

基于云模型的企业数据资产质量评价模型

尤建新, 程敏倩, 徐涛

(同济大学 经济与管理学院, 上海 200092)

摘要: 结合云模型与最优最劣法(best worst method, BWM)、逼近理想解距离法(TOPSIS)提出一个综合考虑定性与定量指标的数据资产质量评价模型。首先,构建数据资产的质量评价指标体系,并基于云模型的黄金分割法将单个自然语言评价等级值转化为云模型;其次,分别对定性和定量指标计算评价群决策值,并结合基于BWM获取的指标权重计算得到云-TOPSIS决策矩阵;再次,基于TOPSIS对评价方案进行质量排序;最后,将所构建模型应用于阿里云公开数据集的质量评价,验证模型的可行性与有效性。

关键词: 数据资产;质量评价;云模型;逼近理想解距离法(TOPSIS);综合评价模型

中图分类号: C93

文献标志码: A

Enterprise Data Asset Quality Evaluation Model Based on Cloud Model

YOU Jianxin, CHENG Minqian, XU Tao

(School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: A data asset quality evaluation model considering qualitative and quantitative indexes comprehensively was proposed by combining the cloud model with the best worst method (BWM) and technique for order preference by similarity to an ideal solution (TOPSIS). Firstly, the quality evaluation index system of data assets was constructed, and the individual natural language evaluation value was transformed into the cloud model based on the golden section method of cloud model. Secondly, the evaluation group decision value was calculated for qualitative and quantitative indexes respectively, and the cloud-TOPSIS decision matrix was obtained by combining the index weights from BWM. Thirdly, the evaluation schemes were ranked qualitatively

based on TOPSIS. Finally, the constructed model was applied to the quality evaluation of Alibaba Cloud's public datasets for the verification of the feasibility and effectiveness of the model.

Keywords: data assets; quality evaluation; cloud model; technique for order preference by similarity to an ideal solution (TOPSIS); comprehensive evaluation models

随着数字经济的快速发展,数据已经成为企业的重要资源和关键资产,在决策制定、产品创新、风险管理等多个方面都起到了重要作用^[1-2]。在数据价值实现过程中,数据质量至关重要,直接影响到数据的有效性、可信度和利用价值。低质量的数据不仅可能导致决策失误、资源浪费,甚至还可能造成巨大的经济损失^[3]。如何有效评估和提升数据资产质量水平,成为当前企业数据价值实现的关键问题^[4-5]。

国内外学者对数据质量相关问题展开了一定的研究^[6-7]。从现有文献来看,数据质量的评价主要包括定量和定性2个角度。定性研究方法通过用户调查、专家访谈等手段,收集用户对数据的满意度进而评价数据的质量。Feng等^[8]利用用户满意度评价方法,调查电子商务网站数据质量对用户购买决策的影响,从而识别并改进数据质量问题。尤建新等^[4]引入多准则决策方法,结合专家评价,对某商业银行的数据资产质量开展评价。定量研究中,Gudivada等^[9]提出一种基于错误率的数据质量度量方法,通过对数据的错误率进行计算和分析,评价数据的质量状况。Mandal^[10]则结合统计分析方法对数据质量进行量化评价,揭示数据的分布、相关性和异常情况,从而识别和改进数据质量问题。

收稿日期: 2024-11-01

基金项目: 教育部人文社会科学研究青年基金(25YJC630158);上海市科技创新行动软科学研究项目(24692111300)

第一作者: 尤建新,教授,博士生导师,管理学博士,主要研究方向为管理理论与工业工程。

E-mail: yjx2256@vip.sina.com

通信作者: 徐涛,管理学博士,主要研究方向为数据质量管理。E-mail: xutao0709@yeah.net



论文
拓展
介绍

从现有研究来看,定性分析主要依赖专家经验和知识,存在一定的主观性;定量分析具有客观性,但难以反映复杂多变的实际情况。因此,仅依靠单一的定性或定量方法来评估数据资产质量,难以全面反映数据的实际质量水平。综上所述,本文拟引入云模型理论^[11-12],旨在构建一种能够同时考虑定性

称为云,每一个 x 称为一个云滴。正态云模型是云中最基本的一种类型^[13],有3个重要的数字特征,用来共同表征定性概念 C 的数学性质:①期望 E_x 表示云滴在定量空间中的中心值或位置,是反映定性概念 C 的典型代表;②熵 E_N 表示概念的不确定性程度,反映随机性和模糊性的综合程度,值越大表明系统越不确定;③超熵 H_e 表示熵的波动程度,也称为二次不确定性,反映系统不确定性的稳定性。

性质1 设论域 U 中存在 i 个云模型 $C_i(E_{x_i}, E_{N_i}, H_{e_i}) (i=1, 2, \dots, n)$,可以得到其叠加合成的综合云 $C(E_x, E_N, H_e)$:

$$E_x = \sum_{i=1}^n \omega_i E_{x_i} \quad (2)$$

$$E_N = \sqrt{\sum_{i=1}^n (\omega_i E_{N_i})^2} \quad (3)$$

$$H_e = \sqrt{\sum_{i=1}^n (\omega_i H_{e_i})^2} \quad (4)$$

式中, ω_i 为第 i 个云模型 C_i 在综合云模型 C 中所占的权重。

性质2 本文基于徐士东等^[12]提出的一种改进的云距离测度方法来计算2朵云之间的距离和相似度,具体算法见图1。

1 理论基础

假设 $U=\{x\}$ 为一个用精确值表示的定量论域, C 为 U 的一个定性概念,集合 U 中的元素 x 可以理解为定性概念 C 的一次随机实现。 x 对 C 存在一个隶属度 $\mu(x)$, $\mu(x)$ 是取值为 $[0, 1]$ 的具有稳定倾向的随机数,云可以理解为论域 U 到区间 $[0, 1]$ 上的一个映射,即:

$$\begin{aligned} \mu: U &\rightarrow [0, 1] \\ \forall x \in U, x &\mapsto \mu(x) \end{aligned} \quad (1)$$

