文章编号:1003－207(2024)07－0028－09
DOI:10.16381/j.cnki.issn1003－207x.2022.0562

需求预测视角下的医疗数据价值
——基于沙普利值方法

赵 越,王衍之,宋 洁,何 璇

(北京大学工学院工业工程与管理系 北京 100871)

摘 要:数字技术的飞速发展推进了全球数字化进程。数据作为推动经济发展的重要资源,其价值评估尚未形成统一范式。随着数据规模增长,数据质量良莠不齐、数据失真也是亟待解决的问题。数据价值评估有利于评估数据质量、筛选失真数据、推进数据融通,其重要性不言而喻。本文基于在线医疗平台背景,利用沙普利值方法评估医患匹配场景下的数据价值,借助评估结果提高平台中的数据价值利用效率。首先,平台在事前利用医生历史问诊特征,使用XGBoost模型对不同医生的未来一年的问诊需求进行预测。进一步,根据医疗平台线上服务模式构建医院根据预测结果的运营收益函数,形成数据－模型－收益的价值链条。最后,使用沙普利值方法评估数据价值,验证价值评估方法的有效性,分析平台高价值数据特征。研究发现:沙普利值能根据任务更有目的性地捕获数据价值;平台依据沙普利值进行数据筛选能实现模型的降本增效,提升数据价值的释放效率;同时沙普利值也能捕捉平台中的虚假数据,改善数据质量。因此,对数据价值进行适当评估有助于平台的业务理解,提升数据价值释放效率。

关键词:数据价值;在线医疗平台;沙普利值
中图分类号:F272.3 **文献标识码**:A

1 引言

随着信息技术的发展,数据的存储空间和计算效率显著提高,推动人们经济生活日益数字化。数据和算法的相互交融,使得经济中的生产和管理效率大幅提升,同时催生了数字化平台等新型商业模式的兴起。为了适应这一趋势,医疗等传统行业积极地寻求数字化转型。2022年中共中央和国务院发布的《关于构建数据基础制度 更好发挥数据要素作用的意见》指出,充分实现数据要素价值、促进全体人民共享数字经济发展红利。但是,在数据要素发挥巨大作用的同时,数据的价值却暂时难以准确衡量,限制了数据要素活力的激发。

数据价值的评估宏观上有助于厘清数据资产价值,精准核算数字经济;微观上则有助于数据治理,提升数据使用效率。例如,随着数字化平台规模扩张,平台生态中出现了恶意刷单、诱导不真实评价等行为。这些行为造成了数据失真,影响数据的的价值发挥。数据价值的评估有助于有助于识别高/低质量的数据。根据质量进行筛选得到数据集可以更好地服务大数据算法和模型,进而积极地发掘数据要素潜在的价值属性。

本文将以在线医疗平台为背景,探讨互联网医疗中提升数据价值和利用效率的可能性。作为传统医疗数字化的产物,在线平台的引入有助于打破问诊的地域与时间的局限性,便于群众就医[1],在提升了医疗系统的运行效率的同时,也提高了医患双方的效用[2]。平台收集医生与患者的特征和问诊数据,并据此建立推荐匹配机制以提升在线问诊效率。换句话说,平台的良好运营是基于医生患者之间信息的对称性和透明性。任意一方引起的信息失真都会造成医患之间的信息不对称,进而增加平台的匹配误差并降低精准推荐的效率,最终导致在线问诊效率和体验上的下降[3]。所以,对于在线医疗平台而言,能够精准地评估数据价值十分重要,以避免失真数据对平台匹配决策带来的干扰。如何更高效地进行数据治理,尽可能地挖掘数据内部的价值,对于在线医疗平台,乃至所有需要数据运营的在线平台而言,都是一个亟需解决的问题。

数据价值的评估是目前的研究热点,总体有三种思路,即成本法、市场法和收入法[4,5]。成本法以数据收集成本为其价值;市场法参考数据市场中相

收稿日期:2022－03－21; **修订日期**:2023－01－22
基金项目:国家自然科学基金重点项目(72131001,71731006);国家自然科学基金应急管理项目(72241420)
通讯作者简介:宋洁(1982－),女(瑶族),湖南长沙人,北京大学工学院,党委书记,长江学者特聘教授,工业工程与管理系,博士生导师,研究方向:运筹优化、数据价值,E-mail:songjie@coe.pku.edu.cn.

似的定价模式对数据价值进行评估或者通过数据供需模型寻找数据均衡价格;收入法则通过加总全生命周期数据产生的贡献流评估价值[6-10]。

相比之下,数字平台更强调平台内数据的价值创造规律,图1展示了平台中数据创造价值的过程。其中,左端是拥有大量用户数据的互联网平台,右端是平台利用数据提供定制化服务的对象,中间则为平台的内部运转、决策过程。大量用户数据首先汇聚到平台中,平台整合数据并训练模型,利用模型提供的信息指导决策,最终服务用户。整体来看,收入法的思路比较适合数据平台的价值评估场景,但从收入流中拆分数据的贡献较为困难,因此本文选择从机器学习算法入手对平台中的数据价值进行研究。

