文章编号：2095-4980(2023)10-1257-06

电力客户需求高适配性关联抽取算法

潘　晖，赵　岩，李　麟，徐　可，李景顺

(广西电网有限责任公司 南宁供电局，广西 南宁 532000)

摘　要：为了准确、高效地分析电力客户需求，从而降低电力企业成本，提高电力服务的产品附加值，基于层次分析法，计算条件属性重要度，构建优先关系矩阵，结合模糊关系判断尺度，确定电力客户需求权重。度量决策树节点纯度，分别对离散型节点变量与连续型节点变量进行指标分析，判断电力客户需求权重的准确性。建立电力客户需求关联抽取模型，获取电力客户需求用户画像，将信息区分值作为区分变量能力强弱的指标，计算不同变量之间的相关系数，设计关联抽取算法，得到电力客户关联结果。该方法在高、中、低3种频率中，虽其平均绝对百分比误差(MAPE)值不断升高，且随着关联层次的增加而逐渐递增，但整体依旧较低，判断电力客户需求权重的准确性较高。

关键词：层次分析；决策树算法；电力客户需求分析；高适配性；关联抽取

中图分类号：TB497　　**文献标志码：**A　　**doi：**10.11805/TKYDA2023055

High adaptability association extraction method of power customer demand based on analytic hierarchy process and decision tree

PAN Hui，ZHAO Yan，LI Lin，XU Ke，LI Jingshun

(Nanning Power Supply Bureau of Guangxi Power Grid Co.，Ltd，Nanning Guangxi 532000，China)

Abstract：In order to accurately and efficiently analyze the needs of power customers，thereby reducing the costs of power enterprises and increasing the added value of power service products，based on Analytic Hierarchy Process(AHP)，the importance of conditional attributes is calculated，a priority relationship matrix is constructed，and the weight of power customer demand is determined by combining with fuzzy relationship judgment scales. The purity of decision tree nodes is measured and the indicator analysis is conducted on discrete and continuous node variables to determine the accuracy of power customer demand weights. A correlation extraction model is established for power customer demand，and a user profile is obtained. Taking the information differentiation values as the indicators of variable differentiation ability，the correlation coefficients between different variables are calculated. By designing correlation extraction algorithms，the power customer correlation results are obtained，and a user profile is got. Taking the information differentiation values as the indicators of variable differentiation ability，the correlation coefficients between different variables are calculated. By designing correlation extraction algorithms，the power customer correlation results are obtained. Among high，intermediate and low frequencies，the Mean Absolute Percentage Error(MAPE) values of this method are 87.3%，71.9%，and 54.1%，respectively. In intermediate-frequency customer data，the MAPE of this method is increased from 62.1% to 71.9%；in low-frequency customer data，MAPE is increased from 42.2% to 54.1%. This method has a good correlation effect.

Keywords：Analytic Hierarchy Process；decision tree algorithm；power customer demand analysis；high adaptability；association extraction

通过智能用电项目，实现节能减排的目标，；不断优化电力企业对区域内的供电服务质量，保证电网运营高效率发展。智能用电项目取得了阶段性成效，提倡以用户为中心，将电力服务作为产品的附加值，在提升队内

收稿日期：2023-03-04；**修回日期：**2023-07-21

业务支撑的同时，利用电力用户的画像与标签体系，构造不同类型的用电负荷群体特征，总结用电规律。文献[1]结合不同时段的居民用电情况，构建了一种电力用户行为画像的方法，并使用可视化的方式将其实现。文献[2]首先建立了一种新的画像模型评价指标，并在该指标体系的基础上，针对用户价值等级无法被准确预测等问题，设计了一种优化的数据驱动画像方法，即基于相似性和聚类的推荐系统模型(Similarity-based and Clustering-based Recommender System model，SC-RS)。联合网格化的搜索策略，针对粗糙集理论，集成已有的三维规则框架，基于系统依赖度完成对规则的挖掘与整理。文献[3]以电力用户画像技术为最优聚类，提出了一种基于信息增益的用户画像技术，结合相关系数，在特征集适应度评价分析算法下，得到迭代解。

