

大数据挖掘技术对数字经济特征的识别优化



李延超*

广东科技学院计算机学院, 广东东莞 523083

摘要: 大数据技术在各个领域的应用比较广泛, 使得数字经济特征识别问题成为研究热点。传统的数据分析无法解决数字经济的特征识别问题, 而且特征识别能力较弱。由此, 本文提出数据挖掘技术联合大数据的算法, 构建特征识别的判断模型。首先, 采用传统数据分析方法对大数据进行分类, 按照特征进行集合划分, 降低不同特征数据的处理复杂性, 削弱多维度数字经济信息对特征识别的影响。然后, 传统数据分析方法将数据分类, 形成不同维度的子特征识别域, 并对不同维度子特征识别进行综合判断。借助 MATLAB 仿真, 在识别等级一定的情况下, 本文构建算法的准确性、计算时间均优于传统数据分析方法、大数据挖掘技术。在初始特征识别标准和预期效果的条件下, 本文构建的算法可以对特征识别进行综合判断, 符合数字经济特征识别的需求。

关键词: 挖掘技术; 大数据; 特征识别; 数字经济

DOI: [10.57237/j.cst.2023.03.002](https://doi.org/10.57237/j.cst.2023.03.002)

Identification and Optimization of Big Data Mining Technology for Digital Economy Characteristics

Yanchao Li*

College of Computer, Guangdong University of Science and Technology, Dongguan 523083, China

Abstract: Big data technology is widely used in various fields, making the problem of digital economy feature recognition become a research hotspot. Traditional data analysis cannot solve the problem of feature recognition in the digital economy, and the feature recognition ability is weak. Therefore, this paper proposes an algorithm of data mining technology combined with Big data to build a judgment model for feature recognition. First of all, traditional data analysis methods are used to classify Big data, and Partition of a set is carried out according to characteristics to reduce the processing complexity of data with different characteristics and weaken the impact of multi-dimensional digital economic information on feature recognition. Then, traditional data analysis methods classify the data to form sub feature recognition domains of different dimensions, and make comprehensive judgments on sub feature recognition of different dimensions. With the help of MATLAB simulation, the accuracy and calculation time of the algorithm built in this paper are better than traditional data analysis methods and Big data mining technology under the condition of a certain recognition level. Under the conditions of initial feature recognition standards and expected effects, the algorithm constructed in this article can comprehensively judge feature recognition, which meets the requirements of digital economy feature recognition.

*通信作者: 李延超, liyanchao@gdust.edu.cn

Keywords: Mining Technology; Big Data; Feature Recognition; Digital Economy

1 引言

国家工信部提出特征数据处理数字经济化的发展战略，国内外学者十分重视特征识别的准确性。依据2022年调查数据显示，数字经济、特征数据、大数据挖掘技术研究占特征数据研究的85.5%，致使特征识别的需求日益增强[1]。目前，人工数字经济、大数据、数据挖掘技术的研究较多，并对上述研究领域进行特征数据分析，为数字经济良好发展提供保证[2]。由于数字经济特征存在大量的半结构化数据，而且数据之间呈现非线性关系，存在数据混乱等问题。数据分析方法具有较高的半结构化分析能力，将多维度的特征数据映射到二维平面，并通过循环计算[3]，实现分析数据的预处理。有学者提出一种数据分析方法，对国内数字经济数据进行分析。结果证明，数据分析方法可以提高特征识别的准确性，而且各项指标显著优于传统方法[4]。

传统数据分析方法属于一种综合的方法，融合大特征数据处理方法的分类优势，对数字经济数据进行模糊分析[5-7]。各种数字经济算法对特征数据的识别存在各自优势，而且在数据挖掘深度、效果方面存在不足，所以要利用大数据方法对特征数据进行处理。故本文以传统数据分析方法为依托，结合数据特征对数字经济数据进行分析，主要在以下几方面进行改进，1) 调整特征数据的参数。大数据的预处理与传统数据分析方法的结合，能增加单次特征数据的处理量；2) 与其他特征识别方法比较，具有较高的优势。将大数据融入人工数字经济中，对数字经济数据进行分类，依据特征权重设定相应的处理等级，提高特征数据的判断速度。3) 与大数据、传统数据分析方法进行比较，利用传统数据分析方法对非结构化特征数据进行处理，实现半结构、非结构的分类，提高特征数据识别的准确性。

本文对传统数据分析方法进行改进，构建特征识别判断模型，为特征识别提供综合性的评价方法，并验证对数字经济的特征识别效果，促进数字经济的良好发展。

2 大数据挖掘技术的相关概念

2.1 大数据挖掘方法

大数据方法是一种网络综合分析方法，结合统计、

网络信息技术、通讯技术，以及计算机数字经济算法等技术，解决特征数据处理、数据分类、数学描述和数字经济发展等问题，广泛应用于信息、计算、金融等领域[8]。大数据挖掘技术优势明显，可以进行数字经济的数据分析，具有持续性优势。大数据技术对非结构数据数字经济分析，但要符合以下假设。