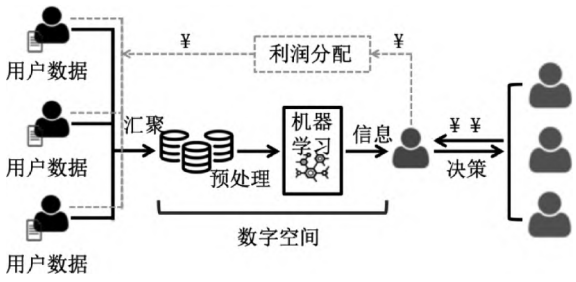

图1　平台中数据创造价值的过程

机器学习视角下的数据价值通过评估数据点或特征对模型的贡献来定义价值。例如value of information (VoI)起源于统计决策理论的研究[11-13],基于多臂老虎机(MAB)问题中以减少不确定性为目标,量化得到单一数据样本为模型带来的收益[14],已经被广泛应用于医疗保健领域,尤其在新药和新治疗方法的应用决策中[15-17]。除此之外,影响函数(influence function,IF)是一种来自稳健统计的经典技术,其利用机器学习优化问题中的梯度信息,近似计算出删减每一个数据点对当前模型效果的影响[18]。另一方面,机器学习与博弈论的有机结合也为数据价值的计算提供了可能性。shapley value (SV)是合作博弈论中经典的收益分配方法,针对收益分配场景,其满足现实世界中一系列公平性原则[19-20]。将每个数据贡献者看作联合博弈中的参与者,通过效用函数来表征任意数据子集的贡献。为了减少遍历计算带来的复杂度,一些学者使用蒙特卡洛模拟,利用随机抽样减少模型重复训练的截断蒙特卡洛沙普利值算法[21];一些学者利用非参数的K近邻思想,使用邻域内数据点的预测结果代替目标数据点的预测结果,规避模型的重复训练的KNN-Shapley算法[22]。

除此之外,一些学者将机器学习模型看作强化学习中的环境,利用沙普利值的"边际效用"思想,在强化学习的策略中确定数据的价值[23];一些学者将SV推广到针对一组数据分布中,开发出了Distributional-SV算法,使其应用的范围更广泛,并且可以针对数据点所处的分布进行价值分析[24]。由于SV计算复杂度较高,一些学者对其复杂度进行了研究[25],在相对简单的效用函数与学习模型中,可以通过模型的理论推导降低数值实验的计算开销,从而降低计算复杂度[26]。也有一些学者将SV算法应用于一些实际应用场景中,如联邦学习的利益分配[27]、持续学习中的经验回放[28]、数据市场的构建[29]等。

针对数据失真处理,目前文献将噪声数据失真一般分为两类:一类为特征噪声;另一类为属性噪声。特征噪声影响定量特征的观测值,属性噪声则改变一个实例的分类标签[30-32]。针对噪声标签的处理方法可以分为三类:第一类是使用鲁棒性更高的模型,使其不易受噪声影响,如集成学习中的LogitBoost[33]、BrownBoost[34]。第二类采用训练前事前过滤的方式删除噪声数据,如计算每条实例的熵过滤低于阈值的数据[35],使用神经网络建立标签预测模型[36],删除误分类数据。第三类是建立一个同时学习标签与噪声信息的模型,比如SVM模型通过损失函数容忍错分类[37],使用贝叶斯方法为错误标签数据建模[38]。

本文希望解决在线医疗平台中数据价值评估的相关问题,即能否在不影响算法效果的前提下提升数据的计算效率?能否基于真实数据的累积来实现平台可持续发展的生态建设?上述问题的解决也有助于平台实际运营中实现数据融通和平台优化治理。

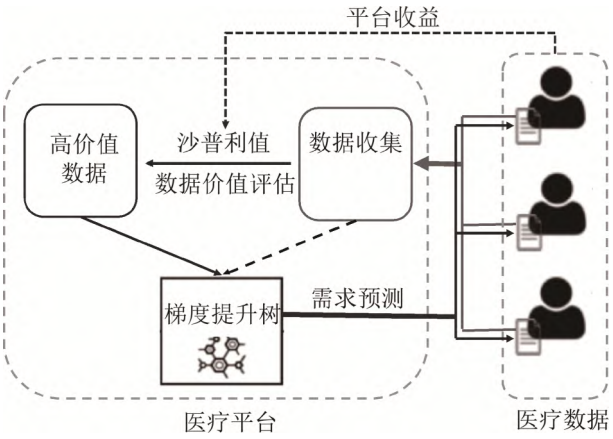

图2　研究框架

本文将在线医疗平台中的医患匹配与基于模型的数据价值评估分析相结合,在平台场景下使用沙普利值评估数据价值,利用数据价值评估进一步

提升平台数据的利用效率，最大化发挥数据要素在在线医疗平台的价值作用。如图 2 所示，本文基于在线医疗平台运营框架，建立从数据到收益的全链路分析框架，对平台使用的数据进行价值分析。本文引入沙普利值对这一场景下的数据进行测算，分析高价值数据特点，通过数据筛选优化平台数据使用效率。本文余下部分安排如下：第 2 部分将介绍平台背景、需求预测模型、数据价值评估的理论方法。第 3 部分基于某在线医疗平台数据，使用 XGBoost 模型建立平台医生需求预测模型，使用沙普利值进行数据价值评估，验证使用方法的有效性。第 4 部分对全文进行总结，并展望未来研究方向。

2 问题背景与模型假设

2.1 问题背景

在比较国内在线问诊平台的医生体量、信息完善程度等相关方面后，本文选择拥有超过 20 万注册医生的“好大夫在线”作为研究对象。“好大夫在线”是中国第一家创建互联网医生数据库、专家门诊以及网络分诊的互联网医疗平台，经过 17 年的功能优化以及用户沉淀，已经达到了中国的最大规模，其作为在线问诊研究对象具有代表性。

在线医疗平台整合全国医生资源，患者根据自身症状，在平台上快速检索医生，结合医生主页展示数据，选择心仪的医生进行问诊。如图 3 所示，左图为患者以“高血压”为关键字搜索，网页给出的推荐医生，网页在医生主页（右图）展示医生信息如职称、擅长疾病、“心意礼物”等评价数据，辅助患者进行决策。

图 3 好大夫在线呈现的信息

为优化医生诊疗资源配置，平衡医患供需，平台基于用户行为数据对各个医生的需求进行预测。另一方面，在线平台中存在参与者恶意刷单行为，部分数据失真，其质量较低。因此，数据对于平台并非多多益善，通过筛选高质量数据实现预测的降本增效，对平台是更优的选择。针对上述问题，本文将使用好大夫在线平台数据建立医生需求预测模型，进而使用数据沙普利值法对数据价值进行评估，分析高价值数据特点，验证该方法的可靠性，对平台数据管理提出建议。表 1 列出了本文所使用的变量。