本文结合上述文献，梳理了不同类型电力客户的用电特征，设计了一种基于层次分析与决策树的电力客户需求高适配性关联抽取方法。

1 基于层次分析法确定电力客户需求权重

想要获取电力客户的关联指标，需要首先使用层次分析法确定不同电力客户在信息系统内的权重值。以粗糙集属性为基础，删除其中的冗余条件，确定客户需求决策[4-5]。在决策表 $B=(K,D)$ 中，K 表示决策项目，D 表示客户的满意度。计算条件属性重要度：

$$Z_m(p)=\frac{|Y_m(p)|}{|K|} \tag{1}$$

式中：$Y_m(p)$ 为客户满意度在客户需求子项目中的正域解；$Z_m(p)$ 为条件属性重要度。在经过归一化处理后，可以获取客户需求的对应条件[6]。应用层次分析方法，获取电力客户的定量系数，使其可以获取具备包容效果的影响因子，同时构建优先关系矩阵 $\boldsymbol{H}$：

$$\boldsymbol{H}=\left(h_{ij}\right)_{m\times n}=\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nm} \end{bmatrix} \tag{2}$$

式中：$\left(h_{ij}\right)_{m\times n}$ 为 $\boldsymbol{H}$ 的 $m\times n$ 项矩阵；且满足 $0\leqslant h_{ij}\leqslant 1$，$h_{ij}+h_{ji}=1$ 两个条件[7]。在此基础上，建立客户需求的模糊隶属关系判断尺度，如表 1 所示。

表1 模糊关系判断尺度表

Table1 Scale table of fuzzy relationship judgment

serial number	scale	guideline
1	0.100	compared to element *i*, element *j* is extremely important
2	0.138	compared to element *i*, element *j* is very important
3	0.325	compared to element *i*, element *j* is more important
4	0.439	compared to element *i*, element *j* is somewhat important
5	0.500	the element *i* is equally important as the element *j*
6	0.561	compared to element *j*, element *i* is somewhat important
7	0.675	compared to element *j*, element *i* is more important
8	0.862	compared to element *j*, element *i* is very important
9	0.900	compared to element *j*, element *i* is extremely important

与表 1 中的判断尺度相比，需要进一步构造模糊判断一致性矩阵。设二者之间的互补求和特性为：

$$f_{\mathrm{p}}=\sum_{i=1}^{n}h_{ij} \tag{3}$$

式中：f_{p} 为矩阵求和数值；h_{ij} 为矩阵中某一元素的值[8-9]。在模糊一致性矩阵的基础上，对各行各列进行排序，并得到向量 $\boldsymbol{\omega}=[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n]^{\mathrm{T}}$，则第 i 个向量满足：

$$\boldsymbol{\omega}_i=\frac{\sum_{i=1}^{n}h_{ji}+\dfrac{n-2}{2}}{n(n-1)} \tag{4}$$

式中：$\boldsymbol{\omega}_i$ 为第 i 个向量的值；n 为矩阵中元素的个数。根据式(4)可将 $\boldsymbol{\omega}_i^2$ 作为电力客户的需求权重。

2　度量决策树节点纯度

为了保证电力客户的需求权重具备可信度，可以在决策树算法中，进一步判断分支节点的所有样本是否属于同一类别。在验证时，需要保证所有变量属性具备唯一性，其一致性越高，各分支节点的纯度就越高，其节点部位的电力客户需求适配性就越强[10]。因此可以建立离散型节点变量的分类指标，定义纯度降低的差额：

$$\Delta p = p(0) - \left[\frac{n_1}{n_0} p(1) + \frac{n_2}{n_0} p(2) + \cdots + \frac{n_i}{n_0} p(i) \right] \tag{5}$$

