



电力系统自动化

Automation of Electric Power Systems

ISSN 1000-1026, CN 32-1180/TP

《电力系统自动化》网络首发论文

题目：面向负荷特征分析的地理分布式协同聚类方法
作者：刘家丞，吴江，刘鹏远，徐占伯，李晓鹏，管晓宏
收稿日期：2021-07-08
网络首发日期：2022-05-17
引用格式：刘家丞，吴江，刘鹏远，徐占伯，李晓鹏，管晓宏. 面向负荷特征分析的地理分布式协同聚类方法[J/OL]. 电力系统自动化.
<https://kns.cnki.net/kcms/detail/32.1180.tp.20220516.2156.006.html>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

面向负荷特征分析的地理分布式协同聚类方法

刘家丞¹, 吴江¹, 刘鹏远^{1,2}, 徐占伯¹, 李晓鹏¹, 管晓宏¹

(1. 智能网络与网络安全教育部重点实验室(西安交通大学), 陕西省西安市 710049;

2. 国网陕西省电力公司西安供电公司, 陕西省西安市 710048)

摘要: 能源互联网架构下, 用户数据传播延迟和电力公司管理规定促使电力数据中心建立在全国各地, 电力数据因而天生呈现地理分布式, 对此研究了在地理分布式情景下的用户负荷特征聚类算法。首先, 面对地理节点内用户负荷的特征分析, 在采用主成分分析与负荷指标特征加权组合算法的基础上, 提出了考虑密度峰值信息的K-means算法, 并为地理节点间的信息共识提供了支撑。其次, 针对用户需求的地理分布式网络化感知结构, 构建了考虑特征迁移的分布式聚类模型框架, 提出将节点局部信息通过参数共识得到全局聚类模型的分布式K-means算法, 在节点间仅传递公示信息的前提下, 实现了用户特征的全局聚类。最后, 通过爱尔兰、中国北方部分城市的负荷数据对模型及算法进行验证。结果显示, 分布式K-means能利用全局信息、考虑不同区域的差异来更好地识别典型用电曲线, 并且算法具有较好的可迁移性。

关键词: 分布式聚类; 负荷特征分析; 参数共识; 密度峰值聚类; 迁移学习

0 引言

近年来, 能源互联网架构下的电力数据资源急剧增长并达到了大数据的程度^[1]。对于电力大数据的分析具有从用电客户精确定位, 到电力生产反馈指导, 再到国民经济精准还原的全方位价值^[1-2]。大数据技术在智能电网中的应用主要集中在需求响应^[3-4]、负荷预测^[5-7]、故障诊断^[8-9]、异常用电检测^[10-11], 还可用于研究电动汽车充电站部署^[12]、光伏设备技术性能分析^[13]等新兴负荷问题。

电力系统天生呈现分布式特征, 在供给侧和需求侧都有所体现。供给侧方面, 电网接入了众多分布在全国各地的微电网和发电站^[14]; 需求侧方面, 诸如居民用电、商业用电多离散分布在各城市中, 其用电数据通常存储在不同地理位置的电力数据中心^[15]。

在区域电网背景下, 应用机器学习进行电力大数据分析, 若采用独立式训练会由于样本数据量少导致结果差; 而集中式训练对于具有分布式属性的电力大数据, 则需进行跨中心的数据调度。因此当数据量急剧增加时会产生以下3个问题^[16], 使得传

统的集中式机器学习在分布式大数据环境下不可行。

1) 隐私保护问题

在智能电网中存在许多隐私保护的问题^[17], 例如智能电表收集到的负荷数据可以用于监测电网状态^[18], 然而截获数据的人可以从中辨别出用户的活动, 使得用户隐私受到威胁^[19]。电网通常对传输的数据进行加密来保证其安全性, 但仍有密钥管理等问题^[20]。

2) 数据时延问题

数据在网络中传播需要经过多个转发节点, 导致数据时延, 短时间进行大量数据的传输更会加重时延。满足使用需求的拓扑设计模型已经被提出^[21], 但数据越大, 传输时延越大, 可能会造成电网的控制性能恶化, 带来更大的成本^[19]。

3) 传输成本问题

电网大数据进行跨数据中心的传输会占据大量的稀缺带宽资源^[22], 并且跨数据中心传输的成本远超过在一个数据中心内进行传输的成本^[19]。

为避免以上问题, 同时实现区域电网下多地理节点的负荷特性分析, 本文研究了地理分布式情景下的负荷特征聚类算法。针对隐私保护问题, 采用基于主成分分析(PCA)-负荷指标的特征加权组合算法, 提取原始数据的抽象特征, 实现用户数据脱敏。考虑地理节点之间的拓扑关系, 设计基于参数

