伪造，使得各区块数据具备能够安全、有时间先后、在系统内易验证的特性。

区块链原理可归纳为数学上的拜占庭将军问题 [2-3]，其内涵为：在各个节点均不可信任的情况下，分布在网络中的各节点如何达成共识，推选出可信任的节点获取记账权。因此，在区块链系统中，共识算法具有重要的作用，它不仅协调各个节点保持数据一致，同时还对代币发行、攻击防范具有一定的功能 [4]。然而，区块链目前的共识算法存在很多不足，导致区块链在效率、能耗等方面远远不能满足许多应用场景的性能需求，共识算法成为制约区块链发展的瓶颈。因此，研究高性能的共识算法对于区块链的发展具有重要意义。

1 区块链及其相关技术

■ 1.1 区块链结构

区块链系统一般包含五个基本关键因素：分布式点对点系统、分布式账本、共识算法、一定的激励机制及可实现的编程应用。区块链具有六层结构，分别是数据层、网络层、共识层、激励层、合约层和应用层，如图 1 所示。

从图 1 中我们可以得出，区块链各个层次含有的主要内容，下面是对区块链各个层次的结构的介绍。

数据层是区块链最底层的技术，将数据区块及其相关加密技术、时间戳等进行封装，主要实现两个方面的功能：数据存储、账户交易实现的安全性 [5]。

图 1　区块链结构

网络层包括分布式组网机制、数据传播机制和数据验证机制，因此在点对点交互的过程中，用户信息不容易被攻击成功。

共识层主要封装网络节点的各类共识算法。公式算法体现的共识机制在去中心化的思想上一定程度地解决了节点间互相信任的问题，例如工作量证明 POW、权益证明 POS。

激励层通过经济平衡来鼓励节点维护区块链的安全运行，该层将节点自身利益最大化与区块链系统安全有效运行相吻合，从而防止总账本被篡改。

合约层可以理解为自定义电子合同，也称智能合约，它由区块链系统的脚本、算法等组成，当达到一定约束条件时，可自动触发执行，它是区块链实现灵活操作数据的基础。高效的共识算法是限制区块链发展的瓶颈，如何在分布式系统中达成共识是分布式计算领域中重要的课题。

应用层涵盖了区块链的各种应用场景，类似日常的网站、APP 等，如在以太网上搭建区块链应用，在将来，可编程金融及电子商务也会部署在应用层。

■ 1.2 共识算法的比较

共识算法是协调全网中所有数据一致性的算法协议，很多的论文已经对共识算法做了大量的研究，本文对应用到区块链的几种共识算法进行研究，算法如下所示。

(1)POW

POW 全称是工作量证明算法，它主要应用在比特币生成的算法，它采用的是 hash 的运算得到一个值，它可以抵御 DDOS 攻击。然而，它能耗太大，并且产生区块时间较长，不适用于具体的场景中 [2]，比特币的工作量证明就是对当前区块的头部数据做双重 hash，具体的算式如下所示。

SHA256(SHA256(nVersion)+hashPrevBlock+hashMerkleRoot+nTime+nBits+nNonce))<Target

该算式中各个参数的含义为：

nVersion：版本号，记录版本信息

hashPrevBlock：父区块哈希

hashMerkleRoot：当前区块所包含交易的梅克尔根

nTime：时间戳

nBits：当前的挖矿难度

nNonce：随机值

Target：目标值，可以由难度值计算出

PoW 以算力为代价达成共识获取奖励，要求节点一直高速在做毫无意义的 SHA256 运算，此过程也成“挖矿”，事实证明，PoW 每年的耗电量高达数十亿美元。然而，如此高的算力确实目前公有链中最好的共识算法，可以将攻击者的算力攻击成功率降到最低。

(2)POS

POS 算法也是股权证算法，它的基本思想是持有股权最多的获得记账权的激励越大，其相对于 POW 算法的优势是不浪费算力，缺点是建立账号股份多少的机制是极度不公平的，且挖矿成本相对于 POW 算法低，易造成网络攻击。

(3)Paxos

Paxos 算法是一种分布式系统中解决一致性问题的最重要的算法 [3]，该算法在分布式系统中的应用场景是存在故障、但不存在恶意节点的场景。Paxos 是第一个被证明的共识算法，其原理基于两阶段提交并进行扩展 [6]。

