

贫数据条件下民用航空发动机制造成本估算建模

梁晶晶

中国航发商用航空发动机有限责任公司, 上海

收稿日期: 2022年8月17日; 录用日期: 2022年8月27日; 发布日期: 2022年9月13日

摘要

民用航空发动机的设计方案一旦冻结, 其制造成本就已基本确定。若要提高发动机的经济性, 就要在设计阶段就估算出发动机的制造成本并改进设计方案。然而, 我国的民用航空发动机行业尚处于起步阶段, 在成本估算工作中可参考的历史成本数据很少。为了克服国内民用航空发动机行业积累的制造成本数据量少对成本估算工作精度的制约问题, 本文提出了一种虚拟数据生成方法, 用于在贫数据条件下对航空发动机制造成本数据集进行扩充。然后, 将提出的虚拟数据生成方法与BP神经网络结合运用, 建立了一个制造成本估算模型, 并且采用贝叶斯优化算法对模型的超参数进行了优化。算例分析结果表明, 提出的虚拟数据生成方法可以显著提高制造成本估算模型的精度, 建立的成本估算模型的性能也显著优于常见的参数法估算模型。研究成果可为民用航空发动机设计方案优化中的制造成本控制工作提供方法支撑, 还能够为航空发动机的产品定价和市场营销等决策提供依据。

关键词

成本估算, 民用航空发动机, 虚拟数据生成, BP神经网络, 贝叶斯优化

Manufacturing Cost Estimation Modeling of Civil Aero-Engine with Few Data

Jingjing Liang

AECC Commercial Aircraft Engine Co., Ltd., Shanghai

Received: Aug. 17th, 2022; accepted: Aug. 27th, 2022; published: Sep. 13th, 2022

Abstract

Once the design scheme of a civil aviation engine is frozen, its manufacturing cost has been basi-

cally determined. To improve the economy of the engine, it is necessary to estimate the manufacturing cost of the engine and improve the design at the design stage. However, the civil aviation engine industry of China is still in its infancy, and there are few historical cost data that can be referenced in cost estimation. To overcome the restriction on the accuracy of cost estimation caused by the small amount of manufacturing cost data accumulated in domestic civil aeroengine industry, a virtual data generation method is proposed to expand the aeroengine manufacturing cost data set under the condition of few data. Then, a manufacturing cost estimation model is established by combining the proposed virtual data generation method with BP neural network, and the hyper parameters of the model are optimized by Bayesian optimization algorithm. The results of example analysis show that the proposed virtual data generation method can significantly improve the accuracy of manufacturing cost estimation model, and the performance of the established cost estimation model is also significantly better than the common parameter estimation model. The research results can not only provide method support for manufacturing cost control in civil aeroengine design scheme optimization, but also provide the basis for aeroengine product pricing and marketing decision-making.

Keywords

Cost Estimation, Civil Aero-Engine, Virtual Data Generation, BP Neural Network, Bayesian Optimization

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

民用航空发动机的制造成本大约占其全生命周期成本的百分之四十，制造成本的高低直接决定了发动机制造商的市场竞争力和利润空间。有关分析表明：民用航空发动机百分之八十五以上的制造成本通常在设计方案冻结之前就已确定，发动机制造商若想对制造成本进行有效控制，就必须在设计方案冻结之前就尽可能准确地估算出批产阶段的制造成本，从而及时优化设计方案。然而，我国民用航空发动机行业尚处于起步阶段，可参考的历史成本数据非常少。对于民用航空发动机制造商而言，在贫数据条件下开展发动机的制造成本估算工作具有非常大的挑战性。

早在 20 世纪 80 年代初，世界航空发动机主制造商就启动了基于参数法的发动机成本估算模型的研究工作，并将研究成果应用于优化发动机的设计方案之中。普惠公司开发了一套发动机全生命周期成本分析模型(Engine/Airframe Generalized Life-cycle-cost Evaluator, EAGLE)，可在初步设计阶段估算出发动机在全生命周期各个阶段的成本。GE 航空公司开发了一套辅助成本估算模型(Auxiliary cost estimation, ACE)，模型中嵌入了发动机各部件的成本估算关系式(Cost Estimation Relationships, CERs)。随后，GE 公司持续改进 ACE 并形成了 COMPEATTM 模型，其中囊括了 34 类零部件的技术和成本数据。罗罗公司也自主开发了一套发动机成本估算模型，其中的制造成本估算模块是内嵌在设计软件 Genesis 之内的工作包，估算原理同样是发动机各部件的 CERs。根据以上标杆企业成本估算模型的发展经验可知，基于参数法建立 CERs 是一种能够快速实现制造成本估算的方式。然而，以上标杆企业在初期都是基于传统的参数法(例如，多元线性回归模型、指数回归模型)拟合得到 CERs，估算精度并不高。当前，他们正在通过自主研发或与第三方机构合作等方式探索新的成本估算模型以提高估算精度。

