基于集成机器学习的政府数据定价模型及系统设计

武少杰
重庆交通大学经济与管理学院　400000

摘　要：近几年我国数字经济快速发展，数据发掘和应用领域在政府加持下飞速更新成长。其中政府数据是一种特殊的公共产品，涵盖不同领域，跨越多个层次，具有较强的公益性和社会性，其价值有着不同维度，受多种因素影响。本文研究分析了传统数据价值评估的方法及其特点，指出目前政府数据资产价值评估的局限性，并构建出政府数据资产定价模型和系统，根据系统功能进行若干模块的划分，应用集成机器学习的算法来评估政府数据资产价值。
关键词：机器学习；政府数据资产；价值评估；定价系统

引言

目前，我国已形成政府大数据、互联网大数据、行业大数据三分天下的格局，但“目前我国信息数据资源，80%以上掌握在各级政府部门手里”。因此，政府大数据的开发利用是大数据开发的重头戏。然而，尽管政府数据包含着巨大的政治、经济和社会价值，在向社会开放后还将创造更大的公共价值，政府数据价值的构成、体现和测度指标尚未公论，价值测度欠缺科学适用的方法，也没有规范化，导致政府数据的管理工作“心中无数”，在开展统计核算、市场化、资产化等工作中缺乏价值依据。

一、文献综述

对于开放的政府数据价值评估，要求对政府数据的资产化和交换交易要求明确数据的价值构成和定价机理，并选择适当的定价方法予以科学估值。通过研究近年来国内外学者和进行数据买卖的企业的市场研究，概括而言，可归纳为如下三个方面：

1.对于政府数据的研究

付熙雯和郑磊曾在一篇综述文章中对政府数据的早期研究做了说明，特别厘清了政府数据开放、政府信息公开和政府信息资源增值利用的区别和联系。[8]在此基础上，孙凡和杨周南研究了政府数据作为公共信息产品向社会提供的策略，并分析了其特殊性。Attard等认为政府因这三方面的理由，应该对外开放数据，也即透明度、释放社会和商业价值、参与治理。夏义堃和管茜认为应强调数据的资产属性，探索了把政府数据作为资产进行管理的框架和运作模式。除此之外，研究者也注意到政府数据价值管理的需要，提出数据的价值应被纳入国民经济核算体系。门理想等也从公共价值的视角分析了政府数据开放的价值内涵问题，但这方面的研究仍比较欠缺。

2.政府数据价值维度及其影响因素的研究

数据价值的高低受到多种因素影响。康旗等认为，数据作为资产的价值需要从五个维度进行考虑：规模、活性、多维度、关联性和颗粒度，并应区分不同的行业和业务特征分别灵活描述。张志刚等从数据资产的成本和应用两方面分析了数据的价值构成及影响因素。其中数据成本由建设费用和运维费用组成，而数据应用包括数据分类、使用次数、使用对象、使用效果评价等。赵子瑞提出应将人力指标、物力指标、交易佣金指标作为大数据交易的基本价格指标。[6]Heckman等则进一步把影响数据价值的因素分为三个方面详细阐述，其中第一方面是基于价值的参数，包括数据在节约时间和金钱上的价值、数据使用者获得的收益、数据是否具有排他性、数据的权属范围等；第二方面是基于质量的参数，包括数据的年份、可信度、准确性等；第三方面是固定成本参数，包括收集、存储和分析数据的成本。他们认为数据质量是影响数据资产价值的关键因素。

3.政府数据价值测度方法的研究

最早的数据价值的测度方法是随着“数据作为一种服务”的理念而提出的，是一种基于订阅的测量模式，即数据服务提供商向用户提供数据产品订阅而收取的费用，这种方法缺乏对数据价值的客观评价，也未考虑数据的资产估值。随着数据市场和数据管理的发展，数据提供商又提出了一种基于数据类型的价值测度方法，即

根据数据的类型或属性来确定数据价值，以体现顾客对于数据更为具体的要求。

此外，研究还发现数据价值有着双重的不确定性，因使用场景（刘朝阳，2016）和使用对象（Lusch和Nambisan，2015）而不同。这些工作帮助研究者提出了一些数据价值的评价指标，也开发了多种针对数据产品和数据资产的价值测度方法。最早有基于数据类型的方法，即根据数据的类型或属性来确定数据价值。然而，尽管已有研究者设计了政府数据价值评估的方法框架，提出了政府数据价值发现的概念模型和实施路径，[7]也分析了政府数据价值的特征和既有测量方法，并对各类方法的适用性进行了讨论，但目前的研究欠缺对于政府数据价值的全面把握，尚未针对性地建立政府数据价值的测度指标体系、开发测度方法，距离科学、完整地对政府数据价值进行统计测度尚一定距离。

