DOI:10.20033/j.1003-7241.(2025)09-0124-05

基于主成分聚类分析的电力用户信息自动集成方法

董俐君，岳 恒，李俊峰
（国网汇通（北京）信息科技有限公司，北京 100053）

摘 要：为了提高电力用户信息的集成效果，促进电力行业的数字化转型和智能化升级，研究提出一种基于主成分聚类分析的电力用户信息自动集成方法。首先通过主成分分析技术对用户信息进行降维处理，然后利用聚类分析算法对用户信息进行聚类，最后通过集成算法将各个类别的用户信息进行整合。研究在Glass数据集上进行实验评估。实验结果表明，基于主成分聚类分析的电力用户信息自动集成方法在Glass数据集的查准率、查全率和F_1值分别为81.27%、92.34%、92.55%，传输速率约为10 MB/s。该方法能够有效地提高电力用户信息的集成效果，为电力行业的信息管理提供新的思路和方法，推动电力行业的数字化转型和智能化升级。

关键词：主成分分析；聚类分析；电力用户信息；集成算法；电力信息管理；数字化转型
中图分类号：TP311.13 **文献标识码：**A **文章编号：**1003-7241(2025)09-0124-05

Automatic Integration Method of Power User Information Based on Principal Component Cluster Analysis

DONG Lijun, YUE Heng, LI Junfeng
(State Grid Huitongjincai (Beijing) Information Technology Co., Ltd., Beijing 100053, China)

Abstract: In order to improve the integration effect of power user information and promote the digital transformation and intelligent upgrade of the power industry, an automatic integration method of power user information based on principal component cluster analysis is proposed. Firstly, the dimensionality of user information is reduced by principal component analysis, then the user information is clustered by cluster analysis algorithm, and finally the user information of various categories is integrated by integration algorithm. The study is evaluated experimentally on the Glass dataset. The experimental results show that the accuracy, recall and F_1 values of the automatic integration method based on principal component cluster analysis in Glass dataset are 81.27%, 92.34% and 92.55%, respectively, and the transmission rate is about 10 MB/s. This method can effectively improve the integration effect of power user information, provide new ideas and methods for the information management of the power industry, and promote the digital transformation and intelligent upgrading of the power industry.

Keywords: principal component analysis; cluster analysis; power user information; integration algorithm; information management of power; digital transformation

0 引言

随着电力行业的发展和智能电网的建设，电力用户信息的筛选、整合和利用变得越来越重要。然而，由于电力用户信息的来源多样、格式各异，如何自动地进行信息集成成为一个挑战。传统的手工集成方法需要大量的人力和时间成本，且容易受到人为因素的影响。因此，需要一种自动的电力用户信息集成方法来提高效率和准确性[1-2]。主成分分析（principal component analysis，PCA）是一种常用的数据降维技术，通过线性变换将高维数据映射到低维空间。聚类分析是一种将数据对象划分为相似组别的技术。集成算法是一种通过组合多个模型来提高预测准确性的技术。陶永辉等提出了一种基于主成分分析的信息自动集成方法，研究结果表明，该方法能够有效地进行用户信息进行集成和分析[3]。陈龙谭等介绍了一种基于聚类算法的电力用户信息集成方法，研究发现，该方法可以对电力用户信息进行自动集成和分析，为电力公司提供有价值的用户行为特征[4]。Kleshchenko等提出了一种基于PCA的电力用户信息管理方法，研究结果表明该方法能够有效地提取电力用户信息中的关键特征，并对其进行分类，提高了电力用户信息的管理效率[5]。在本研究中，将结合这些技术来实现电力用户信息的自动集成。基于主成分聚类分析的电力用户信息自动集成方法是一种利用主成分分析和聚类分析相结合的方法，用于自动整合和分析电力用户的相关信息。研究利用主成分分析对电力用户信息进行降维，提取出最具代表性的