我国虽然早在 20 世纪 60 年代就开始了航空发动机研制，但却长期处于测绘、仿制、改进改型的发

展状态, 自主研制发动机的时间较短, 将成本作为独立变量融入到发动机研制设计之中的观念还未成为主流。在航空发动机的成本估算方面, 国内关注的较多的是航空发动机的研制成本和维修成本的估算问题。刘若泠[1]在对弹用发动机研制成本和生产成本的详细分解和分析的基础上, 采用了灰色系统参数建模方法建立了弹用发动机的研制成本估算模型, 并且使用国外型号发动机成本和性能数据进行了算例分析。刘芳和张海涛[2]研究了小样本理论中的偏最小二乘法在航空发动机研制费估算中的适用性, 形成了基于主要性能指标的军用航空发动机研制费参数估算模型。韩朋[3]利用偏最小二成回归法建立了发动机性能与维修成本关系模型, 并且根据维修费用和航空发动机性能恢复的程度进行费用-效益分析。赵曼[4]则对国内外采用的参数法发动机维修成本估算模型进行了梳理, 给出了具体的估算关系式和涉及的模型参数。以上研究虽然可以为民用航空发动机的制造成本估算工作提供一定的方法和经验借鉴, 但仍然不能解决在民用航空发动机成本可用数据量不足对成本估算精度的制约问题。

增加数据集的大小是提升成本估算精度的重要途径, 但收集数据往往缺乏渠道或者十分耗时、耗力。虚拟数据生成方法是实现数据集扩充的一种有效途径, 常被用于机器学习中对训练数据的预处理过程。虚拟数据生成的基本原理是: 基于提取出的原数据自身固有的特征、数据之间的关联关系等信息, 人工合成或采用计算机生成具有相同特征和关联关系的新数据。这种方法在图像识别、大气环境检测等领域已获得广泛地应用, 尤其是在提升深度学习模型的精度和训练速度方面, 采用虚拟数据生成方法可以获得很好的效果[5][6]。

本文将因子分析方法与灰色预测模型结合运用, 提出了一种虚拟数据生成方法[1][7][8]。它可以生成任意数量的与原始数据集有相似特征的虚拟数据, 显著增加可用于训练成本估算模型的样本数量, 从而在一定程度上克服成本数据的贫乏对制造成本模型估算精度的制约问题。算例分析结果表明, 本文提出的虚拟数据生成方法可以显著提高模型的估算精度, 建立的成本估算模型的性能也显著优于常用的参数法估算模型。该项研究成果可为航空发动机设计方案优化中的制造成本控制工作提供方法支撑, 以及为航空发动机的产品定价和市场营销等决策提供依据。

2. 虚拟数据生成方法

假设要生成一个样本量为 H 的虚拟数据集, 相应的实施步骤如下:

1) **提取数据特征, 重构数据:** ① 将用于生成虚拟数据的数据集记为 $\text{DATA1}=(\mathbf{Y}, \mathbf{X})$, 其中, \mathbf{Y} 是 $n \times 1$ 维矩阵, 代表收集到的 n 个不同型号发动机的制造成本值, \mathbf{X} 是 $n \times m$ 维的矩阵, 代表与 n 个不同型号发动机相对应的 m 种性能指标的取值。② 计算 \mathbf{X} 的相关系数矩阵 \mathbf{R} , 以及 \mathbf{X} 的特征值 λ 和特征向量 μ , 计算因子载荷矩阵:

$$\mathbf{A} = (\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \dots, \sqrt{\lambda_{m'}}u_{m'}),$$

其中, m' 是选取的数据特征的个数并且 $m' \leq m$ 。③ 对因子载荷矩阵进行最大方差正交旋转, 得到旋转后的因子载荷矩阵 $\tilde{\mathbf{A}}$ 。④ 计算数据特征的取值: $\mathbf{F} = \mathbf{X} \text{pinv}(\mathbf{R})\tilde{\mathbf{A}}$, 得到重构的数据集 $\text{DATA2}=(\mathbf{Y}, \mathbf{F})$ 。

