文章编号: 1671-5896(2022)02-0269-06

电力物联网用户侧数据深度挖掘方法研究

颜远海
(广州华商学院 数据科学学院, 广州 511300)

摘要: 针对在电力物联网中用户侧数据处于相对孤立的位置, 导致数据关联规则挖掘难度增加的问题, 提出了基于关联规则映射的电力物联网用户侧数据深度挖掘方法。该方法以用户侧数据网状拓扑的有向图结构为基础, 根据关联属性组分析数据集的关联映射关系, 利用相互关系矩阵挖掘数据集的关联规则。引入极值规范化策略与径向基函数神经网络, 构建无量纲方法与离散聚类方法, 通过隐藏层神经元网络中心获取与连接权重计算等训练阶段, 按照 K 均值聚类流程完成数据预处理, 根据显性与隐性的不同用户侧数据类型以及用户-项目评分矩阵与兴趣度矩阵, 实现数据深度挖掘。实验结果表明, 该方法可以用较短的时间完成挖掘任务, 不同规模数据集处理效果更好, 且能在较小的内存空间内完成数据深度挖掘。

关键词: 关联规则; 关联映射; 电力物联网; 用户侧; 数据挖掘
中图分类号: TP391.44 **文献标识码:** A
DOI:10.19292/j.cnki.jdxxp.2022.02.003

Research on Deep Mining Method of User Side Data for Power Internet of Things

YAN Yuanhai
(College of Data Science, Guangzhou Huashang College, Guangzhou 511300, China)

Abstract: In the power Internet of Things, user-side data is in a relatively isolated position, which makes it more difficult to mine data association rules. Therefore, a deep mining method of user-side data in the power Internet of Things based on association rule mapping is proposed. Based on the directed graph structure of user-side data mesh topology, the association mapping relationship among data sets is analyzed according to the association attribute group. And the association rules among data sets are mined using the correlation matrix. Extreme value standardization strategy and radial basis function neural network are introduced, and the dimensionless method and discrete clustering method are built. Through the hidden layer neural network is obtained. According to the K-means clustering process, data preprocessing, data types according to the different users of dominant and recessive side matrix, score matrix and users-project scale, deep data mining is realized. Experimental results show that this method can complete the mining task in a relatively short time, the processing effect of different data sets is better, and the data depth mining can be completed in a small memory space.

Key words: association rules; association mapping; power Internet of things; user side; data mining

0 引 言

电力数据中蕴含着大量的有用信息[1], 随着电网信息化水平的不断提升, 在巨大规模数据集里挖掘有价值信息, 逐渐成为电网规划建设、增容改建、可靠运行的主要手段。计算机技术与信息技术高速发展, 推动了数据挖掘技术[2]在电力领域中的普及与应用, 使深层次的电网数据挖掘成为可能。

收稿日期: 2021-07-09
基金项目: 广东省普通高校人文社科 2017 年"创新强校工程"基金资助项目(2017KQNCX266)
作者简介: 颜远海(1985—), 男, 江西吉安人, 广州华商学院讲师, 主要从事数据可视化和数据分析算法研究, (Tel) 86-18924273591(E-mail) yan85028@163.com。

目前人们主要以确保电网平稳、安全运行进行数据挖掘研究，例如高翔等[3]与浦雨婷等[4]分别针对电力信息系统网络安全态势与电压暂降的评估问题，以数据挖掘为技术支撑，各自引入支持向量机与靶心度优化算法，建立了电力信息系统网络安全态势评估模型与电压暂降严重度评估模型，确保电网安全稳定运行；而郭阳等[5]则从电力企业内部管理角度出发，利用大数据挖掘技术，提取电力企业的影响因素历史特征，经聚类分析，构建了电力企业评价体系，为决策者提供可行的管理建议。

互联网技术的革新推动了电力系统与物联网的耦合发展，电网用户规模日益庞大使用户侧数据量持续上升，因此笔者针对电力物联网，设计一种新的用户侧数据深度挖掘方法。关联规则是较为普及且有效的一种数据规律发现策略，可使无法确切描述的信息都实现清晰展示，且可提升挖掘精准度。

