DOI：10.13334/j.0258-8013.pcsee.191563 文章编号：0258-8013 (2020) 22-7296-10 中图分类号：TM 715

基于深度嵌入聚类的水光荷不确定性源场景生成方法

杨晶显[1]，刘俊勇[1]，韩晓言[2]，刘继春[1*]，丁理杰[3]，张帅[1]，胡帅[1]

(1．四川大学电气工程学院，四川省 成都市 610065；2．国网四川省电力公司，四川省 成都市 610041；
3．国网四川省电科院，四川省 成都市 610041)

An Uncertain Hydro/PV/Load Typical Scenarios Generation Method Based on Deep Embedding for Clustering

YANG Jingxian[1], LIU Junyong[1], HAN Xiaoyan[2], LIU Jichun[1*], DING Lijie[3], ZHANG Shuai[1], HU Shuai[1]

(1. School of Electrical Engineering, Sichuan University, Chengdu 610065, Sichuan Province, China;
2. State Grid Sichuan Electric Power Company, Chengdu 610041, Sichuan Province, China;
3. State Grid Sichuan Electric Power Research Institute, Chengdu 610041, Sichuan Province, China)

ABSTRACT: With the increasing wide application of integrated hydro-photovoltaic (PV) system, how to model the integration correlation of the uncertainty of hydro/PV and load growth is an important research on the operation scheduling and planning of hybrid hydro/PV system. Therefore, the scenario generation method is a way to solve the problem. Since the traditional probability modeling based on historical data and generating scenes through sampling and cutting has high computational complexity and low accuracy, this paper proposed a scenario generation method based on deep embedding for clustering, which learned the initial feature representations through stacked auto-encoder(SAE) and reduce dimension. Then the encoder structure was optimized though Kullback-Leibler (KL) divergence clustering objective, and adaptive moment estimation (Adam) optimization algorithm was applied to adjust parameters. Thus, optimal scenarios can be generated though optimizing the embedding feature iteratively to capture the spatial and temporal correlation characteristics between the uncertain hydro/PV/load variables. The presented method was applied on a certain power grid database, and the effectiveness of the proposed algorithm was verified by comparing with other methods through SSE, SIL, CHI indexes.

KEY WORDS: scenario generation; hybrid hydro/PV power system; deep embedding for clustering; auto-encoder; KL divergence

摘要：随着水、光互补发电系统的应用越来越广泛，如何对水、光出力及负荷增长变化的不确定的融合特性建模对电网的运行调度及规划愈加重要。典型场景生成是解决该问题的主要方法之一，由于传统采用历史数据概率建模，抽样并削减生成场景的方法计算复杂度高、准确率低，且无法有效处理高维多变量数据，该文提出一种基于深度嵌入聚类的水光荷不确定性源场景生成方法。首先利用堆栈自编码(stacked auto-encoder, SAE)网络提取水光荷不确定变量的初始特征，降低数据维度；然后，利用 KL(Kullback-Leibler)散度优化聚类分配目标对自编码网络进行调整，采用自适应矩估计(adaptive moment estimation，Adam)优化算法得到模型最佳参数，通过对编码所嵌入的特征向量不断迭代优化，得到水光荷不确定性变量间的时空依赖关系，从而生成典型场景。算例分析以某地区电网实际采集数据为研究对象，利用误差平方和(sum of squared error，SSE)、SIL、CHI 指标对比传统聚类方法，验证了所提算法的有效性。

关键词：场景生成；水光互补发电；深度嵌入聚类；自编码；KL 散度

0 引言

随着能源转型，清洁能源上网比例不断增加，光伏发电由于低成本和零污染等优良特性得到了大力发展[1]。但是由于光伏具有“随机性、间歇性和波动性”的特点，使其对电力系统的规划、安全、调度和控制等方面带来了挑战[2]。水电是一种清洁可再生的能源，利用其出力具有快速调节的优良性能点来平衡光伏出力的波动，可提高光伏的消纳能力。随着世界上第一座水光互补电站于 2009 年在青海玉树建成，水光互补发电系统引起了全世界的

基金项目：国家重点研发计划项目(2018YFB0905200)。
National Key R&D Program of China(2018YFB0905200).

关注，在中国得到了大力推广[3-4]。

水电、光伏出力和电力负荷随着时间的变化呈现一定的季节或日周期性[5]。电力系统规划和运行需要基于水电、光伏出力及负荷的日特性和年特性，进行场景分析，将水/光出力及负荷增长变化的不确定性转化为多个确定性的场景，为后续水光互补系统的优化运行和规划提供基础[6]。

目前，有 3 种场景分析方法，即时序模拟法、典型日法和场景聚类方法。文献[7]采用时序模拟法，考虑风电出力特性，利用蒙特卡洛采样模拟获取系统风电、负荷场景，模拟了全年的时间序列，虽然具有工程应用价值，但计算效率低。典型日法由于将一年某一天的水、光出力及负荷消纳作为典型场景，不能完全体现全年水、光出力特性及负荷的变化情况。聚类通过场景削减技术对大规模生成场景缩减，通过一定的相似度度量将具有相似时空特性的水、光出力及负荷消纳聚为一类，形成了典型的场景，已有文献证明基于聚类形成的典型场景与与时序仿真法所得结果大体一致，但计算时间大大减小，所得的结果能够准确体现场景特征，相比典型日法准确率大大提高[8]。

海量的发电数据将电网推向了大数据时代，聚类作为一种无监督的机器学习算法，能够有效辨识典型场景，已广泛应用于数据挖掘领域[9]。已有 K-means、K-modes、FCM、DTW、GMM 和谱聚类等聚类方法应用于电力系统领域[10-12]，但是传统的聚类方法针对的对象单一，不能捕捉水、光互补耦合的依赖关系，随着数据维度的增加，计算复杂度也大大增加。随着深度学习在电力行业的发展及应用，已有相关技术为场景生成提供了新的解决思路[13-14]。

为了降低高维多变量水、光、荷时序数据的时空复杂度，一般先采用主成分分析[15]、线性判别分析[16]、小波分析、符号聚合近似等变换方法进行降维和特征提取，再对嵌入的特征进行聚类得到典型场景，但这些变换都会丢失原始水、光、荷数据的信息，虽然降低数据的维度，但不能确保场景生成的精度[17]。自编码机作为一种新型的机器学习模型，利用多层神经网络将原始数据信息非线性嵌入到低维空间，通过无损重构得到原始数据特征，由于准确提取特征的特性已在图像识别，风/光场景生成等领域得到了应用[14,18-19]。但是自编码机不能保证非线性嵌入的特征空间是最合适的聚类空间，是否满足聚类的标准，为此，文献[20]提出了深度嵌入聚类(deep embedding for clustering，DEC)方法，通过聚类最优分配目标对自编码网络不断优化，得到最优特征空间，从而得到最优分类结果。文中利用 DEC 方法对手写体进行识别，其类别个数已知，在设置聚类优化目标时直接计算聚类分配概率，实质是用聚类的手段解决分类问题。