此时, x 在论域 U 上的分布称为云模型^[11],简

Algorithm:	
Input	云模型 $C_1(E_{x_1}, E_{N_1}, H_{e_1})$ 和 $C_2(E_{x_2}, E_{N_2}, H_{e_2})$
Process	需要生成的云滴数量 n
(1)	用正向云发生器对2朵云各生成 n 个云滴,这 $2n$ 个云滴在云图中的位置分别记为 $(x_{1i}, \mu(x_{1i}))$ 、 $(x_{2i}, \mu(x_{2i})) (i=1, 2, \dots, n)$;
(2)	分别对2组云滴按照 x_{1i} 和 x_{2i} 的大小进行升序排列;
(3)	对2朵云分别筛选出云滴位置在 $E_{x_i} \pm 3E_{N_i} (i=1, 2)$ 范围内的 n_1 和 n_2 个云滴,生成2朵新的云 C'_1 和 C'_2 ;
(4)	If $n_1 \leq n_2$: 对第2朵云 C'_2 随机保留 n_1 个云滴,其中共有 $C_{n_2}^{n_1}$ 种组合,生成一朵新云 $C'_{2i} (i=1, 2, \dots, C_{n_2}^{n_1})$,对这 n_1 个云滴进行横坐标升序排列,分别将云 C'_1 和新云 C'_{2i} 的各 n_1 个云滴保存在集合 $drop_1$ 和 $drop_2$ 中; Else: 对第1朵云随机保留 n_2 个云滴,其中共有 $C_{n_1}^{n_2}$ 种组合,生成1朵新的云 $C'_{1i} (i=1, 2, \dots, C_{n_1}^{n_2})$,对这 n_2 个云滴进行横坐标升序排列,分别将云 C'_2 和新云 C'_{1i} 的各 n_2 个云滴保存在集合 $drop_1$ 和 $drop_2$ 中;
(5)	End if
(6)	计算云 C_1 和 C_2 之间的距离 $d(C_1, C_2)$
Output	2朵云之间的距离

图1 云距离的计算方法

Fig.1 Method for cloud distance calculation

图1中,

$$d(C_1, C_2) = d(drop_1, drop_2) = \frac{\sum_{j=1}^{n_1} \sqrt{(x_{1j} - x_{2j})^2 + (\mu(x_{1j}) - \mu(x_{2j}))^2}}{n_1}, n_1 \leq n_2 \quad (5)$$

或

$$d(C_1, C_2) = d(drop_1, drop_2) = \frac{\sum_{j=1}^{n_2} \sqrt{(x_{1j} - x_{2j})^2 + (\mu(x_{1j}) - \mu(x_{2j}))^2}}{n_2}, n_1 > n_2 \quad (6)$$

2 模型构建

本文引入云模型理论,结合BWM和TOPSIS构建数据资产的质量评价模型。云模型能够处理模糊性和随机性的不确定性特征,在复杂的评价问题中具有较好的适应性。TOPSIS通过计算各对象与正理想解和负理想解之间的距离评估对象的相对优劣,适用于对多个评价指标同时进行分析的复杂场景^[14]。通过将定性评价中的模糊信息转化为量化数据,云模型能够在保持评价信息完整性的同时对多个待评价的数据资产进行排序。BWM通过对最优和最劣指标的直接比较,减少了评估过程中的主观偏差,使得权重分配更加合理^[15]。

2.1 指标体系构建

指标体系的构建是数据资产质量评价的基础。针对数据质量的评估,国内外学者主要从多个维度展开探讨,包括准确性、完整性、规范性、一致性、时效性和可访问性等方面^[16-17]。这些维度为衡量数据质量提供了科学依据。此外,考虑到数据作为一种新型资产的独特属性,指标体系的构建应涵盖更广泛的评价维度,本文进一步引入了收益性、风险性和成本性等指标^[4],以全面评估数据的经济价值和潜在风险。

根据中国信息技术标准化技术委员会发布的《信息技术:数据质量评价指标》规范,数据质量的评价可以在准确性、完整性、规范性、一致性、时效性和可访问性等基础指标上进行进一步扩展和量化。为确保评价的科学性与精确性,将上述质量维度确立为定量指标,通过明确的数值评估数据质量。鉴于数据作为资产的价值评估涉及复杂计算与多维度考量,将收益性、风险性和成本性设定为定性指标。所构建的最终指标体系见表1。

2.2 单个评价语言值与对应云模型的转换

假设存在 M 个待评价的数据资产、 N 个评价指标,其中 p 个定性评价指标, q 个定量评价指标,共 l 个对指标进行评价的专家。设各方案的语言值评价等级数为 a (a 一般约定为奇数,取值在3~7之间),论域 $U = \{x|x \in [X_{\min}, X_{\max}]\}$ 由决策专家指定。参照改进的黄金分割法生成对应评价等级云模型^[18],其遵循的基本原理是在传统黄金分割法的基础上,将中间云 C_0 和最后一朵半升(降)云的期望值作为