1 用户侧数据关联规则映射

为降低用户侧数据特征的挖掘复杂度，以用户侧数据间的关联映射为依据，分析数据集的关联规则，以便快速、准确地挖掘出用户侧的用电量水平走势等相关电力数据，为用户用电特征提取、电力负荷预测提供参考。

假设电力物联网用户侧数据的网状拓扑结构可通过一个有向图[6] $G=(V,E)$ 表示，该图中各数据点及其连接线分别是 V、E，$V_i=(x_{1i},x_{2i},\cdots,x_{mi})$ 是 n 个数据点集合 $V=(V_1,V_2,\cdots,V_n)$ 中的一个数据集，i 是 $1\sim n$ 之间的正整数，表示数据集序号，x_{ji}是数据集 V_i 中第 j 个有效的用户侧数据，j 的取值范围是 $1\sim m$ 之间的正整数。若第 i 个数据集 V_i 与第 k 个数据集 V_k 间存在一定的关联性，其大小关联、语义关联以及种类关联分别为 α_{ik}、β_{ik}和 θ_{ik}，则采用关联属性组$(\alpha_{ik},\beta_{ik},\theta_{ik})$ 界定该关联程度，则 k 取值范围同 i。

基于上述设定条件与关联映射关系，得出以下结论。

1) 数据集 V_i、V_k 的关联属性组$(\alpha_{ik},\beta_{ik},\theta_{ik})$ 可以描述两数据集内任意数据间的关联程度。

2) 对关联属性组$(\alpha_{ik},\beta_{ik},\theta_{ik})$，可通过下列关联系数矩阵形式完成设置，则数据集 V_i 与 V_k 中所有数据间的关联程度均值如下

$$K_1\begin{pmatrix}\alpha_{ik}\\ \beta_{ik}\\ \theta_{ik}\end{pmatrix}=\begin{pmatrix}\alpha_{i1} & \cdots & \alpha_{1k}\\ \vdots & \ddots & \vdots\\ \alpha_{1k} & \cdots & \alpha_{i1}\end{pmatrix}\begin{pmatrix}\beta_{i1} & \cdots & \beta_{1k}\\ \vdots & \ddots & \vdots\\ \beta_{1k} & \cdots & \beta_{i1}\end{pmatrix}\begin{pmatrix}\theta_{i1} & \cdots & \theta_{1k}\\ \vdots & \ddots & \vdots\\ \theta_{1k} & \cdots & \theta_{i1}\end{pmatrix} \tag{1}$$

其中 K_1 是关联系数[7]。

3) 利用各关联属性倒数[8]描述集合区别，即下列差异系数矩阵

$$K_2\begin{pmatrix}\frac{1}{\alpha_{ik}}\\ \frac{1}{\beta_{ik}}\\ \frac{1}{\theta_{ik}}\end{pmatrix}=\begin{pmatrix}\frac{1}{\alpha_{i1}} & \cdots & \frac{1}{\alpha_{1k}}\\ \vdots & \ddots & \vdots\\ \frac{1}{\alpha_{1k}} & \cdots & \frac{1}{\alpha_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\beta_{i1}} & \cdots & \frac{1}{\beta_{1k}}\\ \vdots & \ddots & \vdots\\ \frac{1}{\beta_{1k}} & \cdots & \frac{1}{\beta_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\theta_{i1}} & \cdots & \frac{1}{\theta_{1k}}\\ \vdots & \ddots & \vdots\\ \frac{1}{\theta_{1k}} & \cdots & \frac{1}{\theta_{i1}}\end{pmatrix} \tag{2}$$

其中 K_2 是差异系数。

利用式(2)与关联系数矩阵[9]，推导出两数据集间的关联映射，其表达式如下

$$\begin{bmatrix}x_{1i}\\ x_{2i}\\ \cdots\\ x_{mi}\end{bmatrix}\rightarrow\begin{bmatrix}\frac{K_1}{K_2}x_{1k}\\ \frac{K_1}{K_2}x_{2k}\\ \cdots\\ \frac{K_1}{K_2}x_{mk}\end{bmatrix} \tag{3}$$