假设 A: D_{it} 是数字经济的发展结果，所处的时间为 t 时刻，并构建特征数据集 $\text{set}(D_{it})$ 。其中，任意数据 x 属于 D_{it} ，特征数据的识别结果为 $P_j(x)$ ，计算结果如公式(1)所示。

$$P_j(x) = \sum_{i,t=1}^n x_{it}^j \Leftrightarrow \varphi(k) \quad (1)$$

其中， $k \in (1, \dots, +\infty)$ ， $\varphi(k)$ 属于映射结果。为了提高特征数据的计算精度，融入 $C(x, \zeta)$ ， x 和 ζ 的特征系数。在公式(1)中，如果 $x_{it} = \theta(\rho \tan t)$ ，说明计算精度较低；如果 $x_{it} < \sum_{i,j,t=1}^n x_{it}^j$ ，说明计算精度符合要求。

2.2 特征数据的分类

假设 B: $P_n(x)$ 的 n 维特征数据 $\varphi(x \cdot k)$ 的分析结果，不同维度的特征数据分析结果，如公式(2)所示：

$$P_n(x) = \int_n k f(x) dx \xleftarrow{\zeta} \sum_{j=1}^n f(P_j(x)) \quad (2)$$

其中， $f(x)$ 的不同维度的综合分析函数。

假设 C: 综合分类函数为 $V(x)$ ，该函数符合以下条件， $V(x) < \zeta$ ，..., $V(x)'' < \frac{\Delta \zeta^2}{2}$ ，则数字经济的特征识别结果均的判断如公式(3)所示。

$$V(x) = \frac{\int k V(x)''}{2} \quad (3)$$

假设 D: 任意特征数据点 x_{ij} 处于数字经济发展轴上, 任意特征数据点的导数将代表经济的发展方向, 其计算如公式(4)所示。

$$D(x, f(x))'|p = \alpha \cdot \ln\left(\frac{1}{\sqrt{x}}\right) \quad (4)$$

其中, α 代表数字经济的发展方向。

由上述定理可知, 使用大数据的到时, 可以计算不同经济数据 x 之间的非线性关系, 而且降低 i 维度和 t 时间对经济特征结果的影响。因此, 大数据分类函数为数字经济特征数据处理提供良好的基础, 降低数据结构对特征识别结果的影响。由定理 3 可知, 大数据中数据的多维判断精度为 $\alpha \cdot \ln\left(\frac{1}{\sqrt{x}}\right)$, 说明多维判断精度符合要求, 进一步降低数据结构对结果的影响。

2.3 数字经济的挖掘

本文选择传统数据分析方法来进行模型构建, 该技术属于一种数字经济分类技术, 具有模糊性、自适应性的优势, 可以实现循环计算, 不断修订分类集。传统数据分析方法可以采用 IF 模式进行约束, 形成约束条件 M , 其分类过程如下:

IF: $x_i < d_{ij}$, and $M(x) \wedge C(x_i, x_{i-1})$,

$$\text{then } y \wedge \sum_{i=1}^n y_i \vee \sum_{i=1}^n M(x_i)$$

其中, $\lambda(x_i)$ 为数字经济调节函数, d_{ij} 为数字经济调整结果的集合, $M(x)$ 为约束条件, y 为经济特征的识别结果。传统数据分析方法对数字经济进行预处理, 处理过程采用数据挖掘, 最终得到准确识别结果。因此, 可以用传统数据分析方法推理出输出结果, 并得到综合性的经济特征识别结果。

假设 E: 任意数字经济为 x_i , 经约束条件 M , 分析输入变量 x_i 与输出变量 y_i 的关系, 如公式(5)所示。

$$\alpha \cdot g_{ij} = \sum_{i,j=1}^n x \cdot \left\{ \frac{(x_i \wedge c_{ij})}{b_{ij}} \right\} \quad (5)$$

其中, c_{ij} 、 b_{ij} 不同经济的结果转化; g_{ij} 为大数据的预处理集合。

依据上述经济特征的识别进行计算, 得到经济特征的识别的连续算子, 计算结果如公式(6)所示。

$$g_{ij} = C \cdot \delta \sum_{i,j,k=1}^n g_{ij}^k \wedge (x^2) \quad (6)$$

其中, δ 为经济特征的识别系数, k 为阶层。依据经济特征的识别结果, 可以得到经济特征的识别的输出值, 如公式(7)所示。

$$y = \sum_{i,j,k=1}^n g_{ij}^k(x) \Rightarrow f(P_j(x)) \quad (7)$$

数据分析方法可以缩短数字经济的处理时间, 增加预处理数据量。依据初始数据量, 进行多维度的经济特征识别, 形成连续的特征数据结果。

3 经济特征识别模型的判断

3.1 特征数据的初始化

构建的经济特征识别模型, 可以提高特征数据的特征识别能力, 缩短数据结构优化时间。体现在初始数据量、多维数据量的综合分析方面, 并利用特征识别报警条件, 实现经济特征识别的综合判断, 输出最优的结果。

3.1.1 处理的数字经济特征数据

数字经济数据的结构为非结构式, 离散性分布。如果经济数据量处于非正态分布, 其计算会出现冗余, 降低综合计算的准确性。为了提高数据的计算准确性, 要扩大数据量的规模, 增加数据的种类。本文通过扩大初始数据的数据量数量, 提高数据量的多样性, 结果如图 1 和 2 所示。

图 1 和 2 分别是非结构技术和传统数据分析方法处理的初始大数据, 数据的分段个数均为 120。通过对比发现, 非结构技术产生的初始数据量杂乱, 而且无方向性。传统数据分析方法处理的数据量更集中, 且具有方向性。由传统数据分析方法的定理 1、2 可知, 该算法的结果与空间维数无关, 而且特征数据处理更加准确。每次取点时, 初始数据量的分布效果一样, 数据集的稳定性较高, 所以选取传统数据分析方法进行初始数据量的处理。