表1 数据资产质量评价的指标体系构建

Tab.1 Index system for data asset quality evaluation

一级指标	二级指标
准确性 c_1	数据内容正确性 c_{11}
	数据格式合规性 c_{12}
	数据重复率 c_{13}
	数据唯一性 c_{14}
完整性 c_2	数据元素完整性 c_{21}
	数据记录完整性 c_{22}
规范性 c_3	数据标准规范性 c_{31}
	数据模型规范性 c_{32}
	元数据规范性 c_{33}
	业务规则规范性 c_{34}
	安全规范 c_{35}
一致性 c_4	相同数据一致性 c_{41}
	关联数据一致性 c_{42}
时效性 c_5	基于时间段的正确性 c_{51}
	基于时间点的及时性 c_{52}
	时序性 c_{53}
可访问性 c_6	可访问 c_{61}
	可用性 c_{62}
收益性 c_7	
风险性 c_8	
成本性 c_9	

线段的2个端点,并将这条线段按照剩余半升(降)云的朵数 $(n-3)/2$ 进行均分,随后将均分后每朵云在线段上对应的位置作为约束条件,赋予对应半升云的期望值并将其作为乘数,以解决 E_x 可能超出论域取值范围的问题。

2.3 评价指标群决策值的确定

2.3.1 定性评价指标群决策值的确定

定性指标的评价采用专家语言评价方式,专家评价后需对每个评价对象的单个指标进行汇总,进而得到专家评价的群决策值。设评价等级云 $\widehat{r}_{muk} = C(E_{x_{muk}}, E_{N_{muk}}, H_{e_{muk}})$ 表示第 k ($k=1, 2, \dots, l$)个专家对评价对象 m ($m=1, 2, \dots, M$)的第 u 个定性指标 c_u ($u=1, 2, \dots, p$)的评价云,根据式(2)~(4)的合成方法对 k 个评价等级云进行合成,得到全体专家组对方案 A_m 的定性评价指标 c_u 的群决策评价云

$$\widehat{r}_{mu} = \sum_{k=1}^l \gamma_k \widehat{r}_{muk} \quad (7)$$

式中, γ_k 表示第 k 个专家在 l 个专家中的权重。所有 \widehat{r}_{mu} 作为矩阵元素共同构成云决策矩阵,表示专家组对于评价对象 m 的第 u 个定性指标的群决策结果,得到

$$R_1 = \left[\widehat{r}_{mu} \right]_{M \times p} \quad (8)$$

2.3.2 定量评价指标群决策值的确定

定量指标可根据数据本身特征计算得出一个确定值,不存在随机性与模糊性,熵和超熵的取值均为零。此时,所有云滴在图上的分布将直接退化为一个点,横坐标等于该指标对应的具体数值,隶属度 $\mu \equiv 1$ 。对于第 c_z ($z=1, 2, \dots, q$) 个定量指标,指标数值计算结果为 d_z ,则专家组对于评价对象 m 的第 z 个定量指标的群决策评价云

$$\widehat{r}_{mz} = C(d_{mz}, 0, 0) \quad (9)$$

所有 \widehat{r}_{mz} 共同构成了云决策矩阵中的元素,得到

$$R_2 = [\widehat{r}_{mz}]_{M \times q} \quad (10)$$

最终,通过融合 R_1 和 R_2 得到全部定性和定量指标的云决策矩阵

$$R = [R_2 | R_1]_{M \times N} \quad (11)$$

2.4 指标权重确定

基于BWM确定各级评级指标的权重,步骤如下:

步骤1 在所有一级评价指标中,由专家分别确定最优(最重要)指标 c_b 和最劣(最不重要)指标 c_w 。

步骤2 由专家分别构建其他一级指标相对于 c_b 和 c_w 的优先级,并以1到9之间的整数对重要程度进行赋值,数值越大代表2个指标之间重要性的差异越大,分别得到比较向量 H_b 和 H_w 。 $H_b = [H_{b1}, H_{b2}, \dots, H_{bN}]$, H_{br} 表示最优指标 c_b 相较于指标 c_r ($r \in 1, 2, \dots, N$) 的重要程度; $H_w = [H_{w1}, H_{w2}, \dots, H_{wN}]$, H_{wr} 表示指标 c_r 相较于最劣指标

c_w 的重要程度。

步骤3 以各一级指标的权重 $W_c = [\omega_{c_1}, \omega_{c_2}, \dots, \omega_{c_N}]$ 作为决策变量求解最优权重,以指标之间的权重比值最接近专家赋值为目标,建立优化模型:

$$\begin{aligned} & \min \xi \quad (12) \\ \text{s.t.} & \left| \frac{\omega_{c_b}}{\omega_{c_r}} - H_{br} \right| \leq \xi, r \in 1, 2, \dots, N \\ & \left| \frac{\omega_{c_r}}{\omega_{c_w}} - H_{wr} \right| \leq \xi, r \in 1, 2, \dots, N \\ & \sum_{r=1}^N \omega_{c_r} = 1, r \in 1, 2, \dots, N \\ & \omega_{c_r} \geq 0, r \in 1, 2, \dots, N \end{aligned}$$

步骤4 针对一级指标对应的二级指标,可重复上述步骤1~3,得到相应二级指标的权重。

2.5 云-TOPSIS决策矩阵构造

根据上述步骤所获得的云决策矩阵 R 和指标权重 W_c ,将指标权重乘以云决策矩阵,可以得到指标加权后的云决策矩阵

$$Y = [\widehat{y}_{mv}]_{M \times N} = [\omega_r \widehat{r}_{mr}]_{M \times N} \quad (13)$$