若要有效区分数据集 V_i 与 V_k，根据该关联映射关系，利用下列相互关系矩阵，联立出两数据集之间的关联规则

$$f(V_i,V_k)=\begin{pmatrix}V_1&\cdots&V_n\\ \vdots&\ddots&\vdots\\ V_n&\cdots&V_1\end{pmatrix}\begin{pmatrix}\alpha_{ik}\\ \beta_{ik}\\ \theta_{ik}\end{pmatrix}+\begin{pmatrix}V_1&\cdots&V_n\\ \vdots&\ddots&\vdots\\ V_n&\cdots&V_1\end{pmatrix}\begin{pmatrix}\frac{1}{\alpha_{ik}}\\ \frac{1}{\beta_{ik}}\\ \frac{1}{\theta_{ik}}\end{pmatrix} \tag{4}$$

针对区分出的数据集 V_i 与 V_k，再次结合关联映射关系，完成两数据集的区分，为后续从电力物联网中深度挖掘用户侧数据提供精准的数据支撑。

2　电力物联网用户侧数据深度挖掘

2.1　用户侧数据预处理

在电力物联网用户侧数据深度挖掘过程中，存在数据量级差异与连续属性不同值过多等不利因素，为解决该问题，引入极值规范化策略与径向基函数神经网络，构建出减小数据量级差异的无量纲方法与离散化处理多个连续属性的聚类方法。

为实现用户侧数据的无量纲处理，采用极值规范化策略线性变换初始的用户侧数据。已知数据属性 A 的极值分别是 m_A、x_A，则通过下列极值规范化表达式，在区间 $[n_{m_A},n_{x_A}]$ 内完成属性 A 值 v 的映射，得到映射值 v'，由此消除数据挖掘时因量级[10]差异而造成的偏差

$$v'=\frac{v-x_A}{m_A-x_A}(n_{m_A}-n_{x_A})+n_{x_A} \tag{5}$$

径向基函数神经网络[11]具有聚类能力，利用该网络分类数据的连续属性，将初始连续值替换为网络中心，令各属性变成离散值，这既有助于维持初始属性间的关联性，还能降低属性的赋值数量，凸显出连续数据属性规律。

径向基函数神经网络中的隐藏层神经元主要用于输入层的特征提取结果的非线性变换[12]，若选择高斯函数作为转换函数[13]，当第 k 个数据在位置是 i 的神经元上时，输出结果为 $x_{k,i}$，此时的数据向量是 $\boldsymbol{H}_k$，神经元的中心与系数为 $\boldsymbol{H}_i$、σ_i，则第 i 个神经元的转换函数表达式如下

$$x_{k,i}=\Gamma\left(\sqrt{(\boldsymbol{H}_k-\boldsymbol{H}_i)^{\mathrm{T}}(\boldsymbol{H}_k-\boldsymbol{H}_i)}\right)=\exp\left\{-\frac{1}{2}\left(\frac{\sqrt{(\boldsymbol{H}_k-\boldsymbol{T}_i)^{\mathrm{T}}(\boldsymbol{H}_k-\boldsymbol{H}_i)}}{\sigma_i}\right)^2\right\} \tag{6}$$

其中 exp 是 Lyapunov 函数[14]，Γ 是 Lyapunov-Krasovskii 泛函形式[15]。

神经网络输出层神经元的聚类过程：已知一个 n 维空间 R^n 中的数据集 V_i，经径向基函数神经网络划分为 n 个类别后，令数据类别及其网络中心之间的间距最短。

该神经网络训练过程由两个阶段构成：获取隐藏层神经元的网络中心；利用最小二乘法[16]求解隐藏层与输出层的连接权重。其中，各数据点的类别标记则通过输出层转换函数实现。由于篇幅限制，笔者仅描述第 1 阶段隐藏层神经元的网络中心获取流程。