为了兼顾水光荷不确定源场景生成的速度和精度，本文对 DEC 进行改进，在场景聚类优化之前，确定初始提取的水光荷特征的最佳聚类个数，在此基础上构建聚类优化目标，并不断迭代训练自编码网络生成典型水光荷不确定性源场景。首先，分析水光发电系统水光荷时间序列数据特点，在数据预处理阶段结合数据分布特性，利用高斯滤波和分段 B-spline 进行数据清洗，保证了数据的完整，减少噪声数据对实验的干扰。其次，利用改进的深度嵌入聚类方法对清洗后的水光荷数据进行聚类，通过堆栈自编码机进行特征提取，并确定嵌入的初始特征的最佳聚类个数，使用 K-means 聚类技术得到初始聚类中心。再次，使用 KL 散度聚类目标不断迭代训练优化水光荷特征空间，使用自适应矩估计(adaptive moment estimation，Adam)优化器进行深度网络参数调整，得到最佳的典型场景。最后，算例仿真通过比较不同聚类方法的有效性评价指标，验证了本文方法的有效性。

1　水/光/荷场景生成总体框架

基于深度嵌入聚类方法的水/光/荷场景生成框架如图 1 所示，主要分为 4 个步骤。

1）利用数据补全和数据校正技术对采集的水/光/荷时序数据进行数据清洗及归一化，完成数据预处理；

2）利用堆栈自编码机技术提取水/光/荷数据的低维特征，对嵌入的低维特征进行聚类个数的选择并由 K-means 聚类得到初始的聚类结果，将此阶段称为水/光/荷初始特征提取；

3）计算场景聚类分配概率，以 KL 散度作为辅助聚类目标函数对初始生成的堆栈编码器进行调优，利用 Adam 优化器进行参数调整，得到适合场景聚类的编码器结构，并由 K-means 聚类反复迭代直到 2 次聚类的结果在一定偏差范围时停止运算，将此过程称为基于聚类目标的水/光/荷嵌入特征调优；

4）将所优化的水/光/荷特征聚类中心进行解码，得到实际生成的水/光/荷场景，并通过聚类指标与其他传统方法进行对比分析；

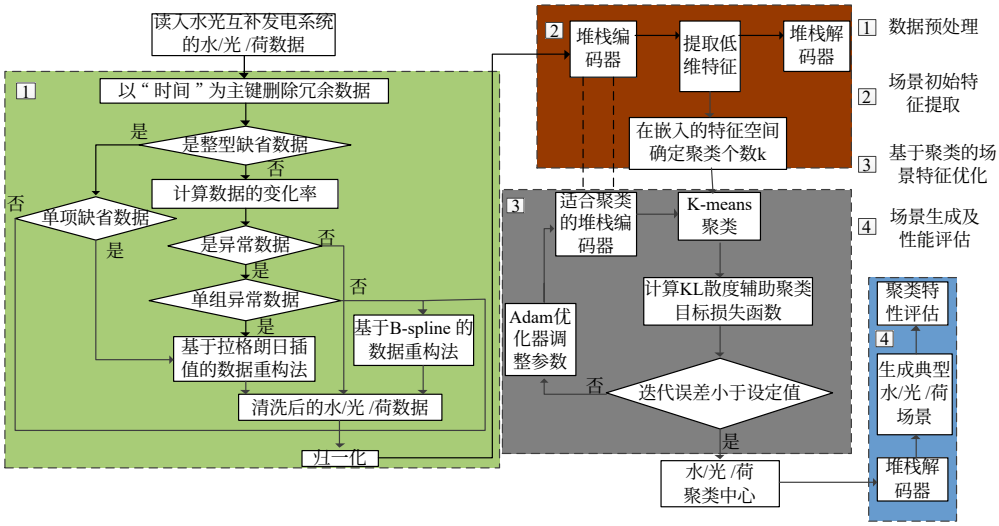


图 1 基于 DEC 的水/光/荷场景生成框架

Fig. 1 Framework structure of hydro/PV/load scenarios generation based on DEC

2 数据预处理

2.1 数据分析

负荷、光伏、水电各自具有不确定性，其间也存在着相关性。对采集的近两年 676 天水/光/荷样本进行皮尔逊相关系数分析得到每天负荷、光伏出力和水电站出力的相关性。计算可得负荷与水电出力的相关性比较集中，大部分集中在区间[0.68 0.82]，绝对平均值为 0.7722，显然负荷与水电出力具有较强相关性，并通过了显著水平 0.05 的单尾检验。同理，得到负荷与光伏出力的相关系数为 0.4912，中度相关。光伏与水电出力的相关系数为 –0.4012，负中度相关。

2.2 数据清洗

数据完整性和准确性是聚类分析的前提，但在数据采集过程中不可避免地存在部分错误数据，可分为冗余数据，异常数据和缺测数据。本文通过检测时间值是否唯一找出冗余数据，采用数据补全技术和数据校正技术对异常值和缺测值进行清洗。清洗流程如图 1 的第 1 部分所示。

由于采用传统回归、贝叶斯等方法建模增加了计算的复杂度，本文采用统计学习和推理方法进行数据清洗，首先对水、光、负荷数据进行描述性统计分析得出大致概率分布，遍历整个实验数据集，利用拉格朗日插值对缺测值进行填充，如式(1)所示。若发现数据有多行缺省时，考虑采用 B-spline 分段拟合进行数据恢复。

$$x=\sum_{j=0}^{23}x_j\cdot\frac{(t-t_1)(t-t_2)\cdots(t-t_{i-1})(t-t_{i+1})\cdots(t-t_{23})}{(t_j-t_1)(t_j-t_2)\cdots(t_j-t_{i-1})(t_j-t_{i+1})\cdots(t_j-t_{23})} \tag{1}$$

式中：x 为一天时刻点 t_i 的缺失的数据；x_j 为当天其余 23 时刻 t_j 已知的数据；t_1, …, t_{23} 为其余的时刻点且不包含 t_j。

若 t 时刻测量的水/光/荷数据的变化率与前一时刻的变化率相比有较大差异，远高于或者低于统计描述的范围时，称为异常值即“离群点”，可以采用高斯滤波的方法进行消噪[10]，也可以将其删除并利用数据补全技术进行插补。为了进行 DEC 深度神经网络训练，对水、光、负荷数据分别进行 MinMaxScaler 归一化到[0,1]，采用离差标准化公式，如下：