2) **生成虚拟数据的特征值:** 设置迭代计数器 k 的取值为 1, 在数据特征矩阵 $\mathbf{F}=(\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_{m'})$ 中每一个数据特征的取值范围 $[L_{m'}, U_{m'}]$ 之内, 随机生成一个 $H \times m'$ 维的数据特征矩阵 $\hat{\mathbf{F}}(k)$, 然后将 $\hat{\mathbf{F}}(k)$ 映射为一个 $H \times m$ 维的矩阵:

$$\hat{\mathbf{X}}(k) = \hat{\mathbf{F}}(k) \cdot \text{pinv}(\text{pinv}(\mathbf{R})\tilde{\mathbf{A}})$$

将矩阵 $\hat{\mathbf{X}}(k)$ 中包含负值或零值的行删除, 并更新矩阵 $\hat{\mathbf{X}}(k)$ 。

3) **虚拟数据特征值的筛选与更新:** 按照第 2) 步的方法生成 $\hat{\mathbf{X}}(k+1)$, 将 $\hat{\mathbf{X}}(k)$ 与 $\hat{\mathbf{X}}(k+1)$ 按行合并作为新的 $\hat{\mathbf{X}}(k+1)$ 。若更新后的 $\hat{\mathbf{X}}(k+1)$ 的行数小于 H , 则令 $k = k+1$, 返回第 2) 步; 否则, 随机筛选出

H 行数据组成最终的 \hat{X} ，转到下一步。

4) **得到虚拟数据集:** 基于重构的数据集 DATA2, 利用灰色预测方法估算出数据集 \hat{X} 对应的成本值 \hat{Y} , 得到一个拥有 N 条数据的虚拟数据集 DATA3 = (\hat{Y}, \hat{X}) 。

在第 4) 步中, 利用灰色预测方法估算出数据集 \hat{X} 对应成本值 \hat{Y} 的过程如下[2]。记原数据集 DATA1 中的成本值为 $Y = (y(1), y(2), \dots, y(n))$, 性能指标序列值为: $X_1, X_2, X_3, \dots, X_k, \dots, X_m$, 其中, $X_k = (x_k(1), x_k(2), \dots, x_k(n))$ 。首先, 按照如下公式计算 Y 与 X_k 之间的灰色关联度 ε_k :

$$\varepsilon_k = \frac{1 + |s_y| + |s_k|}{1 + |s_y| + |s_k| + |s_k - s_y|},$$

其中,

$$|s_y| = \left| \sum_{i=2}^{n-1} y(i) + 0.5y(n) \right|,$$

$$|s_k| = \left| \sum_{i=2}^{n-1} x_k(i) + 0.5x_k(n) \right|,$$

$$|s_k - s_y| = \left| \sum_{i=2}^{n-1} (x_k(i) - y(i)) + 0.5(x_k(n) - y(n)) \right|。$$

取 \hat{X} 中的一条虚拟数据, 令其对应的成本值等于 \hat{y}_p , 组成一条临时成本数据 $d = (\hat{y}_p, x_1, x_2, \dots, x_m)$ 。将 d 插入到原数据集 DATA1 中, 组成一个 $(n+1) \times m$ 维的数据集 Data_{temp}。然后, 选取灰色关联度 ε_k 最高的性能指标值 k' 作为主关键值, 对 Data_{temp} 进行升序排序。记临时数据 $d = (\hat{y}_p, x_1, x_2, \dots, x_m)$ 在排序之后的序号为 p , 按照如下公式更新临时数据 d 的制造成本值 \hat{y}_p :

$$\hat{y}_p = \alpha \cdot y(p-1) + (1-\alpha) \cdot y(p+1),$$

$$1 \leq p \leq n+1。$$

其中, α 称为下界估算值 $y(p-1)$ 的比例系数, 又称为白化参数。白化参数通常基于数据列之间的灰关联性和样本点之间的间隔计算得到。本文采用如下白化参数计算方法: 对任意样本点 $x_{k'}(p)$, 令 $\Delta x_{k'}(p) = x_{k'}(p) - x_{k'}(p-1)$ ($p=1, 2, \dots, n+1$), 表示第 p 个样本点与前一个样本点的间隔, 白化参数则由如下公式计算得到:

$$\alpha = \frac{\Delta x_{k'}(p+1)}{\Delta x_{k'}(p) + \Delta x_{k'}(p+1)}。$$

将更新之后的临时数据 d 从 Data_{temp} 中取出, 放入 DATA3 之中。然后, 从 \hat{X} 取出下一条虚拟数据, 重复上述步骤, 直到全部虚拟数据的制造成本都被更新, 最终得到一个拥有 H 条数据的虚拟数据集 DATA3 = (\hat{Y}, \hat{X}) 。