收稿日期: 2021-07-08; 修回日期: 2022-02-25。

国家重点研发计划资助项目(2016YFB0901900); 国家自然科学基金资助项目(61803297)。

共识的分布式聚类算法,使数据中心之间仅传输极少量的拓扑特征,降低数据时延,并可构建全局聚类模型。针对传输成本问题,搭建考虑特征迁移的迁移学习框架,在原有模型基础上快速构建新模型,减少跨数据中心的交互次数。本文选取爱尔兰电网和中国北方部分城市电网的实际负荷数据进行测试,验证所提地理分布式协同聚类算法的有效性。

1 地理分布式协同聚类模型

能源互联网架构下,电力系统遍布全国千家万户,电力数据也在电力用户与电网的交互中不断产生。为了最大限度减小基础服务设施与电力终端用户之间时延以及方便监管区域电力用户,电力数据中心通常分散建立在不同地理位置的各个城市中,呈现典型的地理分布式属性^[23]。然而,考虑到数据的隐私保护问题,这些数据中心通常相互独立存储、独立维护,彼此间难以相互通信,形成了电力数据孤岛。针对这类处理地理分布数据集的机器学习应用,可以称之为“地理分布机器学习”^[23]。相较于传统分布式针对复杂问题分而划之解决,地理分布式更注重在克服地理隔离的困难下搭建模型。

传统聚类模型的构建,通常需要一次访问多个区域的数据,考虑到隐私保护及传输成本等问题,原始数据难以在各电力数据孤岛之间进行通信,往往仅能传输少量脱敏信息,如模型参数等^[24]。本文针对数据孤岛背景下的数据中心提出了地理分布式协同聚类框架,该框架允许地理分布式数据中心在仅传输少量脱敏参数的情况下独立搭建聚类模型,使得每个数据中心都能生成一个具有全局信息的聚类模型,且不同的聚类模型受不同地理位置影响呈现地理分布式特性,如图1所示。

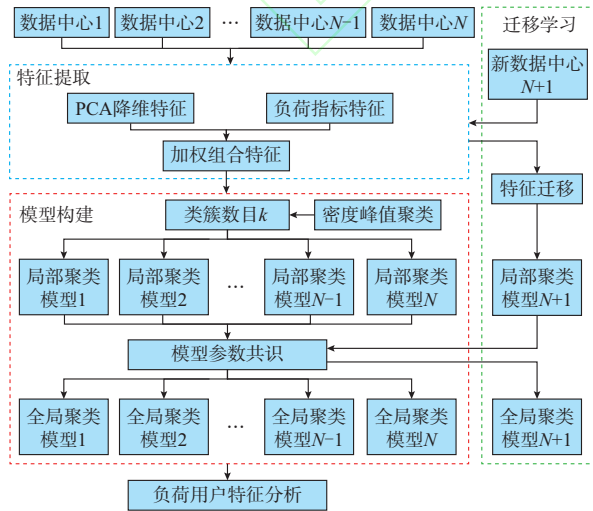


图1 地理分布式协同聚类框架
Fig. 1 Framework of geo-distributed collaborative clustering

第2~4章将分别阐述图1所展示特征提取模块的特征加权组合算法、模型构建模块的考虑密度峰值信息^[25]的分布式聚类算法以及迁移学习模块的特征迁移算法。

2 基于PCA-负荷指标的特征加权组合算法

在聚类模型构建之前,首先需要提取适合模型的负荷特征。特征提取一方面可以将原始的用户负荷数据抽象为难以理解的特征数据,实现数据脱敏、保护用户隐私;另一方面也可以降维压缩庞大的用户数据,大大减少传输成本。

本文采用将PCA特征与负荷指标特征加权组合的方式,针对用户的月负荷数据进行特征降维提取处理,以两电网6月数据为例,该数据原始维度为1440维。PCA降维特征较为抽象,不能很好地说明原始数据的物理意义;负荷指标是先验知识,反映电力负荷的经验特性,但很多重要的抽象特征却无法体现。两类特征反映不同的特性,组合在一起可以获得更全面的负荷信息。

本文记地理节点 v 的原始负荷数据集为 $Y_v = [y_{v,1}, y_{v,2}, \dots, y_{v,i}, \dots, y_{v,n}] \in \mathbb{R}^{1440 \times n}$,其中 $y_{v,i}$ 为节点 v 第 i 个用户特征矢量,维度是 1440×1 , n 为节点 v 的用户总数量。