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

式中：x 为清洗后的水/光/荷数值；x'为归一化后水/光/荷的数值；$x_{\max}$，$x_{\min}$ 分别表示样本的最大值与最小值。

3 场景初始特征提取

基于改进 DEC 技术的水/光/荷场景生成主要包括 2 步：1）通过自编码机进行水/光/荷初始特征提取，以均方误差(mean square error，MSE)作为损失函数训练自编码深度神经网络，将水/光/荷数据非线性嵌入，得到低维空间的特征，对低维特征进行聚类个数确定并聚类得到初始聚类中心；2）将第 1 步所得的特征空间作为第 2 步优化的初始特征空间，在该特征空间不断迭代优化聚类目标，即根据优化聚类目标函数不断训练自编码网络结构得到最适合聚类的水/光/荷特征，通过深度嵌入和聚类的联合优化得到水光发电系统典型水/光/荷场景，其网络结构如图 2 所示。

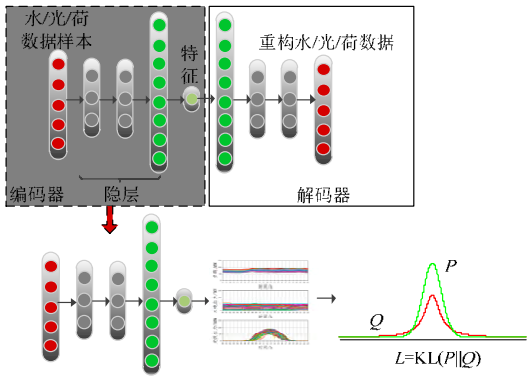


图 2　DEC 网络结构

Fig. 2　Network structure of DEC

3.1　堆栈自编码机

自编码机是一种能够复现输入信号的神经网络，将输入的水/光/荷数据经过非线性映射到隐藏层，即嵌入到另一维度空间，得到原始水/光/荷信号的特征信息，降低数据维度，可以提高计算的效率。训练分为编码和解码两部分，利用反向传播算法，以 MSE 为损失函数不断训练自编码机网络结构参数，得到每一层的权重及偏置，从而得到精确的低维度的水/光/荷特征信息。

堆栈式自编码网络由多个自编码器网络堆叠而成[21]，通过多层非线性神经网络变换，大大降低水/光/荷数据维度，提取水/光/荷特征，完成了原始水/光/荷数据序列的无损压缩，为后续聚类提高精度，并降低计算量。基于堆栈式自编码机水/光/荷数据编码、解码的过程如下所示：

$$f_{\text{encoder}}(\theta): x^{\text{H}} \to z \tag{3}$$

$$f_{\text{decoder}}: z \to x^{\text{H}} \tag{4}$$

式中：f_{encoder}和f_{decoder}分别代表堆栈式自编码机的编码和解码的过程，其中输入 x 为输入的水/光/荷多变量的时间序列数据，显然 H 是高维数；z 为映射到低维度的水/光/荷特征向量，维度远远小于 H；θ 为后续聚类调整所需要训练的自编码深度神经网络的学习参数。解码是编码的相反过程，训练过程中通过损失函数不断调整自编码机的参数使得输出的水/光/荷数据等于原始输入样本，提取精确的场景特征。

3.2　初始特征聚类

水力发电受到自然水文条件的影响，丰水年、枯水年、丰水期、枯水期等年际年内发电量大小变化趋势明显；光伏出力随天气变化出力大小变化明显；负荷需求量也随着时间变化，受温度、实际运行及经济发展的影响，不同年月，工作日和节假日量值变化明显。利用 K-means 进行迭代求解进行特征空间的划分，得到初始的聚类中心 c_i。聚类分析中聚类个数的确定对聚类质量至关重要[22-23]，但传统的 DEC 方法应用在手写体识别，STL-10 等相关类别数目已知的数据集上，没有考虑聚类个数 k 非参的影响，而本文并不知道实际水/光/荷场景类别数目，需要在初始特征空间对 k 进行优化选取来保证场景聚类的质量。因此，本文对 DEC 结构进行了改进，在优化聚类分配目标前，首先在 SAE 提取的初始特征集上确定了参数 k，利用肘部法则优化成本函数，将成本函数明显出现的拐点即肘部所对应的值作为最佳聚类数[24]。

4　基于聚类目标的场景特征优化

将上文所得的水/光/荷嵌入特征结构和估计的 k 个初始聚类中心 c_i 作为聚类的初始化参数。计算辅助聚类目标，并最小化 KL 散度相似度度量损失函数，对原堆栈式自编码网络特征提取的结构不断调整得到最优的水/光/荷场景。

4.1　聚类分布

为进一步提高低维的水/光/荷特征与聚类中心的匹配精度，一般采用高斯分布描述水/光/荷特征到不同场景的分布。但由于高斯分布属于轻尾分布，离群点对模型参数的估计结果影响较大，本文采用更稳健的重尾 t 分布描述嵌入的低维特征与聚类中心的相似程度[25]，如下所示：

$$q_{ij}=\frac{(1+\| z_j-c_i \|^2/\alpha)^{\alpha+1/2}}{\sum_{i'}(1+\| z_j-c_{i'} \|^2/\alpha)^{\alpha+1/2}} \tag{5}$$

式中：q_{ij} 为将低维嵌入的水/光/荷特征 z_j 分类到场景 c_i 的概率；$\| z_j-c_{i'} \|^2$ 代表 z_j 到所有场景 $c_{i'}$ 的距离和；α 为 t 分布的自由度。

4.2　基于 KL 散度优化

将初始特征提取阶段的堆栈式自编码机的解码部分去除，只保留编码结构作为初始的特征提取网络，利用最优场景分配与初始聚类场景分配之间的损失函数不断迭代调整编码网络参数，得到适合聚类的最优低维水/光/荷嵌入特征，对其聚类得到典型场景。

如何构造初始聚类场景分配与最优场景分配目标之间的差异将是设置损失函数的关键。KL 散度，是一种量化 2 种概率分布之间差异的方式，在概率统计中，为简化计算，通常用一个简单的概率分布 P 来近似代替一个复杂的分布 Q，将其近似代

替后的信息损失用 KL 散度度量。KL 散度其实是概率分布 P 和 Q 差别的非对称性度量。本来利用场景聚类优化分布的 KL 散度作为损失函数训练自编码网络。

为提高聚类簇内的紧凑度，规范每个质心的损失贡献，强调场景分配的高置信度数据，进一步提高聚类性能，采用式(6)作为辅助场景聚类目标分布函数。

$$p_{ij}=\frac{q_{ij}^2/f_i}{\sum_{i'}q_{i'j}^2/f_{i'}} \tag{6}$$

$$f_i=\sum_j q_{ij} \tag{7}$$

式中：$q_{i'j}$ 为低维嵌入的水/光/荷特征 z_j 分类到不同场景的概率之和；f_i 和 $f_{i'}$ 分别为所有水/光/荷特征分类到场景 c_i 的概率之和及分类到不同场景的概率之和；p_{ij} 为优化的聚类目标，通过 q_{ij} 的平方项来提高聚类的精度。