二、政府数据资产定价模型建立

开放政府数据价值评估特征多、复杂且冗余度高，使得寻找评估分类器的最优评估特征集成为难点问题。传统的评估方法前期往往需要繁复的特征工程来保证其预测精度，不仅效率低下，模型的精确度很大程度上会受到前期特征工程工作的影响。因此，我们采用混合集成机器学习的算法研究针对这一问题进行了创新改进，采用下列机器学习集成的算法模型进行前期特征工程与数据资产的估值定价：

1.卷积神经网络

卷积神经网络（CNN）是一种复杂的神经网络，它具有强大的预测能力，能够捕捉到大量的信息，并且能够有效地处理复杂的模型。CNN由一个或多个卷积层、一个全连通层、一个关联权重和一个池化层组成，能够准确地预测模型的复杂性。通过这种结构，卷积神经网络可以有效地捕捉输入数据的二维结构。

2.XGBoost算法

Boosting方法通过叠加多个弱分类器的方式来增加总分类器的准确度，XGBoost用改造后的CART树去迭代，用添加新树的方式去不断降低总的损失函数，采用传统决策树的分裂方式，每一步分裂的原则则是让上述损失减小。

3.LGB算法

LGB算法主要是由三个算法构成：第一，与XGBoost的加权分位数草图类似，直方图算法也是在决策树寻找分割点的过程中进行优化，减少大量数据的重复遍历。具体来说，直方图算法首先将每个特征的连续数据分割到K个bin中，即每个bin中分得一定数量的数据，由此原始的连续数据就变成了离散的bin数据。第二，在Histogram算法之上，LightGBM进行进一步的优化。抛弃了大多数GBDT工具使用的按层生长的决策树生长策略，而使用了带有深度限制的按叶子生长算法。[1]XGBoost采用Level-wise的增长策略，该策略遍历一次数据可以同时分裂同一层的叶子，容易进行多线程优化，也好控制模型复杂度，不容易过拟合。第三，Gradient-based One-Side Sampling即单边梯度采样算法。为了降低数据量，笔者将大部分小梯度样本剔除，只使用剩余的样本来计算信息增益，以确保减少数据量和保持精度上的平衡。[2]

4.随机森林

随机森林是Bagging算法的一个拓展，在以决策为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入随机属性选择，具体来说，传统决策树在选择划分属性的时候在当前节点选出一个最有属性；[3]而在随机森林中，对基决策树的每个结点，先从该结点的属性合中随机选择一个包含k个属性的子集，然后再从这个子集中选择一个最优属性用户划分。[4]

三、系统设计及运行流程

采用最新的模块技术，我们的数据估值系统可以有效地改善模型的性能，提高估值效率。该系统由数据输入接口模块、数据采集模块、数据预处理模块、特征工程构建模块、估值定价模块这五个模块组建。

S1数据输入接口模块：针对开放政府数据信息特征类型多的特点，该系统设计数据输入接口模块。该模块分为两个子模块：第一个子模块为数据输入模块，此模块测试数据包括数据大小、数据类型、字段数量、数据条数、采集时间等7项自变量数据的输入。另一个子模块为分类模块，此模块按照产业经济、金融征信、舆情监测、科研技术等10个应用场景，利用文本相似度将第一个子模块中输入的数据划分为不同的可比案例集。[9]

S2数据采集模块：政府数据资产具备量大、高维、数据类型多样等特点，为此，针对性地组建数据采集模块。笔者借助Python进行信息爬取，从而获得包括块数据和API政府数据交易的资产标题、数据大小、数据类型、采集时间、字段数量、数据条数等在内的所有信息。

S3数据预处理模块：为了解决政府数据存在数据类型繁杂、数据源不唯一、数据存在缺失等问题，该系统组建异常数据预处理模块，以提高数据质量。在收集和存储数据后，该模块会从多条数据中提取出历史交易信息，并进行一致性检查、缺失数据处理、回归分析等预

