基于多目标效用优化的分布式数据交易算法

黄小红[1]，张勇[1]，闪德胜[2]，钱叶魁[3]，韩璐[1]，李丹丹[1]，丛群[4]

（1. 北京邮电大学计算机学院（国家示范性软件学院），北京 100876；2. 中国人民解放军 32147 部队，陕西 宝鸡 721000；
3. 陆军炮兵防空兵学院郑州校区，河南 郑州 450052；4. 北京网瑞达科技有限公司技术研发部，北京 100876）

摘 要：传统的集中式数据交易模式不能很好地适用于当前万物互联、数据实时产生的智能时代，为了使产生的数据发挥更大的价值，设计一种有效的数据交易框架至关重要。为此，提出了一种基于联盟区块链的分布式数据交易框架，在不依赖第三方的情况下实现了 P2P 的数据交易。针对已有数据交易模型仅考虑数据本身的因素，而忽略用户任务相关因素的问题，基于数据质量、数据属性、属性的相关性、消费者竞争等多维因素构建了双层多目标优化模型，以优化数据提供者（DP）和数据消费者（DC）的效用。为求解上述模型，提出了一种改进的多目标遗传算法——协作式 NSGAII，通过 DP、DC 以及数据聚合器（AG）的协作进行计算。仿真结果表明，所提算法在 DP 和 DC 的效用方面取得了更好的性能，实现了更有效的数据交易。

关键词：联盟区块链；分布式数据交易；优化匹配模型；多目标遗传算法

中图分类号：TP399

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021034

Distributed data trading algorithm based on multi-objective utility optimization

HUANG Xiaohong[1], ZHANG Yong[1], SHAN Desheng[2], QIAN Yekui[3], HAN Lu[1], LI Dandan[1], CONG Qun[4]

1. School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China
2. The PLA Army of 32147, Baoji 721000, China
3. Zhengzhou Campus, PLA Army Academy of Artillery and Air Defense, Zhengzhou 450052, China
4. Beijing WRD Technology Co., Ltd., Beijing 100876, China

Abstract: The traditional centralized data trading models are not well applicable to the current intelligent era where everything is interconnected and real-time data is generated, and in order to maximize the use of collected data, it is essential to design an effective data trading framework. Therefore, a distributed data trading framework based on consortium blockchain was proposed, which realized P2P data trading without relying on a third party. Aiming at the problem that existing data trading models only consider the factors of the data itself and ignore the factors related to user tasks, a bi-level multi-objective optimization model was constructed based on multi-dimensional factors, such as data quality, data attributes, attribute relevance and consumer competition, to optimize the utilities of data provider (DP) and data consumer (DC).To solve the above model, an improved multi-objective genetic algorithm-collaborative NSGAII was proposed, calculated by the cooperation of DP, DC and data aggregator (AG). The simulation results show that the collaborative NSGAII achieves better performance in terms of the utilities of DP and DC, thus realizing more effective data trading.

Keywords: consortium blockchain, distributed data trading, optimization matching model, multi-objective genetic algorithm

---

收稿日期：2020−08−07；**修回日期**：2020−11−22

通信作者：钱叶魁，qyk1129@163.com

基金项目：国家重点研发计划基金资助项目（No.2020YFE0200500）；北京邮电大学优秀博士生创新基金资助项目（No.CX2019212）

Foundation Items: The National Key Research and Development Program of China (No.2020YFE0200500), The BUPT Excellent Ph.D. Students Foundation (No.CX2019212)

1　引言

随着物联网、车联网、微电网以及移动应用的发展，大数据呈现爆炸式的增长趋势[1-2]。预计到2026 年，数据价值将达到 922 亿美元[3]。目前，大部分数据都存储在公司或组织的数据库中，形成一个个数据孤岛，只有少部分数据真正得到充分利用，而且在使用上也有各种限制[4]。许多 App 开发者和研究人员迫切需要数据来提高产品和研究的质量，并且愿意为此支付一定的经济成本[5-6]。因此，一些数据交易市场应运而生，如 Infochimps、Datacoup、Microsoft Azure Marketplace 等[7]，但这些市场仍处于起步阶段，缺乏适用的交易规则。而且，从经济学的角度来看，市场组织者都是自私的，他们追求的是自身利益的最大化，而不是系统的整体效用。

因此，设计一种有效的数据交易机制成为学者的研究重点[8-9]。Niyato 等[10]描述了一种典型的数据交易市场模型，该模型由 3 种实体组成：数据提供者（DP, data provider）、数据消费者（DC, data consumer）和数据代理（DA, data agent）。Cao 等[11]考虑了数据市场中存在多个 DP、DA 和 DC 的情况，提出了一种迭代拍卖机制来协调交易。Jiao 等[12]考虑了大数据的“无限供给”性，提出了一种基于拍卖的大数据市场模型，通过拍卖机制得到最优的数据交易价格和交易量。Yu 等[13]通过考虑数据质量和数据多版本发布的策略，提出了一个双层数学规划模型来优化数据交易。

当前针对数据交易的研究主要考虑了数据量、数据质量等数据本身的因素，忽视了数据属性、属性的相关性等与用户任务相关的因素。而且，在存在多个 DC 的场景下，这些因素之间是存在竞争关系的，在研究数据交易时也应将这些竞争考虑在内。此外，现有框架一般以 DA 作为中介来实施数据交易，即 DA 从 DP 处购买数据，然后将其出售给 DC。由于 DA 的信任危机和数据产品的低复制成本，这些集中式的解决方案带来了数据泄露方面的问题。

因此，一些学者针对如何提高数据交易的安全性和隐私性进行了研究[14]。Niu 等[15]提出了基于同态加密和身份签名的数据交易市场架构来保护 DP 的隐私。Jung 等[16]研究了数据交易中 DC 的行为，并设计了 AccountTrade 来保证 DC 为其不诚实的行为承担责任。但是，这些基于密码学的解决方案复杂性较高，在时间和空间方面代价昂贵。基于分布式对等网络的区块链技术具有去中心化、不可篡改、可追溯、匿名性和透明性五大特征，这些特征为数据交易的安全问题提供了很好的解决方案[17-18]。Delgado-segura 等[19]提出了一种基于比特币的公平的分布式数据交易协议。Missier 等[20]使用区块链构建了一个分布式的、可信的、开放的物联网数据交易架构，并基于以太坊进行了实现。Sabounchi 等[21]利用区块链技术和契约理论设计了一种基于以太坊的 P2P 数据交易机制。Gao 等[22]提出了一种安全、公平的数据交易方案，并设计了一种新型的比特币脚本来减少交易时延。这些方案采用分布式的架构解决了数据交易中的安全与信任问题，但是在效率方面仍存在不足，比如，比特币的吞吐量为 7 TPS（transaction per second），以太坊的吞吐量为 15 TPS，与实际需求相去甚远。

基于以上研究，本文提出了基于联盟区块链的分布式数据交易框架。在此框架下，通过考虑多维因素，建立了 DP 和 DC 的效用函数，并构建了双层多目标优化模型对他们进行优化。为求解优化模型，提出了协作式 NSGAII，并通过实验对算法的性能进行了评估。本文的主要贡献如下。

1) 针对中心化交易方案容易出现数据泄露的问题，提出基于联盟区块链的分布式数据交易框架，实现了 DP 与 DC 的 P2P 的数据交易。

2) 为了提高数据交易的有效性，通过综合考虑数据本身和用户任务相关的因素，基于数据质量、数据属性、属性的相关性、消费者竞争等多维因素构建双层多目标优化模型，以优化 DP 和 DC 的效用函数。