PCA常用于高维数据的降维处理,针对不同数据集PCA抽取特征数量可能不同,通常借助特征方差占比情况提取数据的主要特征分量。以电网6月负荷数据为例,选取方差占比之和大于90%的前几维特征,最终可压缩成6维PCA特征 $y_{v,i}^{\text{PCA}} = [y_{v,i,1}^{\text{PCA}}, y_{v,i,2}^{\text{PCA}}, \dots, y_{v,i,6}^{\text{PCA}}]$,其中 $y_{v,i}^{\text{PCA}}$ 为PCA降维特征矢量。

用户负荷特征直接反映了各类型用户的用电模式,可用于电力用户分类和分析用户行为。本文采用负荷峰值、平均负荷、均方负荷、负荷率、负荷偏度和负荷散度等六大负荷指标作为特征(表达式见附录A表A1),计算可得6维负荷特征 $y_{v,i}^{\text{Load}} = [y_{v,i,L_1}^{\text{Load}}, y_{v,i,L_2}^{\text{Load}}, y_{v,i,L_3}^{\text{Load}}, y_{v,i,L_4}^{\text{Load}}, y_{v,i,L_5}^{\text{Load}}, y_{v,i,L_6}^{\text{Load}}]$,其中 $y_{v,i}^{\text{Load}}$ 为负荷特征矢量。

进一步地,采用特征加权组合的方法对以上两类特征进行特征处理。对于每一种特征所附加的权重 $w_{v,l}$ 计算表达式为:

$$w_{v,l} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{v,i,l} - \bar{y}_{v,l})^2} \quad (1)$$

式中: $y_{v,i,l}$ 为节点 v 第 i 个用户的第 l 种特征; $\bar{y}_{v,l}$ 为节点 v 第 l 组特征的均值。

加权组合后的节点 v 第 i 个用户特征矢量 $x_{v,i}$ 为:

$$x_{v,i} = [w_v^{\text{PCA}} y_{v,i}^{\text{PCA}}, w_v^{\text{Load}} y_{v,i}^{\text{Load}}] = [w_{v,1} y_{v,i,1}^{\text{PCA}}, w_{v,2} y_{v,i,2}^{\text{PCA}}, \dots, w_{v,6} y_{v,i,6}^{\text{PCA}}, w_{v,7} y_{v,i,L_1}^{\text{Load}}, w_{v,8} y_{v,i,L_2}^{\text{Load}}, \dots, w_{v,12} y_{v,i,L_6}^{\text{Load}}] \quad (2)$$

式中: w_v^{PCA} 和 w_v^{Load} 分别为 PCA 特征和负荷指标特征对应的附加权重矢量。

通过特征加权组合,电网 6 月数据集最终的特征维度可压缩至 12 维,相较于原始数据降低了 1 428 维,极大减小了数据传输量,降低了传输成本。同时,文献[26]证明加权组合的特征聚类效果要显著优于单独使用以上二者时的结果。

3 基于参数共识的分布式聚类算法

各节点通过特征加权组合算法完成特征提取,即可将提取到的特征用于构建分布式聚类模型。分布式聚类算法的第 1 步是在各节点进行密度峰值聚类,并借助多节点结果求众数,共同确定统一的聚类中心数目;第 2 步是一个迭代过程,该过程首先采用 K -means 算法进行一次聚类构建局部聚类模型并获得局部模型参数,接着通过参数共识算法使各节点模型参数交互计算并返回给原节点模型,最后各节点再根据返回的新参数进行一轮模型更新,此时完成一轮迭代。该算法会反复执行从局部聚类到模型更新的过程,直至整个模型收敛,最终每个节点都会得到一个全局聚类模型。接下来将逐步介绍本文模型中的密度峰值聚类、参数共识以及分布式聚类算法。

3.1 密度峰值聚类算法

本文采用密度峰值的快速聚类 (clustering by fast search and density peak, CFSFDP) 算法^[25]预先确定 K -means 算法所需的类簇数目。对于每一个数据,该算法需要计算局部密度和相对距离这两个参数。当数据点局部密度和相对距离都大于其他点时,该数据点被定义为聚类中心,进一步地计算存在多少这样的数据点,从而可以确定类簇数目。

各节点首先采用密度峰值算法确定各自的类簇数目,随后统计所有节点类簇数目的众数,以该数作为之后局部聚类和分布式聚类的聚类数量,可以避免单个节点异常结果,该过程仅需在初始化时进行一次。本文选取一个典型节点绘制“局部密度 ρ -相对距离 δ ”决策图。由图 2 可知,有 4 个点的局部密度和相对距离远大于其他数据点,最终本文确定类簇数目为 4。