通过 KL 散度构造初始聚类分配与辅助聚类目标匹配的损失函数进行自编码网络的训练，其损失函数如下：

$$L=\mathrm{KL}(\mathrm{P}\,\|\,Q)=\sum_j\sum_i p_{ij}\log p_{ij}/q_{ij} \tag{8}$$

利用 KL 散度不断迭代细化聚类，提高聚类分配的置信度。通过不断地训练得到最佳的水/光/荷特征估计，并且得到最优的场景聚类中心，进一步解码生成最佳水/光/荷场景。

4.3 基于 Adam 的 SAE 优化调参

通过梯度下降方法进行 SAE 网络反向传播训练，选择不同的优化算法，可能会得到不同的训练结果。传统神经网络训练采用随机梯度下降(stochastic gradient descent，SGD)进行训练，每一次迭代计算 mini-batch 的梯度，然后对参数进行更新，计算时间不依赖样本数目，计算速度快，但是容易收敛到局部最优，并且在某些情况下可能被困在鞍点[26]。由于 SGD 在高曲率的情况下会发生震荡而迟迟不能接近极小值，引入动量加速学习，在梯度方向不变的维度上速度变快，梯度方向有所改变的维度上的更新速度变慢，从而加快收敛并减小震荡。

Adam 可看作是 RMSprop 和动量的组合,基于动量对损失函数的梯度调整速度，加快收敛[27]。在超参数的设置中，学习率 η 的调整至关重要，Adam 对 η 进行约束，提高了 SGD 的鲁棒性，其动量更新如下所示：

$$g_t=\nabla_{\theta_t}L(\theta_t) \tag{9}$$

$$m_t=\mu m_{t-1}+\eta g_t \tag{10}$$

$$\theta_{t+1}=\theta_t-m_t \tag{11}$$

式中：g_t 为梯度；$L(\theta_t)$ 为误差损失函数；μ 为动量因子，也称为衰减权重；m 为动量向量，初始为 0；θ_{t+1} 为更新后的权重和偏置参数。μ 会随着时间不断调整，当在局部最小值来回震荡时，μ 值增加加速动量相，跳出局部最小值加快收敛。将式(9)、(10)代入式(11)，展开可得：

$$\theta_{t+1}=\theta_t-(\mu m_{t-1}+\eta g_t) \tag{12}$$

Adam 利用梯度的 1 阶矩估计和 2 阶矩估计动态调整每个参数的学习率，经过偏置校正后，每一次迭代学习率都有个确定范围，使得参数比较平稳，其更新如下所示：

$$m_t=\mu m_{t-1}+(1-\mu)g_t \tag{13}$$

$$n_t=\nu n_{t-1}+(1-\nu)g_t^2 \tag{14}$$

$$\hat{m}_t=m_t/1-\mu^t \tag{15}$$

$$\hat{n}_t=n_t/1-\nu^t \tag{16}$$

$$\theta_{t+1}=\theta_t-\frac{\hat{m}_t}{\sqrt{\hat{n}_t}+\varepsilon}\eta \tag{17}$$

式中：m_t，n_t 分别为对梯度的一阶矩估计和二阶矩估计，可以看作对期望 $E|g_t|$，$E|g_t^2|$ 的估计，m_t 在梯度更新时加了一个动量项对梯度进行校正；$\hat{m}_t$，$\hat{n}_t$ 为分别对 m_t，n_t 的校正，可以近似为对期望的无偏估计；μ，ν 为修正参数。将式(13)、(15)代入式(17)，可得：

$$\theta_{t+1}=\theta_t-\frac{\eta}{\sqrt{\hat{n}_t}+\varepsilon}\left(\frac{\mu m_{t-1}}{1-\mu^t}+\frac{1-\mu}{1-\mu^t}\mathrm{g}_t\right) \tag{18}$$

将前一时间步的动量向量的偏差校正估计 $m_{t-1}/1-\mu^t$ 用 $\hat{m}_{t-1}$ 近似代替，可得：

$$\theta_{t+1}=\theta_t-\frac{\eta}{\sqrt{\hat{n}_t}+\varepsilon}\left(\mu\hat{m}_{t-1}+\frac{1-\mu}{1-\mu^t}\mathrm{g}_t\right) \tag{19}$$

可知，式(19)和式(12)很类似，都包含了动量相，除此之外，式(19)通过 $1-\mu/(1-\mu^t)(\sqrt{\hat{n}_t}+\varepsilon)$ 项对学习率形成动态约束，直接影响梯度的更新。

通过最优聚类目标的 KL 散度损失函数不断地迭代训练堆栈自编码神经网络，当迭代误差小于设定值时收敛停止迭代，得到最优聚类结果。

5 算例分析

实验计算机配置为 Ubuntu 18.04 操作系统，处

理器为 Intel(R))Core(TM)i7-4700、CPU@3.40GHz×8, GPU 为 GeForce GTX1070, 内存为 16GB DDR3, 基于 Python3.6 及 Tensorflow1.12 运行环境。

利用改进 DEC 技术生成某电网公司实际水光系统的典型场景，并通过聚类评价指标与其他传统的场景聚类方法进行分析对比。

5.1　数据预处理

算例分析采用某电网公司水光互补示范区近 2 年水电、光伏、负荷实测数据，数据覆盖了丰水期、平水期、枯水期，采集的颗粒度为小时。

首先对采集的数据进行描述性统计分析，得出水、光、负荷时间序列数据的分布，并对响应数据进行均值、标准差、最小值、最大值及 1/4、1/2、3/4 分位数描述，得到数据大致的概率分布。通过高斯滤波消噪剔除异常值，其失真负荷数据高斯滤波消噪结果如图 3 所示。对单点缺测值进行拉格朗日插值，在实验过程中，发现负荷数据有多行大量缺失，对负荷数据进行恢复重构。根据负荷曲线在不同的时刻波动程度不同, 将每天分为 5 个时间段, 在不同时间段使用不同节点个数进行一次 B 样条曲线拟合，根据负荷年月因子对负荷数据重构后，典型日拟合对比结果如图 4 所示。