3) 针对 NSGAII 会泄露用户隐私的问题，提出一种改进算法——协作式 NSGAII，通过 DP、DC 以及数据聚合器（AG, data aggregator）的协作来进行计算。DP 和 DC 的效用函数由用户在本地进行计算，以免泄露隐私。

4) 使用北京的空气质量数据进行了实验，以评估所提算法的性能。结果表明，所提算法在 DP 和 DC 的效用方面可以实现很好的性能。

2　系统模型

本节描述了分布式数据交易的场景，并构建了基于联盟区块链的分布式数据交易框架。

2.1　分布式数据交易框架

图 1 展示了基于联盟区块链的数据交易框架，为确保数据交易的安全性和效率，本节设计了基于联盟区块链的数据交易框架。联盟区块链是一种特殊的区块链，它建立在许多预选的共识节点上，并由这些节点执行共识算法[23]。相比于公链，由于参与共识的节点较少，联盟链多采用实用拜占庭容错机制（PBFT, practical Byzantine fault tolerance）、委托拜占庭容错机制（DBFT, delegated Byzantine fault tolerance）、Raft 等共识机制。这些机制相较于公链所采用的工作量证明（PoW, proof of work）、权益证明（PoS, proof of state）等机制，数据处理速度有很大提升，因此极大地提升了联盟链的效率和吞吐量，对比比特币的 7 TPS 和以太坊的 15 TPS，采用联盟链的 Fabric 可以达到 500 TPS、Quorum 可以达到 800 TPS[24]。

数据交易框架包含 3 种实体，分别是 AG、DP 和 DC。

1) AG 在系统中是数据交易组织者，负责数据交易过程中一些信息的传递以及对 DP 数据的质量评估。每个 AG 负责某种类型数据的交易。例如，在图 1 中，AG_4 负责物联网数据的交易，AG_6 负责互联网数据的交易。在联盟区块链中，AG 充当选定节点来验证交易。

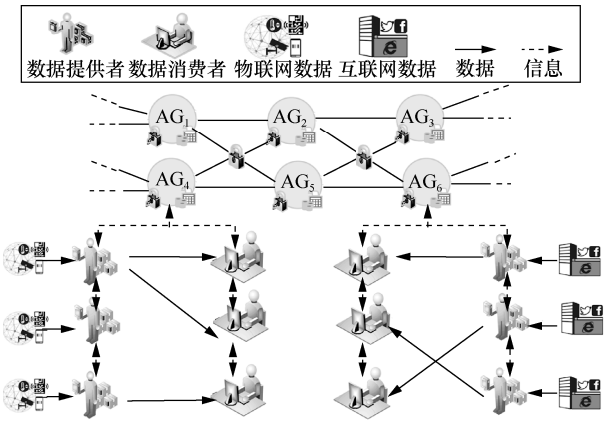

图 1　基于联盟区块链的数据交易框架

2) DP 收集从各种来源（如传感器、移动设备、互联网等）生成的数据，这些数据包含某些属性。

3) DC 是购买数据并对数据具有某些属性需求的最终用户。

每个实体都有自己的账户和钱包。账户用于存储交易记录；钱包用于管理账户中的数字货币。交易完成后，DP 和 DC 使用钱包来进行数字货币的收/付款。AG 使用钱包收取数据质量评估的费用。借助 PBFT，至少 1/3 的 AG 确认交易，才能达成共识。

2.2　数据交易流程

数据交易流程如图 2 所示，其具体步骤如下。

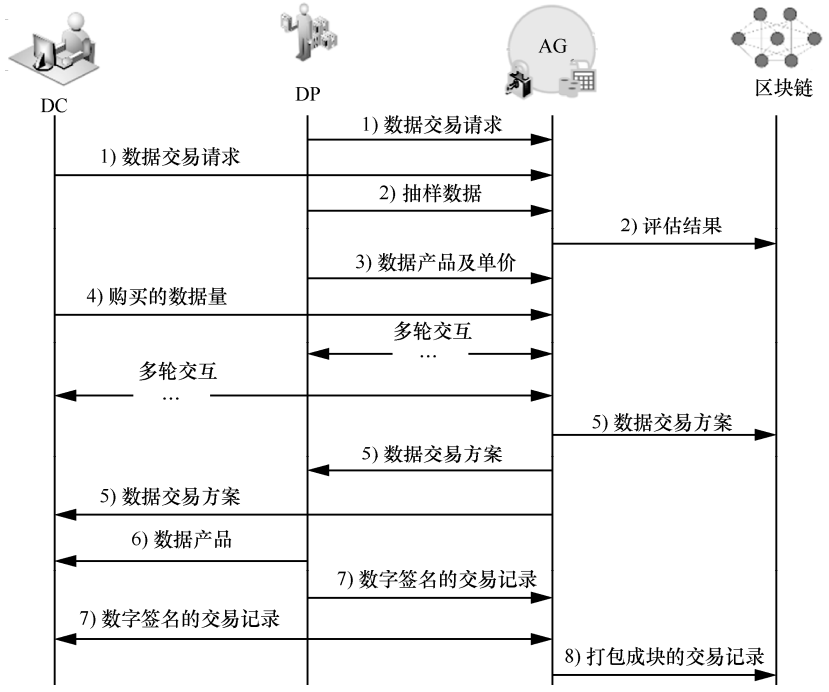

图 2　数据交易流程

1) 初始化。DP 和 DC 使用有效的身份证明（例如居民身份证、企业营业执照等）向 AG 注册成为合法实体，包括公钥、私钥的一对密钥将被分配给他们。DP 和 DC 使用私钥生成钱包地址来参与数据交易。随后，DP 和 DC 根据数据的类型将交易请求发送到相应的 AG，DP 的请求包括其数据的数量和属性信息，DC 的请求包括其属性需求。

2) AG 广播这些数据交易请求并进行数据质量评估。AG 从 DP 处获取一定比例的数据样本，执行质量评估，并将评估结果记录到区块链上。

3) DP 根据 DC 的属性需求为每个 DC 提供相应的数据产品和单价。数据产品的属性是 DP 所能提供的属性与 DC 所需的属性的交集。

4) 根据 DP 所提供的数据产品，DC 以最大化自身的效用为目标，选择最合适的数据产品，并计算购买量。DP 根据 DC 的反馈，进一步调整数据产品的单价。

5) 经过多轮交互，确定最终的交易方案，并由 AG 将其存储在区块链上。根据数据交易方案，AG 为每一笔交易创建相应的智能合约，DC 将数字货币存储在相应的合约地址中。

6) DP 将相应的数据产品发送给 DC。

7) 数据传输完成后，DP 对交易记录进行数字签名，然后发送给 AG。AG 将交易记录发送给相应的 DC。DC 验证交易记录并进行数字签名后回传给 AG。此时，DC 保留在智能合约中的数字货币也将被发送到 DP 的钱包地址。

8) 经过一段时间的收集后，具有排序功能的 AG 将交易记录打包成块，然后通过 PBFT 将它们存储在区块链上。

在交易过程中，AG 负责数据质量评估、信息传输和共识算法执行。交易的数据由 DP 直接传输给 DC，因此是分布式的 P2P 数据交易。这种交易方法可以避免单点故障、信息泄露和其他集中式交易的安全问题。

3　双层多目标数据交易匹配优化模型

本节将描述如何构建数据交易匹配模型，以优化 DP 和 DC 的效用，并给出模型的求解算法。

3.1　问题描述

在描述问题之前，给出所需参数及其含义，如表 1 所示。

如图 1 所示，存在一组 AG，记作 A，$a_i \in A$，$1 \leqslant i \leqslant I$。一段时间内，$a_i$ 收集 DP 和 DC 的数据交易请求。设 P 和 C 分别表示 DP 和 DC 的集合，P 中包含 J 位 DP，$p_j \in P$，$1 \leqslant j \leqslant J$，$p_j$ 向 a_i 提交其数据属性集合 M_j^p 和数据量 N_j；C 中包含 K 位 DC，$c_k \in C$，$1 \leqslant k \leqslant K$，$c_k$ 向 a_i 提交其属性需求集合 M_k^c。