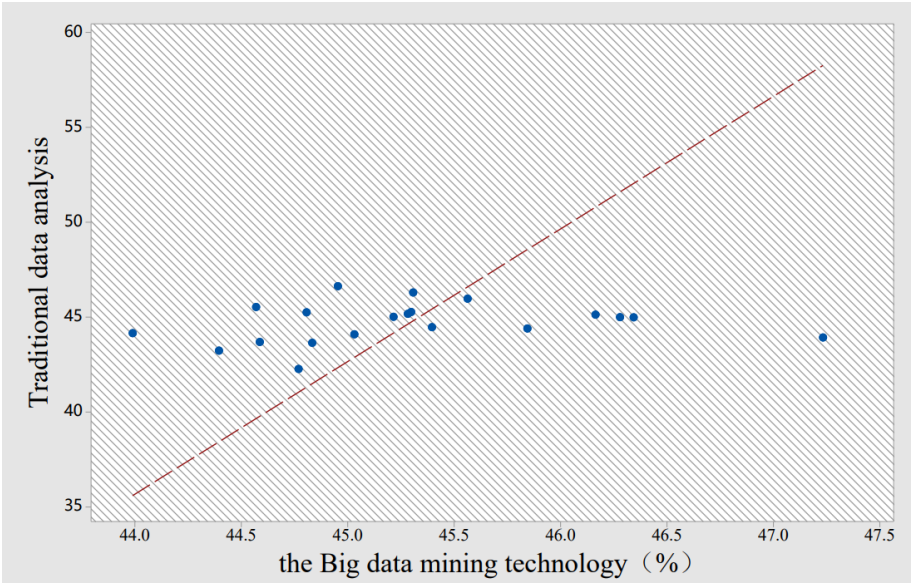


图 1 不同处理方面的特征数据处理结果

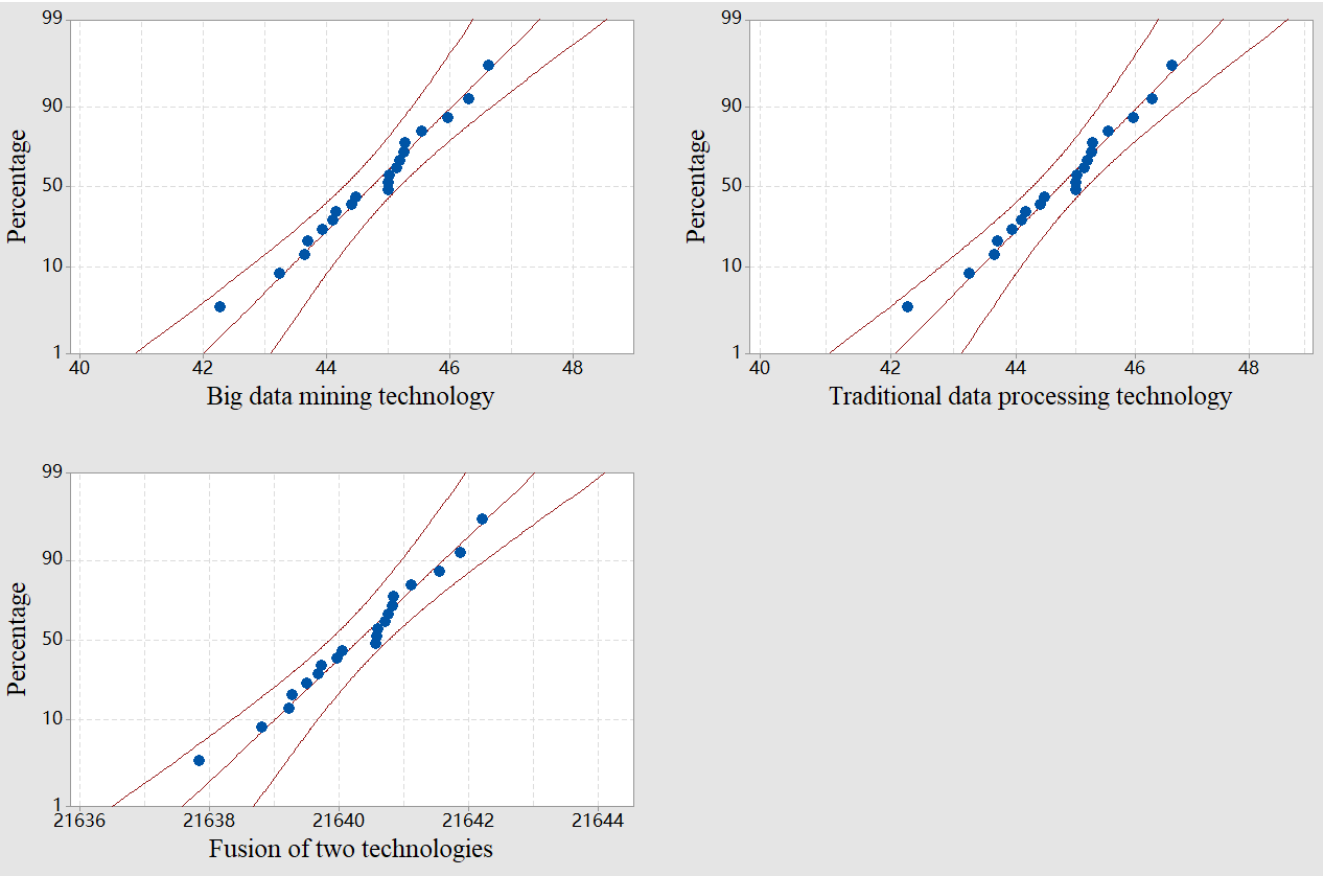


图 2 两种方法的数据处理结果

3.1.2 不同指标间的数据处理过程

特征识别能力是算法的重要衡量标准。在初始阶段，算法注重多维数据量，而后会向单一特征识别递

进[9-11]。为了提高计算准确性，本文引入综合方式，让数据依据不同特征进行特征识别，同时实现综合和单一特征识别。目前，除了标准的传统数据分析方法，还有其他改进算法。