表 1 本文变量

符号	意义
t	$t\in\{1,2,\cdots\}$为平台进行预测－决策的时间点
i	$i\in\{1,2,\cdots,n\}$代表平台的医生
D_{it}	医生 i 在第 t 期的实际需求患者数
$\hat{D}_{it}$	平台预测医生 i 在第 t 期的需求患者数
ΔD_{it}	平台预测的需求误差
S_{it}	平台预测医生 i 在第 t 期的需求时长
w_i	医生工作单位时间的薪资
τ	医生服务一位患者的平均时长
α	临时服务每位患者的调度费
X_t	截止到第 t 期平台收集的历史数据
r_i	医生 i 服务一位患者的平台收益分成
R_{it}	第 i 位医生在第 t 期为平台带来的收益
R_{it}^*	第 i 位医生在第 t 期为平台带来的最大收益
R_t	平台在第 t 期的总收益
V	沙普利价值计算使用的效用函数

2.2 需求预测模型

平台根据进行需求预测－决策的模型如下：设 $t\in\{1,2,\cdots\}$为平台进行预测－决策（资源配置）的决策时间点，即平台需要在此时根据预测结果决定下一期各医生需要安排的问诊时长。具体地，平台根据第 t 期的历史数据 X_{t-1} 决定下一个周期需要雇佣医生 i 工作的时长 S_{it}（平台供给）。而医生工作时长由平台第 t 期预测的患者需求 $\hat{D}_{it}$ 决定：

$$S_{it}=\tau E\left[D_{it}|X_{t-1}\right]=\tau\hat{D}_{it} \tag{1}$$

其中，D_{it} 为医生 i 在第 t 期的需求患者数，为随机变量；τ 为医生服务一位患者时长。在后文的算例中，本文将使用 XGBoost 预测平台未来需求。医生 i 每服务一位患者，平台可以从中获益 r_i，医生 i 单位时间薪资为 w_i。若供给（平台预测）小于需求，平台需要花费额外调度费用寻找医生，临时服务每位患者需要调度费用 α，则综合起来，第 t 期平台的总收益为：

$$R_t=\sum_{i=1}^{n}\left[r_i\cdot\min\left\{\frac{S_{it}}{\tau},D_{it}\right\}-w_iS_{it}-\alpha\left(D_{it}-\min\left\{\frac{S_{it}}{\tau},D_{it}\right\}\right)\right] \tag{2}$$

进一步，$\hat{D}_{it}=\dfrac{S_{it}}{\tau}$，$\Delta D_{it}=\hat{D}_{it}-D_{it}$，$R_{it}=r_i\cdot\min\left\{\dfrac{S_{it}}{\tau},D_{it}\right\}-w_iS_{it}-\alpha\left(D_{it}-\min\left\{\dfrac{S_{it}}{\tau},D_{it}\right\}\right)$，整理（2）式，可以得到：

若 $\dfrac{S_{it}}{\tau}=D_{it}$，即当平台完美预测时，则 $R_{it}=r_iD_{it}-w_iS_{it}=(r_i-w_i\tau)D_{it}$，此时为平台运营的

最优情况,令其为$R_{it}^{*}=(r_i-w_i\tau)D_{it}$。

若$\frac{S_{it}}{\tau}\geqslant D_{it}$,即需求预测较大时,有$R_{it}=r_iD_{it}-w_iS_{it}=R_{it}^{*}-w_i\tau(\Delta D_{it})$。

若$\frac{S_{it}}{\tau}<D_{it}$,即需求预测较小时,有$R_{it}=r_i\cdot\frac{S_{it}}{\tau}-w_iS_{it}-\alpha\left(D_{it}-\frac{S_{it}}{\tau}\right)=R_{it}^{*}-(-r_i+w_i\tau+\alpha)\Delta D_{it}$。

综上,在平台对需求的估计产生偏差时,需要平台处理供需间的错配。当平台的需求预测偏大时,会额外付给医生薪资,造成利润损失;当需求预测偏低,平台被迫进行临时资源调配,弥补需求缺口。假设资源的错配将对平台造成利润损失,即$\forall i, w_i\tau+\alpha<r_i, w_i>0$,平台需要事先做出更准确的预测。

平台的利润损失与预测的绝对误差ΔD_{it}相关,误差来源于平台数据的机器学习模型结果,由此构成了数据—模型—收益的价值链条。根据数据沙普利值,本文对此场景下的不同数据价值进行评估。

2.3　平台数据价值分析

本文使用沙普利值(Shapley value)方法对模型中的数据价值进行评估。沙普利值是劳埃德·沙普里(Lloyd Shapley)为了解决合作博弈中的收益公平分配所提出的[19],依据参与者个体的"边际贡献"计算合作收益。在机器学习场景中,可以将每个数据点z视作合作博弈的参与者,模型评价函数v为合作博弈的收益。计算各个数据点的沙普利值$\varphi_z(v)$可以评估数据的价值,具体公式如下:

$$\varphi_z(v)=\sum_{D\subseteq S}\frac{|D|!(n-|D|)!}{n!}\left[v(D\cup\{z\})-v(D)\right] \tag{3}$$

其中,S为全集,$n=|S|$,$i\in S$,v是效用函数,$\varphi_z(v)$为元素z在效用函数v下的贡献,也就是元素z的沙普利值,这种价值评估方式保留了沙普利值的完备性、公平性、可加性。

在本文的需求预测模型中,平台收益和预测误差高度相关,平台实际预测距离"最优预测"带来的收益差值$R_t^{*}-R_t$与预测的绝对误差$MAE\left(D,\hat{D}\right)$成正比。由此,本文将预测的平均绝对误差($MAE$)的相反数定义为效用函数,进行数据沙普利值的计算,如下所示:

$$V(X_{t-1})=-MAE\left(D,\hat{D}\right)=-\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[D_{it}|X_{t-1}\right]-D_{it}\right| \tag{4}$$