利用 K 均值聚类算法[17]明确神经元网络中心的具体操作步骤描述如下：

1）初始化处理径向基函数神经网络中前几个训练数据的簇中心；

2）以目前簇中心为标准，聚类训练数据，求解各数据点与簇中心之间的欧几里得距离[18]，按最短距离将所有数据点划分至对应簇中心；

3）根据划分结果修正簇中心 T_j，得到更新后的簇中心 T_j'，若归属于簇中心 T_j 的数据点数量为 M_j，则采用

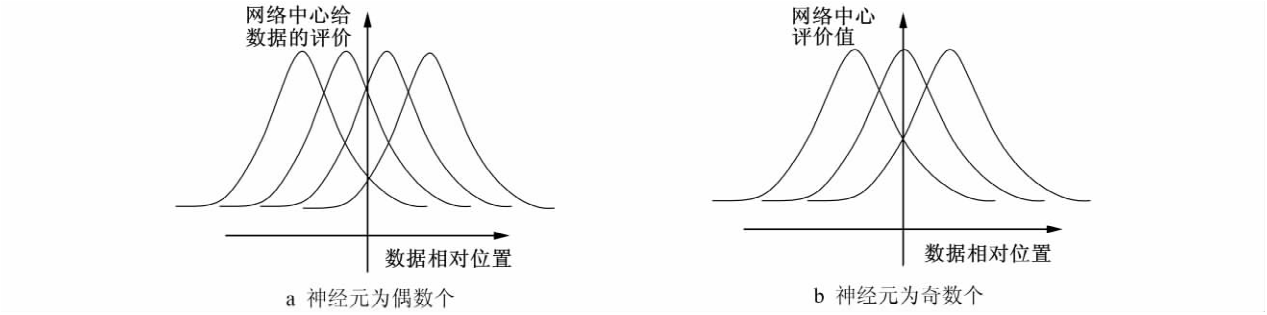
$$T_j'=\frac{\sum x_i}{M_j},\qquad \forall x_i\in T_j \tag{7}$$

实现簇中心更新；

4）返回第 2）步，直到簇中心没有变化后停止循环，得到最终的网络中心。

用户侧连续数据经过 K 均值算法聚类后，实现了属性离散化处理。若神经网络隐藏层的神经元数量

是偶数，则网络的数据评价分类形式如图1a所示；若是奇数，则如图1b所示。从图1可以看出，各数据点均会被分类至给予最高评价的网络中心，以此将数据点值替换为对应的网络中心，以离散化处理连续属性，令数据点取值范围为相应网络中心的集合。



a 神经元为偶数个　　b 神经元为奇数个

图1 数据评价分类示意图

Fig. 1 Schematic diagram of data evaluation classification

2.2 用户侧数据深度挖掘

在电力物联网中，用户侧数据分为显性数据与隐性数据两种。当对用户为交易的电力项目进行评分时，该评分就是用户侧的显性数据，挖掘此类数据有助于分析用户的电力偏好需求，以此设计相关对策提升电力物联网的交易数量。对目标用户及与其拥有相同偏好项目的邻域用户，利用用户-项目评分矩阵[19]，得到任意项目的用户兴趣度。已知 q 类别的项目总数是 N_q，被用户 p 评分的项目数量是 $M_{p,q}$，故兴趣度可由

$$I_{pq} = M_{p,q}\sum_{p=1} N_q \Big/ \left(N_q \sum_{p=1} M_{p,q}\right) \tag{8}$$

求出。为全面了解用户电力需求，需结合用户针对任意电力项目的网页浏览时长等隐性数据，深度挖掘用户侧数据。假设 t 是某项目网页的停留时长阈值，则用户的隐性兴趣度矩阵架构如下

$$\boldsymbol{I}' = \begin{bmatrix} (I_{11},T_{11}) & (I_{12},T_{12}) & \cdots & (I_{1q},T_{1q}) \\ (I_{21},T_{21}) & (I_{22},T_{22}) & \cdots & (I_{2q},T_{2q}) \\ \cdots & \cdots & \cdots & \cdots \\ (I_{p1},T_{p1}) & (I_{p2},T_{p2}) & \cdots & (I_{pq},T_{pq}) \end{bmatrix} \tag{9}$$

其中矩阵的列和行分别是电力项目类型与用户，I_{pq} 与 T_{pq} 是第 q 类项目第 p 个用户的兴趣度与浏览时长[20]。

根据显性与隐性的不同用户侧数据类型以及用户-项目评分矩阵与兴趣度矩阵，实现数据深度挖掘。

3 用户侧数据深度挖掘仿真实验

以由关联规则产生的真实数据集 mushroom_exp. db 与生成数据集 AT60_AP10. db 为基础，采用笔者方法进行电力用户侧数据挖掘仿真实验。

为验证方法时效性，测试不同用户支持度下笔者方法对两种数据集的挖掘效果。真实数据集 mushroom_exp. db 中包括近百个项目与近万条记录，生成数据集 AT50_AP8. db 中含有150多个项目产生的近5万条记录。对不同规模数据集进行10组深度挖掘模拟实验，计算执行时间均值，得到不同支持度下的两种数据集挖掘时长，如图2所示。