由图 4、5 可知，通过分段一次 B 样条技术进行负荷数据恢复，相对误差在 2.2%以内，平均绝对

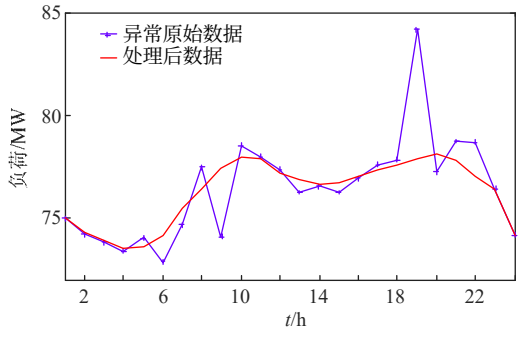


图 3　异常数据处理

Fig. 3　Abnormal data processing

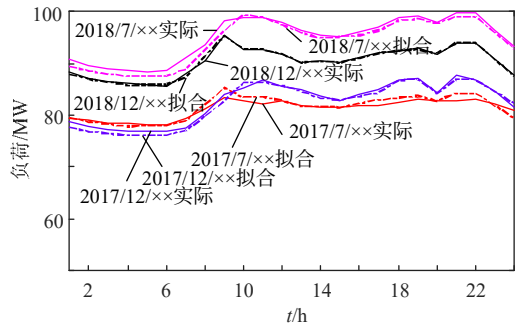


图 4　B 样条分段负荷重构

Fig. 4　B-spline segmental loadreconstruction

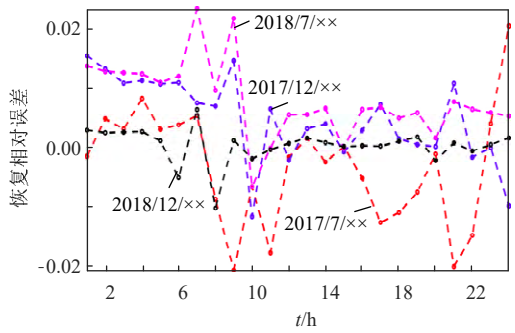


图 5　负荷恢复相对误差

Fig. 5　Relative error of load recovery

百分误差为 4.45%，可以有效地补全负荷数据，完成了水/光/荷数据清洗。

5.2　初始场景生成

在 Keras 框架下，对归一化的水/光/荷样本进行堆栈自编码机训练，编码侧 5 层神经网络，隐藏层单元分别设置为 72、48、48、360、30，解码侧 5 层神经网络，隐层单元分别为 30、360、48、48、72，采用 AdaDelta 优化器，以 MSE 作为损失函数进行编译，初始学习率设置为 0.1，批尺寸设置为 32，epoch 设置为 100。

在嵌入的低维水/光/荷特征空间，对提取的特征向量 z_j 进行 K-means 聚类, 为了保证聚类的质量, 通过 Elbow Method 选择肘值得到最佳聚类个数 k 为 8，如图 6 所示。

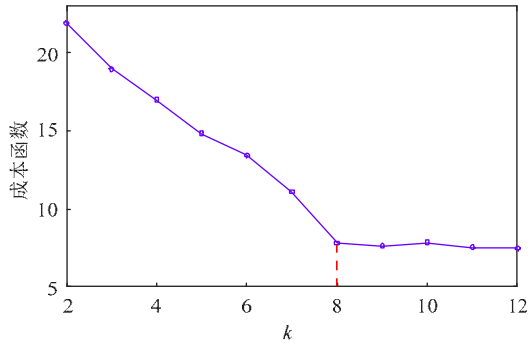


图 6　聚类个数的确定

Fig. 6　Determination of the number of clusters

采用 K-means 提取初始水/光/荷聚类中心，设置 n_init=20，通过 20 次 K-means 不断迭代求解得到最佳的初始场景中心。

5.3　最优场景生成

利用 3.2 节所得的初始聚类中心计算初始 KL 散度, 在优化聚类过程中, 设置初始学习率为 0.01, 利用 Adam 优化算法对 SAE 调参，设置μ=0.9, ν=0.999, $\sum$=10^{-8}，最大迭代次数为 1000，迭代误差设为 1%，为测试 Adam 优化算法的性能，分别采用 SGD 和 Adam 优化算法进行代训练, 其 KL 散

度损失函数的迭代过程如图 7 所示。

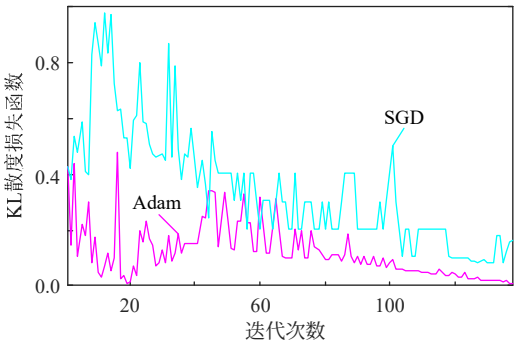


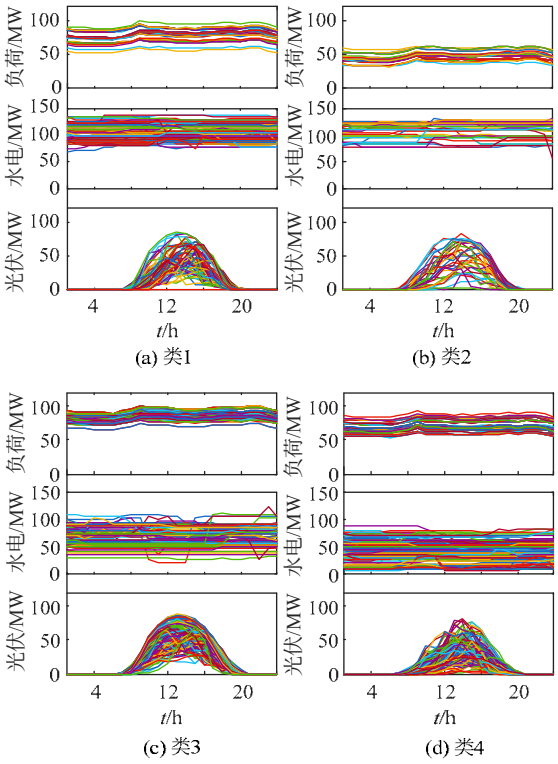
图 7 KL 散度损失误差迭代过程

Fig. 7 KL divergence objective iterative process

当利用 Adam 优化算法进行 SAE 训练时，迭代到 138 代，KL 散度损失误差约为 0.009，小于设定阈值，退出训练，得到嵌入低维空间的水/光/荷场景中心，而利用 SGD 优化算法仍远远大于迭代误差，需要接着不断调参训练，由此可见，Adam 由于能够动量加速和动态地调整学习率，加快收敛。

将调优的聚类中心经过 SAE 解码器后生成实际的水/光/荷典型场景，如图 8 所示。

示范区光伏装机为 100MW，水电装机为 141MW，由图 8 可知，聚类得到了 8 类典型场景，其中第 1、2、3 类处于丰水期，水电出力大致在 90~120MW，达到装机容量的 70%~80%多。第 3 类为晴天大负荷，光伏出力曲线平滑，出力大致在 70~80MW 左右，负荷为典型双峰曲线，大致在 80~90MW 左右，从聚类结果的时间标签看，主要