由于沙普利值的计算复杂度较高,为$O(n!)$,n为参与收益分配的个体数,本文中即为医生数量。而每次计算数据i的边际贡献$v(D)-v(D/\{i\})$都需重新训练模型,因此计算精确沙普利值耗费过大。对此,本文主要采取了两种算法对沙普利值进行近似计算:第一种是截断蒙特卡洛沙普利值法(Truncated Monte Carlo Shapley,TMC—Shapley)[21],该算法利用随机抽样估算每个数据点的边际贡献,并设定抽样停止条件,大幅减少模型重复训练的次数。第二种是K近邻沙普利值算法(KNN Shapley Value)[22],该算法利用了机器学习模型损失函数的连续性,利用K—近邻非参数模型的预测思想对模型预测效果进行非参数估计,避免重新训练完整模型。

3　算例分析

3.1　数据概述

本文使用的好大夫在线平台数据可以分为以下两类:第一类是医生注册信息,包括医生的性别、职称、所在医院级别,反映医生个人特征;第二类是医生平台行为信息,为基于医生平台行为特征,包括"总单量""文章数""诊后患者报到数""患者投票""感谢信""心意礼物""一般等待时长""综合推荐热度"。表2中详细介绍了数据的实际意义。

表2　医生信息变量介绍

变量名称	变量类型	变量介绍
性别	定性变量	医生性别
医院级别	定性变量	所在医院级别
文章数	定量变量	医生在平台上传的有关疾病治疗的文章数量
诊后患者报到数	定量变量	为患者在线上问诊结束后,转到线下进一步问诊的患者数量
患者投票	定量变量	患者根据自己在问诊过程中服务的满意程度,对医生进行网上投票数量
感谢信	定量变量	患者根据自己在问诊过程中服务的满意程度,为医生写的感谢信数量
心意礼物	定量变量	患者根据自己在问诊过程中服务的满意程度,为医生赠送的虚拟礼物数量
一般等待时长	定量变量	医生对于患者问诊需求的响应速度
综合推荐热度	定量变量	平台通过医生的背景信息、患者投票和患者报道综合计算的推荐热度
总单量	定量变量	根据订单详细信息统计,2019—2020年每位医生平台总订单量。模型的预测目标。

如图4所示,对于本文的定量变量,通过散点图矩阵展示其分布与相关性情况,对角线为单一数据的分布情况,上三角部分为两两数据的相关系数,下三角部分为两两数据的散点图分布情况。从分布上看,除综合推荐热度外,其他数据具备长尾分布特征。整体来看,定量数据因其反映医生平台行为与服务能力,彼此均呈现较为明显的正相关关系,而与其余变量的正相关性不显著。

3.2 数据价值评估

本部分对好大夫在线平台的医生数据进行价值评估，主要目标为对不同数据点进行价值分析。数值实验将分为四部分进行：首先，本文使用XGBoost模型使用2018－2019年数据建立医生单量预测模型，即预测公式(1)中的$\hat{D}_{it}$，在A部分给出预测精度与平台收益的关系；进一步，利用沙普利值测算平台中数据点的价值：在B部分将数据集分为大小不同的子集，以此探究数据集大小(包含数据点个数)对其价值的影响；在C部分对数据集中的数据点分别价值评估；最后，在D部分通过去除低价值数据与利用噪声数据识别实验验证数据价值评估方法的有效性。