1) 本地经济特征识别模型, 计算如公式(8)所示

$$y_{ij}(t+1) = \omega \cdot y_{ij}(t) + c_1 \cdot r_1 \cdot \delta \cdot \sum_{i,j,k=1}^n g_{ij}^k \{x(t) \wedge f(P_j[x(t)])\} \quad (8)$$

2) 跨域数字经济特征识别模式, 计算如公式(9)所示

$$y_{ij}(t+1) = \omega \cdot y_{ij}(t) + c_2 \cdot r_2 \cdot f[P_j x(t)] \quad (9)$$

3) 定点专业数字经济的特征识别判断模式, 计算如公式(10)所示

$$y_{ij}(t+1) = n \cdot \delta \cdot \sum_{i,j,k=1}^n g_{ij}^k \{x(t) \rightleftharpoons f(P_j[x(t)])\} \quad (10)$$

4) 综合特征识别模式, 计算如公式(11)所示

$$y_{ij}(t+1) = \omega \cdot y_{ij}(t) + \sum_{i=1}^2 c_i \cdot r_i \cdot g_{ij}^k \{x(t) \cdot f(P_j[x(t)])\} \quad (11)$$

其中, c_1 、 c_2 为综合的协同系数, r_1 、 r_2 为综合的相关系数, ω 综合的平均权重, n 不同策略数。

本文对传统数据分析方法进行两方面的改进, 一方面, 扩大特征识别范围。每一次迭代计算, 就会从上述 5 中形式中, 非结构选择一种大数据形式进行传统数据分析方法计算。为了获得单一最优的计算结果, 需要扩大特征识别范围。非结构数据的选择计算形式, 不仅要保持数据的多样性, 而且要提高单一结果的准确性。另一方面, 通过简化计算过程, 缩短收敛时间。为了平衡多维特征识别和单一特征识别的能力, 加入特征识别因素融入系数 α 、数据结构权重 ω , 以提高计算速度, 计算如公式(12)所示。

$$\alpha = Line_t^d - 1 \quad (12)$$

其中, e 为经济数据放大系数, t 为经济阶段, T 为最大时间, d 为迭代次数, D 为迭代次数。在初始阶段, α 呈现下降趋势, 方便综合特征识别结果的计算。在计算中后期, α 呈现浮动形式, 方便单一最优值计算。数据结构权重的计算如公式(13)所示。

$$w = \frac{(w_{\max} \wedge \sum_{t=1}^T \Delta w_t)}{(\sum_{t=1}^T \Delta w_t \cdot w_{\min})} \quad (13)$$

其中, w_{\max} 和 w_{\min} 分别为最大、最小的数据变化权重, d 、 t 、 D 、 T 的参数与公式(10)一致。

3.1.3 数字经济的综合判断策略

本算法对不同维度的数据采用异构化策略, 并调整相应参数, 实现多维度数字经济信息分布式协同, 以完成综合判断过程。模型中, 将大数据分为 5 个多维, 任意维度代表解空间的子空间。在迭代过程中, 5 个多维度数字经济信息同时进化。迭代计算完成后, 比较不同维度的适应值, 记录各多维度数字经济信息的综合特征识别结果的位置。然后, 各个子多维度数字经济信息均向综合特征识别结果递进学习, 用最简洁的方式获得子多维度数字经济信息最优的位置, 以此来提高特征识别计算的速度和准确性。

3.2 数据分析方法的特征识别判断技术

数据分析方法的基本思想是多时间多维度数字经济信息的综合判断, 对大数据的初始特征识别标准、特征识别报警条件进行调整与优化[12], 以获得最优解, 降低数字经济的特征识别率, 结果如图 3 所示。

步骤 1: 确定大数据的数字经济信息、数据量结构, 依据数据特征及问题解决需求, 确定大数据的数据结构。将整个数据的初始权值、特征识别报警条件作为一个整体, 映射到大数据上, 每个大数据的数字经济信息为权值、特征识别报警条件的乘机。依据实际的应用情况, 确定本文的大数据数字经济信息 $D=433$ 。

步骤 2: 数据初始化。非结构初始化大数据的相关参数。

步骤 3: 生成适度函数。运用传统数据分析方法理论产生大数据的初始数据量, 并将其映射到大数据上。计算每个大数据的准确率, 并将其平方和的绝对值作为适度函数。

步骤 4: 特征识别大数据的最优位置和各子特征识别域的最优位置。将初始大数据非结构分为 5 个子数据量, 求得适应度比值, 记录最优的综合位置和各子数据量的最优位置。