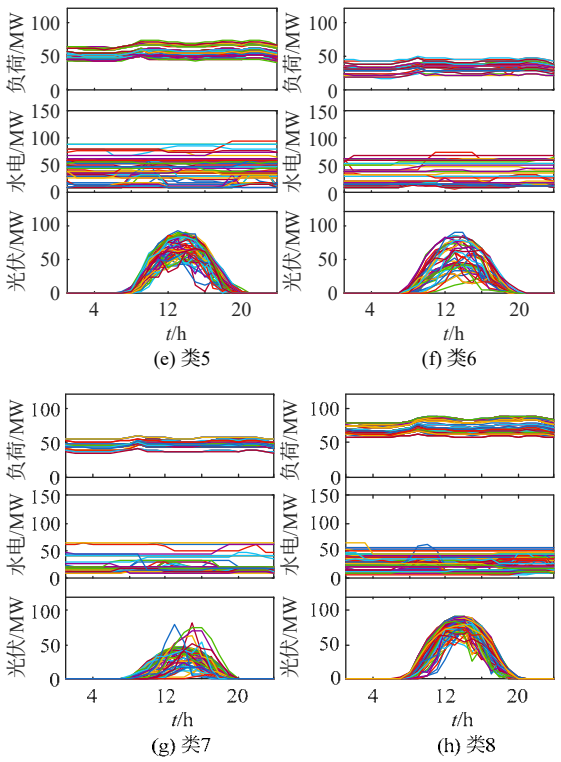


图 8 水/光/荷典型场景

Fig. 8 Typical hydro/PV/load scenes

分布在 18 年 7—9 月工作日，聚类所得光伏与爬虫所得的气象数据大致一致，此时间段正好处于丰水期，且由于工作日工业生产和空调等负荷占比大，与聚类所得的负荷特性一致；第 1 类为多云天气中等负荷水平，光伏出力在 60~70MW 左右，出力曲线变化幅度较大，负荷大致 60MW 左右，主要分布在 18 年 6—9 月休息日，由于休息日减少了某些商业负荷，与本地实际负荷特性基本一致；第 2 类为雨天小负荷，光伏整体出力较小，大致 40MW 左右，主要原因是雨天太阳辐射度小。负荷比较小，大致 40~50MW，时间标签处在 17 年 5、6 月及 18 年 5 月节假日，与聚类所得的水、光、荷场景特性一致。第 4、5 类为平水期，水电出力 50~60MW 左右。第 4 类雨天大负荷，光伏出力整体波动性较大，出力小，负荷为 70~80MW，为典型工作日负荷，主要分布在 18 年 10、11 月工作日；第 5 类为多云中等负荷，光伏出力波动性大，出力约 70MW 左右，负荷 50~60MW 左右，主要分布在 17 年 11 月工作日和 18 年 3 月的休息日，17 年负荷小于 18 年，也反映了负荷的年际增长。第 6、7、8 类为枯水期，水电出力大致在 30MW 左右，第 6 类为多云特小负荷枯水期，光伏出力波动大，负荷为 30MW 左右，从聚类结果的时间标签得到此场景处在 17 年 1、2 月，负荷极小主要由于当地在 3 月份新投产某重工业，

使得之后负荷有极大提高，聚类的结果和当地实际情况一致；第 7 类为雨天小负荷期，光伏整体出力小且负荷为 50MW 左右, 主要分布在 17 年 12 月至 18 年 1、2 月休息日；第 8 类为晴天中等负荷，光伏出力曲线平滑，且负荷大致为 70MW，主要分布在 17 年 12 月至 18 年 1 月的工作日，所生成的场景与气象数据大体一致，聚类所得典型场景基本反映了负荷的逐年增长，及自然来水的变化和光伏依据天气出力的变化情况。

5.4　聚类性能分析

聚类评价指标可以对聚类结果进行定量分析，利用误差平方和(sum of squared error，SSE)指标、SIL(Silhousette)指标、CHI(Calinski-Harabasz)指标对 DEC 场景生成性能分析[17,28]，分别设置聚类数 k 为 4~10，将本文所提的聚类算法和传统的 K-means 和 PCA 特征提取后再 K-means 的结果进行对比分析，其评价结果如图 9—11 所示。

由图 9—11 可知，本文所提算法的聚类性能指标 SIL、CHI 均高于直接 K-means 和 PCA_K-means 方法, 而 SSE 指标均小于 K-means 和 PCA_K-means 算法，由此可见，所提算法的聚类质量要优于直接 K-means 和 PCA_K-means 算法。

采用训练好的 SAE 模型进行 K-means 聚类，耗时约为 43.37s, 而直接 K-means 聚类耗时 124.92s，

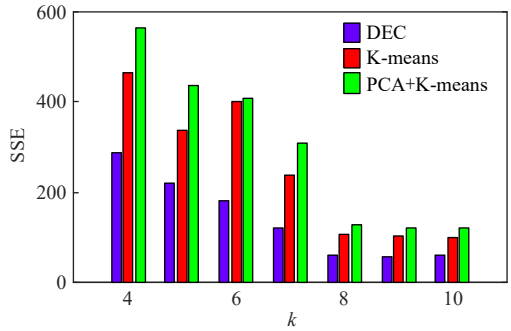


图 9　SSE 指标计算结果

Fig. 9　Computation result of SSE index

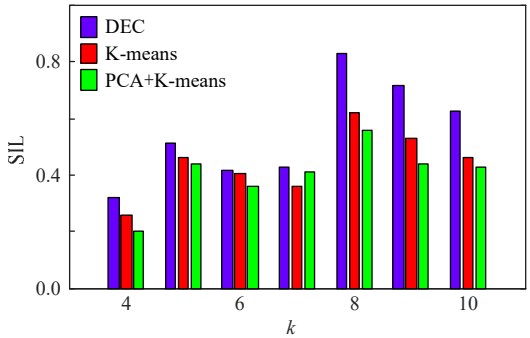


图 10　SIL 指标计算结果

Fig. 10　Computation result of SIL index

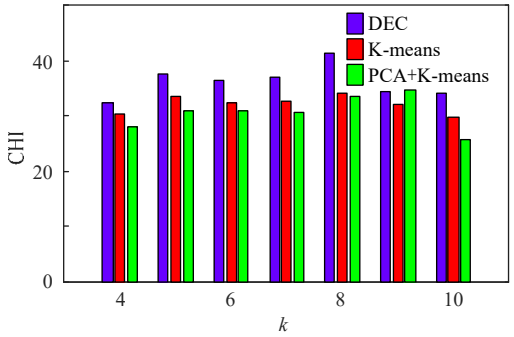


图 11　CHI 指标计算结果

Fig. 11　Computation result of CHI index

FCM 聚类耗时 399.52s，采用 DTW 聚类耗时 3960.86s，可见，由于 SAE 对样本进行了压缩，减小了样本规模，提高了相应聚类的速度。

6　结论

本文提出了一种基于改进深度嵌入聚类的典型场景生成方法，分析本文算例，得出以下结论：

1）相比传统的聚类方法，本文所提的聚类方法获得有效的水、光、荷不同变量之间的耦合关系，能够对多变量的融合特性进行建模；

2）相比传统针对多维数据先降维提取特征再聚类方法，DEC 在提取特征的同时进行聚类，通过 KL 散度作为优化聚类目标不断迭代训练优化所嵌入的特征空间，使得所获取的特征更有利于聚类，提高聚类的性能；

3）利用 Adam 优化的梯度下降算法训练水/光/荷场景特征提取的 SAE 模型, 增强了对学习率进行约束，抑制震荡，加速收敛，提高 SAE 模型的训练性能；

4）本文基于机器学习技术，利用改进 DEC 技术进行场景生成，通过优化聚类分配目标损失对基于 SAE 的特征提取网络进行调优，得到最优的水/光/荷聚类场景，通过算例分析，在聚类有效性等指标上均优于直接 K-means 算法和 PCA_K-means 聚类算法。

下一步工作包括如何利用聚类所形成的典型场景指导水电调控、优化调度和储能配置等工作。

参考文献

[1] Shabani M，Mahmoudimehr J. Techno-economic role of PV tracking technology in a hybrid PV-hydroelectric standalone power system[J]. Applied Energy，2018，212：84-108.