Stage1：

(a)proposer 节点向系统中过半数的 acceptor 节点发送 prepare 消息；

(b)acceptor 节点若正常，则回复 promise 消息。

Stage 2：

(a) 若存在足够多 acceptor 节点回复 promise 消息，则 proposer 节点发送 accept 消息；

(b) 若 acceptor 节点正常，则回复 accepted 消息。

其中，proposer 节点往往有客户端担任，负责提出提案，若大家批准则为结案；acceptor 节点负责对提案进行反馈和投票，往往是服务端担任该角色；一般需要至少 3 个且节点个数为奇数，因为 Paxos 算法最终要产生一个大多数决策者都同意的提议；learner 节点不参与投票过程，但会被告知投票结果，并与投票结果进行统一，该节点可能为客户端也可能是服务端担任。

(4)PBFT

PBFT 是一种状态机副本复制算法，目前在解决拜占庭将军问题上最被普遍使用的一种算法。PBFT 在保证安全性和活性的前提下，提供失效节点不超过 $(n-1)/3$ 的容错保证。此算法采用 C/S 结构，经过三个阶段达成一致性，其过程如图 2 所示。

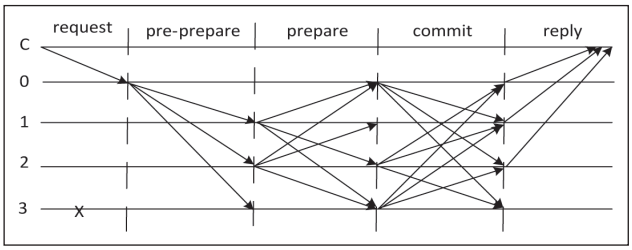

图 2　PBFT 算法过程

主节点 O 收到客户端节点 C 发送的调用服务操作的请求，想备份节点 1、2、3 广播该请求，备份节点收到请求并执行，通过 prepare、commit 等阶段，将最终的阶段反馈给客户端 C，客户端 C 得到从不同节点反馈的结果，若结果在大概率下一致，即使客户端 C 的请求执行结果。

PBFT 算法扩展性较差，适用于节点数量较为固定的联盟链或私有链，且 PBFT 算法要求系统总节点数 $n>(3f+1)$，f 为失效节点数，在容错率上较低。

从 PBFT 的结构分析，PBFT 包含预处理、处理、提交阶段，且在确认过程中，会从节点进行广播，这无疑会造成巨大的浪费，若每次传播的数据量为 BlockSize，那么有 N 个节点，则消耗的总功耗为：

Total=N*(N-1)*BlockSize

（5）共识算法性能的比较

基于对上面各共识算法的分析，本文总结共识算法的从是否拜占庭容错、是否权利集中、是否具备承载更过的交易量的能力、是否具有更快确认速度及是否高效节能等多个指标进行比较，如表 1[2] 所示。

表1　共识算法性能比较

特征	POW	POS	DPOS	Paxos	PBFT
拜占庭容错	√	√	√	×	√
权利集中	√	√	×	×	×
承载更多的交易量	×	×	×	√	√
更快的确认速度	×	×	×	√	√
高效节能	×			√	√
鼓励开发	×	√	√	√	√

从表 1 中可以得出，PBFT 算法具有拜占庭容错机制，且相对于算法在承载更多节点、更快确认速度以及高效节能方面具有优势，然而却存在权力集中的缺点，权利的集中是 FBFT 算法容错率、扩展性差等缺点的原因。本文在 PBFT 共识算法的基础上，提出了信用系数的概念，对 PBFT 实现动态改进，可避免权力集中的缺点，提高其在节点出入链灵活性，提升算法容错率。

2 基于信用系数的动态改进的 PBFT 算法

■ 2.1 信用系数

从 PBFT 的结构分析，其是 C/S 模式，并非严格的点对点的协议，本文引入了类似于 DPOS 的投票机制，我们对节点的按信都分为：信任、待考察、不信任三个层次，该层次是基于信用系数 γ 表示，γ 是基于其他节点的投票 γ_i 来得到的。使用该方法可以避免权力过于集中的问题，FBFT 算法本身节点静态，动态性能略差的问题。