其中,每个元素是云模型 $\widehat{y}_{mv} = C(E_{x,mv}, E_{N,mv}, H_{e,mv})$ ($m=1, 2, \dots, M$ 为评价对象, $v=1, 2, \dots, N$ 为指标)。

根据加权后的云决策矩阵,基于TOPSIS,分别求取正负理想解。由于本文中涉及的云-TOPSIS决策矩阵元素全部为云的三元表达形式,因此需要采用分场景的方式对正负理想解进行计算:

$$\Phi^+ = \left\{ \begin{array}{l} \tilde{\varphi}_v^+ \\ \tilde{\varphi}_v^+ \end{array} \right\} = \left\{ \begin{array}{l} C(\max_{1 \leq m \leq M} E_{x,mv}, \min_{1 \leq m \leq M} E_{N,mv}, \min_{1 \leq m \leq M} H_{e,mv}), v \in J^+ \\ C(\min_{1 \leq m \leq M} E_{x,mv}, \min_{1 \leq m \leq M} E_{N,mv}, \min_{1 \leq m \leq M} H_{e,mv}), v \in J^- \end{array} \right\} \quad (14)$$

$$\Phi^- = \left\{ \begin{array}{l} \tilde{\varphi}_v^- \\ \tilde{\varphi}_v^- \end{array} \right\} = \left\{ \begin{array}{l} C(\min_{1 \leq m \leq M} E_{x,mv}, \max_{1 \leq m \leq M} E_{N,mv}, \max_{1 \leq m \leq M} H_{e,mv}), v \in J^+ \\ C(\max_{1 \leq m \leq M} E_{x,mv}, \max_{1 \leq m \leq M} E_{N,mv}, \max_{1 \leq m \leq M} H_{e,mv}), v \in J^- \end{array} \right\} \quad (15)$$

式中: J^+ 表示效益型指标,即该指标对最终排序结果起到正向作用; J^- 表示成本型指标,即该指标对最终排序结果起到负向作用。正理想解是各指标下“最优云模型”的集合,需同时考虑期望的最优性、熵/超熵的最小性;负理想解是各指标下“最劣云模型”的集合,需同时考虑期望的最劣性、熵/超熵的最大性。

在归一化环节,由于云模型的3个特征值并不适用于一般的归一化处理,因此为了消除不同量纲对结果的影响,在模型应用时可将评价论域指

定在 $[0, 1]$ 之间,以便进行有效的计算和分析。

根据上述方法确定了正、负理想解后,按照式(6)云模型间距离的计算方法,分别对评价对象 m 计算其与正、负理想解之间的距离:

$$D_m^+ = \sum_{v=1}^N d(\tilde{y}_{mv}, \tilde{\varphi}_v^+) \quad (16)$$

$$D_m^- = \sum_{v=1}^N d(\tilde{y}_{mv}, \tilde{\varphi}_v^-) \quad (17)$$

最后,计算各评价对象与正、负理想解之间的贴

近度,作为最终评价对象之间进行比较的参照依据,贴进度越高,则评价对象排序越前。贴进度计算公式为

$$S_m = \frac{D_m^-}{D_m^+ + D_m^-} \quad (18)$$

3 模型应用

将本文所构建的质量评价模型应用于阿里云天池公开的数据集,进行数据资产质量评估,并对评价结果进行讨论。

表2 公开数据集概况

Tab.2 Overview of publicly available datasets

数据集	数据集描述	字段示例
A ₁	某电商企业用户行为数据集,用于隐式反馈推荐问题的研究	用户ID、商品ID、商品类目ID、时间戳等
A ₂	某移动电商平台的脱敏数据	user_id、item_id、behavior_type等
A ₃	来自某推荐算法比赛	event_time、event_type、price等
A ₄	来自某电商企业,根据用户信息预测其购买行为从而进行推荐	user_id、birthday、buy_mount等
A ₅	来自阿里云天池算法竞赛	item_list、item_id、img_data等

(2) 随机改造原表数据。由于很多公开数据集均已经过清洗和加工,很多质量指标的计算结果均为1,因此在原数据基础上借助随机数生成技术,随机选择数据元素/数据行进行改造。随后,邀请3位专家 $S=\{S_1, S_2, S_3\}$ 进行质量评价,3位专家权重 $\omega_s=(0.2, 0.4, 0.4)$ 。

3.2 模型应用

3.2.1 评价指标选择

参照前序所建立的企业数据资产质量评价指标体系,选取共9个一级指标、18个二级指标作为企业数据资产质量的评价准则,其中收益性、风险性和成本性3个一级指标为定性指标,邀请专家进行自然语言评价,采用BWM获取指标体系权重,如表3所示。根据相应规则计算得到的定量指标结果见表4。

3.2.2 单个自然语言评价值云化

将评价标准划分为5个等级,评语集 $V=\{VL, L, M, H, VH\}$,依次表示很差(VL)、差(L)、一般(M)、好(H)、很好(VH),定义有效论域 $U=[X_{\min}, X_{\max}]=[0, 1]$,指定 $H_e=0.01$ 。根据上文所述的改进黄金分割法,最终计算得到如表5所示的5朵评价等级云。5朵评价等级云的云图见图2,云滴分布连贯且云的边界较为清晰,说明上述过程所构建的云较为合理。