*基金项目：国家电网项目（717403219557）
收稿日期：2024-01-22

主要特征。将经过特征提取的数据进行聚类分析，将电力用户按照相似性进行分组。相比现有集成方法，基于主成分聚类分析的电力用户信息自动集成方法能够有效地对高维电力用户信息进行降维处理，从中提取关键特征。这不仅降低了数据处理的复杂度，也能够更精确地反映用户信息的内在结构。同时，主成分聚类分析能够深入挖掘电力用户信息的内在联系，通过聚类分析识别出具有相似用电行为和需求的用户群体，为电力公司的市场策略和服务优化提供数据支持。该方法的创新在于能够自动处理大量的电力用户信息，提取出有代表性的特征，并将用户进行有效分类和集成。通过分析用户群体的特征和行为，电力公司可以更好地了解用户需求，优化供电方案，并提供个性化的服务。同时，该方法也可以帮助电力公司发现潜在的问题和挑战，以便及时采取相应的措施。研究在电力用户信息集成领域具有一定的应用价值和研究意义。

1 基于主成分聚类分析的电力用户信息自动集成

1.1 基于主成分分析的电力用户信息降维

在电力用户信息处理中，主成分分析是一种常用的降维技术。它通过线性变换将原始的高维特征转换为一组互相无关的低维特征，从而减少数据的维度，并保留最重要的信息。通过主成分分析的降维，可以减少电力用户信息的维度，简化数据处理过程，并且保留了主要的信息，有助于提高算法的效率和准确性[6]。PCA的主要目标是寻找一个正交变换，使得第一个主成分具有最大的方差，第二个主成分具有次大的方差，以此类推。这样选取的主成分可以尽可能地保留原始数据中的变异信息，从而达到降维并保留重要信息的目的。通常只需选取更具代表性的数据，这样就能引用原始数据，这就大大降低了数据分析的难度，同时也排除了大量无用的数据[7-8]。PCA的基本原理如图1所示。

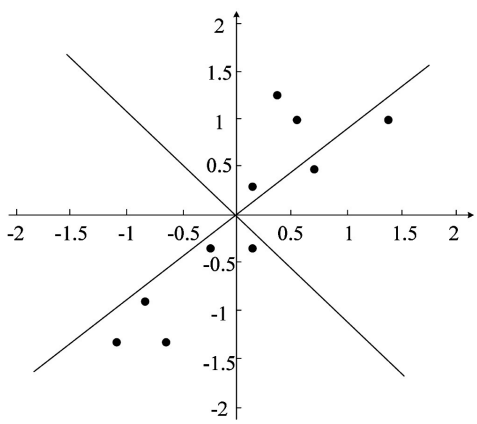


图1 PCA的原理图

假定原始信息由p个指标来表达，然后用$X_1, X_2, \cdots, X_P$来表示初始变量，每个变量都以一维矢量为主体，这时p个变量组成了$\boldsymbol{p}$维向量，则$\boldsymbol{X}=(X_1, X_2, \cdots, X_P)'$，然后利用$\mu$对$\boldsymbol{X}$的平均值来表达，其主要特征是$\boldsymbol{X}$相应的协方差矩阵。$\boldsymbol{X}$展开按需要的线性转换，通常都是用$Y$来表达综合变量的，然后此变量具体如式(1)所示。

$$\begin{cases} Y_1 = \mu_{11}X_1 + \mu_{12}X_2 + ... + \mu_{1P}X_P \\ Y_2 = \mu_{21}X_1 + \mu_{22}X_2 + ... + \mu_{2P}X_P \\ \quad\quad ... \\ Y_P = \mu_{P1}X_1 + \mu_{P2}X_2 + ... + \mu_{PP}X_P \end{cases} \tag{1}$$

式中，通过线性组合可以获得不同的综合变量，因此要使重建效果最好，就需要将$Y_i=\mu_i X$的方差提高到最大，并且每一个Y_i都有比较高的独立性。而对于常数c，具体如式(2)所示。

$$\mathrm{var}(cu_i'\boldsymbol{X}) = cu_i'\sum u_i c = c^2 u_i'\sum u_i \tag{2}$$

假定主成分分析包含N组数据信息，每个数据信息对应的维数是M，然后构造一个由组数据信息组成的矩阵$\boldsymbol{B}$，这时每一行矢量都代表着数据信息，所以可以得出，矩阵$\boldsymbol{B}$的规格是5×4，具体如式(3)所示。

$$\boldsymbol{B} = \begin{bmatrix} b_{0,0} & b_{0,1} & b_{0,2} & b_{0,3} \\ b_{1,0} & b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,0} & b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,0} & b_{3,1} & b_{3,2} & b_{3,3} \\ b_{4,0} & b_{4,1} & b_{4,2} & b_{4,3} \end{bmatrix} \tag{3}$$