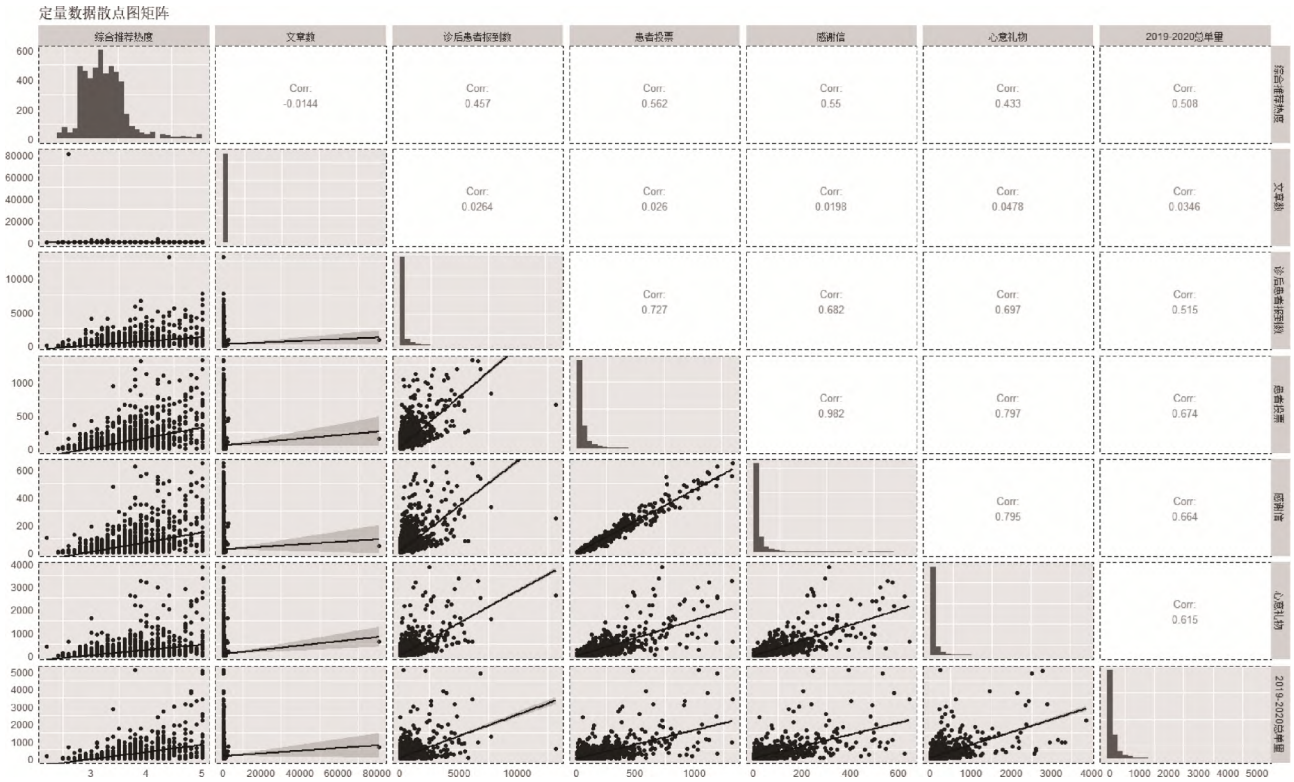

图4 不同医生信息变量之间的散点图矩阵

A. 预测精度与平台收益

本部分根据2019年初的医生行为信息、静态信息(表2除去总单量的其余变量)对2019－2020年医生总单量使用XGBoost进行预测。本文将25000条问诊数据按照8:2的比例划分为训练集和测试集，将训练集等分为50个大小为400的集合，分别训练并计算测试集的预测误差与收益，图5、图6展示了部分训练集的预测误差(MAE)与平台总收益(R_t)的关系。

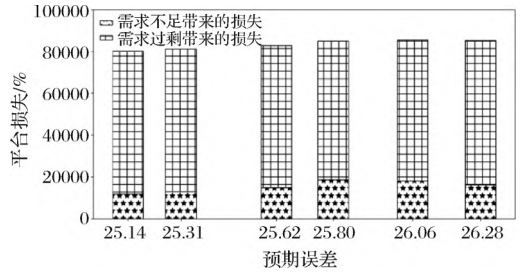

图5 平台额外成本与预测误差

图5中，随着预测误差(MAE)的增大，平台运营成本增大，其中分为两部分：一部分为平台因需求不足所付出的额外雇佣成本；另一部分为平台需求过剩所付出的处理成本。而在图6中，随着预测误差增大，平台总收益也随之减小，是平台数据的价值体现。

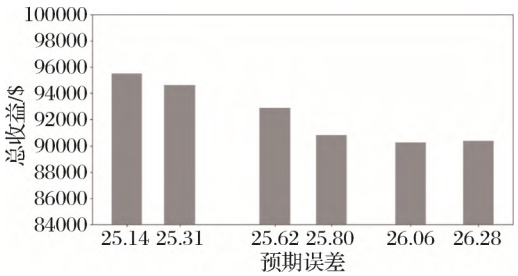

图6 平台收益与预测误差

B. 数据集大小与数据价值

在线医疗平台通过不断积累增加数据量，形成平台数据资产，数据资产的价值通常难以衡量，本部分探索数据集大小与其价值变化的关系。为此，本文将训练集划分为100～900大小不同的数据集，使用TMC沙普利算法计算不同数据集的价值，区分不同规模数据集的价值。与此同时，本文使用数

据集特征的统计量如信息熵(使用 Kozachenko—Leonenko 估计[39])、均值(Mean)、标准差(Std)、变异系数(Cv)等常用统计量辅助对比。如图 7 所示,不同折线图展示了数据集大小与数据价值或统计量变化的趋势,选取“感谢信”特征的统计量进行计算。

从图 7 中可以看出,随着数据集的增大,熵逐渐下降,沙普利值逐渐上升,其余变量逐步趋于稳定。理论上,熵是随机变量的函数,数值由概率分布直接决定,数据集大小将影响数值方法对信息熵的估算。本文采用 Kozachenko—Leonenko 方法对本文所使用连续数据的信息熵进行近似计算[39],数据量的增加使数据分布逐渐逼近理论分布,因此熵呈现出了逐步收敛下降的趋势,而其余统计量则遵循大数定律,逐步稳定收敛。与之相对比,沙普利值对数据规模逐渐增加的价值捕获更为直接。

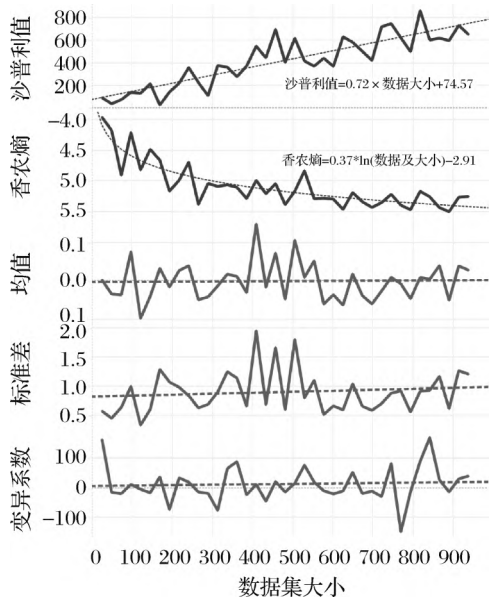

图 7　数据价值与数据集大小比较

进一步,本文比较不同大小数据集的预测精度与沙普利值的关系。如图 8 所示,对平台而言,预测绝对误差(MAE,实线)体现平台效用,随着数据集的增大,预测误差逐步减小,其沙普利值(虚线)逐步增大——即数据的价值增大,且在数据集较小时,呈现正线性相关,由此可见,沙普利值对于数据价值增加的捕捉是可靠的。

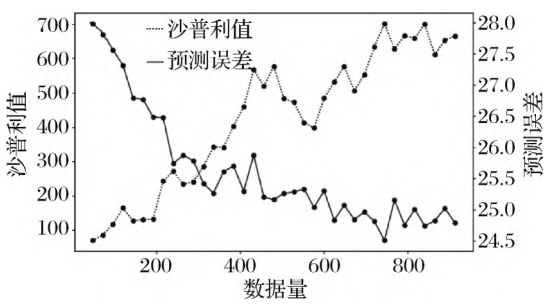

图 8　沙普利值与预测绝对误差(MAE)趋势图

C. 平台数据价值特点分析

从前两个部分的分析可以看出,沙普利值能够利用平台收益辅助数据价值评估,通过其可以建立数据—价值分析链条。本部分将从模型中拆分出每个数据点对平台收益的影响,找出高价值数据的特征,对平台的数据管理提供建议。