步骤 5: 最佳位置与速度的迭代。在 5 种子数据量的进化中, 非结构选择一种进行进化, 如公式(12)和(13)。

步骤 6: 如果迭代次数 d 小于最大迭代次数 D , 则重复步骤 2~6, 否则停止迭代, 并返回特征识别报警条件、权重、最佳位置等结果。

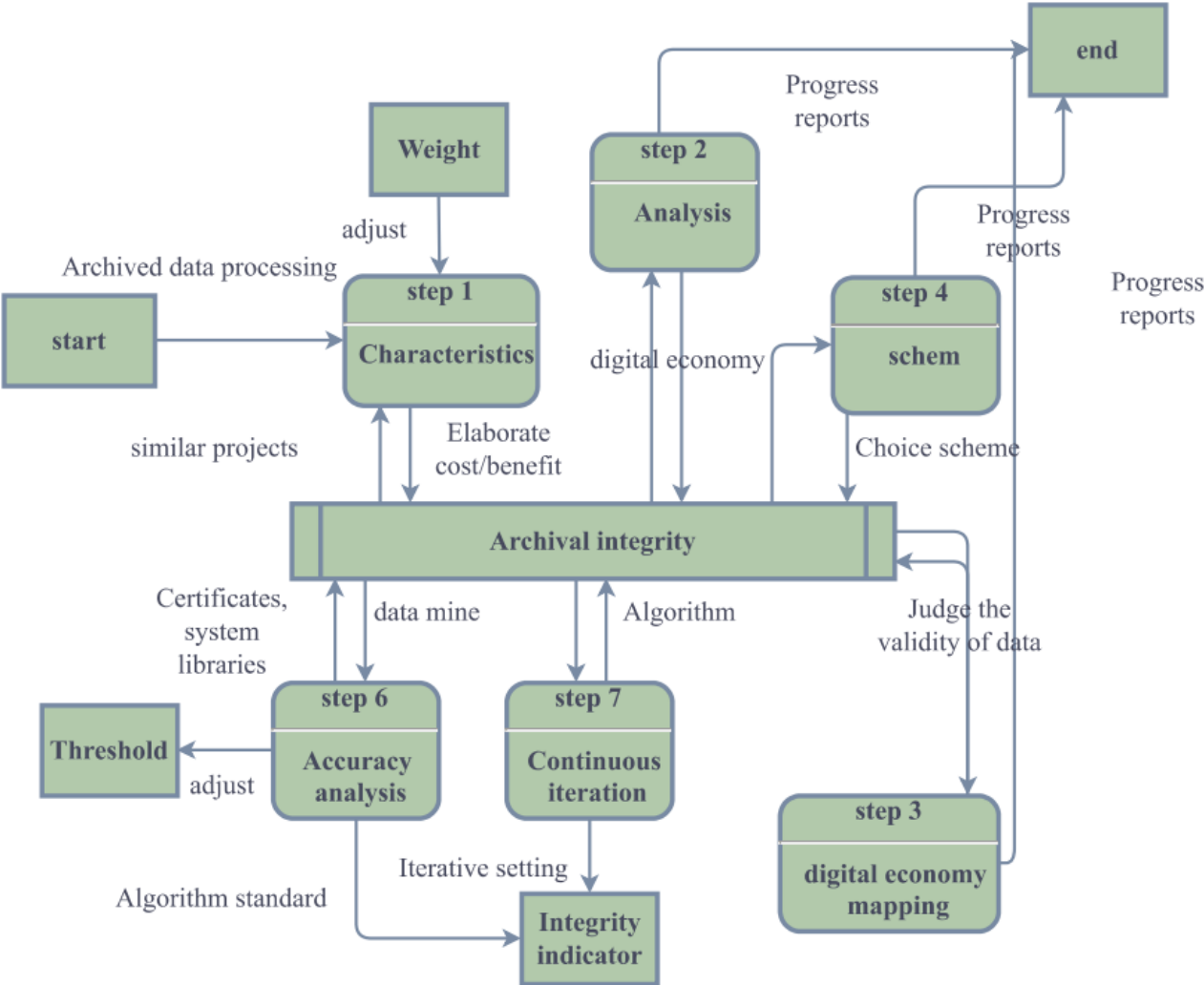


图3 传统数据分析方法与大数据的计算流程图

4 实际案例分析

4.1 模型的性能判断

用单指标识别、多指标识别、多维度指标等指标，测试大数据挖掘技术，以验证本文提出模型的性能。
单指标识别是测试模型综合唯一最小值函数，公式如下：

$$A(x) = \sum_{i=1}^n [x_i^2] \tag{14}$$

多指标识别是通过余弦调制传递函数，频繁产生单一最小值，以验证模型解的实用性，公式如下：

$$B(x) = \sum_{i=1}^n \cos \alpha \bullet e^{x_i^2} \tag{15}$$

多维度指标是多维点进行梯度寻优函数，测试多维数据的计算速度，以检测综合的判断速度，公式如下：

$$C(x) = -\alpha \bullet e^{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}} \tag{16}$$

其中， n 为计算数据的指标总数， x_i 为任意指标数。
为了便于计算，本文的数字经济个数 $n=1200$ 个，最迭代数 $D=30$ 次，最长时间 $T=24$ 月，分别进行上述3个函数测试。为了降低非结构性对结果的影响，取结果的10次平均值，具体计算结果如表1所示。