表 1　参数及其含义

参数	含义
A	AG 的集合
P	a_i 中 DP 的集合
C	a_i 中 DC 的集合
p_j	P 中的第 j 位 DP
c_k	C 中的第 k 位 DC
M_j^p	p_j 数据的属性集合
M_k^c	c_k 的数据属性需求集合
$\mathrm{PC}_{j,k}$	p_j 为 c_k 提供的数据产品
$M_{j,k}$	$\mathrm{PC}_{j,k}$ 的属性集合
N_j	p_j 的数据量
$n_k^{\min}$	c_k 想要购买完全满足其属性需求的最小数据量
$n_k^{\max}$	c_k 想要购买完全满足其属性需求的最大数据量
β	数据质量评估的抽样比例
r_j	由 a_i 给出的 p_j 数据的质量评估
$q_{j,k}$	$\mathrm{PC}_{j,k}$ 对 c_k 的可用性
$x_{j,k}$	c_k 从 p_j 处购买数据的数量
$w_{j,k}$	c_k 从 p_j 处购买数据的单价

接收到 AG 广播的 DC 的属性需求后，DP 将为每个 DC 提供相应的数据产品，数据产品的属性将是 DP 可以提供的属性与 DC 的属性需求的交集。例如，p_j 将为 c_k 提供数据产品 $\mathrm{PC}_{j,k}$，它的属性集合为 $M_{j,k}$，$M_{j,k} = M_j^p \cap M_k^c$，然后根据成本和效用，给出 $\mathrm{PC}_{j,k}$ 的单价 $w_{j,k}$。收到数据产品的信息后，DC 将根据其效用函数确定最合适的数据产品以及购买的数据量。DP 收到 DC 的决定后将进一步调整单价。通过多轮交互，确定最终的交易方案。AG 会将交易方案记录在区块链上，以解决可能发生的争议。

在匹配过程中，需考虑 2 个方面：DP 的效用函数和 DC 的效用函数。

3.2　DP 的效用函数

DP 的效用函数主要考虑数据交易的收入、数据收集的成本以及数据质量评估的成本，计算

式为

$$U(p_j)=R(p_j)-L(p_j)-S(p_j) \tag{1}$$

其中，$R(p_j)$ 表示 p_j 的收入，$L(p_j)$ 表示 p_j 收集数据的成本，$S(p_j)$ 表示 a_i 为质量评估收取的费用。$R(p_j)$ 的计算式为

$$R(p_j)=\sum_{k=1}^{K}x_{j,k}w_{j,k} \tag{2}$$

其中，$x_{j,k}$ 和 $w_{j,k}$ 分别表示 c_k 从 p_j 处购买数据的数量和单价。

对于数据收集成本，不同的属性通常需要不同的收集设备，因此成本也有所不同。将收集第 e 个属性的单位成本表示为 b_e，$1\leqslant e\leqslant\left|M_j^p\right|$，则 p_j 的数据收集成本为

$$L(p_j)=\sum_{e=1}^{\left|M_j^p\right|}b_e\max\left(x_{j,1},x_{j,2},\cdots,x_{j,K}\right) \tag{3}$$

a_i 将对 p_j 的数据采样并进行质量评估，评估费用与采样数据的数量成正比。因此，质量评估的成本为

$$S(p_j)=\chi\beta N_j \tag{4}$$

其中，χ 是单位数据评估的价格，β 是采样比例。

DC 购买数据主要是为了建立数据挖掘模型，其中分类和回归是数据挖掘的 2 个主要类别。a_i 考虑与客户体验相关的绩效指标，使用所有采样数据和历史数据建立模型，以评估 DP 数据的质量。对于分类问题，将分类准确性（即正确预测结果的比例）作为性能指标。对于回归问题，使用基于绝对误差的满意率作为性能指标。以 p_j 为例，a_i 将使用 $(p_1,\cdots,p_{j-1},p_{j+1},\cdots,p_J)$ 的采样数据及其自身的历史数据训练模型，并将 p_j 的属性作为标签依次进行测试。

$$r_j=\frac{\sum_{e=1}^{\left|M_j^p\right|}r_e}{\left|M_j^p\right|} \tag{5}$$

$$r_e\left(y,\hat{y}\right)=\frac{n\left(y_n-\hat{y}_n\right)<\tau}{\beta N_j} \tag{6}$$

其中，y_n、$\hat{y}_n$ 和 $\left|y_n-\hat{y}_n\right|$ 分别表示数据样本的真实值、预测值和绝对预测误差；τ 是预设阈值，表示预测的最大容限；函数 $n(\cdot)$ 用来计算满足条件的数据样本数。

为了确保采样的数据可以代表完整的数据，β 的取值至关重要。为确定 β 的取值，首先定义欺骗。如果 DP 在 L 条数据中包含了 l 条伪造数据，并且 AG 从 DP 处抽取了 d 条数据，但是这些采样数据中不包含伪造数据，那么 DP 成功欺骗了 AG。DP 成功欺骗 AG 的概率为

$$\Omega\left(L,l,d\right)=1-\sum_{i=1}^{\min\{b,L\}}\frac{\binom{d}{i}\binom{L-d}{l-i}}{\binom{L}{d}} \tag{7}$$

其中，$\binom{L}{d}$ 计算了从 L 条数据中选择 d 条数据的采样方法数量；$\binom{d}{i}\binom{L-d}{l-i}$ 计算了选择 d 条数据，其中 i 条数据被伪造的采样方法数量；将所有可能的 i 值相加得出的数量表示至少包含了一个伪造数据的方法数；$\sum_{i=1}^{\min\{b,L\}}\frac{\binom{d}{i}\binom{L-d}{l-i}}{\binom{L}{d}}$ 给出了在采样中选择至少一个伪造数据的可能性，即欺诈的可能性。关于如何确保采样的随机性，可以进一步参考文献[19]。

AG 的公平性对数据质量评估有很大影响，为保证 AG 能够公平地进行质量评估，本文采用如下的评价方式。c_k 购买了 p_j 提供的产品 $\mathrm{PC}_{j,k}$，如果 c_k 使用 $\mathrm{PC}_{j,k}$ 进行训练得到的模型准确率 $r_{j,k}$ 与 a_i 给出的数据质量评估 r_j 之间的差距超过约定的阈值 σ，即 $r_j-r_{j,k}\geqslant\sigma$，则 c_k 可以对 a_i 进行投诉。在一段时间内，如果 a_i 受到的投诉累积量大于阈值 ϑ，则 a_i 会被认为是不公平的，它将会被裁撤。

3.3 DC 的效用函数

DC 从 DP 处购买数据，以满足自己的需求，并为此付费。因此，DC 的效用函数由 2 个部分组成：满意度函数和成本函数，计算式为

$$U(c_k)=\mathrm{ST}(c_k)-\mathrm{CO}(c_k) \tag{8}$$

其中，$\mathrm{ST}(c_k)$ 表示 c_k 的满意度函数，参考文献[17]中的构造方式，其计算式为

$$\begin{aligned}&\mathrm{ST}\left(c_k\right)=(1-\alpha_k)\lambda_k\ln\left(q_{j,k}r_jx_{j,k}-n_k^{\min}+1\right),\\&0\leqslant x_{j,k}\leqslant N_j,\ n_k^{\min}\leqslant q_{j,k}r_jx_{j,k}\leqslant n_k^{\max}\end{aligned} \tag{9}$$