式中：Δp 为离散型节点变量中纯度下降的差值；$p(0)$ 为父节点的纯度值；$p(i)$ 为第 i 个子节点的纯度值；n_0 为父节点的分类系数；n_i 为第 i 个子节点的分类系数[11-12]。在各节点的衡量指标中，信息熵值的变化量是最基本的数值，可以表示为：

$$K\left(p_1, p_2, \cdots, p_n\right) = -\sum_{i=1}^{n} p_i \log_2\left(p_i\right) \tag{6}$$

式中：$K\left(p_1, p_2, \cdots, p_n\right)$ 为 n 个子节点的信息熵值变化量；p_i 为第 i 个分类变量的节点概率。在该数值中，信息的纯度一般与取值大小无关，仅仅与熵值有关[13]。而在连续性节点变量中，节点的纯度降低指标则可以表示为：

$$m(i) = \frac{\sum_{i=1}^{n} \left(h_i - \bar{h}_i\right)^2}{n_i} \tag{7}$$

式中：$m(i)$ 为连续性节点下第 i 个子节点的纯度值；h_i 和 $\bar{h}_i$ 分别为第 i 个节点的实际数值与观测数值；n_i 为第 i 个子节点的观测点数量[14]。通过对上述两类节点的纯度度量，可以确定这些权重指标相似的电力客户是否能够被归类为同一种需求类型。

3　建立电力客户需求关联抽取模型

3.1　电力客户需求用户画像

随着大数据应用技术的发展与创新，很多行业均开始使用挖掘技术，对客户的基础数据进行分析。随着用户数据的应用越来越广泛，很多研究人员开始不再期望数据的挖掘精确度，而是开始将数据进行杂交处理。在数据量越来越多的基础上，从混合的信息中提取有效信息，这就构成了“用户画像”的概念。在构建用户画像的过程中，除必要的数据预处理外，还需要对其进行单变量分析[15]。将信息区分值作为区分变量能力强弱的指标：

$$Q_{\mathrm{iv}} = \sum_{j=1}^{n} \left(\frac{y_j}{y_n} - \frac{p_j}{p_n} \right) \times \ln \left(\frac{\dfrac{y_j}{y_n}}{\dfrac{p_j}{p_n}} \right) \tag{8}$$

式中：Q_{iv} 为信息区分值，该数值越大，证明用户的分布差异越大；y_j 和 y_n 分别为第 j 个变量对应的坏客户数量与坏客户整体数量；p_j、p_n 为第 j 个变量对应的好客户数量以及好客户的整体数量[16]。当 $Q_{\mathrm{iv}} \leqslant 0.03$ 时，该数值无法解释电力客户的变量区分问题；当 $Q_{\mathrm{iv}} \geqslant 0.5$ 时，表示该数据的区分度极高，此时即可选择变量模型。在多变量分析中，也可以假设样本 $\left(m_i, n_i\right)$，并计算不同变量之间的相关系数：

$$per = \frac{\sum_{i=1}^{n} \left(\dfrac{m_i - \bar{m}}{f_x} \right) \left(\dfrac{n_i - \bar{n}}{f_y} \right)}{r - 1} \tag{9}$$

式中：per 为电力客户需求样本之间的相关系数；m_i 和 n_i 分别为 2 个电力客户需求样本；$\bar{m}$ 和 $\bar{n}$ 为变量样本的均值；f_x 和 f_y 为样本方差；r 为样本总量。在上述标准模型的基础上，可以对不同样本进行划分，并得到电力客户的需求画像。