表 1 不同测试函数的检测结果

测试指标	测试函数	Equation parameter	Standard Error	Wald chi-square	95%Confidence interval
单指标识别	传统数据分析方法	1	-0.3168	0.1980	-2.6079~0.1640
	大数据	-0.3168	1	-0.1497	
	挖掘技术	0.1980	-0.1497	1	
多指标识别	传统数据分析方法	-0.0766	-0.3297	0.4752	-0.5460~1.2811
	大数据	1	-0.3168	0.1980	
	挖掘技术	-0.3168	1	-0.1497	
多维度指标	传统数据分析方法	0.98	0.4623	0.7780	3.5760~0.2947
	大数据	0.9027	0.5799	0.2124	
	挖掘技术	0.8532	0.1980	-0.1497	

表 1 中的各数据的收敛图如图 4 所示。

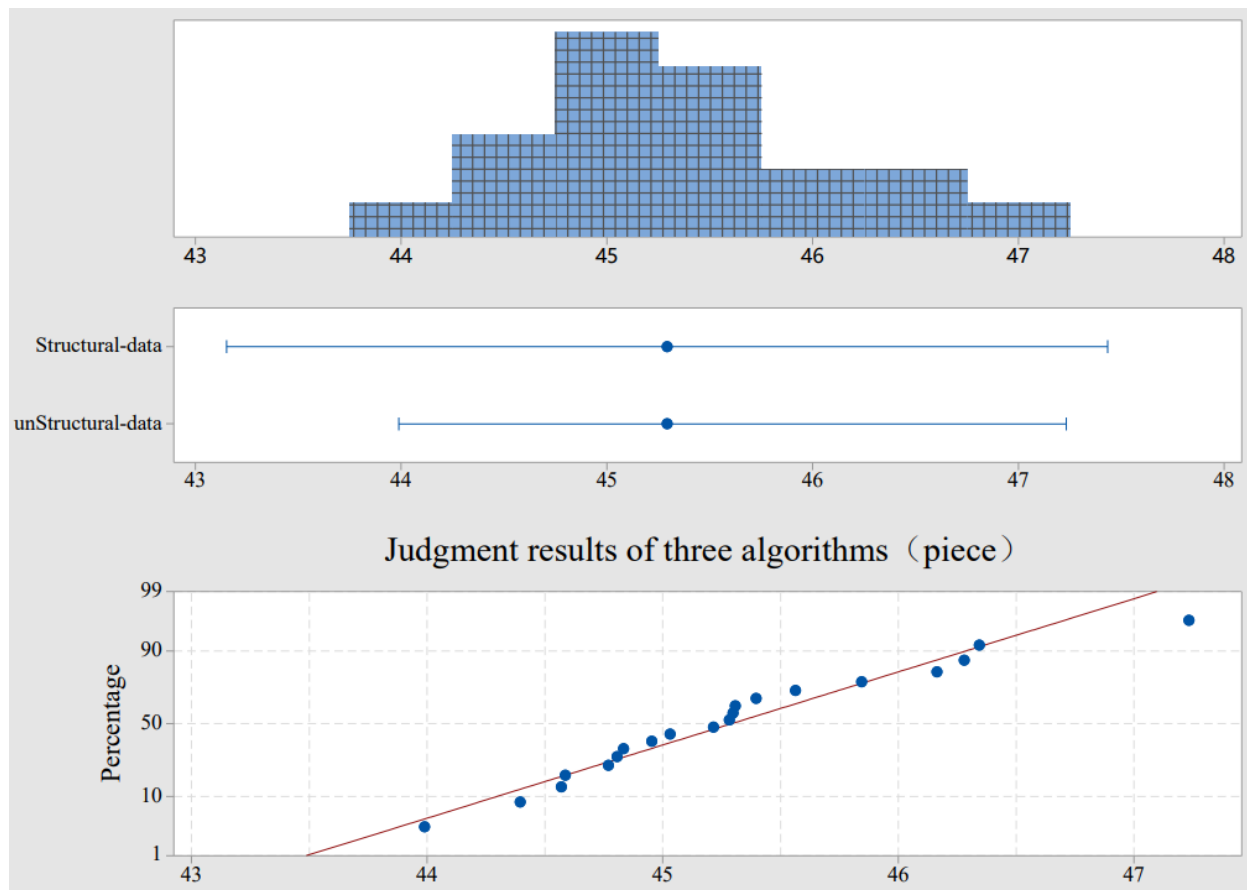


图 4 表 1 中各数据的收敛图

由表 1 可知,与大数据、传统数据分析方法比较,本文提出的挖掘技术更接近综合特征识别结果。在标准差、平均值、取值范围等方面,挖掘技术优于其他两种算法。通过图 4 中的曲线变化可知,挖掘技术的稳定性更优、判断速度更快。因此,挖掘技术的判断速度、识别等级判断、求和稳定性更佳。

4.2 数字经济特征识别案例

特征识别的判断数据集包含数字特征识别、仿生特征识别、自然特征识别、心理特征识别、预期特征识别[13-15]。经过数据的初步预处理,得到 43 行结构化数据、32 行半结构化数据。为了便于数据分析,选

择不同领域的的数据，分别为：金融领域、金融领域、机械领域、电子领域，数据量的处理结果如表 2 所示。

表 2 特征识别等级划分及占比

数据类型	Mean	SD
机械领域	44.6071	0.9847
数字经济	45.1265	0.9027
机械领域	45.3421	0.8532
电子领域	44.7917	1.2428
Test Items	Test value	p-value
-2Ln LR(L ²)	15.5244	0.3433
Pearson chi-square	12.6163	0.5569
Scaled Deviance	15.5212	0.3435
Degrees of freedom=14		