■ 2.2 动态改进的 PBFT 结构

通过对 PBFT 算法结构的分析，该算法在可扩展性、容错率、能耗等方面具有缺陷，因此，本文对其结构进行改变，使之更加适应区块链点对点的模式，从三个方面进行改进：

（1）节点部分主从节点，整个网络的结构为点对点的传输方式。

（2）算法结构分为两个阶段，准备和预准备阶段，去掉确认阶段。

（3）该算法通过信用系数进行判断哪个节点作为记账节点。

3 算法性能比较

通过查阅文献 [2]，我们可以得到 PBFT 算法在吞吐量的性能如图 3 所示。

从图中我们可以得出，PBFT 算法吞吐量在万级以上，通过对 PBFT 算法的分析，我们了解到使用该算法的系统共识节点固定，在拓展性上差，无法动态增删节点，比较适用数目固定的私有链中。

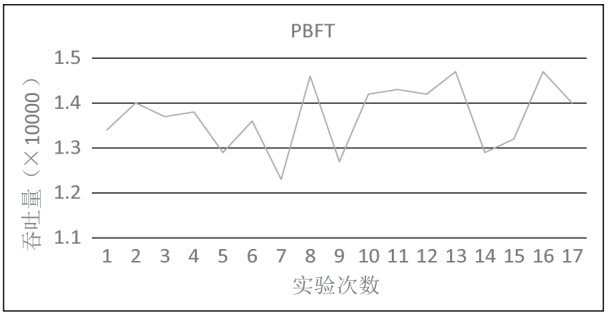

图 3　PBFT 算法吞吐量

本文对改进后的 PBFT 算法性能进行测试，结果如图 4 所示，从图中我们可以得到，改进后的 PBFT 算法的吞吐量较 PBFT 算法平均提高了约 10%。

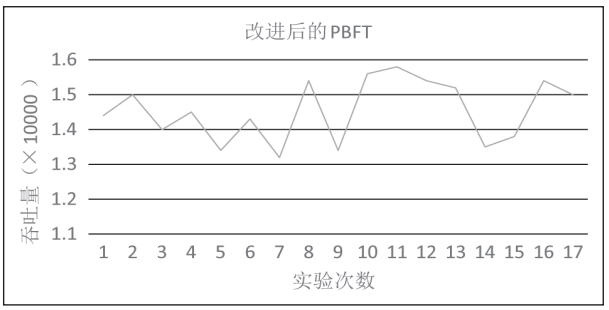

图 4　改进的 PBFT 算法性能

4 总结

共识算法作为区块链的最核心部分，在区块链数据一致性以及性能方面具有至关重要的影响。本文对常见的共识算法进行分析，Pow 耗能较高，PoS 在股权分配上不均等，Paxos 是假设节点存在故障而不存在恶意节点下的算法，无法满足现实存在恶意节点的情况，PBFT 算法是解决拜占庭将军的一种常见算法，该算法数据吞吐量较好，然而比较适合节点固定的情况，因此本文对该算法进行动态改进。首先，该算法引入信用系数来评判节点受信任的状态，以此判断该节点成为记账节点的机率，其次，将该算法的 C/S 模式改为点对点模式，可以有效提高算法效率。实验证明，改进后的算法在数据吞吐量上可以提高 10%。

参考文献

＊ [1] 袁勇，王飞跃 . 区块链技术发展现状与展望 [J]. 自动化学报，2016, 42(4):481-494.

＊ [2] 刘肖飞 . 基于动态授权的拜占庭容错共识算法的区块链性能改进研究 [D]. 浙江大学 ,2017.

（下转第 89 页）

各项生理指标以及患病情况，其中生理指标作为数据特征，患病情况作为数据标签。并且验证数据集并不附加标签，只附加有数据特征。通过输入一组训练数据，将数据以坐标的形式呈现在 n 维空间上，并设为 A。然后再输入验证数据，并且以相同的形式生成一个 n 维空间内的坐标，并设为 B。然后计算该坐标 B 与训练数据集中坐标 A 之间的距离，找出与训练数据集坐标 A 距离最近的 k 个坐标，并提取出 k 个训练数据中占大多数的标签。最后将不带标签的验证数据集 B 进行分类，最后将 B 所分类别的标签与其原本属于的标签进行对比，并计算吻合率。在此过程中不断地将 k 值在一个区间范围内进行变换，直到找到一个吻合率最高的 k 值，以此作为预测糖尿病模型的最合适准确的 k 值。