3. 结合 BP 神经网络建立制造成本估算模型

3.1. 建模过程

在基于参数法的成本估算建模相关研究中, 常用的参数法有: 多元线性回归、偏最小二乘回归、指数回归以及 BP (Back Propagation) 神经网络。与前三种方法相比, BP 神经网络可以隐式地学习和存贮任何“输入-输出”之间的映射关系, 而无需事前提供显式的数学公式来描述这种映射关系, 并且具有较好的泛化能力和容错能力[9]。因此, 本文选择采用 BP 神经网络方法进行制造成本估算建模。应用 BP 神经网络进行建模的基本原理是: 将收集到的原始数据集随机划分为两个子集: 训练集和测试集, 利用训练

集的数据训练出一个 BP 神经网络模型，然后用测试集中的数据评估模型的预测能力。

1) 数据集的划分与预处理

记收集到的民用航空发动机的制造成本数据集为 $D_0 = (Y_{N \times 1}, X_{N \times m})$ ，其中， $Y_{N \times 1}$ 代表 N 个不同型号民用发动机的制造成本值， $X_{N \times m}$ 则代表 N 个不同型号民用发动机的 m 个性能指标取值。首先，将数据集 D_0 随机划分为两个子集 $D_1 = (Y_{n \times 1}, X_{n \times m})$ 和 $D_2 = (Y_{(N-n) \times 1}, X_{(N-n) \times m})$ 。其次，采用上一节中所述的虚拟数据生成方法，基于训练数据集 D_1 生成一个数据条数为 H 的虚拟数据集 D_3 。然后，将 D_3 作为训练数据集，将 D_2 作为测试数据集。

2) BP 神经网络的基本结构

BP 神经网络的训练是通过输入层到输出层的计算来完成的。BP 神经网络通常由 1 层输入层、多层隐藏层和 1 层输出层组成。通常来说，增加隐藏层的层数可以提升模型的训练效果，但也会增加模型的训练时间。假设共有 X 层隐藏层，第 $x(1 \leq x \leq X)$ 层隐藏层中的神经元个数表示为 hl_x ，则 BP 神经网络的结构可以表示为： $m - \dots - hl_x - \dots - 1$ 。其中，模型的输入是航空发动机的 m 个性能指标，因此，第 1 层输入层中有 m 个神经元，最后 1 层中只有 1 个神经元，代表模型的输出，即发动机的制造成本值。

3) BP 神经网络的超参数优化

除了上述结构参数之外，BP 神经网络中还需要设置目标函数与优化算法所需的参数，例如：学习速率 lr 、激活函数 $g(x)$ 、最大训练次数 T_{\max} 、训练算法，等等[10] [11]。以上参数被统称为 BP 神经网络的超参数。为了提升 BP 神经网络模型的性能，在应用中通常需要反复试验去获得较好的超参数组合，整个过程会耗费大量的时间和人力成本。本文将采用贝叶斯优化算法实现 BP 神经网络超参数的自动优化，以使用较低的成本来快速提升 BP 神经网络模型的估算精度。在本文建立的 BP 神经网络中，将对隐藏层的层数 X 、每个隐藏层中的神经元数量 hl_x 以及学习速率 lr 这三类超参数进行优化。其他参数则设置如下：激活函数选择 ReLU 函数： $g(x) = \max(0, x)$ ，最大训练次数取 1000 次，模型训练算法采用 Nesterov Momentum 优化算法，未提及的参数均采用默认值。

3.2. 模型的评估标准

用于评估成本估算模型性能的指标通常有两种：拟合优度 R^2 和估算精度 ε 。用 y_i 表示样本的实际值， \hat{y}_i 表示样本的估算值， \bar{y} 表示样本实际值的均值， n 表示样本数据集的大小，这两种指标的计算方法可以表示为：

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$\varepsilon = \frac{|\hat{y}_i - y_i|}{n}.$$

为了客观地评估快速成本估算模型的性能，本文将分别计算训练数据集和测试数据集上这两种指标。

4. 算例分析

4.1. 收集数据

通过对《世界航空发动机手册》(1996 年)、《世界中小型航空发动机手册》(2006 年)、《国外航空发动机简明手册》(2014 年)以及公开的网络资源等信息的整理，收集到了 34 条国外涡扇航空发动机的成本与性能数据，详见表 1 [12] [13] [14]。表 1 中的制造成本是在综合考虑报价年份、基准利率和货币单位之后，按照 65% 的目录价格折算得到的，折算的基准年份为 2020 年，统一的货币单位是美元。此外，出于