4.3 测试结果

为了验证本文提出的挖掘技术，与传统数据分析方法、大数据进行结果比较，结果如图 5 所示。

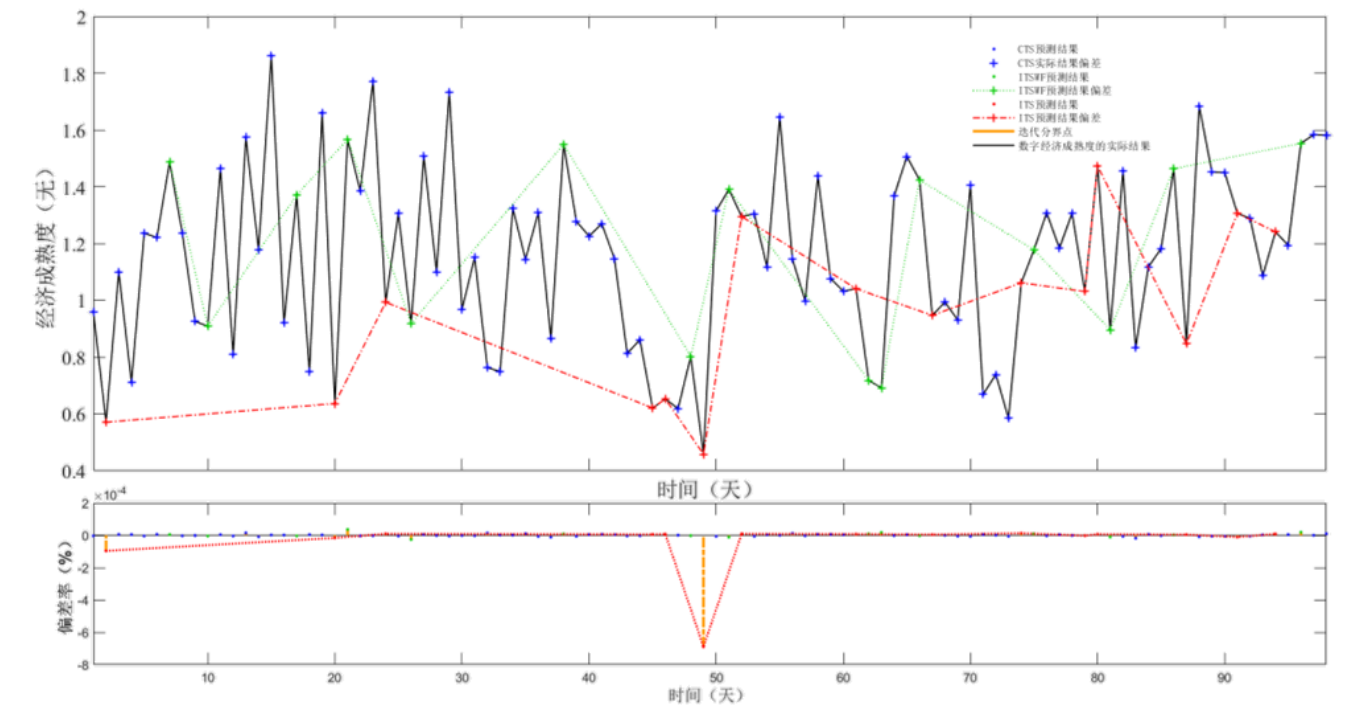


图 5 不同算法的测试结果

由图 5 可知，挖掘技术的精准度高于传统数据分析方法、大数据，误差率却较低，说明传统数据分析方法、大数据的计算比较稳定，而传统数据分析方法、大数据计算参差不齐。上述三种算法的平均结果如表 3 所示。

表 3 不同等级的判断准确性对比

算法	Size of samples	Mean R	Se	99%Confidence interval	P-value	Accuracy
大数据挖掘技术	673	0.5008	0.0111	0.4722~0.5294	0.0710	0.9434
传统数据处理技术	679	0.4985	0.0111	0.4700~0.5270	0.1326	0.8945
挖掘技术	675	0.5013	0.0111	0.4727~0.5299	0.1174	0.9066

由表 3 可知, 单一的大数据挖掘技术和传统数据分析方法在不同等级的特征识别判断方面, 存在精准度不足, 计算结果变化幅度大的问题。相对于大数据挖掘技术来说, 本文构建算法的精准度显著提高。同

时, 本文构建算法与大数据挖掘技术的准确度比较接近, 均大于 80%, 优于传统数据分析方法。为了进一步验证挖掘技术的优越性, 对不同算法的最优适应度值进行比较, 结果如图 6 所示。