距离的计算类型选择的是欧式距离，以下是 n 维空间欧式距离的算法推导结论：

二维度内点 $A(x_1,y_1)$ 与 $B(x_2,y_2)$ 之间的欧氏距离

$$d_{AB}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

三维度内点 A(x1,y1,z1) 与 B(x2,y2,z2) 之间的欧氏距离

$$d_{AB}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

我们可以由此类推出 n 维空间内两点间的欧式距离公式。

接下来是有关数据预处理的问题，对于一些值域范围很大的数据，它们会干扰到对距离的计算，从而会使得算法模型的准确性大大降低。例如如果样本数据集中的“静止血压”这一特征数据的数量级是“糖尿病谱系功能”这一特征数量级的 100 倍甚至更多倍，那么“静止血压”这一特征对于距离计算的干扰强度就要远远大于“糖尿病谱系功能”这一特征的干扰强度。为了防止某些数据特征对距离的计算产生决定意义上的干扰，作者认为可以利用数据标准化或归一化等处理方法来解决。例如将不同值域范围内的数据通过映射的手段到同一个较小范围的区间内。我们通过这种方式来统一数量级，这样就可以大幅减少计算量，大幅提升模型精度，并且还可以提升算法的计算速度，规避某些数据特征对算法模型的误导性干扰，为测试者进行更好更准确地预测的预测。

标准化运用如下公式所示，其中 $\bar{x}$ 为特征数据的平均值，σ 为特征数据的标准差。

$$x'=\frac{x-\bar{x}}{\sigma}$$

3 结论与展望

本论文首先通过对数据的代入完成了对 K 近邻算法建模过程。进而通过对模型的配合使用，对数据集中的数据进行一个系统的预处理后，然后运用多数表决等分类方法继续对数据进行详细的分类，并且作者通过对数据的分析与属性的对比后结合以上模型可以预测潜在患者的患糖尿病的概率。所以通过本研究作者可以一定程度上的帮助潜在患者了解并推测他患糖尿病的可能性，可以对医疗事业的预防方面做出一定的帮助。

但是在本文创作过程中，我在建模完成之后，通过余下数据，对预测结果进行检测，发现测结果存在一些误差。所以作者认为以上模型存在一些不足之处。例如，K 近邻算法存在一些弊端，当样本数据集里的情况不平衡时，比如数据集里的不同类别样本的数目相差较大则会对结果产生影响，易产生误差；当样本数据集里的样本容量太大时又会增加时间成本，降低计算机的运算效率。

没有哪个模型是完美的，也没有那种方法是一成不变的，所以我也将在今后的学习生活中学习了解更多模型以便于我对整个研究的进行，进而进行疾病的预测。我进行这个实验数据的研究以及模型的建立，最终的愿望还是为了我国人民能够具有更好的身体，能尽量减少患糖尿病的可能性，我们希望并由衷地祝愿在这个新的时代我们中国人民能过上更幸福更健康的生活。

参考文献

* [1] 王清 , 马华 , 孙静 , 韩忠东 . 改进的 KNN 算法及其在医学图像处理中的应用 [J]. 泰山医学院学报 ,2006(06):564-566.
* [2] 陈治平 , 王雷 . 基于自学习 K 近邻的垃圾邮件过滤算法 [J]. 计算机应用 ,2005(S1):7-8.
* [3] 尹航 , 常桂然 , 王兴伟 . 采用聚类算法优化的 K 近邻协同过滤算法 [J]. 小型微型计算机系统 ,2013,(4):806-809.
* [4] 张涛 , 陈先 , 谢美萍 , 张玥杰 . 基于 K 近邻非参数回归的短时交通流预测方法 [J]. 系统工程理论与实践 ,2010,30(02):376-384.

（上接第 48 页）

* [3] 杨革 , 徐虹 .Paxos 算法的研究与改进 [J]. 科技创新与应用 , 2017(7):25-26.
* [4] 段希楠 , 延志伟 , 耿光刚 , 等 . 区块链共识算法研究与趋势分析 [J]. 科研信息化技术与应用 ,2017,8(6):43-51.
* [5]Huang X J,Gan T.Analysis on the Application of Block Chain in the Field of Supply Chain Finance[J].Journal of Changchun Finance College,2018.
* [6]Natoli C,Gramoli V.The Blockchain Anomaly[C]//IEEE,International Symposium on Network Computing and Applications. IEEE,2016:310-317.