[2] Li Fangfang，Qiu Jun. Multi-objective optimization for integrated hydro-photovoltaic power system[J]. Applied

Energy，2016，167：377-384.

[3] 安源，方伟，黄强，等. 水-光互补协调运行的理论与方法初探[J]. 太阳能学报，2016，37(8)：1985-1992.
An Yuan，Fang Wei，Huang Qiang，et al. Preliminary research of theory and method of hydro/solar complementary operation[J]. Acta Energiae Solaris Sinica，2016，37(8)：1985-1992(in Chinese).

[4] 明波，黄强，王义民，等. 水-光电联合运行短期调度可行性分析[J]. 太阳能学报，2015，36(11)：2731-2737.
Ming Bo, Huang Qiang, Wang Yimin, et al. The feasibility analysis of short-term scheduling for joint operation of hydropower and photoelectric[J]. Acta Energiae Solaris Sinica，2015，36(11)：2731-2737(in Chinese).

[5] 王群，董文略，杨莉. 基于 Wasserstein 距离和改进 K-medoids 聚类的风电/光伏经典场景集生成算法[J]. 中国电机工程学报，2015，35(11)：2654-2661.
Wang Qun，Dong Wenlue，Yang Li. A wind power/photovoltaic typical scenario set generation algorithm based on Wasserstein distance metric and revised K-medoids cluster[J]. Proceedings of the CSEE，2015，35(11)：2654-2661(in Chinese).

[6] Hu Wei，Zhang Hongxuan，Yu Dong，et al. Short-term optimal operation of hydro-wind-solar hybrid system with improved generative adversarial networks[J]. Applied Energy，2019，250：389-403.

[7] 刘挺坚，刘友波，刘若凡，等. 风电外送断面极限输电能力的非参数回归估计[J]. 电网技术，2017，41(11)：3514-3522.
Liu Tingjian, Liu Youbo, Liu Ruofan, et al. Nonparametric regression estimation for total transfer capability of transmission interface considering centralized wind power integration[J]. Power System Technology，2017，41(11)：3514-3522(in Chinese).

[8] 丁明，解蛟龙，刘新宇，等. 面向风电接纳能力评价的风资源/负荷典型场景集生成方法与应用[J]. 中国电机工程学报，2016，36(15)：4064-4071.
Ding Ming, Xie Jiaolong, Liu Xinyu, et al. The generation method and application of wind resources/load typical scenario set for evaluation of wind power grid integration [J]. Proceedings of the CSEE, 2016, 36(15)：4064-4071(in Chinese).

[9] 朱文俊，王毅，罗敏，等. 面向海量用户用电特性感知的分布式聚类算法[J]. 电力系统自动化，2016，40(12)：21-27.
Zhu Wenjun，Wang Yi，Luo Min，et al. Distributed clustering algorithm for awareness of electricity consumption characteristics of massive consumers [J]. Automation of Electric Power Systems, 2016, 40(12)：21-27(in Chinese).

[10] 李阳，刘友波，刘俊勇，等. 基于形态距离的日负荷数据自适应稳健聚类算法[J]. 中国电机工程学报，2019，39(12)：3409-3419.
Li Yang, Liu Youbo, Liu Junyong, et al. Self-adaptive and robust clustering algorithm for daily load profiles based on morphological distance[J]. Proceedings of the CSEE，2019，39(12)：3409-3419(in Chinese).

[11] Lai Chunsing，Jia Youwei，McCulloch M D，et al. Daily clearness index profiles cluster analysis for photovoltaic system[J]. IEEE Transactions on Industrial informatics，2017，13(5)：2322-2332.

[12] Li Ran，Li Furong，Smith N D. Load characterization and low-order approximation for smart metering data in the spectral domain[J]. IEEE Transactions on Industrial Informatics，2017，13(3)：976-984.

[13] Chen Yize，Wang Yishen，Kirschen D，et al. Model-free renewable scenario generation using generative adversarial networks[J]. IEEE Transactions on Power Systems，2018，33(3)：3265-3275.

[14] 王守相，陈海文，李小平，等. 风电和光伏随机场景生成的条件变分自动编码器方法[J]. 电网技术，2018，42(6)：1860-1867.
Wang Shouxiang，Chen Haiwen，Li Xiaoping，et al. Conditional variational automatic encoder method for stochastic scenario generation of wind power and photovoltaic system[J]. Power System Technology，2018，42(6)：1860-1867(in Chinese).

[15] Motlagh O，Berry A，O'Neil L. Clustering of residential electricity customers using load time series[J]. Applied Energy，2019，237：11-24.

[16] Dizaji K G, Herandi A, Deng Cheng, et al. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization[C]//2017 IEEE International Conference on Computer Vision. Venice，Italy：IEEE，2017.

[17] 王潇笛，刘俊勇，刘友波，等. 采用自适应分段聚合近似的典型负荷曲线形态聚类算法[J]. 电力系统自动化，2019，43(1)：110-118.
Wang Xiaodi，Liu Junyong，Liu Youbo，et al. Shape clustering algorithm of typical load curves based on adaptive piecewise aggregate approximation [J]. Automation of Electric Power Systems，2019，43(1)：110-118(in Chinese).

[18] Ye Xulun，Zhao Jieyu. Multi-manifold clustering：a graph-constrained deep nonparametric method[J]. Pattern Recognition，2019，93：215-227.

[19] Hinton G E，Salakhutdinov R R，et al. Reducing the dimensionality of data with neural networks[J]. Science，2006，313(5786)：504-507.

[20] Xie Junyuan，Girshick R，Farhadi A. Unsupervised deep embedding for clustering analysis[J]. arXiv preprint

arXiv：1511.06335，2016.

[21] Huang Peihao，Huang Yan，Wang Wei，et al. Deep embedding network for clustering[C]//Proceedings of 2014 22nd International Conference on Pattern Recognition. Stockholm，Sweden：IEEE，2014.