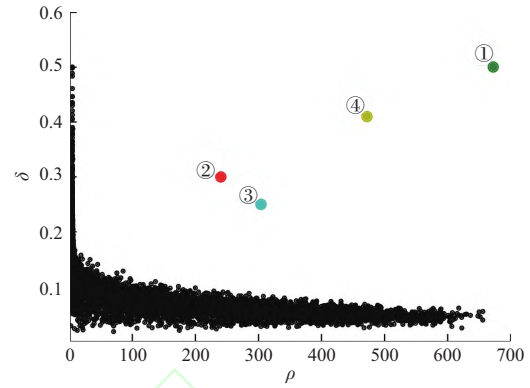


图 2 CFSFDP 算法决策图
Fig. 2 Decision graph of CFSFDP algorithm

3.2 参数共识

考虑到单个数据节点独立聚类时,所用到的用户数量较少、类别较为单调,难以获得理想的聚类效果。参数共识算法可以将每个节点得到的模型参数进行整合,构建得到拥有所有节点特征的全局模型。该算法仅利用抽象的模型参数,实现了数据脱敏,同时极大压缩了传输数据,降低了数据时延。

参数共识是多个参与节点在预设规则下,通过节点信息交互,从而获得对各节点均适用参数的过程。该共识问题的数学表述如下:记加权无向图为 $G=(V, E, A)$,其边集和顶点集分别为 E, V ,边的加权邻接矩阵为 $A=(a_{uv})$ 。定义与节点 v 直接相连节点所组成的集合为 $U_v=\{u \in V: (u, v) \in E\}$, z_v 为节点 v 的观测值。若对于节点 v 与 u ,存在 $z_v=z_u$,则称 v 与 u 共识。进一步地,若图中所有节点 v 与 j ,均存在 $z_v=z_u$,就称图 G 达到共识状态^[27]。

本文中各节点采用的共识策略为平均共识。该算法获得的共识结果是各节点的参数均值,其中的加权邻接矩阵需为双随机矩阵,本文采用文献[28]给定的一种方式构造的该矩阵,具体公式描述可见 3.3 节。

3.3 分布式 K -means 算法

分布式聚类模型主要基于 K -means 算法,由局部聚类、参数共识、全局更新 3 个阶段组成。 K -means 算法的基本思想是将数据集中的所有数据划分为 K 个类别,使得不同类别的数据呈现较大差异,而同一类别中的数据表现相似。本文采用欧氏距离作为评价数据相似度的指标。

局部聚类阶段,各节点首先构建各自的局部聚类模型。记 t 时刻节点 v 的第 k 个聚类中心为 $c_{v,k}(t)$, $C_v(t)=[c_{v,1}(t), c_{v,2}(t), \dots, c_{v,K}(t)]$ 为 t 时刻节点 v 的类簇中心集。基于传统 K -means 算法的局部聚类模型在多地理节点背景下的表达式为^[29]:

$$\mu_{v,i,k}(t+1) = \arg \min_{\mu_{v,i,k} \in I_v} \sum_{i=1}^n \sum_{k=1}^K \mu_{v,i,k}(t) \|x_{v,i} - c_{v,k}(t)\|^2 \quad (3)$$

式中: K 为聚类中心数目, 依赖于数据的分布, 本文采用 3.1 节的密度峰值聚类算法事先确定各节点统一的 K 值; I_v 为节点 v 的所有数据点所属类别的信息集合; $\mu_{v,i,k}$ 为节点 v 用户 i 数据点在类簇 k 中的归属情况, 表达式如式(4)所示。

$$\mu_{v,i,k} = \begin{cases} 1 & x_{v,i} \in I_{v,k} \\ 0 & x_{v,i} \notin I_{v,k} \end{cases} \quad (4)$$

式中: $I_{v,k}$ 为节点 v 中属于类簇 k 的数据点集合。根据以上表达式即可完成局部聚类模型搭建。

参数共识阶段, 首先需要在局部聚类模型的基础上计算得到数据总数与特征矢量之和这两个模型参数。记节点 v 在 $t+1$ 时刻属于类簇 k 的数据总数为 $P_{v,k}(t+1)$, 矢量之和为 $Q_{v,k}(t+1)$, 表达式为:

$$P_{v,k}(t+1) = \sum_{i=1}^n \mu_{v,i,k}(t+1) \quad (5)$$

$$Q_{v,k}(t+1) = \sum_{i=1}^n \mu_{v,i,k}(t+1) x_{v,i} \quad (6)$$

随后, 各节点与其邻居节点通过下列两式进行共识:

$$P_{v,k}(\phi+1, t+1) = \sum_{u \in U_v, a_{uv} \in A} a_{uv} P_{u,k}(\phi, t+1) \quad (7)$$

$$Q_{v,k}(\phi+1, t+1) = \sum_{u \in U_v, a_{uv} \in A} a_{uv} Q_{u,k}(\phi, t+1) \quad (8)$$