ST(c_k) 的构造考虑了数据购买量、数据的属性及质量、c_k 的个人偏好、DC 之间的竞争性等方面的因素。首先，ST 是数据购买量的单调递增函数，并且随着购买数量的增长，ST 的增长速度越来越慢，这是因为数据中所包含的信息熵增长越来越慢[25]。DP 所提供的数据具有不同的属性和质量，属性越丰富，数据质量越高，越能够带来有效的数据量。数据属性的丰富程度用属性相关性 $q_{j,k}$ 来度量，属性越丰富，$q_{j,k}$ 越高。数据质量用 r_j 来表示。此外，c_k 的个人偏好也会对 ST(c_k) 产生影响，用 λ_k 来表示 c_k 的个人偏好程度，λ_k 可以根据 c_k 的需求、习惯等来定义。例如，经常出差的人更喜欢天气预报服务。DC 之间的竞争性采用属性需求的相似度来度量，2 个 DC 的数据属性需求越相似，他们之间具有竞争业务的可能性就越大，从而对 DC 满意度的负面影响也越大。利用 Jaccard 相似系数，定义其他 DC 对 c_k 的影响参数 a_k 为

$$\alpha_k=\frac{\sum_{i=1,i\neq k}^{K}\frac{\left|M_i^c\cap M_k^c\right|}{\left|M_i^c\cup M_k^c\right|}}{K-1} \tag{10}$$

计算属性相似度实际上是计算 2 个集合之间的相似度，因此采用了 Jaccard 相似系数进行计算。

消费者 c_k 具有属性要求 M_k^c，但是在某些情况下，DP 可能无法完全满足 DC 的需求。当属性需求未完全满足时，数据仍然可用，但需要更多数据才能达到相同的效果，$q_{j,k}$ 表示 p_j 提供的数据对 c_k 的可用性。当 c_k 对数据属性与其任务的相关性没有一定的了解，或者认为数据的每个属性的相关性一致时，$q_{j,k}$ 可以表示为

$$q_{j,k}=\begin{cases}\dfrac{\left|M_{j,k}\right|}{\left|M_k^c\right|},\ \dfrac{\left|M_{j,k}\right|}{\left|M_k^c\right|}\geqslant\theta_k\\ 0,\ \dfrac{\left|M_{j,k}\right|}{\left|M_k^c\right|}<\theta_k\end{cases} \tag{11}$$

其中，θ_k 是 c_k 根据相关知识或经验设置的阈值。

在实际的数据挖掘过程中，数据属性的相关性是不一致的。如果 c_k 曾经执行过相关的数据挖掘任务，他将对数据属性的相关性有一定的了解。设第 f 个属性的相关性为 δ_f，$1\leqslant f\leqslant\left|M_k^c\right|$，$\sum_{f=1}^{\left|M_k^c\right|}\delta_f=1$。$q_{j,k}$ 可更新为

$$q_{j,k}=\begin{cases}\sum_{f=1}^{\left|M_{j,k}\right|}\delta_f,\ \sum_{f=1}^{\left|M_{j,k}\right|}\delta_f\geqslant\theta_k\\ 0,\ \sum_{f=1}^{\left|M_{j,k}\right|}\delta_f<\theta_k\end{cases} \tag{12}$$

CO(c_k) 表示购买数据的成本，计算式为

$$\mathrm{CO}(c_k)=x_{j,k}w_{j,k},\quad 0\leqslant x_{j,k}\leqslant N_j \tag{13}$$

上述描述定义了 DP 和 DC 的效用函数，在数据交易过程中他们都以最大化自身的效用为目标。

3.4　数据交易匹配模型

在交易过程中，DP 首先提供产品和价格，可以看作交易的领导者，随后，DC 根据 DP 所提供的信息做出自己的购买决策，可以看作交易的跟随者。因此，为了优化 DP 和 DC 的效用函数，建立了一个涉及多个领导者（所有 DP）和多个跟随者（所有 DC）的双层优化问题（BLPP, bi-level programming problem）。在上层，DP 根据 DC 的属性需求提供相应的数据产品，并给出单价，以最大化其效用。在下层，DC 以最大化自身的效用为目标，通过自我选择过程做出购买决定。上述问题的模型可以表示如下。

1) 上层：DP 的决策

$$\max\left[U\left(p_j\right),w_{j,k}\right]=$$
$$\max_{w_{j,k}}\left(\sum_{k=1}^{K}x_{j,k}w_{j,k}-\sum_{e=1}^{\left|M_j^p\right|}b_e\max\left(x_{j,1},x_{j,2},\cdots,x_{j,K}\right)-a\beta N_j\right)$$
$$\text{s.t. C}_1:0\leqslant x_{j,k}\leqslant N_j \tag{14}$$

DP 的决策变量是为每个 DC 提供的数据产品的单价。

2) 下层：DC 的决策

$$\max[U\left(c_k\right),x_{j,k}]=$$
$$\max_{x_{j,k}}\left(\alpha_k\lambda_k\ln\left(q_{j,k}r_kx_{j,k}-n_k^{\min}+1\right)-x_{j,k}w_{j,k}\right)$$
$$\text{s.t. C}_1:n_k^{\min}\leqslant q_{j,k}r_kx_{j,k}\leqslant n_k^{\max} \tag{15}$$

DC 的决策变量是从某个 DP 处购买的数据量。

3.5　求解算法

BLPP 描述了现实世界中自然发生的层次结构，但是很难求解，甚至最简单的线性双层规划问题也被证明是 NP 难的[26]。由于所提模型涉及许多变量和非线性的效用函数，且下层目标函数是非凸的，因此很难使用经典算法求解[27]。在下层，DC

遍历 DP 的产品，可以得出最优的决策。在得到 DC 的决策后，BLPP 问题就退化为单个多目标优化问题。因此，本文提出了基于 NSGAII 的 BLPP 求解方案。

NSGAII 算法是一种中心化的优化算法，由一个节点来计算所有的目标函数，需要知晓 DP 和 DC 的完整参数（如 b_e、$n_k^{\min}$、$n_k^{\max}$、λ_k、$q_{j,k}$ 等）来计算 $U(p_j)$ 和 $U(c_k)$。这些信息中包含了 DP 和 DC 的隐私。因此，本文对 NSGAII 进行了改进，提出了协作式 NSGAII 算法，$U(p_j)$ 和 $U(c_k)$ 分别由 DP 和 DC 在本地进行计算，以免泄露隐私。DP 负责种群的产生，即提供给 DC 的产品的价格，以及种群的遗传和变异。DC 从 DP 提供的产品中进行选择，以决策从哪个 DP 购买产品和所购买的数据量。AG 负责非支配的排序，它可以避免 DP 的一些自私行为产生的影响，例如欺诈性排序等，还可以加快算法的收敛速度。T 和 $\text{Gen}^{\max}$ 是需要设置的参数，T 是种群中的个体数；$\text{Gen}^{\max}$ 是最大迭代次数，用来限制算法的迭代次数，当迭代次数达到 $\text{Gen}^{\max}$ 时，算法会终止运行，并输出结果。协作式 NSGAII 如算法 1 所示。

算法 1 协作式 NSGAII

输入 M_j^p，M_k^c，N_j，$n_k^{\min}$，$n_k^{\max}$，b_e，β，r_j，a，θ_k，T，$\text{Gen}^{\max}$