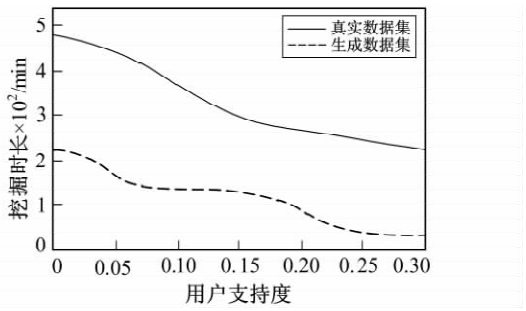


图2 不同数据集的支持度与挖掘时长相关性

Fig. 2 Correlation between the support of different data sets and the mining time

由图2可以看出，挖掘时长均随着用户支持度上涨而不断下降，与数据集规模无关，即使在支持度较小时，该方法仍然可以根据用户侧数据间的关联映射

关系，用较短的时间完成挖掘任务。

设置数据集的记录均长是 60，用户支持度是 0.3。利用 JVisualVM 性能监测工具记录的整体挖掘过程如图 3 所示。图 3 显示了笔者方法在挖掘不同规模数据集时内存占用、垃圾回收活跃度以及总堆大小、已用堆大小等情况。

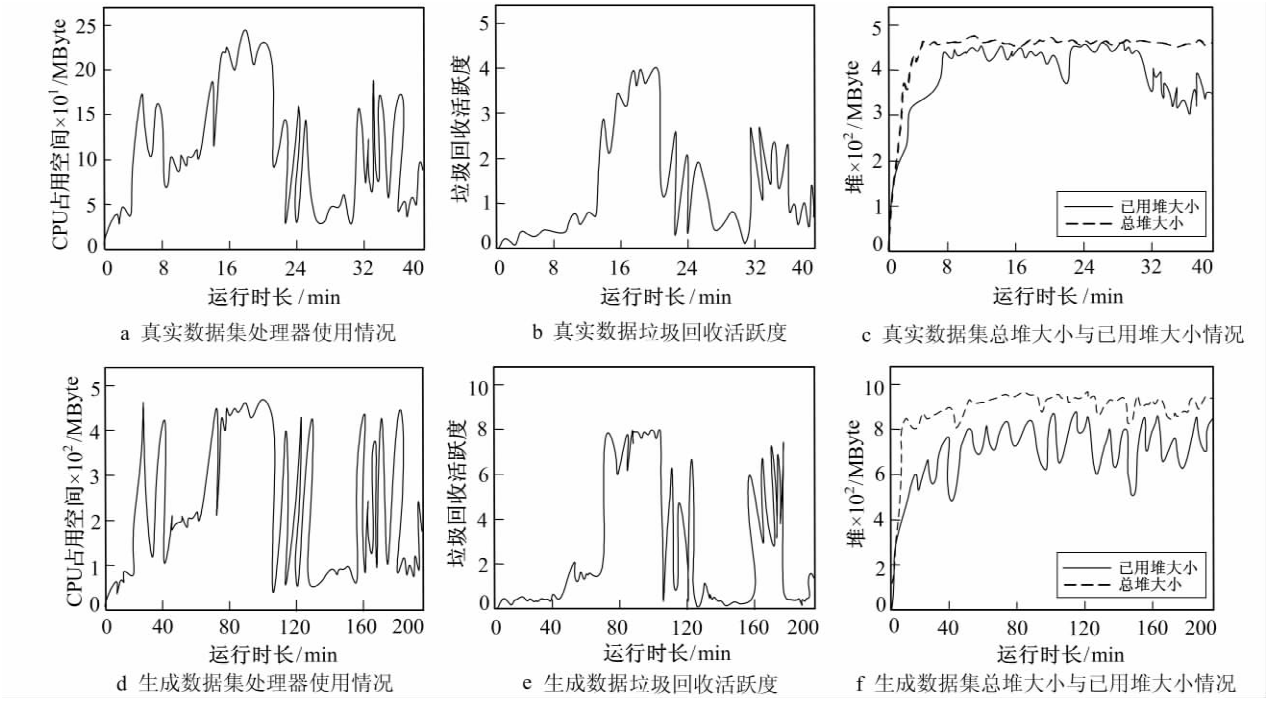


图 3　不同数据集的挖掘效果

Fig. 3　The mining effect of different data sets

根据图 3a ~ 图 3d 中的指标曲线走势可以看出，对处理器使用情况与垃圾回收活跃度而言，不管是两数据集的相同指标还是单一数据集的不同指标，曲线趋势均趋于近似。这表明数据量越大 CPU 占用空间越大，同时垃圾回收活跃度也越高。尽管在挖掘较大规模的生成数据集过程中，内存应用结果多次趋近于可用极值 512 MByte，但结合图 3f 堆的使用情况可知，已用堆大小并未超过总堆极值，这表明所提挖掘方法可以较为理想地处理不同规模数据集，且能在较小的内存空间内完成用户侧数据深度挖掘。这是因为笔者在设计过程中引入极值规范化策略与径向基函数神经网络，利用构建的无量纲方法与聚类方法，减小了数据量级差异，离散化处理了多个连续属性，且由于笔者方法建立了关联规则映射关系，防止数据集被过早分解，提升了数据处理性能。

4　结　语

随着智能电网日益普及，电力物联网日趋成熟，因此根据用户需求分析结果为用户提供优质的用电服务具有重要的意义。笔者根据数据集的关联规则映射相关性，深度挖掘电力物联网用户侧数据，以期为供电管理奠定决策依据。以优化数据挖掘方法性能为目标，可从以下几个方面做进一步探究：通过汇总大量更详实、更准确的用户侧数据；以此建立自动挖掘系统，实现挖掘自动化与数据可视化，进一步优化电网领域经济性。

参考文献：

[1] GUO P. Detection of Power Data Tampering Attack Based on Gradient Boosting Decision Tree [J]. Journal of Physics Conference Series, 2021, 1846(1): 012057-012069.