3.1 应用说明

选取阿里云天池提供的5个电商推荐领域公开数据集 $Q=\{A_1, A_2, A_3, A_4, A_5\}$ (见表2),并对每个数据集各选取一张数据表进行分析。公开数据已经过清洗和加工,为了使数据资产质量评价更贴近企业实际,对选取的数据进行模拟与改造。具体包括:

(1) 生成模拟数据。由于某些指标依赖具体条件才能判断是否符合要求,因此本文在原表基础上借助Python的Faker库中生成模拟数据的方法新增字段并赋予判断规则,以判断相应数据是否符合指标定义。

表3 企业数据资产质量评价指标体系与权重

Tab.3 Enterprise data asset quality evaluation index system and weights

一级指标	权重	二级指标	权重
准确性	0.210	数据内容正确性	0.284
		数据格式合规性	0.129
		数据重复率	0.183
		数据唯一性	0.404
完整性	0.086	数据元素完整性	0.358
		数据记录完整性	0.642
规范性	0.047	数据标准规范性	0.159
		数据模型规范性	0.280
		元数据规范性	0.103
		业务规则规范性	0.126
一致性	0.083	安全规范	0.332
		相同数据一致性	0.542
时效性	0.109	关联数据一致性	0.458
		基于时间段的正确性	0.457
		基于时间点的及时性	0.186
可访问性	0.042	时序性	0.357
		可访问	0.365
收益性	0.181	可用性	0.635
		风险性	0.116
成本性	0.126		

3.2.3 评价指标群决策值的确定

对于3个定性评价指标,通过邀请3位专家 $S=\{S_1, S_2, S_3\}$ 对上述5个企业数据资产

表 4 评价方案定量指标计算结果

Tab.4 Calculation results of quantitative index for the evaluation programme

二级指标	各公开数据集评价指标值				
	A ₁	A ₂	A ₃	A ₄	A ₅
c ₁₁	0.983	0.975	0.996	0.928	0.990
c ₁₂	0.820	0.831	0.902	0.844	0.893
c ₁₃	0.997	1.000	1.000	0.992	1.000
c ₁₄	1.000	1.000	1.000	1.000	1.000
c ₂₁	0.986	0.980	0.993	0.945	0.988
c ₂₂	1.000	1.000	1.000	1.000	1.000
c ₃₁	0.985	0.988	1.000	0.993	1.000
c ₃₂	1.000	1.000	1.000	1.000	1.000
c ₃₃	1.000	1.000	1.000	1.000	1.000
c ₃₄	0.992	0.998	1.000	0.994	1.000
c ₃₅	0.930	0.908	0.921	0.983	0.993
c ₄₁	1.000	1.000	1.000	1.000	1.000
c ₄₂	1.000	1.000	1.000	1.000	1.000
c ₅₁	0.993	0.972	0.958	0.990	0.977
c ₅₂	0.971	0.995	0.990	0.982	0.971
c ₅₃	1.000	1.000	1.000	1.000	1.000
c ₆₁	1.000	1.000	1.000	1.000	1.000
c ₆₂	1.000	1.000	1.000	1.000	1.000

表 5 自然评价语言与评价等级云的转化

Tab.5 Transformation of the natural evaluation language to the evaluation rating cloud

评价语言	评价等级云
很差(VL)	C(0, 0, 103, 0, 026)
差(L)	C(0, 309, 0, 064, 0, 016)
一般(M)	C(0, 500, 0, 039, 0, 010)
好(H)	C(0, 691, 0, 064, 0, 016)
很好(VH)	C(1, 000, 0, 103, 0, 026)

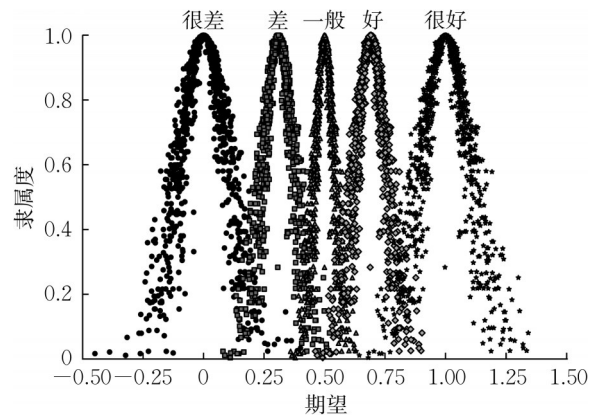


图 2 5 朵评价等级云的云图展示

Fig.2 Cloud diagram displaying the five evaluation level clouds

Q={A₁, A₂, A₃, A₄, A₅} 的质量进行定性评价, 并转换成评价等级云模型。3 位专家权重假设为(0.2, 0.4, 0.4), 结合云合成算法和专家权重计算得到定性评价指标云决策矩阵为:

$$R_1 = \begin{bmatrix} C(0.3854, 0.0326, 0.0082) & C(0.4764, 0.0527, 0.0133) & C(0.5000, 0.0370, 0.0093) \\ C(0.7000, 0.0447, 0.0113) & C(0.5382, 0.0384, 0.0096) & C(0.2000, 0.0486, 0.0123) \\ C(0.2236, 0.0491, 0.0124) & C(0.6764, 0.0527, 0.0133) & C(0.4618, 0.0384, 0.0096) \\ C(0.8146, 0.0502, 0.0126) & C(0.5764, 0.0310, 0.0078) & C(0.4000, 0.0302, 0.0077) \\ C(0.3000, 0.0447, 0.0113) & C(0.3472, 0.0370, 0.0093) & C(0.3854, 0.0326, 0.0082) \end{bmatrix}$$