对矩阵$\boldsymbol{B}$各列的平均值进行分析，具体如式(4)所示。

$$u(m) = \frac{1}{n}\sum_{n=1}^{N} X(m,n) \tag{4}$$

则矩阵$\boldsymbol{E}$的表达式具体如式(5)所示。

$$\boldsymbol{E} = \boldsymbol{B} - \boldsymbol{U} \tag{5}$$

式(5)中$\boldsymbol{U}$的组成具体如式(6)所示。

$$\boldsymbol{U} = \begin{pmatrix} u(1) & \cdots & u(1) \\ \vdots & \cdots & \vdots \\ u(M) & \cdots & u(M) \end{pmatrix}_{M\times N} \tag{6}$$

共变异数矩阵$\boldsymbol{C}$具体如式(7)所示

$$\boldsymbol{C} = \frac{1}{N}\sum \boldsymbol{E}\boldsymbol{E}^T \tag{7}$$

式中，共变异数矩阵$\boldsymbol{C}$的规格为$M\times N$。对于共变异数矩阵$\boldsymbol{C}$满足下列关系式具体如式(8)所示。

$$\left|\lambda \boldsymbol{I} - \boldsymbol{C}\right| = 0 \tag{8}$$

式中，$\lambda=\lambda_1, \lambda_2, \cdots, \lambda_m$，且$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$。

对特征向量$\boldsymbol{D}$进行分析，计算过程具体如式(9)所示。

$$\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{V} = 0 \tag{9}$$

式中，矩阵$\boldsymbol{V}$表示特征向量矩阵。特征向量$\boldsymbol{D}$的表达式具体如式(10)所示。

$$\boldsymbol{D}=\begin{bmatrix}\lambda_1 & 0 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0 & 0\\ \vdots & 0 & \ddots & 0 & 0\\ 0 & 0 & 0 & \lambda_{M-1} & 0\\ 0 & 0 & 0 & 0 & \lambda_M\end{bmatrix}_{M\times N} \tag{10}$$

式中，λ_m表示第m行对应的特征值。选择最适用的特征向量作为基底可以选取前W个特征向量，但必须满足$W\leqslant M$。将原始的N个数据投影到指定的基底上，并取前W个投影值。

1.2 基于K-means算法的电力用户信息聚类

研究在对电力用户信息降维以后，再对电力用户信息进行聚类处理，选择K-means算法进行聚类。基于K-means算法的电力用户信息聚类是一种常用的无监督学习方法，它将用户信息根据特征相似性进行划分，将相似的用户归为同一类别，它通过对数据集进行训练，无监督地输出K个聚类中心[9-10]。首先随机选择K个数据点作为初始的聚类中心，然后根据数据点与聚类中心的距离，将每个数据点分配到最近的聚类中心所在的类别中。接着，重新计算每个聚类的中心点，更新聚类中心的位置，具体如式(11)所示。

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-u_i\right\|_2^2 \tag{11}$$

式中，E为样本聚类所得簇划分的最小化平方误差。$\boldsymbol{\mu_i}$的表达式如式(12)所示。

$$\boldsymbol{u}_i=\frac{1}{|C_i|}\sum\nolimits_{x\in C_i^x} \tag{12}$$

式中，$\boldsymbol{\mu}_i$表示簇C_i的均值向量。

如果簇的均值向量较小，说明该簇内的样本距离聚类中心的距离较近，即簇内样本更加紧密地围绕聚类中心分布。K-means中所用最重要方法是求点群中心的算法，即欧氏距离，以维数据为例具体如式(13)所示。