式中: a_{uv} 为节点 v 与节点 u 的邻接权重; ϕ 为参数共识算法的迭代次数; A 为拓扑图结构映射的邻接矩阵, 表征地理节点之间的连接关系, 该矩阵借助随机数在以下两个条件的约束下计算构建^[27]: 一是 A 为双随机矩阵, 其各行、各列之和皆为 1; 二是对于邻居节点 u 和 v , 需满足 $a_{uv} \geq \xi$, $a_{vu} \geq \xi$, 其中 ξ 为任意小的正数。经过快速迭代收敛, 各节点的模型参数可以达到共识状态。

全局更新阶段, 各节点获取参数共识后的模型参数, 并通过该参数计算新的类簇中心。记共识停止步骤为 Φ , 并通过下式计算类簇中心 $c_{v,k}(t+1)$:

$$c_{v,k}(t+1) = \frac{Q_{v,k}(\Phi, t+1)}{P_{v,k}(\Phi, t+1)} \quad k=1, 2, \dots, K \quad (9)$$

经过上述 3 个阶段, 节点 v 完成了分布式聚类算法的一次迭代, 之后算法会再次进行局部聚类到参数共识再到全局更新的整个过程。当算法达到规定迭代次数或 $c_{v,k}(t+1)$ 收敛时, 该分布式聚类算法停止。实验证明分布式 K -means 是可以收敛的^[29]。

该算法在节点之间仅传输模型参数, 即使被不法分子截取也无法获得用户信息, 实现了数据脱敏, 解决了用户隐私保护问题。此外, 相较于原本需要

在节点之间传输千万条数据, 该算法仅需传输 2 个参数数据, 大大降低了数据时延。但是算法中的共识步骤增加了节点之间的交互更新次数, 模型收敛速度有所降低, 传输成本问题仍需进一步改善。

综上, 本文所提分布式聚类方法首先采用 PCA-负荷指标的加权组合算法对各节点用户负荷数据进行特征提取, 随后通过密度峰值聚类确定统一的类簇数目。在此基础上, 分布式聚类模型基于 K -means 算法, 以聚类中心作为共识参数, 通过局部聚类、参数共识、全局更新 3 个阶段反复迭代更新, 直至模型收敛, 最终各节点都可构建出适用的模型且模型的聚类中心一致。

4 基于 TCA 的迁移学习

当数据分布发生变化时, 传统机器学习需重新进行建模, 迁移学习则能够解决传统机器学习无法适用于训练集与测试集属于不同特征空间的问题^[30]。本文采用迁移成分分析 (transfer component analysis, TCA) 算法进行迁移学习, 使得新模型可快速迭代收敛, 减少了各节点之间的数据交互次数, 从而降低了传输成本。

TCA 算法是一种基于特征的迁移学习^[31], 可以解决源域与目标域数据分布不同的问题, 其目的是将源域与目标域的特征变换到同一特征空间下使得二者近似服从相同的分布进行学习。在迁移学习之前, 采用最大平均差异 (maximum mean discrepancy, MMD) 距离来量度源域与目标域之间的可迁移性。该距离最小为 0, 表征源域与目标域分布完全相同; 距离大于 1 表示可迁移性较差, 易产生负迁移。TCA 算法中存在超参数优化问题, 不同的超参数最终获得的结果也会有所不同。本文采用网格化搜索的方式, 比对选取最好的迁移学习结果, 从而确定合适的超参数^[32]。

针对分布式聚类的迁移学习, 当有新的数据中心融入该拓扑图中, 原拓扑结构将发生改变。一方面, 对于新节点的数据, 采用数据规约方法将其转换为与原有数据相似的范围, 并通过与原有节点相同的特征提取方法获得新节点特征, 随后计算新节点与原有节点之间的 MMD 距离, 当距离小于 1 时判断源域与目标域之间具备可迁移性, 并通过 TCA 算法使新节点与原有节点的特征近似服从相同的分布; 另一方面, 新节点将获得并采纳原先构建全局模型的模型参数, 随后通过参数共识算法与其余节点重新进行迭代共识, 直至新模型收敛。

5 算例分析

本文选取 2010 年爱尔兰电网 (commission for

energy regulation, CER)^[33-34]和2019年中国北方部分城市电网负荷数据为研究对象,验证本文所提分布式聚类算法的有效性。其中爱尔兰电网数据分属6个独立的数据中心,共6 085个用户,时间粒度为30 min;中国北方部分城市电网数据分属4个独立的数据中心,共219个用户,时间粒度为1 h。