对于 6 个主观评价指标, 结合基于 BWM 计算结果和指标权重, 得到定量评价指标云决策矩阵

$$R_2 = \begin{bmatrix} C(0.971, 0, 0) & C(0.995, 0, 0) & C(0.973, 0, 0) & C(1.000, 0, 0) & C(0.991, 0, 0) & C(1.000, 0, 0) \\ C(0.971, 0, 0) & C(0.993, 0, 0) & C(0.967, 0, 0) & C(1.000, 0, 0) & C(0.986, 0, 0) & C(1.000, 0, 0) \\ C(0.986, 0, 0) & C(0.997, 0, 0) & C(0.974, 0, 0) & C(1.000, 0, 0) & C(0.979, 0, 0) & C(1.000, 0, 0) \\ C(0.958, 0, 0) & C(0.980, 0, 0) & C(0.992, 0, 0) & C(1.000, 0, 0) & C(0.992, 0, 0) & C(1.000, 0, 0) \\ C(0.983, 0, 0) & C(0.996, 0, 0) & C(0.998, 0, 0) & C(1.000, 0, 0) & C(0.984, 0, 0) & C(1.000, 0, 0) \end{bmatrix}$$

将一级指标权重 W=(ω₁, ω₂, ..., ω₉) 乘以矩阵 [R₂|R₁] 得到最终的云-TOPSIS 决策矩阵, 如表 6 所示。

表6 云-TOPSIS决策矩阵
Tab.6 Cloud-TOPSIS decision matrix

评价对象	c_1	c_2	c_3
A_1	$C(0.2039, 0, 0)$	$C(0.0856, 0, 0)$	$C(0.0457, 0, 0)$
A_2	$C(0.2039, 0, 0)$	$C(0.0854, 0, 0)$	$C(0.0454, 0, 0)$
A_3	$C(0.2071, 0, 0)$	$C(0.0857, 0, 0)$	$C(0.0456, 0, 0)$
A_4	$C(0.2012, 0, 0)$	$C(0.0843, 0, 0)$	$C(0.0466, 0, 0)$
A_5	$C(0.2064, 0, 0)$	$C(0.0857, 0, 0)$	$C(0.0469, 0, 0)$
评价对象	c_4	c_5	c_6
A_1	$C(0.0830, 0, 0)$	$C(0.1080, 0, 0)$	$C(0.0420, 0, 0)$
A_2	$C(0.0830, 0, 0)$	$C(0.1075, 0, 0)$	$C(0.0420, 0, 0)$
A_3	$C(0.0830, 0, 0)$	$C(0.1067, 0, 0)$	$C(0.0420, 0, 0)$
A_4	$C(0.0830, 0, 0)$	$C(0.1081, 0, 0)$	$C(0.0420, 0, 0)$
A_5	$C(0.0830, 0, 0)$	$C(0.1073, 0, 0)$	$C(0.0420, 0, 0)$
评价对象	c_7	c_8	c_9
A_1	$C(0.0698, 0.0059, 0.0015)$	$C(0.0553, 0.0061, 0.0015)$	$C(0.0630, 0.0047, 0.0012)$
A_2	$C(0.1267, 0.0081, 0.0020)$	$C(0.0624, 0.0045, 0.0011)$	$C(0.0252, 0.0061, 0.0015)$
A_3	$C(0.0405, 0.0089, 0.0022)$	$C(0.0785, 0.0061, 0.0015)$	$C(0.0582, 0.0048, 0.0010)$
A_4	$C(0.1474, 0.0091, 0.0023)$	$C(0.0669, 0.0036, 0.0009)$	$C(0.0504, 0.0038, 0.0010)$
A_5	$C(0.0543, 0.0081, 0.0020)$	$C(0.0403, 0.0043, 0.0011)$	$C(0.0452, 0.0041, 0.0010)$

3.2.4 评价方案优劣性排序

基于上述决策矩阵R,求取各评价对象与正、负理想解之间的距离以及相对贴程度,计算结果见表7,数据资产质量的排序为 $A_4 > A_2 > A_1 > A_3 > A_5$ 。

表7 云-TOPSIS各评价对象与正负理想解间距离和相对贴程度

Tab.7 Distance and relative closeness degree between each cloud-TOPSIS evaluation object and positive and negative ideal solutions

评价对象	D_m^+	D_m^-	S_m
A_1	0.3877	0.3545	0.4776
A_2	0.3237	0.3425	0.5141
A_3	0.3618	0.3222	0.4711
A_4	0.3356	0.3918	0.5386
A_5	0.3816	0.3251	0.4600

3.3 结果讨论

为验证本文方法的有效性,依据主观-TOPSIS计算得到各评价对象与正负理想解之间的距离以及相对贴程度,计算结果见表8。数据资产质量的排序为 $A_4 > A_1 > A_3 > A_2 > A_5$ 。

由以上计算结果可知,云-TOPSIS对数据资产质量的优劣性排序为 $A_4 > A_2 > A_1 > A_3 > A_5$,主观-TOPSIS方法对数据资产质量的优劣性排序为 $A_4 > A_1 > A_3 > A_2 > A_5$ 。2种评价方法相对贴程度计算结果对比见图3。通过对比分析可知,2种方法均是以数据资产 A_4 的质量为最优,而以数据资产 A_5 的质量为最劣。在针对数据资产 A_1 、 A_2 、 A_3 的质量进行排序时结果存在差异。这是由于云模型考虑到了评价自