根据式(4)的效用函数,本文计算出训练集中每个数据点的沙普利值,并使用 2.3 中的截断蒙特卡洛沙普利值法(下简称 TMC 沙普利值法)与 K 近邻沙普利值法(下简称 KNN 沙普利值法)进行计算。由于医生数据中含有定性变量(性别、医院级别),导致 KNN 沙普利值法计算数据点间距离时误差较大,算法表现较差。因此,本文使用 TMC 沙普利值作为分析的标准。如表 3 所示,本文将数据点的沙普利值从低到高排序,并按照 20% 的分位数划分为五组(SV 分位数),0%～20% 代表沙普利值最低的一组,80%～100% 为最高的一组。表 3 展示了每组内的样本特征均值,以反映不同组数据的差异情况。

观察价值最高与最低的两组数据,与其他组数据相比,价值最高的一组(80%～100%)数据具备礼物、感谢信较多且总单量较高的特点,即符合在图 3 中展示的变量间的高相关性,这一部分数据对模型贡献度较高。但价值最低的数据(0%～20%)中,礼物、感谢信等数据较少,单量偏高,与数据分布中礼物数据与单量的正相关性相反,可能存在数据失真,对模型的贡献较低。

表 3　沙普利值均值与数据分布

SV 分位数	心意礼物	感谢信	综合推荐热度	诊后患者报到	文章数	总单量
0% ~ 20%	−0.036	−0.066	−0.052	−0.037	0.067	34.66
20% ~ 40%	−0.040	−0.041	−0.057	−0.046	−0.022	26.03
40% ~ 60%	−0.058	−0.041	−0.049	−0.059	−0.019	25.62
60% ~ 80%	0.093	0.052	0.024	0.084	−0.014	37.13
80% ~ 100%	0.040	0.095	0.134	0.058	−0.011	48.52

整体来看五组数据,20%～40%、40%～60% 两组数据均值较为接近,总体单量偏低,60%～100% 的数据为单量较多医生的数据,观察到这部分数据单量与感谢信等指标显著的正相关性,机器学习算法能从这部分数据中提取更准确的规律,价值较高。反映于实际业务中,平台能根据规律甄别数据,规避刷单数据的干扰。

在上述算例中,价值较高的数据能成功反映数据分布中感谢信、礼物等定量变量与单量的正相关关系,平台一方面能够从数据分布特征中提取高价值数据特点,建立平台内部事前数据价值评估体系;另一方面能根据评估结果留存高价值数据,识别平台中数据潜在的失真,提升数据使用效率。

D. 低价值数据筛选与噪声数据识别

在线平台中的恶意刷单、刷点评是普遍存在的问题。在好大夫在线平台，医生在平台积累的荣誉如文章数、心意礼物等是吸引患者的重要因素，存在刻意刷高数据吸引患者的可能，从而导致数据失真、模型失效，降低平台的运营效率。从数据对模型贡献的角度而言，恶意刷单行为对模型造成负面影响，沙普利值较低甚至为负。根据这一点，平台删除低价值的数据点，尤其是负价值数据，可以提高平台使用数据的质量，改进模型效果。

在本部分的实验中，本文对删除低价值数据的模型效果进行验证：将训练集中的数据点按沙普利值的大小从低到高删除，使用部分数据集重新训练模型，观察模型效果（MAE，预测与实际的绝对误差）的变化，并与随机顺序删去数据点进行对比。实验结果如图 9 所示，本实验用 TMC 沙普利值法与 K 近邻沙普利值法两种方法计算数据价值，其中纵轴为 MAE，横轴为删去数据点的比例。

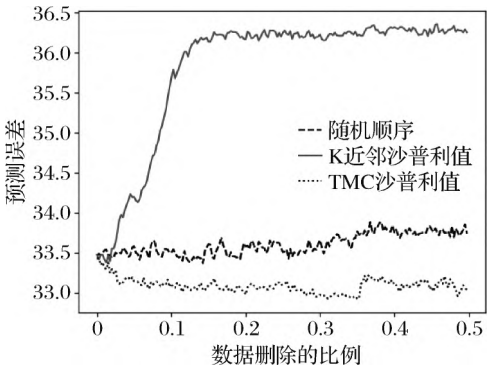

图 9 按价值顺序删去数据的模型表现变化

三种顺序中，TMC 顺序的曲线呈现先下降后上升的趋势，体现 TMC 沙普利值有效捕捉低价值数据点，模型因“去噪”而受益，即模型效果先因数据价值提升而提高，后随正价值数据被删除而降低。同为沙普利值近似计算的 KNN 沙普利方法由于定性变量，导致数据点之间距离计算出现误差，影响沙普利估算，因此算法的表现不佳，整体曲线高于随机顺序去除数据点。

进一步，本文在数据集中选择部分数据进行扰动，以模拟平台医生数据的失真情况，进一步使用沙普利值进行噪声识别。具体地，在训练集中对预测目标“总单量”施加扰动 ϵ。其中，ϵ 服从集合 E 上的离散均匀分布，$E=\{-20,-19,\cdots,-10,10,11,\cdots,20\}$。由于单量数据的均值为 33.6，方差为 184.7，因此，一方面，设定扰动下界为 10，使噪声足够大，模拟数据失真；另一方面，控制数据的扰动范围，故设定扰动上限为 20。

本文计算了扰动后数据的沙普利值，将其依照从低到高的顺序从数据集中移除，观察被移除数据中噪声数据的比例。由图 10 所示，与随机顺序相比，依据低价值沙普利值移除数据的噪声比例是更高的，即平台能够利用沙普利值找出更多噪声数据，提升平台数据的质量，这也从侧面说明了使用沙普利值评估数据价值的有效性。