输出 $x_{j,k}$，$w_{j,k}$

1) for $j=1:J$

2) p_j 生成大小为 T 的种群 $\text{PS}_j(1)$，种群中每个个体包含 K 个元素（$\boldsymbol{w}_j=\{w_{j,1},w_{j,2},\cdots,w_{j,K}\}$），将 $\text{PS}_j(1)$ 发送给 a_i，a_i 获取所有的 $\text{PS}_j(1)$ 后，会对它们进行重构生成 $\text{PS}_j(1)$，种群大小为 T，每个个体包含 J 个元素（$\boldsymbol{w}_k=\{w_{1,k},w_{2,k},\cdots,w_{J,k}\}$）

3) end for

4) 执行子程序

5) for $j=1:J$

6) p_j 获取 $x_{j,k}$，并根据式(1)计算适应度函数（即效用函数），然后将计算结果发给 a_i

7) end for

8) a_i 执行非支配排序，并将 p_j 的排序结果传输给 p_j

9) for $j=1:J$

10) p_j 执行遗传、交叉和变异操作以生成子群体 $\text{CS}_j(1)$

11) end for

12) 定义 Gen=1

13) for $j=1:J$

14) p_j 将 $\text{PS}_j(\text{Gen})$ 与 $\text{CS}_j(\text{Gen})$ 合并，然后发送给 a_i

15) end for

16) a_i 重构种群，并执行子程序

17) for $j=1:J$

18) p_j 获取 $x_{j,k}$ 并根据式(1)计算适应度函数，发送计算结果给 a_i

19) end for

20) a_i 执行非支配排序，并将 p_j 的排序结果传输给 p_j

21) for $j=1:J$

22) p_j 计算每个非主导层中个体的拥挤度

23) p_j 根据非主导关系和拥挤度选择合适的个体来组成 $\text{PS}_j(\text{Gen+1})$

24) p_j 执行遗传、交叉和变异操作以生成 $\text{CS}_j(\text{Gen+1})$

25) end for

26) 判定 Gen 是否等于 $\text{Gen}^{\max}$

27) 如果否，则设定 Gen=Gen+1，并返回步骤 14)

28) 如果是，则输出最高等级的 PS(Gen) 子程序

1) for $k=1:K$

2) c_k 从 a_i 处接收 PS(Gen)

3) 对于种群中的每个个体，c_k 计算 $x_{j,k}=\arg\limits_{x_{j,k}} U(c_k)$，并将 $x_{j,k}$ 发送回 a_i，对于没有选取的数据产品，返回 $x_{j,k}=0$

4) end for

5) 结束子程序

本文所提交易匹配算法中，每次迭代内的计算复杂度是 $O(JT^2)$，迭代进行的最大次数是 $\text{Gen}^{\max}$，因此匹配算法的计算复杂度是 $O(\text{Gen}^{\max}JT^2)$。

4 安全性分析和实验结果

4.1 安全性分析

本文所提基于联盟区块链的数据交易框架通

过加密算法、数字签名等技术，可以满足以下与区块链相关的安全要求。

1) 避免不可信第三方造成的数据泄露。在基于区块链的交易框架中，DP 会将数据产品直接传输给 DC，避免了不可信第三方可能造成的数据泄露。

2) 用户身份隐私保护。链上的交易记录中记录的都是 DP 和 DC 的钱包地址，而通过钱包地址很难挖掘出用户的真实身份。

3) 数据不可篡改。联盟区块链的分散性质与数字签名相结合，确保了区块链上的数据不会被篡改。

4) 没有双花。数字货币依靠数字签名来证明所有权以及公开的交易历史，可以防止双花。

同时，交易框架中一些机制的设定，也可以进一步保证交易的安全性与公平性。

1) 交易过程中，DP 仅需提供数据的属性和数据量信息，DC 仅需提供属性需求信息，这些信息中并不包含可能泄露用户身份的隐私数据，这一设定有助于保护用户的身份隐私。

2) AG 对 DP 数据的质量评估需要采样 DP 的数据，这部分数据可能被泄露。为保证采样数据能够代表完整的数据，并且尽量少地泄露 DP 的数据，需要设定合适的采样比例。为此，本文计算了不同采样比例下成功欺骗的概率。图 3 展示了不同伪造数据占比的情况下，成功欺骗的概率随采样比例的变化。在伪造数据占比为 0.2%、1%和 2%的情况下，很难通过低的采样率来得到较低的欺骗成功率，但此时伪造数据对数据集的影响也很小。当伪造数据的占比达到 5%、采样比例为 5%时，成功欺骗的概率只有 0.72%，这能够满足 AG 的需求，而且 5%的数据泄露对 DP 来说也是可接受的[19]。

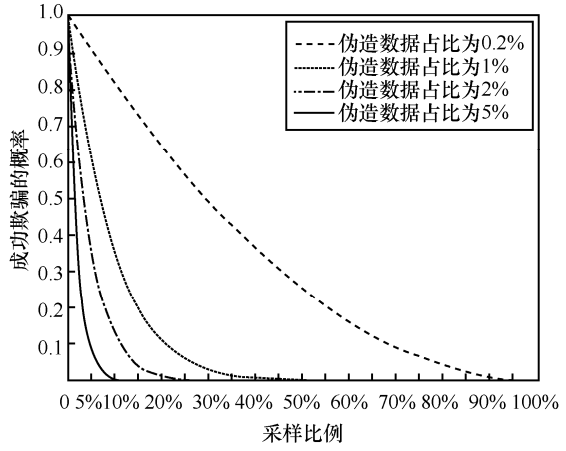

图 3　不同伪造数据占比的情况下，成功欺骗的概率随采样比例的变化

3) 在执行交易匹配算法时，由 AG 进行非支配排序，相比于选定某一个 DP 或 DC 进行排序，可以保证排序的公平性，从而维护所有 DP 与 DC 的效用。为此设计了以下实验，分别由 AG 和随机选取的某个 DP(实验中选取了 DP_1)进行非支配排序，DP 在排序时会进行欺诈排序，即将自己的出价排在上层，实验结果如图 4 所示。

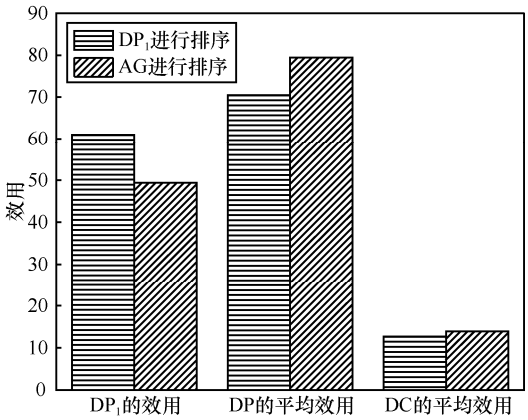

图 4　AG 和 DP_1 分别进行排序时效用函数的对比

从图 4 中可以看出，当 DP_1 进行排序时，可以提高自己的效用，但会使 DP 和 DC 的平均效用下降，即损失了其他 DP 和 DC 的效用。因此，采用 AG 进行排序有助于维护所有 DP 与 DC 的效用。

4.2　实验结果

实验使用北京的空气质量数据来提供空气质量预测服务[28]。该数据集包括 12 个空气质量监测站点的空气污染物数据，每个监测站点被视为一个 DP。空气质量数据来自北京市环境监测中心。每个空气质量监测站点的气象数据均与最近的气象站相匹配。时间段是从 2013 年 3 月 1 日到 2017 年 2 月 28 日。每个数据集都包含 35 064 条数据样本。每条样本包含时间、PM2.5、PM10、SO_2 含量等 17 种属性。为了使这 12 个 DP 的数据产生显著差异，实验中删除了某些 DP 数据的一些属性，并给数据集加入了不用程度的噪声。在实验过程中，设 AG 的采样比例为 5%，即 $\beta = 5\%$。AG 使用经典的机器学习算法(随机森林回归算法)进行数据质量评估。实验虚拟了 100 个 DC，即 $K = 100$。在 DC 中，一半 DC 是经验丰富的，具有属性相关性的知识，利用式(12)进行计算；另一半 DC 没有经验，利用式(11)进行计算。设定 1 000 条数据样本为单位数据。