表8 主观-TOPSIS各评价对象与正负理想解间距离和相对贴程度

Tab.8 Distance and relative closeness degree between each subjective-TOPSIS evaluation object and positive and negative ideal solutions

评价对象	$D_i^{+'}$	$D_i^{-'}$	S_i'
A_1	0.4234	0.4310	0.5044
A_2	0.6084	0.4460	0.4230
A_3	0.4291	0.4253	0.4978
A_4	0.1507	0.7037	0.8290
A_5	0.6549	0.1995	0.2335

然语言的模糊性和表达隶属度的随机性,并不是将某一评价等级直接转化为具体数值,而是基于云分布的数字特征进行语言值的转换和后续计算。同时,由于主观-TOPSIS没有考虑隶属度较低的云滴分布情况,而是直接将评价语言映射成具体数值,因此相对于云-TOPSIS模型,其对各方案相对贴程度的计算结果之间具有较大差异。

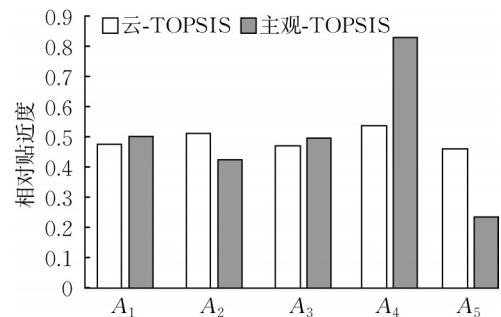


图3 2种评价方法的相对贴程度对比
Fig.3 Comparison of relative closeness degree between two evaluation methods

需要进一步说明的是,在模型应用部分,为了全面展示模型的应用情况,考虑并计算了全部一、二级指标,并通过构造模拟数据资产数据集进行评估。然而,值得注意的是,数据资产质量评价是具有场景性的,数据资产质量的标准和指标重要性可能随应用场景的不同而有所变化,质量评价也需要针对特定的业务需求进行。例如,在客户关系管理系统中,数据的完整性和一致性至关重要,而在供应链管理中往往更加注重数据的及时性和准确性。因此,在实际评价过程中,某些指标可能并不适用于当前的评价场景,或者并不需要进行计算。在模型的实际应用过程中可以综合考虑所有指标,根据实际需求进行选择 and 取舍,并请相关领域专家根据具体需求为指标重要性打分从而获取指标权重,以最大程度地贴合评价目标 and 需求。本文所构建的模型具有一定的灵活性和可扩展性,能够较好地支持上述过程。

4 结语

本文提出的基于云模型、BWM和TOPSIS的数据资产质量评价模型,为数据资产的综合评价提供了模型参考。结合定性 with 定量指标,将改进的云模型与BWM和TOPSIS相结合,不仅增强了评价体系的完整性和科学性,还提升了模型对模糊性和不确定性的适应性。在实际应用中,将模型应用于电商领域互联网企业的公开数据集,有效验证了模型在真实场景中的适用性和可靠性,具备一定的推广价值。本研究仍存在进一步拓展的空间。首先,指标体系的构建在一定程度上依赖相关领域专家的经验,未来研究可结合企业对于数据资产的质量需求,进一步拓展和改进质量评价指标体系;其次,本研究选用的测试数据集仅限于电商领域,尚未在更广泛的行业中验证其通用性,未来的研究可进一步拓展模型的应用范围;最后,还可结合机器学习方法自动优化评价指标的权重分配,以降低主观性并提升评价模型的精准性与效率。

作者贡献声明:

尤建新:研究选题策划,思路设计,论文指导。

程敏倩:数据处理,模型计算分析,论文初稿撰写。

徐涛:模型构建,方法论证,论文修改审阅。

参考文献:

- [1] 徐涛,尤建新,曾彩霞,等.企业数据资产化实践探索与理论模型构建[J].外国经济与管理,2022,44(6):3.
XU Tao, YOU Jianxin, ZENG Caixia, *et al.* Exploration of enterprise data assetisation practice and theoretical model construction [J]. Foreign Economy and Management, 2022, 44 (6): 3.
- [2] 陈国青,曾大军,魏强,等.大数据环境下的决策范式转变与使能创新[J].管理世界,2020,36(2):95.
CHEN Guoqing, ZENG Dajun, WEI Qiang, *et al.* Decision-making paradigm shift and enabling innovation in big data environment [J]. Management World, 2020, 36(2): 95.
- [3] 尤建新.大数据时代呼唤数据质量治理[J].上海质量,2019,37(10):20.
YOU Jianxin. Big data era calls for data quality governance [J]. Shanghai Quality, 2019, 37(10): 20.
- [4] 尤建新,徐涛.基于多准则决策方法的数据资产质量评价模型[J].同济大学学报(自然科学版),2021,49(4):585.
YOU Jianxin, XU Tao. A data asset quality evaluation model based on multi-criteria decision-making method [J]. Journal of Tongji University (Natural Science), 2021, 49(4): 585.
- [5] MUNAPPY A R, BOSCH J, OLSSON H H, *et al.* Data management for production quality deep learning models: challenges and solutions [J]. Journal of Systems and Software, 2022, 191(9): 111359.
- [6] 刘桂锋,聂云贝,刘琼.数据质量评价对象、体系、方法与技术研究进展[J].情报科学,2021,39(11):13.
LIU Guifeng, NIE Yunbei, LIU Qiong. Research progress on objects, systems, methods and techniques of data quality evaluation [J]. Intelligence Science, 2021, 39(11): 13.
- [7] ALTENDEITERING M, GUGGENBERGER T M, MILLER F A. A design theory for data quality tools in data ecosystems: findings from three industry cases [J]. Data & Knowledge Engineering, 2024, 153(9): 102333.
- [8] FENG Z T, CHEN M. Platformance-based cross-border import retail e-commerce service quality evaluation using an artificial neural network analysis [J]. Journal of Global Information Management, 2022, 30(11): 1.
- [9] GUDIVADA V N, APON A, DING J. Data quality considerations for big data and machine learning: going beyond data cleaning and transformations [J]. International Journal on Advances in Software, 2017, 10(1): 1.
- [10] MANDAL P. Data quality in statistical process control [J]. Total Quality Management & Business Excellence, 2010, 15 (1): 89.
- [11] 叶琼,李绍稳,张友华,等.云模型及应用综述[J].计算机工程与设计,2011,32(12):4198.
YE Qiong, LI Shaowen, ZHANG Youhua, *et al.* A review of cloud models and applications [J]. Computer Engineering and Design, 2011, 32(12): 4198.
- [12] 徐士东,耿秀丽.云模型与TOPSIS相结合的多属性群决策