信息保密的需要，表格中隐去了具体的发动机型号信息。

取上述数据集中的前 30 条数据组成训练数据集 D_1 ，取后 4 条数据组成测试数据集 D_2 。在训练数据集上应用本文提出的虚拟数据生成方法生成一定数量的虚拟数据，从而得到一个虚拟数据集 D_{virtual} 。将虚拟数据集 D_{virtual} 作为输入 BP 神经网络模型的训练数据，训练出 BP 神经网络模型之后，用测试数据集 D_2 来测试模型的预测能力。BP 神经网络中的隐藏层层数 X 的取值范围是 $\{1, 2, 3\}$ ，每层隐藏层中神经元的个数 hl_x 的取值范围是 $\{1, 2, \dots, 20\}$ ，学习速率 lr 的取值范围是 $[0, 1]$ 。模型的损失函数采用平方损失函数，如下：

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2,$$

其中， θ 代表 BP 神经网络模型中各层神经元的阈值和连接权等参数组成的向量， $h_{\theta}(x^i)$ 代表采用当前参数下的模型根据输入的 x^i 计算得到的估计值， y^i 则是相应的实际值， i 代表数据的编号， m 则代表数据集中的数据条数。

Table 1. Example of aeroengine cost data set

表 1. 航空发动机成本数据集示例

编号	制造成本 (万美元)	最大起飞 推力/kN	涵道 比	总增 压比	涡轮进口 温度/ $^{\circ}\text{C}$	巡航耗油率 /kg/(kgf·h)	空气流 量/(kg/s)	风扇 直径/m	长度 /m	干重 /kg
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
E-1	1822.93	512.9	7.1	42	1482	0.52	1641	3.26	7.29	8761
E-2	49.85	86.73	0.31	27.5	1426	1.79	69	0.89	4.04	1050
E-3	712.08	104.5	6	25.2	1372.85	0.67	322.05	1.6	2.36	1951
E-4	447	91.67	5.4	22.7	1354.85	0.63	311.2	1.65	2.51	2366
E-5	58.03	8.45	3.3	12.7	1017.85	0.46	29.58	0.53	1.19	203
E-6	712.08	138.7	6.6	31.5	1359.85	0.55	483.07	1.95	2.62	2585
E-7	88.3	8.9	0.55	21	1445	0.645	29.8	0.59	1.21	315
E-8	62.31	10.23	2.58	12.8	1017.85	0.61	28.71	0.56	1.2	231
E-9	80.42	13.55	2	13.1	1014.85	0.552	34.02	0.52	1.6	284
E-10	94.47	15.57	2.79	13	963	0.82	51.25	1	1.27	283
E-11	712.08	111.1	6	26.5	1263.85	0.6	386.46	1.74	2.42	2266
E-12	126.67	19.15	3.15	19.4	1104.85	0.8	64.86	0.75	1.65	401
E-13	128.82	18.9	2.9	22	1022	0.73	65.77	0.72	1.55	406
E-14	509	106.8	5.3	26	1354.85	0.63	342	1.65	2.51	2366
E-15	148.57	22.24	3.9	22	1022	0.73	84.82	0.78	2.08	448
E-16	145.62	26.33	5.3	35	1369.85	0.65	95.3	0.9	2.51	551
E-17	196.3	26.69	0.8	11.3	1139.85	0.8	47.7	0.58	1.96	603
E-18	244.4	42.58	5	23	921	0.625	110.2	0.98	2.92	719.4
E-19	151.15	35.76	0.8	11	1139.85	0.78	47.6	0.76	2.9	713
E-20	207.69	28	0.4	19.4	1344	0.78	42.6	0.91	2.59	521.6
E-21	301.56	56.4	4.9	27	1021	0.68	200	1.3	3.3	1090
E-22	604	133.5	5.5	32	1286.85	0.6	427.7	1.83	2.6	2381