3.2　设计关联抽取算法

结合上文中的内容，可以获取一种基于层次分析与决策树的电力客户需求高适配性关联抽取算法，算法流

程如图 1 所示。首先需要建立数据样本集，该样本集中收录了大量的电力客户信息。计算不同电力客户的需求权重，选择特征属性，建立决策树训练集。在训练集内设置一个最大的估计值，并对其进行分类处理：

$$y_n = \arg\max R(x,y) \tag{10}$$

式中：y_n 为分类结果；$R(x,y)$ 为训练器的估计值。根据该分类器得到的结果，对任意数据集进行估计，并得到局部与整体的最优关联抽取结果。

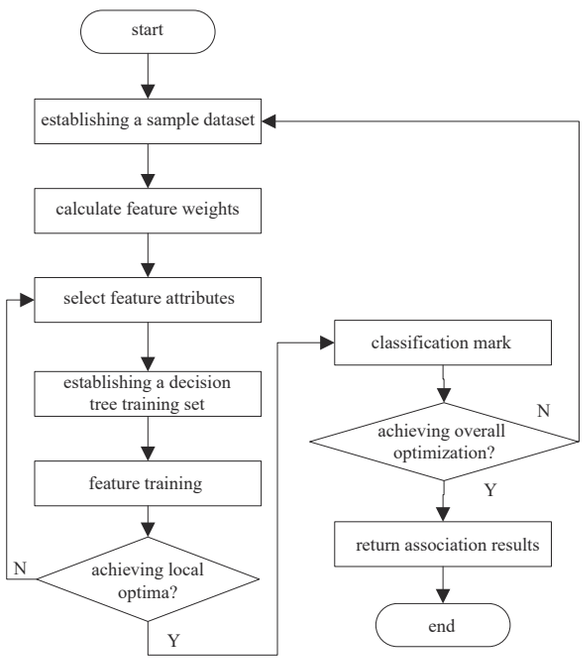


Fig.1 Power customer association algorithm

图 1 电力客户关联算法

4 实验研究

4.1 数据预处理

为测试上文中设计的基于层次分析与决策树的电力客户需求高适配性关联抽取方法的有效性，设计如下实验进行验证。选择某居民区的电力数据作为实验材料，在这些数据的基础上，首先需要建立数据挖掘库，结合电力数据分析电力客户的用电规律，并使用上文中所述的方法，对被测对象进行深入挖掘与分析。在数据挖掘库中，采集与抽取电能数据，包括电力负荷数据、电力营销数据等，这些数据均可以通过数据接口获取。一些无法从公共渠道得到的数据，也可以从 Excel 表格中导入进来。在数据的预处理过程中，首先需要进行数据清洗工作。数据清洗一般为对异常数据的处理，在整理历史资料的过程中，经常会发现此类异常数据。有些异常数据是人为造成的异常，有些数据是突发事故或其他因素导致的非规律性变化。对于前者，需要予以清除，对于后者，则需要在与其他数据对比之后进行保存。对于遗漏数据的清除与处理，需要使用手工填补的方式，对其进行补全。利用该研究区域内用电客户的数据指标，对产业类型进行分类，经过分析与整理后，取平均值。在处理噪声数据和不一致数据时，可以进行针对性的来源查找，重点核查有价值的基础数据，对于无关紧要的数据，则需要直接清除。在数据集成过程中，需要考虑到模式集成、数据冗余、数据冲突等多方面的问题。对于数据源不同的数据，需要对其尽快进行实体识别，经过多个系统的编码组合之后，确定同一实体计量单位的资产，并将各个系统的元数据进行同步抽取与整合，保证数据库中数据属性的一致与完整。如果其中一个属性的数据呈现出溢出的状态，则该属性可以通过供电量与售电量之间的差值获取。经过核查，可以针对不同数据的存储格式，获取影响系统内的用电信息，并在结构化查询语言(Structured Query Language，SQL)的查询下，保证系统数据的唯一性。在进行数据转换的过程中，最重要的一个步骤就是数据的规格化处理，其具体的映射公式为：

$$f' = \frac{f_k - \min P}{\max P - \min P}\left(New_{\max P_a} - New_{\min P_a}\right) \tag{11}$$