$$\mathrm{dist}(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{13}$$

K=2时，K-means算法的简单示例如图2所示。

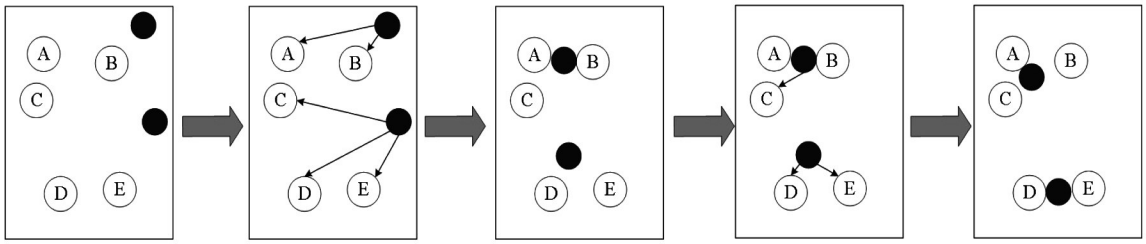


图2 K-means算法的简单示例

K-means算法流程，首先需要选定一个正整数K，表示将数据集分成K个类别。从数据集中随机选择K个数据点作为初始的质心。对数据集中的每个数据点，计算它与K个质心之间的距离，然后将该数据点分配到距离它最近的质心的类别中。对于每个类别，重新计算其质心，新的质心位置是该类别中所有数据点的平均值。如果新的质心与旧的质心的距离小于某个设置的阈值，或者迭代次数达到预设的上限，则算法终止。否则，继续进行迭代。最后，算法将返回K个聚类，以及每个聚类的质心位置。K-means算法可以为电力用户信息提供聚类分析、用户分群和用电特征分析等应用，帮助电力公司了解用户行为和需求，制定相应的策略和服务。

1.3 基于随机森林算法的电力用户信息集成

基于随机森林算法的电力用户信息集成首先采用数据预处理技术对各个数据源的用户信息进行清洗和标准化处理，以确保数据的一致性和可比性。然后，利用随机森林集成算法对清洗后的用户信息进行整合。集成算法通过构建多个基分类器，并将它们的结果集成来得到最终的整合结果。在集成过程中，可以采用投票、加权平均等方法对基分类器的结果进行整合。此外，为了提高整合结果的准确性，还可以引入特征选择算法，选择最具有代表性和区分性的特征进行整合。对于给定的训练数据集，随机森林通过有放回地抽样形成多个不同的子样本集，作为决策树的训练集。对于每个决策树的节点，随机森林只考虑随机选择的一部分特征进行分裂，防止过于依赖某个特定的特征。对于每个子样本集，随机森林使用决策树算法构建决策树模型，直到达到预设的停止条件。在预测过程中，采用随机森林法对各决策树的预测值进行表决或求平均值，从而得出最后的预测值。随机森林是一种在融合多个决策树模型的基础上建立起来的一种机器学习算法[11-12]。假设电力用户信息初始数据集为T，其中含有N个样本，其中X_{i1}，X_{i2}，…，X_{iM}表示该数据的属性或特性，Y表示一个类别标记。在K个属性集中，选取最佳类别节点，选取最佳类别特征，将此属性作为下一阶段划分的主要依据，若属性为连续性，则选取对应的分割点。在分析了相关的信息熵概念之后，可以看出，数据的纯净程度越高，其值也就越大，具体如式(14)所示。

$$Entropy(T)=-\sum_{i=1}^{c}p_i\log_2 p_i \quad (14)$$

式中，p_i 为电力用户信息集成类别样本占总样本的比例。采用决策树构建随机森林模型，并在表决结果的基础上，以少数服从多数的原则来确定最后的集成结果。随机森林算法还可以用于电力用户信息的分析和预测。收集电力用户的相关信息，包括用户的个人信息、用电历史数据、用电行为特征等。根据实际情况选择合适的特征来预测目标变量，可以使用相关性分析、主成分分析等方法进行特征选择。将数据集划分为训练集和测试集，用训练集来建立随机森林模型，用测试集来评估模型的性能。使用随机森林算法建立模型，通过集合多个决策树的结果来提高预测准确性。使用训练好的随机森林模型来预测未知数据的目标变量，可以预测电力用户的用电量、用电行为等。总之，随机森林算法可以为电力用户信息提供预测、分类和异常检测等应用，帮助电力公司进行合理的供电规划和用电管理。基于主成分聚类分析的电力用户信息自动集成流程如图 3 所示。