Continued

E-23	568.75	106.8	5.9	25.4	1354.85	0.66	370.6	1.73	2.43	2186
E-24	544	117.48	5.1	27.9	1354.85	0.63	355.2	1.65	2.51	2366
E-25	620	137.95	5.5	32.9	1286.85	0.6	433.6	1.83	2.6	2381
E-26	636	142.4	5.4	33.7	1286.85	0.6	433.6	1.83	2.6	2381
E-27	555	120.15	5.7	29.3	1286.85	0.6	406.9	1.83	2.6	2381
E-28	471	97.9	6	24.5	1286.85	0.6	371	1.83	2.6	2381
E-29	497	104.58	5.9	25.9	1286.85	0.6	382.8	1.83	2.6	2381
E-30	428	86.78	5.5	21.7	1354.85	0.63	302.5	1.65	2.51	2366
E-31	111.64	15.57	2.66	13.6	963	0.73	66.22	0.72	1.55	401
E-32	145.38	35.76	0.8	11	1139.85	0.95	47.6	0.76	1.96	544
E-33	315.69	38.8	6.3	21	1377	0.69	150.6	1.12	2.62	758
E-34	560	121.49	5.1	28.9	1354.85	0.63	354.7	1.65	2.51	2366

4.2. 模型估算性能的影响因素分析

模型参数的选取是影响模型估算性能的重要因素。在上述建模过程中，共有两类模型参数：虚拟数据生成过程中的参数和 BP 神经网络模型的超参数。其中，虚拟数据生成过程中的参数主要是指虚拟数据生成的样本数量 H ，BP 神经网络模型的超参数则主要包括：隐藏层的层数 X 、每个隐藏层中的神经元数量 hl_x 以及学习速率 lr 。下面，将分别对这两类参数对模型估算性能的影响进行算例分析。

4.2.1. 虚拟数据生成的样本数量

为了评估虚拟生成样本数量的多少对模型估算性能的影响，在保持 BP 神经网络模型的隐藏层层数 X 为 2，两层隐藏层中的神经元个数依次为： $hl_1 = 6$ ， $hl_2 = 8$ ，学习速率取 0.01，最大迭代次数为 1000 次的前提下，分别记录了虚拟样本数量等于 50 条、100 条、300 条、500 条、1000 条、3000 条、5000 条和 10,000 条时模型的估算性能指标取值，见表 2。其中，模型对测试集数据的估算误差如图 1 所示。观察发现，随着虚拟生成的样本数量的增加，模型对测试集的估算误差得到了明显的降低，但在样本数量大于 3000 条之后，估算误差基本维持在 20%~25%。上述结果表明：增大虚拟生成的样本数量有利于提升模型的估算性能，但是这种提升效果会随着样本数量的增大而逐步减弱，虚拟生成的样本数量达到一定数值之后，性能提升效果将在维持在一定的水平(上下波动)。此外，由于 BP 神经网络的超参数采用人工随机选取，即使采用了虚拟数据生成方法来增大可用的数据量，模型的性能仍然不尽人意。如表 2 中所示，模型在训练集上的性能表现尚可，但在测试集上的估算误差却只能降到 22.92%，拟合优度也只能达到 55.32%。由此可见，模型还需要进一步通过优化超参数来提高估算性能。

Table 2. The performance of BP neural network model under different number of virtual data generation samples

表 2. 不同虚拟数据生成样本数量下的 BP 神经网络模型的性能

虚拟生成的样本数量(条)	50	100	300	500	1000	3000	5000	10000
训练集误差	0.13%	0.12%	0.25%	0.01%	0.10%	0.00%	0.00%	0.00%
测试集误差	41.16%	38.70%	30.63%	25.75%	25.22%	23.57%	22.92%	23.36%
训练集拟合优度	99.52%	99.47%	99.43%	99.37%	99.38%	99.48%	99.47%	99.48%
测试集拟合优度	34.60%	37.87%	47.46%	53.47%	53.67%	54.50%	55.32%	54.78%

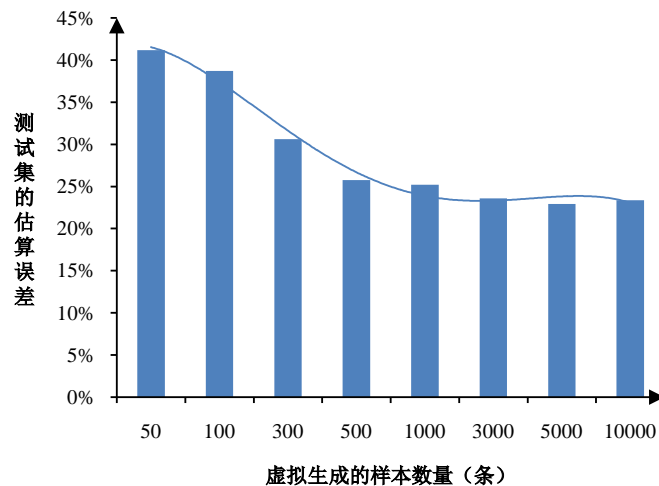


Figure 1. The improvement effect of different numbers of virtual generated samples on model performance