式中：f' 为数据规格化处理的映射函数；f_k 为原始数据；P 为规格属性；$New_{\max P_a}$ 和 $New_{\min P_a}$ 分别为映射后属性集合内数据的最大值与最小值。除规格化处理外，还需要使用零均值的方式，获取映射值，公式为：

$$f' = \frac{f_d - \bar{P}_d}{\lambda_d} \tag{12}$$

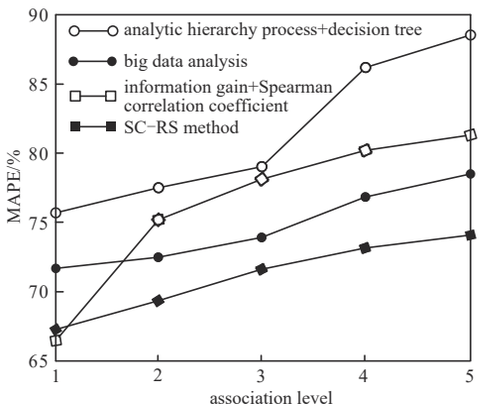
式中：f_d 为数据集 D 中的原始电力客户数据；$\bar{P}_d$ 为映射属性的均值；λ_d 为映射方差。使用上述方法获取点负荷数据的平均值后，还需要进行属性构造，获取不同数据之间的联系。在数据冲突消减过程中，可以通过立方体聚集的方式，对不同类别的用电类型进行统计与规范。此时可以使用维度消减、数据压缩等方式，采用主元素分析的方式，对数据进行压缩处理。依据用户阈值，去除重要性相对较低的共轭向量。并将多维数据转换成两维数据。

4.2 不同集合层次上客户关联建模性能

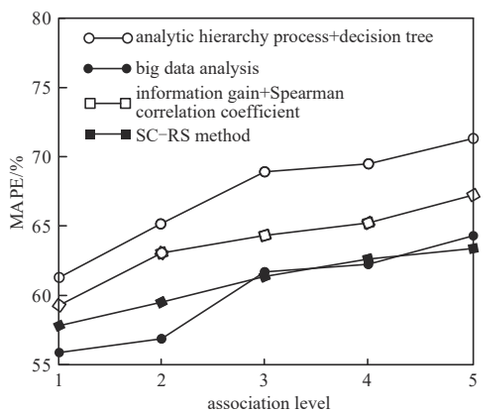
将不同类型的客户进行关联适配，达成用户需求响应的配电特征，从而实现高效率的能源利用，为用户提

供具备差异化的电力服务。将电力客户的需求层次细分为5个不同的阶段，并作为X轴坐标。将平均MAPE值作为Y轴坐标，表示客户关联的准确率，针对3个不同频率的客户进行划分。并将其与3种传统的客户关联匹配算法进行对比，关联模型性能的结果如图2所示。

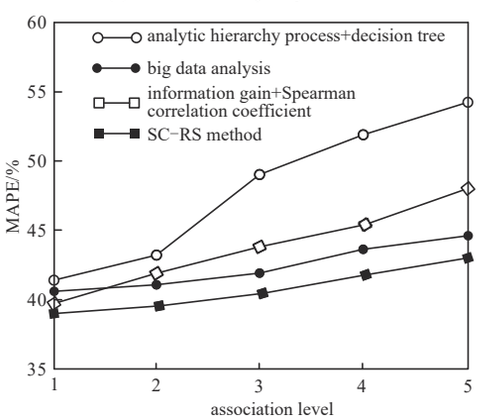
从图中可以看出，在3种不同频率的客户类型中，4类电力客户需求关联抽取方法的准确率均随着关联层次的增加而逐渐递增。在高频率客户中，4种算法在第一层关联中的MAPE分别为75.7%、71.9%、66.8%、67.4%，当关联层次为5时，4种算法的MAPE增长至87.3%、76.9%、71.6%、73.1%；在中频率客户数据中，本文算法的MAPE由62.1%增长至71.9%；在低频率客户中，本文算法的MAPE则由42.2%增长至54.1%。在上述3种数据类型中，本文方法的关联效果均优于其他3种对比方法，可见其关联结果优越。