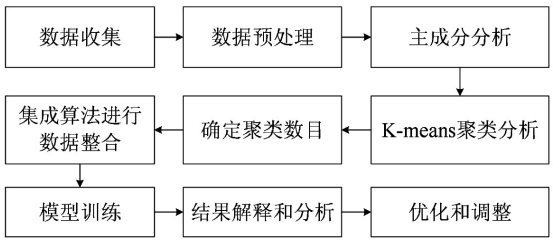


图 3　基于主成分聚类分析的电力用户信息自动集成流程图

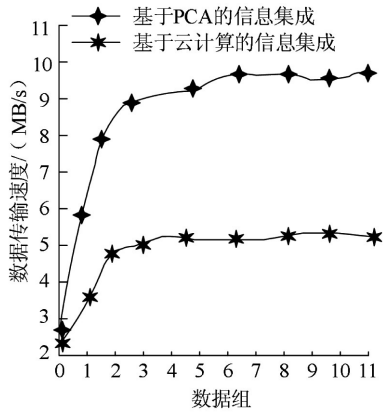
2　基于主成分聚类分析的电力用户信息自动集成性能和效果评估

为验证基于主成分聚类分析的电力用户信息自动集成性能和效果，研究选取查准率、查全率、F_1 值、精准率、召回率、准确度和平均精度来作为评价指标。研究采用的数据集是 UCI 数据集，研究在 UCI 数据集中筛选了五个具有代表性的样本数据集。实验采用信息增益法筛选特征维数 2 000，k 值取 30。研究将基于主成分聚类分析的电力用户信息自动集成方法进行实验，结果如表 1 所示。

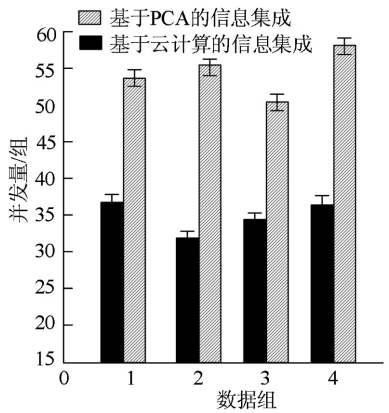
表 1　不同数据集的不同指标结果

数据集	Wpbe	DataR2	Sona r	Par kinsons	Glass	平均值
查准率/%	84.26	83.68	86.11	89.34	81.27	84.93
查全率/%	83.41	82.38	87.25	89.58	92.34	86.992
F_1 值/%	88.59	86.30	86.25	90.38	92.55	88.814
精准率/%	88.19	94.25	87.38	89.27	91.36	90.09
召回率/%	90.54	91.67	89.71	88.26	92.34	90.504
准确度/%	88.17	87.31	84.15	92.64	93.55	89.164
平均精度/%	86.51	88.24	89.26	91.65	92.57	89.646

表 1 中，基于主成分聚类分析的电力用户信息自动集成方法在五种数据集上都表现良好，其中在 Glass 数据集的查准率、查全率、F_1 值、精准率、召回率、准确度和平均精度分别为 81.27%、92.34%、92.55%、91.36%、92.34%、93.55%、92.57%。说明了基于主成分聚类分析的电力用户信息自动集成方法的信息集成准确率高，性能优良，能够满足对电力分析的需求，有效地提高电力用户信息的集成效果。为了进一步分析基于主成分聚类分析的电力用户信息自动集成方法的性能，研究将其与基于云计算的信息集成方法进行对比，两种方法的传输速率与并发率对比如图 4 所示。



(a)　两种数据集成方法的传输速率

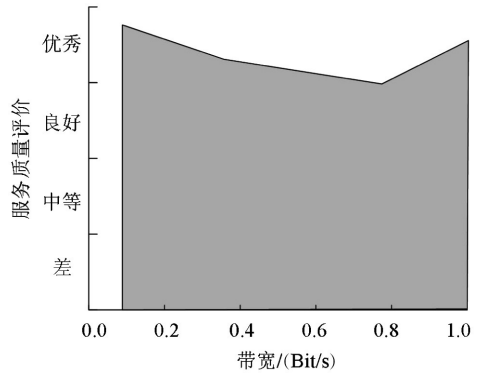


(b)　两种数据集成方法的并发量

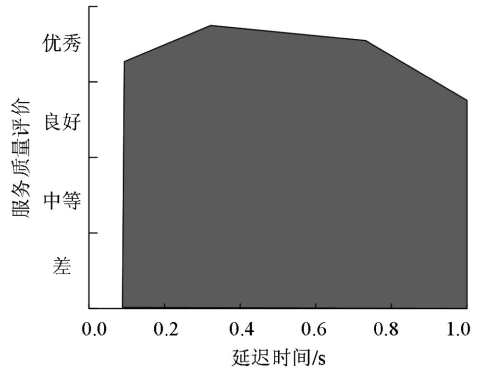
图 4　两种方法的传输速率与并发率对比

图 4 中，当信息数据组数量增长至 2 组后，两种信息集成方法的传输速率骤增。参与数据组大于 5 组后，传输速度数值趋向平缓。基于云计算的信息集成方法的最终传输速度约为 5 MB/s，传输速率相对较低。而基于主成分聚类分析的电力用户信息自动集成方法的传输速率约为 10 MB/s，传输速率较高。基于云计算的信息集成方法的并发运行的数据组均值为 35 组，并发量结果较差。相比之下，基于主成分聚类分析的电力用户信息自动集成方法可以处理并发任务量达到 55 组，具有较好的运行可行性，能够处理大部分的数据组。研究还进一步分析测