为了展示协作式 NSGAII 的收敛性，图 5 和图 6 分别给出了 DP 和 DC 的效用随迭代次数的变化。

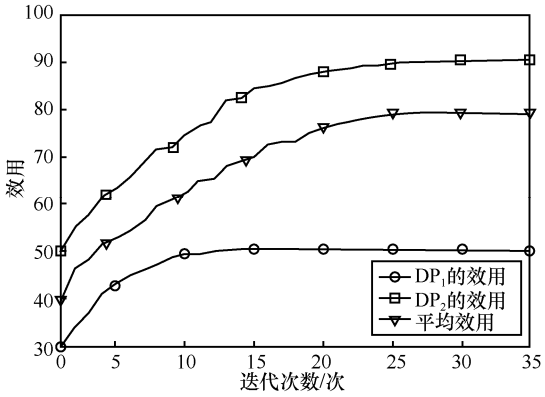

图 5 DP 的效用随迭代次数的变化

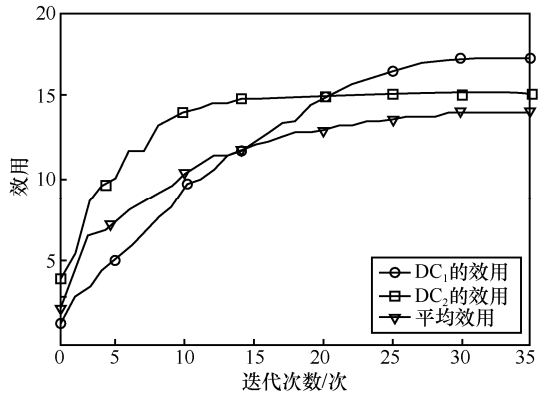

图 6 DC 的效用随迭代次数的变化

从图 5 和图 6 可以看出，随着迭代次数的增多，DP 和 DC 的效用都逐渐增长。当达到一定的迭代次数后，DP 和 DC 的效用逐渐趋于稳定。从所有 DP 和 DC 的平均效用来看，算法在 30 次之内可以达到收敛。就迭代次数而言，30 次是满足用户需求可接受的数量。

为了体现效用函数的灵敏性，实验中对 DP 和 DC 进行了差异化处理。针对 DP，对数据的属性和数量进行了调整，有些 DP 只能提供一部分属性，并且在数据集中增加了不同程度的噪声。为模拟不同 DC 的需求，实验中设定了多样化的属性需求、数据量需求、偏好以及对属性相关性是否熟悉等参数。DP 和 DC 的效用函数分别如图 7 和图 8 所示。

从图 7 和图 8 可以看出，实验中对 DP 和 DC 的差异化处理在他们的效用上得到了体现，这说明效用函数对 DP 和 DC 的参数具有灵敏性。

为了验证协作式 NSGAII 的可扩展性，图 9 和图 10 分别给出了迭代次数随 DP 和 DC 数量增长的变化。

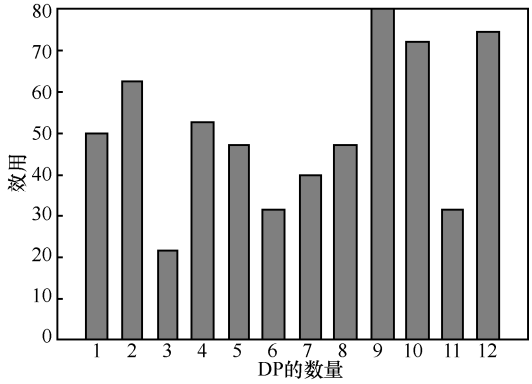

图 7 DP 的效用函数

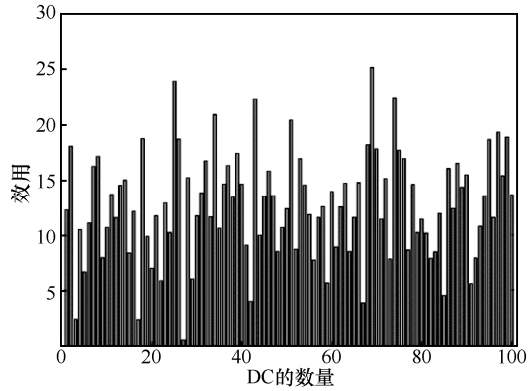

图 8 DC 的效用函数

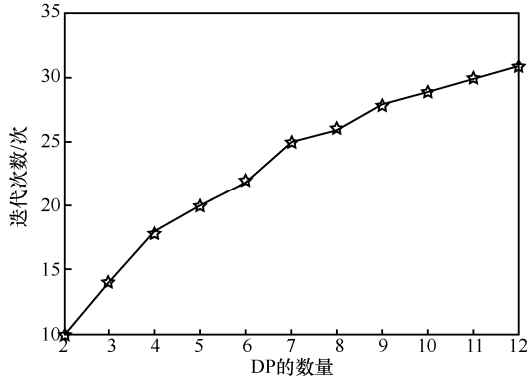

图 9 迭代次数随 DP 数量增长的变化

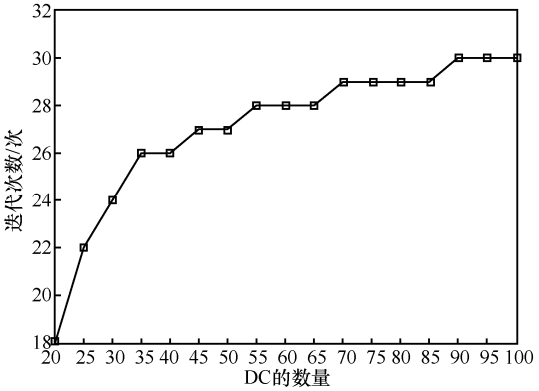

图 10 迭代次数随 DC 数量增长的变化

从图 9 可以看出，随着 DP 数量的增加，迭代次数也随之增加，并且两者近似成正比。从图 10 可以看出，随着 DC 数量的增加，迭代次数的增长速度逐渐减慢并最终趋于稳定。根据对数据交易市场的调研发现，对于特定类型的数据，数据市场基本是寡头市场的形式，DP 的数量不会太多，因此不会导致迭代次数的大幅增加。而对于 DC 的增加，该算法具有很好的收敛性。这表明协作式 NSGAII 是可扩展的，可以应用于大规模的用户。

在模型的构建中，考虑了数据属性的影响。当不能完全满足 DC 属性需求或完全满足属性需求的数据价格过高时，DC 可以通过使用较低的价格来购买较少属性的数据以提升自身的效用。具有较少属性的 DP 可以通过使用较低的价格将数据卖给更多的 DC，从而提升效用。为了验证这种考虑是否有助于提高 DP 和 DC 的效用，图 11 和图 12 分别给出了 DP 和 DC 的效用对比。