[2] ABUBAKER M. Data Mining Applications in Understanding Electricity Consumers´ Behavior: A Case Study of Tulkarm District, Palestine [J]. Energies, 2019, 12(22): 4287-4287.

[3] 高翔，陈贵凤，赵宏雷. 基于数据挖掘的电力信息系统网络安全态势评估[J]. 电测与仪表，2019，56(19)：102-106.
GAO Xiang，CHEN Guifeng，ZHAO Honglei. Power Information System Network Security Situation Assessment Based on Data Mining[J]. Electrical Measurement and Instrumentation，2019，56(19)：102-106.

[4] 浦雨婷，杨洪耕，马晓阳. 基于数据挖掘与改进灰靶的电压暂降严重度分析与评估[J]. 电力系统自动化，2020，44(2)：198-206.
PU Yuting，YANG Honggeng，MA Xiaoyang. Analysis and Evaluation of Voltage Sag Severity Based on Data Mining and Improved Gray Target[J]. Automation of Electric Power Systems，2020，44(2)：198-206.

[5] 郭阳，孔文佳，冯和宁，等. 大数据挖掘用于电力企业评价体系构建[J]. 数学的实践与认识，2019，49(8)：117-123.
GUO Yang，KONG Wenjia，FENG Hening，et al. Big Data Mining Used in the Construction of Evaluation System for Electric Power Enterprises[J]. Mathematics in Practice and Knowledge，2019，49(8)：117-123.

[6] 宋炎侃，陈颖，于智同，等. 基于同构有向图的电网多场景仿真GPU批量并行加速方法[J]. 电工电能新技术，2020，39(3)：17-23.
SONG Yankan，CHEN Ying，YU Zhitong，et al. GPU Batch Parallel Acceleration Method for Multi-Scenario Simulation of Power Grid Based on Isomorphic Directed Graph[J]. New Technology of Electrical Engineering and Energy，2020，39(3)：17-23.