5.1 算法收敛性验证

为更好验证算法的可行性,本文以全局数据点与类簇中心之间距离的误差平方和(sum of squared errors, SSE)作为算法收敛的判断条件之一,对比展示传统集中式K-means聚类、无特征加权组合分布式算法与本文所提分布式K-means算法聚类收敛情况, SSE的计算表达式为:

$$e_{SSE} = \sum_{v \in V} \sum_{i=1}^n \sum_{k=1}^K \mu_{v,i,k} \|x_{v,i} - c_{v,k}\|^2 \quad (10)$$

以爱尔兰电网6月负荷数据为例,各类算法收敛情况如图3所示。

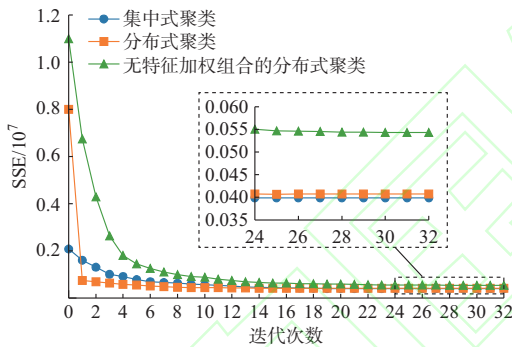


图3 算法收敛结果对比

Fig. 3 Comparison of algorithm convergence results

由图3可见,3种聚类算法均能在迭代一定次数后收敛,表明了分布式聚类算法的可收敛性。无特征加权组合分布式聚类的SSE相较于分布式和集中式聚类收敛更慢,且收敛效果更差,验证了特征加权组合算法的有效性。同时,分布式初始的SSE相较于集中式聚类算法大,但经过1次迭代后能够断崖式收敛到与集中式聚类相近甚至相同的位置。这是因为多个独立分布的数据节点由于局部信息不同,初始化类簇中心时也会有非常大的差距,经过1次参数共识,各节点利用大区域特征达到共识状态,从而实现加速收敛。

5.2 不同方法聚类效果对比分析

为验证分布式聚类算法的有效性,本文对比分析了不同算法聚类结果情况,算法包括集中式聚类、分布式聚类、独立式聚类、基于密度的有噪空间聚类(density-based spatial clustering of applications with noise, DBSCAN)、层次聚类和分布式密度聚类

(density based distributed clustering, DBDC)^[36],其中独立式聚类算法指各个节点仅进行局部聚类而不进行参数共识。考虑到不同类型电力用户的负荷偏度和负荷散度之间具有较大差异,根据这两个指标可以更直观地区分不同用户^[26],因而图4选取了负荷数据中负荷偏度和散度作为横纵坐标,以爱尔兰电网2010年6月与中国北方城市电网2019年6月负荷数据为例,绘制3种算法对用户负荷的分类情况。爱尔兰电网结果如图4和附录B图B1所示,中国北方电网结果见附录B图B2和图B3。

由图4可以看出,集中式、分布式和层次聚类算法都能很好地将用户负荷分为4种类型(I、II、III、IV型),且4种类型之间的界限较为明显;DBSCAN和DBDC算法能将用户有层次地分为4种类型,但类型界限不清晰且噪声点过多;而独立式聚类结果非常差,类簇之间差距很小,难以看出用户分类情况。分析其原因在于DBSCAN和DBDC算法根据样本间距和样本密度逐步寻找类簇,而电力用户数据样本密度不均匀且间距差较大,导致很多高耗能的工商业用户易被识别为噪声点。DBDC算法更是由于地理节点样本数量少、间距大且节点之间没有进行有效通信,难以进行有效聚类。而独立式聚类中,各个节点仅利用各自的局部信息进行分类,而局部信息本身由于地理分布不同,导致各节点局部信息之间差异较大,进一步影响了聚类结果。对比图4与附录B图B1至图B3可以看到,分布式聚类得到的结果和集中式基本相同,且分布式聚类中各节点模型基本一致,是因为分布式聚类采用了参数共识策略,各节点之间能够有效传递不同的区域特征,使得每个节点最终都享有全局信息,并能收敛得到很好地聚类模型。为了更量化地反映各个聚类模型所得结果的差异性,本文对6种聚类模型采用轮廓系数(silhouette coefficient, SC)、戴维森堡丁指数(Davies-Bouldin index, DBI)、CH(Calinski-Harabasz, CH)指标和邓恩指数(Dunn validity index, DVI)衡量对负荷用户分类的效果。SC综合了内聚度和分离度两种系数,其值范围在 $[-1, 1]$,越接近于1,效果越好;DBI计算类簇内平均距离和类簇之间最小距离的比值,该值越小,聚类效果越好;CH指标计算类簇内各点与类簇中心的距离平方和来量度类内的紧密程度,该值越大代表类簇自身越紧密;DVI综合衡量簇内和簇间距离,其值越大说明聚类效果越好。以爱尔兰电网2010年6月与中国北方部分城市电网2019年6月负荷数据为例,进行20次实验并取各指标的均值,不同算法聚类结果的性能指标和计算时长见附录B表B1。