图 1. 不同虚拟生成样本数量对模型性能的提升效果

4.2.2. BP 神经网络的超参数

为了评估 BP 神经网络的超参数选择对模型估算性能的影响, 本文保持虚拟数据生成的样本数量为 5000 条, 分别在 BP 神经网络的不同隐藏层数 X 的情况下进行了超参数优化, 超参数优化的对象是学习速率 lr 和每个隐藏层中的神经元数量 hl_x 。BP 神经网络的相关研究表明, 隐藏层数越多模型的性能越好, 但模型的训练难度和训练时间也会大幅提高。因此, 隐藏层数 X 的取值通常在 3 以内。本文分别令 X 等于 1、2 和 3, 模型的性能优化结果见表 3。观察表 3 可知, 在不同的隐藏层数下, 模型对训练集的估算误差都可以达到 0.00%, 对测试集的估算误差也都能控制在 15% 以内, 对两类数据集的拟合优度均能达到 90% 以上。由此可见, 对 BP 神经网络的超参数进行优化可以显著提高模型的估算性能。当隐藏层层数 X 等于 3 时, 测试集上的估算误差最小, 但程序运行的时间也最长。因此, 在实际的应用中, 应当根据用户可接受的估算时间和估算误差来选择最适合的模型超参数。

Table 3. Bayesian parameter optimization results

表 3. 贝叶斯参数优化结果

隐藏层层数	1 层	2 层	3 层
最优 BP 网络结构	1-1-1	1-3-1-1	1-2-16-17-1
最优学习率 lr	0.11336	0.02069	0.23793
训练集估算误差 ε_1	0.00%	0.00%	0.00%
测试集估算误差 ε_2	13.39%	9.06%	7.92%
训练集拟合优度 R_1^2	91.43%	91.71%	90.11%
测试集拟合优度 R_2^2	94.39%	94.68%	94.32%
运行时间(秒)	16.33	19.62	31.27

4.2.3. 估算结果的比较

根据上述分析, 当 BP 神经网络的隐藏层层数 X 取 3 时, 隐藏层最优的神经元个数为: 2-16-17, 最优的学习率为 0.23793。基于以上参数设置, 模型在训练集和测试集上的平方损失随迭代次数的变化曲线

见图 2。观察可知，随着训练过程中迭代次数的增加，模型的平方损失越来越小，当迭代次数大于 100 之后，模型已趋于收敛。这表明 BP 神经网络的训练速度较快，估算效率较高。此外，为了评估建立的 BP 神经网络模型的估算性能，本文还采用了其他三种常用的参数法成本估算模型进行了估算，分别是：多元线性回归模型、偏最小二乘回归模型和指数回归模型。这四种不同的成本估算模型的估算性能结果见表 4。对比可知，本文结合虚拟数据生成方法建立的 BP 神经网络模型的估算误差是最小的，其在测试集上的估算误差可以控制被到 7.92%，拟合优度高达 94.32%。

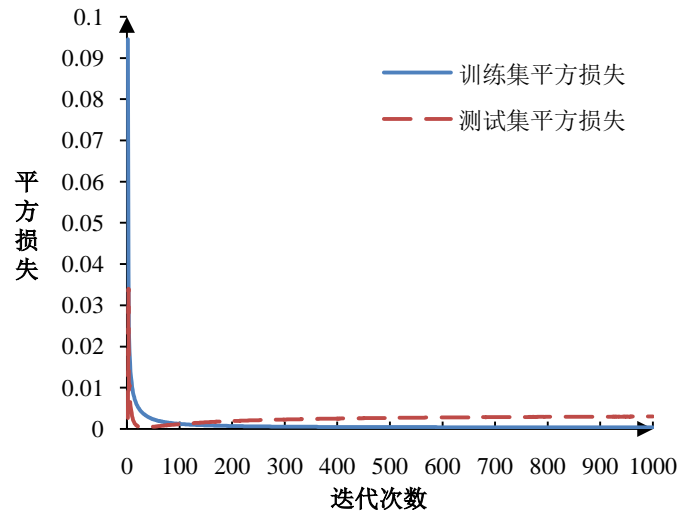


Figure 2. Variation curve of square loss in BP neural network model with the number of iterations

图 2. BP 神经网络模型中的平方损失随迭代次数的变化曲线

Table 4. Comparison of performance with other three commonly used parameter method models