[7] KHAN Z A，ADIL M，JAVAID N，et al. Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data[J]. Sustainability，2020，12(19)：1-25.

[8] SCHEUBEL C，MATTHUS D，FRIEDL G. The Impact of Industrial Self-Supply on Bavaria's Electricity System-Effects on Supply Security and Market Prices[J]. International Journal of Energy Sector Management，2019，13(2)：450-466.

[9] 林松，尹长明，孙晗. 计数数据广义估计方程相关系数矩阵估计的相合性[J]. 数学的实践与认识，2019，49(2)：298-303.
LIN Song，YIN Changming，SUN Han. The Correlation Coefficient Matrix Estimation of the Generalized Estimator of the Data is Estimated[J]. Mathematics in Practice and Theory，2019，49(2)：298-303.

[10] 杨懿男，齐林海，王红，等. 基于生成对抗网络的小样本数据生成技术研究[J]. 电力建设，2019，40(5)：71-77.
YANG Yinan，QI Linhai，WANG Hong，et al. Research on Small Sample Data Generation Technology Based on Generative Confrontation Network[J]. Electric Power Construction，2019，40(5)：71-77.

[11] FANG S，ZHANG Z，CHEN W，et al. 3D Crosswell Electromagnetic Inversion Based on Radial Basis Function Neural Network[J]. Acta Geophysica，2020，68(7)：711-721.

[12] MERINO A，GARCIA-ALVAREZ D，SAINZ-PALMERO G I，et al. Knowledge Based Recursive Non-Linear Partial Least Squares (RNPLS)[J]. ISA Transactions，2020，100(5)：481-494.

[13] 琚垚，祁林，刘帅. 基于改进乌鸦算法和ESN神经网络的短期风电功率预测[J]. 电力系统保护与控制，2019，47(4)：58-64.
JU Yao，QI Lin，LIU Shuai. Short-Term Wind Power Prediction Based on Improved Crow Algorithm and ESN Neural Network[J]. Power System Protection and Control，2019，47(4)：58-64.

[14] YOUNSI L E，BENZAOUIA A，HAJJAJI A E. Decentralized Control Design for Switching Fuzzy Large-Scale T-S Systems by Switched Lyapunov Function with H_∞ Performance[J]. International Journal of Fuzzy Systems，2019，21(4)：1104-1116.

[15] HUA C，WANG Y，WU S. Stability Analysis of Neural Networks with Time-Varying Delay Using a New Augmented Lyapunov-Krasovskii Functional[J]. Neurocomputing，2019，332(7)：1-9.

[16] LI R，CHEN X，BALEZENTIS T，et al. Multi-Step Least Squares Support Vector Machine Modeling Approach for Forecasting Short-Term Electricity Demand with Application[J]. Neural Computing and Applications，2021，33(1)：301-320.

[17] 陈洪涛，蔡慧，李熊，等. 基于k-means聚类算法的低压台区线损异常辨别方法[J]. 南方电网技术，2019，13(2)：2-6.
CHEN Hongtao，CAI Hui，LI Xiong，et al. Anomaly Identification Method of Low-Voltage Station Area Line Loss Based on k-Means Clustering Algorithm[J]. Southern Power Grid Technology，2019，13(2)：2-6.

[18] LI K，YANG R J，ROBINSON D，et al. An Agglomerative Hierarchical Clustering-Based Strategy Using Shared Nearest Neighbours and Multiple Dissimilarity Measures to Identify Typical Daily Electricity Usage Profiles of University Library Buildings[J]. Energy，2019，174(1)：735-748.

[19] 向小东，邱梓咸. 基于slope-one算法改进评分矩阵填充的协同过滤算法研究[J]. 计算机应用研究，2019，36(4)：1064-1067.
XIANG Xiaodong，QIU Zixian. Research on Collaborative Filtering Algorithm Based on Slope-One Algorithm to Improve the Filling of Score Matrix[J]. Application Research of Computers，2019，36(4)：1064-1067.

[20] ZHANG Y，MENG K，KONG W，et al. Bayesian Hybrid Collaborative Filtering-Based Residential Electricity Plan Recommender System[J]. IEEE Transactions on Industrial Informatics，2019，15(8)：4731-4741.

(责任编辑：刘东亮)