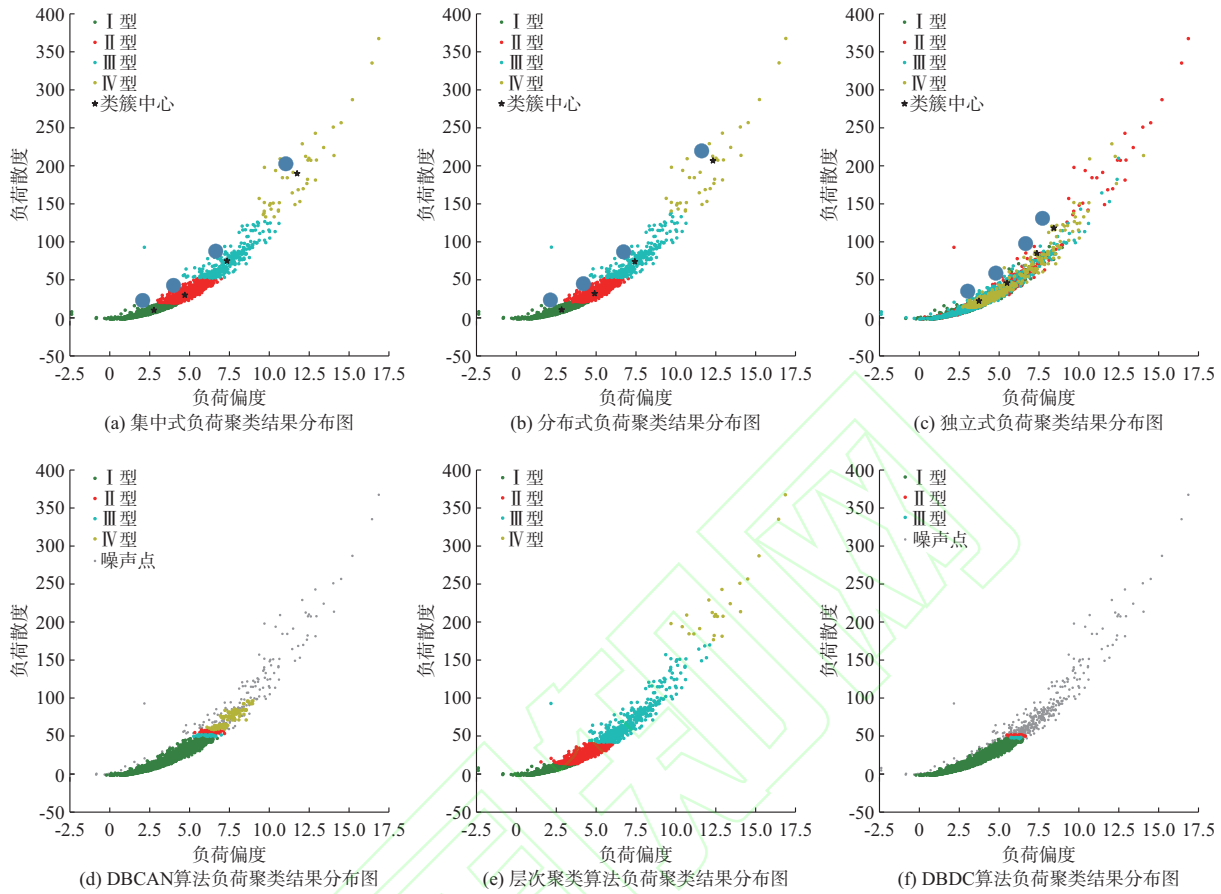


图4 爱尔兰电网聚类结果
Fig. 4 Clustering results of CER

由附录B表B1可知,本文所提分布式聚类多数指标都能达到最佳,集中式与分布式协同聚类算法结果相近,层次聚类法各项指标居中,而独立式聚类、DBSCAN和DBDC结果最差,证明分布式聚类算法可以通过参数共识步骤实现数据集中训练并获得较好的效果。从DBI簇间指标来看,层次聚类能够实现非常好的不同簇间划分,分布式和集中式聚类的簇间划分次之;从CH簇内指标来看,分布式聚类簇内划分最好,集中式划分次之,层次聚类法簇内划分稍差;从SC、DVI综合指标来看,分布式和集中式聚类能够很好地平衡簇内和簇间距离,实现优质分类,层次聚类综合而言没有分布式K-means算法好,而独立式聚类、DBSCAN和DBDC在样本数据量小、样本间距较大的情况下表现最差。从算法耗时来看,分布式聚类的计算时长远小于集中式聚类和DBDC算法,略大于独立式、层次聚类法和DBSCAN算法。集中式聚类由于所用数据量较大,多次迭代计算耗时也较大;DBDC算法为了实现分布式聚类进行了多次划分导致计算耗时增大;层次聚类和DBSCAN仅进行一次聚类或划分,算法复杂度小,因而耗时最少;分布式和独立式聚类的单节点

数据量小,计算耗时也较小。

进一步对比集中式和分布式聚类算法数据传输量和时间的不同,如附录B表B2所示。可以很明显看到,分布式聚类的数据传输量约是集中式的千分之一,传输速度比集中式快约3 000倍。因为分布式聚类在各节点之间仅传输两个脱敏的模型参数,实现了用户数据隐私保护的同时也大大降低了数据的传输时延。但相较于集中式仅需传输一次数据,分布式聚类由于共识算法需要多次传输迭代,增加了模型构建时间。然而分布式仅比独立式聚类慢约1 s,也证明了参数共识步骤耗时极小。

综上结果可以看出,在地理分布式的背景下,分布式聚类算法综合簇内和簇间的划分最好,实现了数据脱敏并减少了传输成本,且计算速度远快于集中式聚类和DBDC算法。