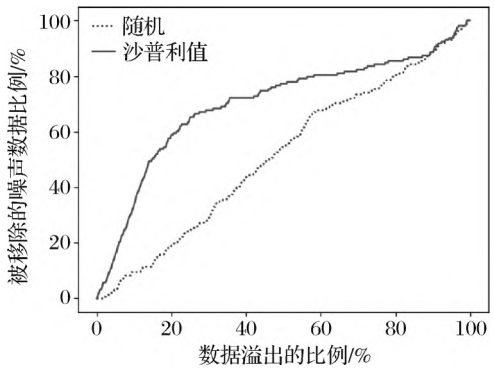

图 10 利用沙普利值进行噪声数据识别

4 结语

数据对数字经济的作用毋庸置疑，平台是数据最大的拥有者，测算平台中的数据价值有助于平台提升数据的使用效率。本文基于“好大夫在线”的问诊数据，使用 XGBoost 模型对平台医生的需求进行预测，建立数据－模型－收益的价值链条；在使用数据沙普利法对平台数据集价值进行评估时，本文发现沙普利值能够捕捉数据量增长带来的价值提升，并且呈线性正相关关系。而对于单一数据点价值，本文发现低价值数据点往往存在礼物与单量的解耦现象，此类数据点可能含有潜在的刷单行为。最后，利用噪声识别实验发现，沙普利值对于噪声识别是比较有效的，并且从全体数据集中剔除低价值数据有助于提升平台的收益，这为医疗平台的数据管理提供了精简数据集的思路，也验证了沙普利值方法识别数据价值的有效性。

随着数量的逐渐增加，数据的价值也会随之增长，数据带来的价值愈发被重视，宏观上促进了数字平台的发展。从本文的算例可以看出，数据积累对于在线医疗平台至关重要，同理于其他数字化平台。对于在线平台而言，如何有效管理数据至关重要。在本文的数值实验中，沙普利值较低的数据点降低了模型效果。这说明，虽然大样本下，数据多多益善，但数据集中也存在对平台有害无益的样本。低价值数据往往与破坏平台生态的行为关联，适当删除一些低价值的数据点可以使平台数据更准确地反映现实情况，降低信息不对称。

另一方面，高价值的医生数据的代表性特征能帮助平台制定更可靠的数据驱动学习策略。具体

到平台的数据管理,沙普利的思想为平台提供了数据治理的新思路:平台可以对数据价值进行多时刻、多场景评价,筛选出高价值数据,减少使用全量数据的计算成本。

本文的研究仍存在一些不足之处:首先,数据沙普利值的计算复杂度较高,本文使用的TMC算法仍然不太适用于大规模数据集,计算复杂度更低的数值算法是沙普利值相关研究的重要方向;其次,本文的评估方法需要先计算模型效果,需要评估方拥有完整数据,而针对数据交易等需要事前价值评估的场景契合度较低,需要将内部数据的价值评估方法函数化,以应对更多情形;最后,本文所提出的数据价值评估方法距离落地仍有一定距离,需要提出适应平台数据实时更新特点的评估－预测框架,以形成稳定的数据治理模式。

参考文献:

[1] 郭薇,薛澜. 互联网医疗的现实定位与未来发展[J]. 探索,2016(6):142－148.
Guo W, Xue L. The realistic positioning and future development of internet healthcare[J]. Probe, 2016(6): 142－148.

[2] 陈惠芳, 徐卫国. 价值共创视角下互联网医疗服务模式研究[J]. 现代管理科学, 2016(3): 30－32.
Chen H F, Xu W G. Research on internet medical service model from the perspective of value co－creation [J]. Modern Management Science, 2016(3): 30－32.

[3] Sillence E, Briggs P, Harris P, et al. A framework for understanding trust factors in web－based health advice [J]. International Journal of Human－Computer Studies, 2006, 64(8): 697－713.

[4] Slotin J. What do we know about the value of data[EB/OL]. (2018－05－03)[2018－05－03]. https://www.data4sdgs. org/blog/what－do－we－know－about－value－data.

[5] EC, IMF, OECD, UN and WB. Systems of national accounts 2008[M]. New York: United Nations, 2009.

[6] Akred J, Samani A. Your data is worth more than you think [EB/OL]. (2018－12－17) [2018－12－17]. https://sloanreview. mit. edu/article/your－data－is－worth－more－than－you－think/.

[7] Li W, Nirei M, Yamana K. Value of data: There's no such thing as a free lunch in the digital economy [R]. Working Papers, Bureau of Economic Analysis, 2018.

[8] Mawer C. Valuing data is hard[EB/OL]. (2015－11－10)[2015－11－10]. https://www. svds. com/valuing－data－is－hard/.

[9] Elia G, Polimeno G, Solazzo G, et al. A multi－dimension framework for value creation through big data[J]. Industrial Marketing Management, 2020, 90: 617－632.

[10] 许宪春, 张美慧. 中国数字经济规模测算研究——基于国际比较的视角[J]. 中国工业经济,2020(5): 23－41.
Xu X C, Zhang M H. Research on the scale measurement of China's digital economy: Based on the perspective of international comparison [J]. China Industrial Economics, 2020(5): 23－41.

[11] Pratt J W, Raiffa H, Schlaifer R. Introduction to statistical decision theory[M]. Cambridge, Massachusetts: MIT Press, 1995.

[12] Anscombe F J . Probability and statistics for business decisions: An introduction to managerial economics under uncertainty[J]. Publications of the American Statistical Association, 1959, 54(288):813－814.

[13] Raiffa H, Schlaifer R. Applied statistical decision theory [M]. USA:Wiley, 1961.

[14] Wilson E C F. A practical guide to value of information analysis[J]. Pharmacoeconomics, 2015,33(2): 105－121.

[15] Claxton K. The irrelevance of inference: A decision－making approach to the stochastic evaluation of health care technologies [J]. Journal of Health Economics, 1999, 18(3): 341－364.