(a) high frequency customer data



(b) intermediate frequency customer data



(c) low frequency customer data

Fig.2 Correlation effect prediction curves at different set levels
图2 不同集合层次上关联效果预测曲线

5　结论

本文设计了一种基于层次分析与决策树算法的电力客户需求高适配性关联抽取方法。该方法可以实现对电力客户需求的分类建模，并设计相应的关联抽取算法。在需求关联模型中，不同集合层次的关联效果会逐渐增加，关联抽取方法的应用效果最优。

本文算法还存在一些亟待完善的地方，在下一步的研究中，可以结合用户标签，对变量模糊的地方进行明确划分，以便进一步提高用电客户的需求关联性。

参考文献：

[1] 王永明,陈宇星,殷自力,等. 基于大数据分析的电力用户行为画像构建方法研究[J]. 高压电器, 2022,58(10):173-179,187. (WANG Yongming, CHEN Yuxing, YIN Zili, et al. Research on construction method of power user behavior portrait based on big data analysis[J]. High Voltage Apparatus, 2022,58(10):173-179,187.) doi:10.13296/j.1001.

[2] 余顺坤,闫泓序,杜诗悦,等. 基于SC-RS的我国工业电力用户价值画像模型研究[J]. 中国管理科学, 2022, 30(3): 106-116. (YU Shunkun, YAN Hongxu, DU Shiyue, et al. Research on the customer value portrait model of industrial power enterprise in China based on spectral clustering technology and Rough set theory[J]. Chinese Journal of Management Science, 2022,30(3):106-116.) doi:10.16381/j.cnki.issn1003-207x.2021.1324.

[3] 王圆圆,白宏坤,王世谦,等. 基于信息增益与Spearman相关系数的电力用户行为画像[J]. 电力工程技术, 2022,41(4):220-228. (WANG Yuanyuan, BAI Hongkun, WANG Shiqian, et al. Power users' behavior portrait based on information gain and Spearman correlation coefficient[J]. Electric Power Engineering Technology, 2022,41(4):220-228.)

[4] 张瑞,耿泉峰,吕云彤,等. 基于互联网+背景下的智慧电力营业厅[J]. 太赫兹科学与电子信息学报, 2021,19(6):1114-1119. (ZHANG Rui, GENG Quanfeng, LYU Yuntong, et al. Intelligent electric power business hall based on the Internet[J]. Journal of Terahertz Science and Electronic Information Technology, 2021,19(6):1114-1119.)

[5] 张建华,杨俊晓,曹子傲,等. 增强隐性知识外显案例适配度的优化方法[J]. 科技管理研究, 2022,42(18):136-143. (ZHANG Jianhua, YANG Junxiao, CAO Ziao, et al. Optimization method enhancing adaptation degree of tacit knowledge explicit cases[J]. Science and Technology Management Research, 2022,42(18):136-143.)

[6] 王磊,刘洋,李文峰,等. 基于用电行为数字特征画像的电力用户两阶段分类方法[J]. 电力建设, 2022,43(2):70-80. (WANG Lei, LIU Yang, LI Wenfeng, et al. Two-stage power user classification method based on digital feature portraits of power

consumption behavior[J]. Electric Power Construction,2022,43(2):70-80.) doi:10.12204/j.issn.1000-7229.2022.02.009.

[7] 刘永红,严妍,王海宁. 基于层次分析和加权灰色关联分析的真无线耳机适配度评价[J]. 包装工程, 2021,42(24):153-160. (LIU Yonghong, YAN Yan, WANG Haining. True wireless stereo earphone fit design based on analytic hierarchy process and grey relational analysis[J]. Packaging Engineering, 2021,42(24):153-160.)

[8] 杨建平,向月,刘俊勇. 面向配电网投资决策的小样本关联规则自适应迁移学习方法[J]. 中国电机工程学报, 2022,42(16):5823-5834,6159. (YANG Jianping, XIANG Yue, LIU Junyong. Adaptive transfer learning of small sample correlation rules for distribution network investment decision[J]. Proceedings of the Chinese Society for Electrical Engineering, 2022,42(16):5823-5834,6159.)