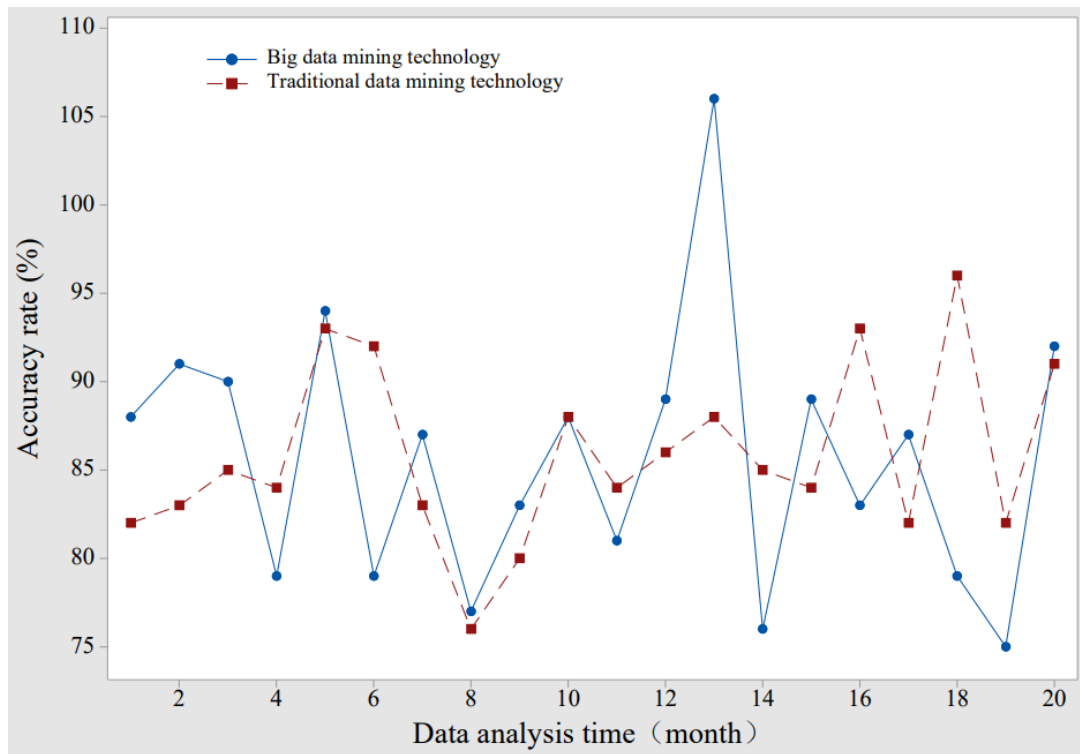


图 6 特征值的识别过程

由图 6 可知, 大数据挖掘技术的结果显著优于传统数据挖掘技术, 究其原因, 挖掘技术增加了不同维度的协同系数、改进权重、收敛因子。

5 结束语

本文提出了传统数据分析方法理论, 并结合多维综合判断技术, 对大数据挖掘技术进行改进。同时, 对大数据挖掘技术进行特征数据识别条件、特征识别标准, 构建特征识别分析模型。结果表明, 相对于经典大数据挖掘技术、传统数据分析方法来说, 传统数据分析方法联合大数据挖掘技术的特征识别判断精度、收敛性更优, 可以对特征识别进行判断。但是, 本模型中, 综合判断策略过于注重特征识别能力, 致使单一特征识别能力相对下降, 降低最优解的计算速度。因此, 在未来的研究中, 将会增加平衡系数函数, 以提高模型的求解能力。

参考文献

- [1] 詹韵秋, 王军, 孙小宁. 数字经济对家庭消费行为的影响研究——基于中国家庭金融调查的经验分析 [J]. 当代经济管理, 2023, 45 (2): 89-96.
- [2] 王钰, 张自然. 中国人口结构特征与经济效率, 经济转型——基于 1992-2017 年中国分地区面板数据的分析 [J]. 商业研究, 2019 (12): 10.
- [3] S. S. Akimov, Business process modeling within the digital economy development framework [J]. AEBMR-Advances in Economics Business and Management Research, 2019, 5 (6): 262-267.
- [4] 王彦军, 孙军. 数字经济发展对中国经济增长促进作用研究——基于中国省域面板数据的实证分析 [J]. 江苏商论, 2022 (2): 5.
- [5] 石月. 数字经济环境下的跨境数据流动管理 [J]. 信息安全与通信保密, 2015 (10): 101-103.

- [6] 王千. 数字经济时代如何通过数据挖掘实现用户增长--基于 Growth Hacking 的病毒式营销路径分析 [J]. 河南大学学报: 社会科学版, 2019, 59 (1): 37-42.
- [7] 丁国峰. 数字经济时代数据竞争反垄断问题探究 [J]. 经济法研究, 2022 (1): 212.
- [8] T. V. Miroyubova, T. V. Karlina, and R. S. Nikolaev, Digital Economy: Identification and Measurements Problems in Regional Economy [J]. Ekonomika Regiona-Economy of Region, 2020, 16 (2): 377-390.
- [9] 孙妍, 胡龙, 冯雪玲. 基于变换匹配层融合的双模态生物特征识别方法[J].计算机工程, 2023, 49 (5): 269-276.
- [10] 李宇轩, 陈壹华, 温兴, 等. 改进 Point-Voxel 特征提取的 3D 小目标检测 [J]. 微电子学与计算机, 2023, 40 (2): 50-58.
- [11] 尹艳芳. 论大数据统计分析方法在经济管理领域中的运用 [J]. 商业文化, 2022 (7): 32-33.
- [12] 谢鸿飞. 中国经济周期波动特征及拐点识别研究 [D]. 华中科技大学, 2011.
- [13] 张彬, 陈双, 马雯. 我国区域数字鸿沟静态与动态综合测度 [J]. 中国通信: 英文版, 2010 (1): 124-130.
- [14] R. Sari, F. L. Gaol, H. Prabowo, F. F. Hastiadi, The General Factors Mapping between Digital Economy and Sharing Economy [J]. 2020, 9 (12): 621-625.
- [15] R. Chinoracky, and T. Corejova, HOW TO EVALUATE THE DIGITAL ECONOMY SCALE AND POTENTIAL? [J]. Entrepreneurship and Sustainability Issues, 2021, 8 (4): 536-552.