表 4. 与其他三种常用参数法模型性能的对比如

模型种类	训练数据集		测试数据集	
	估算误差 ε	拟合优度 R^2	估算误差 ε	拟合优度 R^2
虚拟数据生成 + BP 神经网络	0.00%	90.11%	7.92%	94.32%
多元线性回归模型	18.46%	99.12%	19.71%	62.41%
偏最小二乘回归模型	20.40%	98.04%	15.90%	63.25%
指数回归模型	17.11%	93.00%	20.35%	60.53%

5. 结论与展望

航空发动机的设计方案一旦定型，其制造成本就已基本确定。因此，发动机主制造商需要在设计阶段就估算出发动机的制造成本，以便于有针对性地改进设计方案。然而，我国的民用航空发动机行业尚处于起步阶段，在成本估算工作中可参考的历史成本数据很少。本文首先提出了一种虚拟数据生成方法，用于在贫数据条件下对成本数据集进行扩充，从而克服了成本数据量少对成本估算工作的制约问题。然后，将提出的虚拟数据生成方法与 BP 神经网络相结合，建立了一个制造成本估算模型并采用贝叶斯优化算法对模型的超参数进行了优化。算例分析结果表明：本文提出的虚拟数据生成方法可以显著提高模型的估算精度，建立的估算模型的性能也显著优于三种常用的参数法模型。经过超参数优化之后，模型

对测试数据集的估算误差为 7.92%，拟合优度为 94.32%。本文的研究成果可为民用航空发动机设计方案优化中的制造成本控制工作提供方法支撑，还能够为航空发动机的产品定价和市场营销等决策提供依据。在民用航空发动机成本估算的实际工作中，发动机的构型、材料选用等因素也会对成本造成较大影响，在后续的研究中，还需要增加考虑发动机性能参数之外的因素，使成本估算模型更便于航空发动机制造企业直接应用。

基金项目

本文受航空动力基础研究项目资助。

参考文献

- [1] 刘若泠. 弹用发动机研制、生产成本分析及估算方法研究[D]: [硕士学位论文]. 南京: 南京航空航天大学, 2013.
- [2] 刘芳, 张海涛. 小样本理论在航空发动机研制费估算中的应用[J]. 北京航空航天大学学报, 2014, 40(11): 1518-1525.
- [3] 韩朋. 基于航空发动机性能的维修成本管理与控制研究[D]: [硕士学位论文]. 南京: 南京航空航天大学, 2009.
- [4] 赵曼. 飞机设计中的制造与维修成本方法研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2012.
- [5] 张珊珊. 基于深度学习的虚拟大气环境数据生成方法[D]: [硕士学位论文]. 哈尔滨工业大学, 2020.
- [6] 杨壮. 面向 Bin Picking 的虚拟数据集构建及智能识别方法的研究[D]: [硕士学位论文]. 华东理工大学, 2019.
- [7] Bandalos, D.L. and Finney, S.J. (2018) Factor Analysis: Exploratory and Confirmatory: the Reviewer's Guide to Quantitative Methods in the Social Sciences. Routledge, London, 98-122. <https://doi.org/10.4324/9781315755649-8>
- [8] Zhou, Q.H., Wu, G.Z., Wang, Z.X., et al. (2021) Analysis and Prediction of the Width of Spreading Carbon Fiber Tow Based on Gray System Theory. *Journal of Applied Polymer Science*, **138**, Article ID: 50069. <https://doi.org/10.1002/app.50069>
- [9] Li, J., Cheng, J.-H., Shi, J.-Y., et al. (2012) Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. *Advances in Computer Science and Information Engineering*. Springer, Berlin, Heidelberg, 553-558. https://doi.org/10.1007/978-3-642-30223-7_87
- [10] Cui, K. and Jing, X. (2019) Research on Prediction Model of Geotechnical Parameters Based on BP Neural Network. *Neural Computing and Applications*, **31**, 8205-8215. <https://doi.org/10.1007/s00521-018-3902-6>
- [11] Sharma, S., Sharma, S. and Athaiya, A. (2017) Activation Functions in Neural Networks. *Towards Data Science*, **6**, 310-316. <https://doi.org/10.33564/IJEAST.2020.v04i12.054>
- [12] 方昌德. 世界航空发动机手册[M]. 北京: 航空工业出版社, 1996.
- [13] 胡晓煜. 世界中小型航空发动机手册[M]. 北京: 航空工业出版社, 2006.
- [14] 黄维娜, 李中祥. 国外航空发动机简明手册[M]. 西安: 西北工业大学出版社, 2014.