[9] 陈勃昊,薛城,任俊,等. 基于德尔菲法的公共卫生事件中家庭服务需求指标体系构建[J]. 复旦学报(医学版), 2022,49(1):60-65,72. (CHEN Bohao, XUE Cheng, REN Jun, et al. The construction of the Family Service Needs Index System(FSNIS) in public health events based on Delphi method[J]. Fudan University Journal of Medical Sciences, 2022,49(1):60-65,72.) doi:10.3969/j.issn.1672-8467.2022.01.008.

[10] 陆晓,徐春雷,冷钊莹,等. 基于数据驱动方法的疫情阶段电力用户负荷特性画像模型[J]. 电力建设, 2021,42(2):93-106. (LU Xiao, XU Chunlei, LENG Zhaoying, et al. Load characteristic portrait model of power users in epidemic stage applying data-driven method[J]. Electric Power Construction, 2021,42(2):93-106.) doi:10.12204/j.issn.1000-7229.2021.02.012.

[11] 闫泓序,余顺坤,林依青. 我国工业电力用户价值画像模型构建与应用研究[J]. 中国管理科学, 2021,29(10):224-235. (YAN Hongxu, YU Shunkun, LIN Yiqing. Research on the construction and application of the customer value portrait model of industrial power enterprise in China[J]. Chinese Journal of Management Science, 2021,29(10):224-235.)

[12] 胡珊,刘晶,王雨晴,等. 基于用户动态需求的产品迭代创新设计方法研究[J]. 现代制造工程, 2020(12):41-48. (HU Shan, LIU Jing, WANG Yuqing, et al. Research on product iterative innovation design method based on user dynamic demand[J]. Modern Manufacturing Engineering, 2020(12):41-48.)

[13] 王阳,裘乐淼,刘晓健,等. 基于关联约束网络的定制产品隐式需求激活技术[J]. 计算机集成制造系统, 2020,26(7):1855-1867. (WANG Yang, QIU Lemiao, LIU Xiaojian, et al. Implicit requirement activation for customized products based on association constraint network[J]. Computer Integrated Manufacturing Systems, 2020,26(7):1855-1867.) doi:10.13196/j.cims.2020.07.014.

[14] 杨勤,李尚泽,王卫星,等. 基于优化 Kano 分析的产品定位设计决策[J]. 机械设计, 2020,37(6):129-133. (YANG Qin, LI Shangze, WANG Weixing, et al. Product positioning design decision based on optimal Kano analysis[J]. Journal of Machine Design, 2020,37(6):129-133.)

[15] 王利利,张琳娟,许长清,等. 能源互联网背景下园区用户画像及成熟度评价模型研究[J]. 中国电力, 2020,53(8):19-28. (WANG Lili, ZHANG Linjuan, XU Changqing, et al. Research on park users portrait and maturity evaluation model under the background of energy internet[J]. Electric Power, 2020,53(8):19-28.) doi:10.11930/j.issn.1004-9649.201912066.

[16] 李娜,唐东芳. 基于客户需求的个性化相册印制路径的优化[J]. 数字印刷, 2020(2):58-62,89. (LI Na, TANG Dongfang. Optimization of personalized photo album printing path based on customer demand[J]. Digital Printing, 2020(2):58-62,89.)

作者简介:

潘 晖(1978-),男,硕士,高级工程师,主要研究方向为控制工程.emial:Pan_Hui8791@tom.com.

赵 岩(1985-),男,硕士,高级工程师,主要研究方向为电力系统及自动化.

李 麟(1990-),男,学士,工程师,主要研究方向为工程管理.

徐 可(1991-),女,学士,工程师,主要研究方向为电气工程及其自动化.

李景顺(1985-),男,硕士,工程师,主要研究方向为电力系统及自动化.