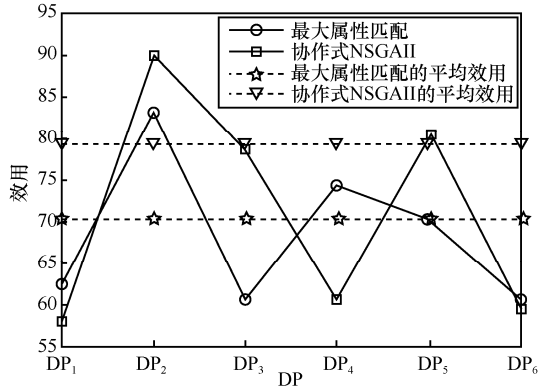

图 11　DP 的效用对比

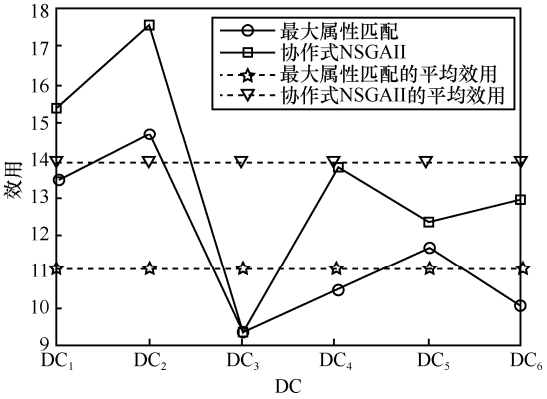

图 12　DC 的效用对比

图 11 中，DP_1～DP_6 是从 12 个 DP 中选取的 6 个，他们在数据属性、数据数量及添加的噪声方面存在区别，具体参数如表 2 所示。

表 2　　DP 的具体参数

DP	$\lvert M_j^p \rvert$	N_j /单元	数据噪声等级
DP_1	10	17	Ⅳ
DP_2	17	35	Ⅰ
DP_3	15	30	Ⅱ
DP_4	12	30	Ⅲ
DP_5	14	30	Ⅱ
DP_6	10	20	Ⅳ

表 2 中，每单元包含 1 000 条数据，增加的高斯白噪声分别为Ⅰ、Ⅱ、Ⅲ和Ⅳ这 4 个等级，对应的噪声量分别为 0～2%、4%～6%、8%～10%和 12%～14%。

图 12 中，DC_1～DC_6 是从 100 个 DC 中选取的 6 个，他们在属性需求、数据量需求、个人偏好等方面存在区别，具体参数如表 3 所示。

表 3　　DC 的具体参数

DC	$\lvert M_k^c \rvert$	$n_k^{\min}$ /单元	λ_k	$q_{j,k}$ 是否熟悉
DC_1	13	27	0.85	是
DC_2	10	25	0.90	是
DC_3	15	30	0.55	否
DC_4	12	27	0.82	是
DC_5	15	29	0.62	否
DC_6	16	26	0.73	否

从实验结果可以看出，协作式 NSGAII 取得了很好的效果。最大属性匹配算法是 DC 将购买最能满足其属性需求的数据产品。在应用最大属性匹配算法的情况下，从 DC 的角度来看，他们的产品选择要小得多，因此他们可能无法实现最高的效用。从 DP 的角度来看，具有丰富属性的 DP 将出售更多的数据，而具有较少属性的 DP 则难以出售数据，最终降低了总体效用。因此，考虑数据属性在数据匹配过程中的影响，对 DP 和 DC 的效用函数具有一定的改善作用。

在实际的数据挖掘中，数据属性与任务的相关性对最终的模型具有很大的影响。在某些实验中，甚至使用 PCA 等算法过滤掉一些属性。相关性的考虑可以促使 DC 选择具有更大相关性的属性，使用较少的属性来获得更好的数据挖掘结果，从而降低成本并提高效用。为了显示这种效果，图 13 给出了是否考虑相关性的情况下 DC 效用的对比。

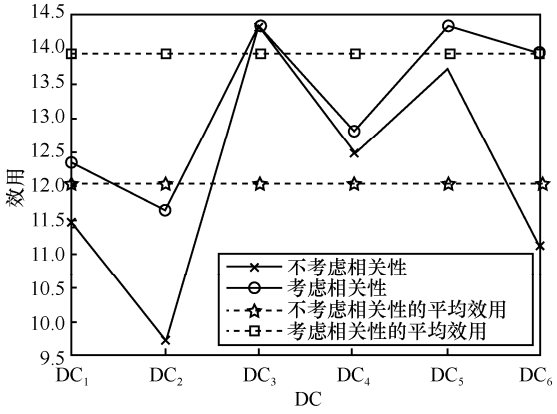

图 13 是否考虑相关性的情况下 DC 效用的对比

从图 13 可以看出，通过考虑属性的相关性，DC 的效用得到了改善。

为了进一步展示协作式 NSGAII 的有效性，将它与其他几种算法进行了比较。时间匹配算法是区块链系统中常见的基于时间戳的匹配方案。数据质量匹配算法是 Yu 等[13]提出的基于数据质量的匹配算法。社会福利最大化是 Cao 等[11]提出的一种社会福利最大化的数据交易算法。利润最大化是 Jiao 等[12]提出的一种以最大化代理利润为目标的匹配算法。DP 和 DC 的平均效用在不同算法上的对比分别如图 14 和图 15 所示。

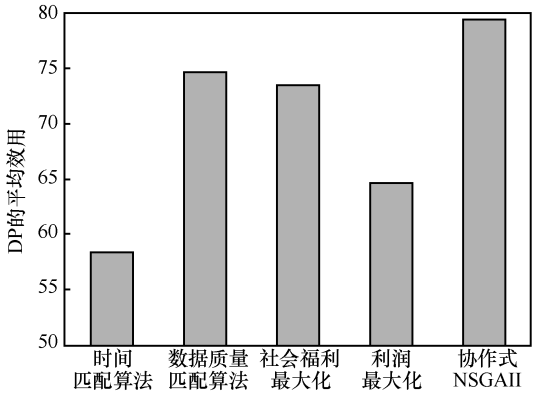

图 14 DP 的平均效用在不同算法上的对比

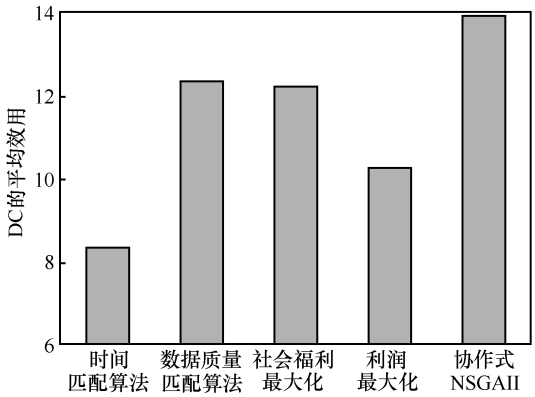

图 15 DC 的平均效用在不同算法上的对比

从图 14 和图 15 中可以看出，协作式 NSGAII 取得了最佳效果，它综合考虑了数据质量、数据属性、属性相关性以及消费者竞争的影响，使 DC 和 DP 能够根据自身的情况更细粒度地执行数据交易，从而提升了他们的效用。

为进一步展示协作式 NSGAII 的效果，将它的实验结果与 Gurobi 优化工具箱的结果进行了对比，结果如图 16 所示。

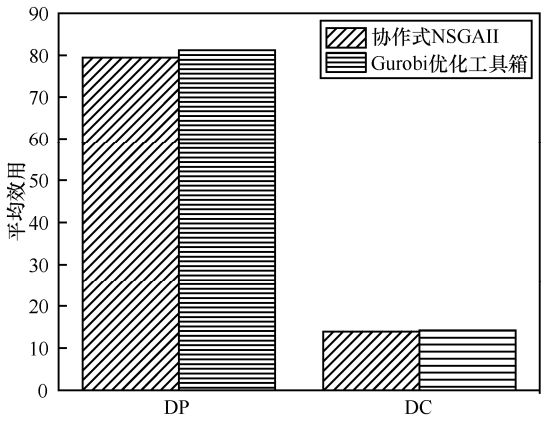

图 16 协作式 NSGAII 与 Gurobi 的对比

Gurobi 优化工具箱是一种应用广泛的优化问题求解工具箱，它的计算结果可以认为已经接近全局最优解。从实验结果可以看出，协作式 NSGAII 的实验结果与 Gurobi 的结果极其相近，这在某种程度上说明了协作式 NSGAII 的实验结果很接近全局最优解。