[16] Eckermann S, Willan A R. Expected value of information and decision making in HTA[J]. Health Economics, 2007, 16(2): 195－209.

[17] Willan A R . Statistical analysis of cost－effectiveness data[J]. Expert Review of Pharmacoeconomics & Outcomes Research, 2006, 6(3):337－346.

[18] Koh P W, Liang P. Understanding black－box predictions via influence functions [C]//Proceedings of the 34th International Conference on Machine Learning－Volume 70, Sydney, Australia, August 06－11, ML Research Press, 2017: 1885－1894.

[19] Shapley L S. A value for n－person games[M/OL]. Princeton: Princeton University Press, 1953: 307－318. https://doi. org/10. 1515/9781400881970－018.

[20] Gul F. Bargaining foundations of Shapley value [J]. Econometrica: Journal of the Econometric Society, 1989,57(1):81－95.

[21] Ghorbani A, Zou J. Data shapley: Equitable valuation of data for machine learning[C]//Proceedings of International Conference on Machine Learning. PMLR, Long Beach, California, USA, June 9－15, ML Research Press, 2019: 2242－2251.

[22] Jia R X, Wu F, Sun X H, et al. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, June 20－25, IEEE, 2021: 8239－8247.

[23] Yoon J, Arik S, Pfister T. Data valuation using reinforcement learning [C]//Proceedings of International Conference on Machine Learning. PMLR, Virtual, July 13－18, ML Research Press, 2020: 10842－10851.

[24] Ghorbani A, Kim M, Zou J. A distributional framework for data valuation [C]//Proceesdings of International Conference on Machine Learning. PMLR, Virtual, July 13－18, ML Research Press, 2020: 3535－3544.

[25] Jia R, Dao D, Wang B, et al. Towards efficient data valuation based on the shapley value [C]//Proceedings of The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, Naha, Okinawa, Japan, April 16－18, ML Research Press, 2019: 1167－1176.

[26] Kwon Y, Rivas M A, Zou J. Efficient computation and analysis of distributional shapley values [C]//Proceedings of International Conference on Artificial Intelligence and Statistics. PMLR, Virtual, April 13－15, ML

Research Press, 2021: 793－801.
[27] Wei S, Tong Y, Zhou Z, et al. Efficient and fair data valuation for horizontal federated learning[J]. Federated Learning: Privacy and Incentive, 2020: 139－152.
[28] Shim D, Mai Z, Jeong J, et al. Online class－incremental continual learning with adversarial shapley value [C]//Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, February2－9, ML Research Press, 2021, 35(11): 9630－9638.
[29] Agarwal A, Dahleh M, Sarkar T. A marketplace for data: An algorithmic solution[C]//Proceedings of the 2019 ACM Conference on Economics and Computation, Phoenix, AZ, USA, June 24－28, Association for Computing Machinery, 2019: 701－726.
[30] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986(1): 81－106.
[31] Zhu X Q, Wu X D. Class noise vs. attribute noise: A quantitative study[J]. The Artificial Intelligence Review, 2004, 22(3): 177－210.
[32] Wu X D. Knowledge acquisition from databases[M]. Norwood, NJ, USA: Ablex Publishing Corp, 1996.
[33] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The Annals of statistics, 2000, 28(2): 337－407.
[34] Freund Y. An adaptive version of the boost by majority algorithm[C]//Proceedings of the Twelfth Annual Conference on Computational learning theory, Santa Cruz, California, USA, July7－9, Association for Computing Machinery, 1999: 102－113.
[35] Sun J, Zhao F, Wang C, et al. Identifying and correcting mislabeled training instances[C]//Proceedings of Future Generation Communication and Networking (FGCN 2007), Jeju Island, Korea, December 6－8, IEEE, 2007, 1: 244－250.
[36] Jeatrakul P, Wong K W, Fung C C. Data cleaning for classification using misclassification analysis[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2010, 14(3): 297－302.
[37] Manwani N, Sastry P S. Noise tolerance under risk minimization[J]. IEEE Transactions on Cybernetics, 2013, 43(3): 1146－1151.
[38] Gaba A, Winkler R L. Implications of errors in survey data: A Bayesian model[J]. Management Science, 1992, 38(7): 913－925.
[39] Kozachenko L F, Leonenko N N. Sample estimate of the entropy of a random vector[J]. Problemy Peredachi Informatsii, 1987, 23(2): 9－16.

Research on Data Value Based on Demand Forecast of Online Medical Platform

Zhao Yue, Wang Yanzhi, Song Jie, He Xuan

(Department of Industrial Engineering and Management, Peking University, Beijing 100871, China)

Abstract: The rapid development of digital technology has advanced global digitalization. As a vital resource for driving economic growth, the value assessment of data has not yet formed a unified paradigm. With the expansion of data scale, the data quality disparity and data distortion urgently need to be addressed. The valuation of data is considered beneficial for evaluating data quality, filtering out distorted data, and promoting data trade, making its importance self-evident. In this paper, an online medical platform is used as the context, and the Shapley value method is employed to evaluate the value of data in doctor-patient matching. The efficiency of data value on the platform is intended to be improved through the data valuation results. First, the demand of doctors for the next year is predicted through the historical characteristics of doctors and the XGBoost model. Second, an operational profit function of hospitals based on the prediction results is constructed according to the online service mode of the medical platform, forming a value chain of data, model, and benefit. Finally, the Shapley value is used to assess data value, the effectiveness of the value assessment method is validated, and the characteristics of high-value data on the platform are analyzed. It is found that data value can be captured according to the task by the Shapley value. Cost reduction and efficiency increase of the model can be realized by filtering data based on the Shapley value, then enhancing the efficiency of application of data. Additionally, false data on the platform can be find, improving data quality. Therefore, appropriate valuation of data is seen as beneficial for better business understanding of the platform and enhancement of the efficiency of data.

Key words: data valuation; online medical platform; Shapley value