- 方法[J]. 计算机应用研究, 2017, 34(10): 2964.
- XU Shidong, GENG Xiuli. A multi-attribute group decision-making method combining cloud modelling and TOPSIS [J]. Computer Application Research, 2017, 34(10): 2964.
- [13] 杨洁, 王国胤, 刘群, 等. 正态云模型研究回顾与展望[J]. 计算机学报, 2018, 41(3): 724.
- YANG Jie, WANG Guoyin, LIU Qun, *et al.* Review and prospect of normal cloud modelling research [J]. Journal of Computing, 2018, 41(3): 724.
- [14] LIN S S, ZHOU A, SHEN S L. Safety assessment of excavation system via TOPSIS-based MCDM modelling in fuzzy environment [J]. Applied Soft Computing, 2023, 138(5): 110206.
- [15] BADI I, BALLEM M. Supplier selection using the rough BWM-MAIRCA model: a case study in pharmaceutical supplying in Libya [J]. Decision Making: Applications in Management and Engineering, 2018, 1(2): 16.
- [16] ZUO W J, YU D J, HU Q, *et al.* A big data quality evaluation method based on group heterogeneity rationality perception information fusion [J]. Computers & Industrial Engineering, 2024, 190(4): 110009.
- [17] MASHOUFI M, AYATOLLAHI H, ZAVARAH D K, *et al.* Data quality assessment in emergency medical services: an objective approach [J]. BMC Emergency Medicine, 2023, 23(1): 1.
- [18] 任剑. 基于云模型的语言随机多准则决策方法[J]. 计算机集成制造系统, 2012, 18(12): 2792.
- REN Jian. A stochastic multi-criteria decision-making method for languages based on cloud models [J]. Computer Integrated Manufacturing Systems, 2012, 18(12): 2792.

(上接第284页)

- (1): 24.
- [22] PETERSSON B, PLUNT J. On effective mobilities in the prediction of structure-borne sound transmission between a source structure and a receiving structure, part I: theoretical background and basic experimental studies[J]. Journal of Sound and Vibration, 1982, 82(4): 517.
- [23] MASHAYEKHI M J, BEHDINAN K. Analytical transmissibility based transfer path analysis for multi-energy-domain systems using four-pole parameter theory [J]. Mechanical Systems and Signal Processing, 2017, 95: 122.
- [24] MOLLOY C T. Use of four-pole parameters in vibration calculations [J]. The Journal of the Acoustical Society of America, 1957, 29(7): 842.
- [25] 曾国锋, 韩紫平, 刘鸣博, 等. 电磁悬浮型高速磁浮车-岔垂向动力响应[J]. 同济大学学报(自然科学版), 2023, 51(3): 303.
- ZENG Guofeng, HAN Ziping, LIU Mingbo, *et al.* Vertical dynamic response of electromagnetic levitation type high-speed maglev vehicle at turnouts [J]. Journal of Tongji University (Natural Science), 2023, 51(3): 303.
- [26] 李云钢, 常文森, 龙志强. EMS磁浮列车的轨道共振和悬浮控制系统设计[J]. 国防科技大学学报, 1999(2): 96.
- LI Yungang, CHANG Wensen, LONG Zhiqiang. Track resonance analysis and levitation control system design for EMS maglev trains [J]. Journal of National University of Defense Technology, 1999(2): 96.
- [27] 李钦. 基于柔性悬浮理论的高速磁悬浮列车整体控制方案研究[D]. 上海: 同济大学, 2022.
- LI Qin. Research on global control scheme of high-speed maglev trains based on flexible levitation theory [D]. Shanghai: Tongji University, 2022.
- [28] 洪小波. 高速磁浮轨道不平顺检测系统的研究[D]. 长沙: 国防科技大学, 2021.
- HONG Xiaobo. Research on detection system of track irregularity for high-speed maglev [D]. Changsha: National University of Defense Technology, 2021.
- [29] 高速磁浮交通设计标准: CJJ/T 310—2021 [S]. 北京: 中国建筑工业出版社, 2021.
- Design standard for high-speed maglev transportation: CJJ/T 310—2021 [S]. Beijing: China Architecture & Building Press, 2021.
- [30] LIU Mingbo, YE Feng, ZENG Guofeng. Research of the track irregularity spectrum of Shanghai High-speed Transrapid Demonstration Line [J]. Vehicle System Dynamics, 2024, 62(8): 2054.