5 结束语

为了提高数据交易的有效性，本文提出了一种基于联盟区块链的分布式交易框架。该框架可在不依赖第三方的情况下实现 DP 和 DC 之间 P2P 的数据交易。在此框架下，本文还提出了双层多目标优化模型，以优化 DP 和 DC 的效用函数，模型构建过程中考虑了数据属性、数据质量、属性相关性以及消费者竞争的影响。为求解此模型，提出了一种协作式 NSGAII，通过 DC、DP 和 AG 的合作来得到该模型的解。最后，基于北京空气质量数据的实验表明，所提算法在 DP 和 DC 的效用函数方面可以实现更好的性能。

参考文献：

[1] DAI W, DAIA C, CHOO K K R, et al. SDTE: a secure blockchain-based data trading ecosystem[J]. IEEE Transactions on Information Forensics and Security, 2019, 15(99): 725-737.

[2] 付钰, 俞艺涵, 吴晓平. 大数据环境下差分隐私保护技术及应用[J]. 通信学报, 2019, 40(10): 157-168.

FU Y, YU Y H, WU X P. Differential privacy protection technology and its application in big data environment[J]. Journal on Communications, 2019, 40(10): 157-168.
[3] SPOERRY C. A marketplace approach to trade-related risk data[C]//2019 IEEE World Congress on Services. Piscataway: IEEE Press, 2019: 230-233.
[4] MCKELVIE S J, MURPHY E E S, DICKSON M J, et al. Islands of data: U.S. Patent Application 12/492,283[P]. (2010-12-30) [2020-08-07].
[5] KONG W, QIAO F, WU Q. Real-manufacturing-oriented big data analysis and data value evaluation with domain knowledge[J]. Computational Statistics, 2019(2): 1-24.
[6] LU Y, HUANG X, DAI Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186.
[7] NIU C, ZHENG Z, WU F, et al. Trading data in good faith: integrating truthfulness and privacy preservation in data markets[C]//2017 IEEE 33rd International Conference on Data Engineering. Piscataway: IEEE Press, 2017: 223-226.
[8] TIAN L, LI J, LI W, et al. Optimal contract-based mechanisms for online data trading markets[J]. IEEE Internet of Things Journal, 2019, 6(5): 7800-7810.
[9] 郭艺, 叶剑, 张鹏. 基于偏差约减的大数据交易模型分析与修复方法[J]. 电子学报, 2018, 46(7): 1754-1761.
GUO Y, YE J, ZHANG P. Analysis and repair of big data transaction model based on deviation reduction[J]. Acta Electronica Sinica, 2018, 46(7): 1754-1761.
[10] NIYATO D, ALSHEIKH M A, WANG P, et al. Market model and optimal pricing scheme of big data and Internet of things (IoT)[C]//2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.
[11] CAO X, CHEN Y, LIU K J R. Data trading with multiple owners, collectors, and users: An iterative auction mechanism[J]. IEEE Transactions on Signal and Information Processing over Networks, 2017, 3(2): 268-281.
[12] JIAO Y, WANG P, NIYATO D, et al. Profit maximization auction and data management in big data markets[C]//2017 IEEE Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2017: 1-6.
[13] YU H, ZHANG M. Data pricing strategy based on data quality[J]. Computers & Industrial Engineering, 2017, 112: 1-10.
[14] GAO W, YU W, LIANG F, et al. Privacy-preserving auction for big data trading using homomorphic encryption[J]. IEEE Transactions on Network Science and Engineering, 2020. 7(2): 776-791.
[15] NIU C, ZHENG Z, WU F, et al. Achieving data truthfulness and privacy preservation in data markets[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(1): 105-119.
[16] JUNG T, LI X Y, HUANG W, et al. Account Trade: accountability against dishonest big data buyers and sellers[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(1): 223-234.
[17] CHEN C, WU J, LIN H, et al. A secure and efficient blockchain-based data trading approach for Internet of vehicles[J]. IEEE Transactions on Vehicular Technology, 2019, 68(9): 9110-9121.
[18] WEBER T, PRINZ W. Trading user data: a blockchain based approach[C]//2019 Sixth International Conference on Internet of Things: Systems, Management and Security. Piscataway: IEEE Press, 2019: 547-554.
[19] DELGADO-SEGURA S, PÉREZ-SOLÀ C, NAVARRO-ARRIBAS G, et al. A fair protocol for data trading based on Bitcoin transactions[J]. Future Generation Computer Systems, 2020, 107: 832-840.
[20] MISSIER P, BAJOUDAH S, CAPOSSELE A, et al. Mind my value: a decentralized infrastructure for fair and trusted iot data trading[C]//Proceedings of the Seventh International Conference on the Internet of Things. New York: ACM Press, 2017: 1-8.
[21] SABOUNCHI M, WEI J, ROCHE R. Blockchain-enabled peer-to-peer data trading mechanism[C]//2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). Piscataway: IEEE Press, 2018, doi: 10.1109/Cybermatics_ 2018.2018.00241.
[22] GAO J, WU T, LI X. Secure, fair and instant data trading scheme based on bitcoin[J]. Journal of Information Security and Applications, 2020, 53: 102511.
[23] HUANG X, ZHANG Y, LI D, et al. An optimal scheduling algorithm for hybrid EV charging scenario using consortium blockchains[J]. Future Generation Computer Systems, 2019, 91: 555-562.
[24] VUKOLIĆ M. The quest for scalable blockchain fabric: proof-of-work vs. BFT replication[C]//International Workshop on Open Problems in Network Security. Berlin: Springer, 2016: 112-125.
[25] LI X J, YAO J G, LIU X, et al. A first look at information entropy-based data pricing[C]//2017 IEEE 37th International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2017: 2053-2060.
[26] HANSEN P, JAUMARD B, SAVARD G. New branch-and-bound rules for linear bilevel programming[J]. SIAM Journal on scientific and Statistical Computing, 1992, 13(5): 1194-1217.
[27] BISWAS A, HOYLE C. A literature review: solving constrained non-linear bi-level optimization problems with classical methods[C]//45th Design Automation Conference (DAC), IDETC, ASME. [S.n.:s.l.], 2019: 1-12.
[28] ZHANG S, GUO B, DONG A, et al. Cautionary tales on air-quality improvement in Beijing[J]. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2017, 473(2205): 20170457.

[作者简介]

黄小红（1979− ），女，广州佛山人，博士，北京邮电大学教授，主要研究方向为计算机网络应用、下一代互联网和网络安全等。

张勇（1990− ），男，河北衡水人，北京邮电大学博士生，主要研究方向为区块链、大数据交易和数据隐私保护等。

闪德胜（1963− ），男，河南孟县人，中国人民解放军 32147 部队高级工程师，主要研究方向为网络测量、网络安全等。

钱叶魁（1980− ），男，安徽枞阳人，博士，陆军炮兵防空兵学院教授、硕士生导师，主要研究方向为网络测量、网络安全等。

韩璐（1991− ），女，蒙古族，内蒙古赤峰人，北京邮电大学博士生，主要研究方向为安全多方计算、联邦学习等。

李丹丹（1987− ），女，河南平顶山人，博士，北京邮电大学讲师，主要研究方向为网络安全、密码学等。

丛群（1980− ），男，辽宁盘锦人，北京网瑞达科技有限公司高级工程师，主要研究方向为网络管理。