试了电力用户的服务质量评价，基于主成分聚类分析的电力用户信息自动集成方法的服务质量评价如图5所示。



(a) 带宽质量评价



(b) 延迟时间质量评价

图5 服务质量评价

图5中，基于主成分聚类分析的电力用户信息自动集成方法的服务质量评价较高，能带更宽，延迟时间更短。因为研究以电力用户服务质量满意度为研究对象，利用k-means聚类算法，对所抽取的用户行为进行聚类，具有更大的带宽、更高的服务品质，能够更快地从用户的行为中挖掘出更多的用户行为，从而达到更好的服务品质。

3 结束语

随着智能电网的快速发展，电力用户信息呈现出海量、异构的特点，对信息的自动集成和性能评估提出了更高的要求。传统的电力用户信息集成方法主要基于规则或模板进行匹配和整合，难以处理大规模、异构的信息。因此，研究提出了一种基于主成分聚类分析的电力用户信息自动集成方法。实验结果表明，基于主成分聚类分析的电力用户信息自动集成方法的传输速率约为10 MB/s，传输速率较高。处理并发任务量达到55组，具有较好的运行可行性，能够处理大部分的数据组。该方法能够通过降维、聚类和集成的步骤，有效地提高电力用户信息的集成效果。但是研究效率和准确率还有提升空间，未来的研究可以进一步优化算法，提高集成效率和准确性。此外，还可以考虑引入其他数据挖掘技术，进一步挖掘电力用户信息中的隐藏知识。

参考文献：

[1] 王丽丽. 基于移动终端的大学生心理健康教育测评系统设计[J]. 自动化技术与应用，2023，42(1)：159-162.

[2] ZHOU Y，LI P，YE Z，et al. Building information modeling-based 3D recons-truction and coverage planning enabled automatic painting of interior walls using a novel painting robot in construction [J]. Journal of Field Robotics，2022，39(8)：1178-1204.

[3] 陶永辉，王勇. 基于初始聚类中心选取的改进K-means算法[J]. 国外电子测量技术，2022，41(9)：54-59.

[4] 陈龙谭，于虹，祁兵，等. 主成分分析与随机森林算法融合的变压器故障诊断方法[J]. 变压器，2022，59(7)：23-28.

[5] KLESHCHENKO A D，SAVITSKAYA O V. Estimation of winter wheat yield using the principal component analysis based on the integration of satellite and ground information[J]. Russian Meteorology and Hydrology，2021，46(12)：881-887.

[6] 莊芳芳. 基于SOM聚类和宽度学习系统的财务危机预测方法[J]. 微型电脑应用，2022，38(3)：169-172.

[7] 刘昌，何正磊，朱小林，等. 基于机器学习的造纸用能负荷特征日获取模型[J]. 造纸科学与技术，2023，42(2)：6-12.

[8] 张弛，王广民，许会博，等. 基于PCA和优化参数SVM的智能变电站故障诊断方法[J]. 计算机应用与软件，2022，39(7)：80-88.

[9] 李海杰，苗蕾，聂磊，等. 基于关联规则和主成分分析的高铁旅客购票行为特征研究[J]. 铁道科学与工程学报，2023，20(6)：2013-2025.

[10] 李伟娟，李雨龙，张磊. 基于多维高斯贝叶斯的造纸机械设备故障自动检测系统[J]. 造纸科学与技术，2022，41(1)：33-39.

[11] 刘佳宁. 基于互联网智能人单合一技术的任务画像算法模型[J]. 自动化技术与应用，2023，42(10)：88-90，124.

[12] 卞悦旭，倪伟，王展旭. 基于大数据聚类的移动机器人运动跟踪控制系统设计[J]. 计算机测量与控制，2022，30(4)：86-120.

作者简介：董俐君(1983—)，女，硕士，高级工程师，研究方向：智能用电技术、电力交费系统建设等。