[22] Zhou Shibing，Xu Zhenyuan，Liu Fei. Method for determining the optimal number of clusters based on agglomerative hierarchical clustering[J] . IEEE Transactions on Neural Networks and Learning Systems，2017，28(12)：3007-3017.

[23] Liang Jie，Yang Jufeng，Cheng Mingming，et al. Simultaneous subspace clustering and cluster number estimating based on triplet relationship[J] . IEEE Transactions on Image Processing，2019，28(8)：3973-3985.

[24] Thorndike R L. Who belongs in the family? [J]. Psychometrika，1953，18(4)：267-276.

[25] 李巍，董明利，吕乃光，等. 基于T分布混合模型的多光谱人脸图像配准[J]. 光学学报, 2019, 39(7)：0710001.
Li Wei，Dong Mingli，Lü Naiguang，et al. Multispectral face image registration based on T-distribution mixture model[J]. Acta Optica Sinica，2019，39(7)：0710001(in Chinese).

[26] Qian Ning. On the momentum term in gradient descent learning algorithms[J]. Neural Networks，1999，12(1)：145-151.

[27] 杨智宇，刘俊勇，刘友波，等. 基于自适应深度信念网络的变电站负荷预测[J]. 中国电机工程报，2019，39(14)：4049-4060.
Yang Zhiyu，Liu Junyong，Liu Youbo，et al. Transformer load forecasting based on adaptive deep belief network [J]. Proceedings of the CSEE，2019，39(14)：4049-4060(in Chinese).

[28] Hong Juhua，Xiang Yue，Liu Youbo，et al. Development of EV charging templates：an improved K-prototypes method[J]. IET Generation，Transmission & Distribution，2018，12(20)：4361-4367.

杨晶显

在线出版日期：2020-02-27。
收稿日期：2019-10-11。

作者简介：

杨晶显(1984)，女，博士研究生，研究方向为电力系统数据挖掘与分析，yangjxian12@163.com；

刘俊勇(1963)，男，博士生导师，研究方向为电力系统数据挖掘与分析，电力市场；

韩晓言(1965)，男，博士，副总工程师，研究方向为电力系统运行及规划；

*通信作者：刘继春(1975)，男，教授，博士生导师，IEEE 高级会员，研究方向为电力系统数据挖掘、稳定与控制，电力市场，jichunliu@scu.edu.cn。

(责任编辑　乔宝榆)

 DOI：10.13334/j.0258-8013.pcsee.191563

An Uncertain Hydro/PV/Load Typical Scenarios Generation Method Based on Deep Embedding for Clustering

YANG Jingxian[1], LIU Junyong[1], HAN Xiaoyan[2], LIU Jichun[1], DING Lijie[3], ZHANG Shuai[1], HU Shuai[1]

(1. Sichuan University; 2.State Grid Sichuan Electric Power Company; 3.State Grid Sichuan Electric Power Research Institute)

KEY WORDS: scenario generation; hybrid hydro/PV power system; deep embedding for clustering; Auto-encoder; KL divergence

As the randomness, intermittency and fluctuating nature of photovoltaic (PV) generation, hydropower which can be regulated and integrated to compensate the fluctuation of PV power. With the increasing wide application of hybrid hydro-photovoltaic system, how to model the integration correlation of the uncertainty of hydro/PV and load growth is an important research on the grid connection, operation and planning.

The scenario generation method provides a way to solve the problem. Since the traditional probability modeling based on historical data and generating scenes through sampling and cutting has high computational complexity and low accuracy, this paper proposes a scenario generation method based on deep embedding (DEC) for clustering. The process of scenario generation is described as:

$$\begin{cases} f_{\text{encoder}}(\theta): x^{\text{H}} \to z \\ f_{\text{decoder}}: z \to x^{\text{H}} \\ L = \text{KL}(\text{P}\|Q) = \sum_j \sum_i p_{ij} \log p_{ij}/q_{ij} \\ q_{ij} = \dfrac{(1+\| z_j - c_i \|^2 / \alpha)^{\alpha+1/2}}{\sum_{i'} (1+\| z_j - c_{i'} \|^2 / \alpha)^{\alpha+1/2}} \\ p_{ij} = \dfrac{q_{ij}^2 / f_i}{\sum_{i'} q_{ij}^2 / f_{i'}} \\ f_i = \sum_j q_{ij} \end{cases} \tag{1}$$

The method learns the initial hydro/PV/load feature representations through stacked auto-encoder and reduce dimension of hydro/PV/load data. Then the encoder structure is optimized with the KL divergence clustering objective, and thus, optimal scenarios can be generated through optimizing the embedding feature iteratively to capture the spatial-temporal correlation characteristics between the uncertain hydro/PV/load variables.

To ensure the accuracy of clustering, hydro/PV/load data is cleaned and normalized firstly. Then an encoder structure is established to map the hydro/PV/load data into low-dimension space and extract the hydro/PV/load features. Simultaneously, the parameters of SAE network are fine tuning through KL divergence optimization objective function to obtain the optimal structure. The Adam optimization function is applied to get higher performance for the SAE training model.

The proposed approach is validated using the real hydro/PV/load data provided on a certain power grid database, and eight scenarios are clustered through DEC method and categorized distinctly; each pattern is consistent with meteorological conditions basically. representing a type of hydro/PV/load scenario. The result is shown as Fig. 1.

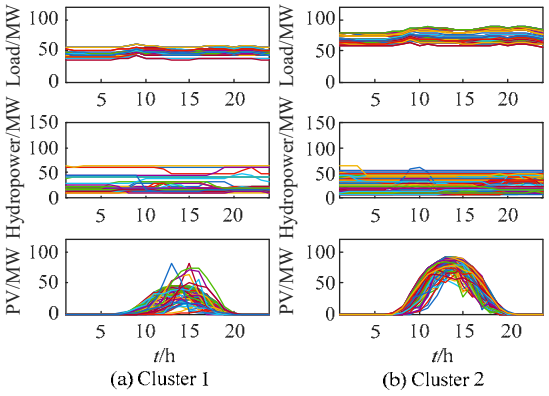


Fig. 1 Typical hydro/PV/load scenes

Taking cluster 1 (medium load on cloudy days in wet season) and cluster 2 (small load on rainy days in wet season) as examples, the hydropower of the two clusters is about 120MW accounting for 80% of the installed hydropower capacity. For cluster 1 and 2, PV power output is about 60~70MW and 40~50MW respectively, cluster 1 fluctuates largely, representing cloudy days while the overall output of cluster 2 is small due to the solar irradiance. Load in cluster 1 is about 60MW while 40~50MW in cluster 2, so cluster 1 mainly distributed at weekends from June to September in 2018 and cluster 2 represents holidays in May, 2017 and 2018.

We can use these clustering scenarios to know more details about hybrid hydro-PV power generation system. Compared with K-means and PCA_K-means methods, the proposed approach in this paper has the maximum value of CHI and SIL index, and the minimum value of SSE index. The effectiveness of the proposed algorithm is verified.