处理操作，从而有效地提升数据的准确性和可靠性。

S4特征工程构建模块：特征工程构建模块可以有效地收集和分析开放政府数据中的相关信息，并将其转换为可用于建立特定模型的特征数据，以此来提升模型的准确性和可靠性，其具体实现了：

①通过文本分析，可以将政府开放数据资产标题的文本转换成词向量，然后使用CNN卷积神经网络进行特征降维，从而更好地理解和处理这些标题；[9]

②One-hot编码是一种有效的数据类型提取和分类方法，它将离散特征的值扩展到欧几里德空间，从而有效地计算出特征之间的距离，从而更好地满足分类需求；

③对于数据大小、采集时间、字段数量、字段条数这些变量采用归一化和标准化处理，从而抑制不同的量纲差异。

S5估值定价模块：估值定价模块输入特征工程提取的特征信息，输出待估政府数据资产的参考定价，具体实现了：

①OGD价值测度理论模型：将可比实例和政府数据资产特征数据进行编码，以构建一个m×n的特征价格矩阵，作为可比的基础，从而评估其价值；

②OGD价值测度算法系统：即基于机器学习算法建立政府数据定价模型。通过使用多种机器学习算法，如XGboost、LGB和CB，我们可以将政府数据中的特征信息与OGD价值测量理论模型相结合，从而构建一个具有良好回归能力和鲁棒性的集成模型。通过对比不同的案例，我们可以更精确地估算政府数据资产的价值。

四、系统运行效果与展望

本文所构建的政府数据资产估值模型，针对现阶段政府数据价值的构成、体现和测度指标尚无公论，价值测度欠缺科学适用的方法，也没有严谨的规范，导致政府数据的管理工作“心中无数”，在开展统计核算、市场化、资产化等工作中缺乏价值依据，本文结合特征工程提取的特征信息，组建数据输入接口模块、数据采集模块、数据预处理模块、特征工程构建模块、估值定价模块这五项模块，实现对大数据资产价值的估值，具有较高的估价精度、良好的鲁棒性、较少的干预度和不错的市场可推广性。该系统可以实现如下的预期功能：

1.数据价格信息透视化。通过用户输入数据标题、数据类型、数据集大小、采集时间、数据条数等信息以及选择数据标签，使用集成机器学习模型对数据资产价格进行估价以及可以进行批量数据定价，通过导入csv、text、xlsx格式文件对数据进行有效预测。合理平衡交易双方因信息不对称产生的一些列困扰，形成公平交易的原则。

2.有效管理数据资产。全球数据呈现爆发增长、海量聚集的特点。通过构建一个完整的元数据管理流程，从需求收集、获取、处理、分析到维护，可以有效地收集和管理不同业务领域的数据，并确保数据处理的准确性和可靠性，从而实现系统的有效运行，最终形成一套完善的元数据管理体系。

3.为有关部门提供政府数据价值的统计测度和核算工具，加强政府数据的管理工作。

4.为政府数据的市场化、资产化工作和交易活动提供价值测度的依据，是建设数据政府、构建数据市场的坚实基础。

参考文献

[1]慕钢，张宏烈，党佳俊等.基于LightGBM模型的二手房房价预测研究[J].高师理科学刊，2020，40（12）：27-31.

[2]张月飞，王伟，代伟.重介分选过程产品指标在线预测方法研究[J].煤炭工程，2021，53（S1）：108-111.

[3]张溶芳，许丹丹，王元光等.机器学习在物联网虚假用户识别中的运用[J].电信科学，2019，35（07）：136-144.

[4]王帅，岳鹏飞，董晗睿等.基于机器学习的织物疵点检测[J].纺织科技进展，2020，（10）：25-30.

[5]宋曦，高文鹏.基于AlphaPose与改进LightGBM算法的触电跌倒检测方法[J].电力信息与通信技术，2023，21（04）：44-50.

[6]王卫，张梦君，王晶.大数据交易业务流程中的风险因素识别研究[J].情报理论与实践，2019，42（09）：80-85.

[7]朱晟.审计视域下企业数据资产增值与防控管理的探索和实践[J].投资与创业，2022，33（10）：207-209.

[8]周斌，雷挺.大数据时代政府治理的研究热点与发展趋势——基于CiteSpace的文献计量学分析[J].岭南学刊，2020，（03）：66-75.

[9]任建宇.基于集成机器学习的数据资产定价模型及系统设计[J].中国管理信息化，2022，25（14）：80-82.