5.3 模型可迁移性验证

为验证算法的可迁移性,选取爱尔兰电网和中国北方部分城市电网负荷数据分别进行分布式聚类并将聚类结果简单拼接到同一坐标中,如附录C图C1所示。可以看出,两地电网用户分布不同,具有不同的区域地理特征。以爱尔兰电网数据作为源

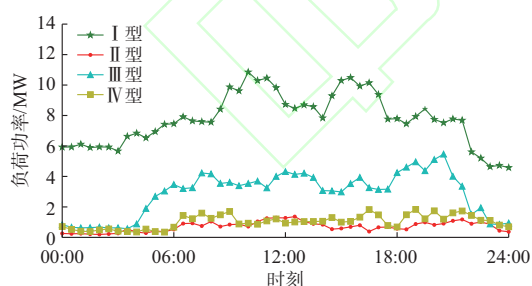
域,中国北方部分城市电网数据作为目标域,计算两者之间的MMD距离为0.112,证明两地数据特征分布有所不同且具备可迁移性。接下来将对两地数据集采用TCA算法进一步验证模型可迁移性。

选取爱尔兰电网负荷数据先构建包含6个地理节点的分布式聚类模型,随后将中国北方城市电网4个地理节点的负荷数据依次加入拓扑图中作为新加入的地理节点,分布式聚类模型类簇中心移动迁移情况如附录C图C1(d)所示,迁移前后评价指标见表C1。

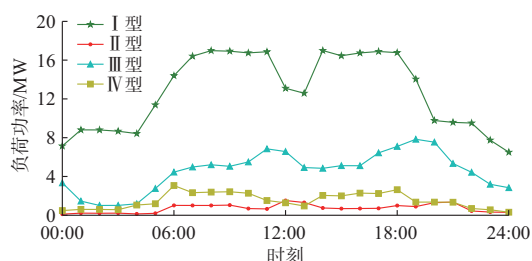
对比图C1(c)和(d)可以看到,经过特征迁移后两地电网中均有部分用户的归类发生偏移且类簇中心也有轻微偏移,最终可以收敛得到容纳两地数据的新聚类模型,说明分布式聚类迁移学习具有很好的效果。根据附录C表C1,特征迁移后DBI、CH、SC指标相较于迁移前略差,是由于两电网用户数据分布不同,经过特征迁移类簇中心发生移动,原有的小部分用户被重新分配导致指标略差。而MMD距离迁移后趋近于0,说明两电网特征分布近乎相同,也证明本文所提分布式聚类算法能够在原有节点的基础上融入新节点实现在线快速聚类,可以较好地应用在迁移学习框架下。

5.4 聚类用户负荷特征分析

采用分布式聚类模型进行聚类,标记所得聚类结果每一类的类簇中心作为典型用电负荷用户,根据标记抽取原始负荷数据集的典型用户负荷数据,可以绘制出对应4类典型负荷曲线进行分析,结果见图5和附录D图D1。



(a) 爱尔兰电网典型用电负荷曲线



(b) 中国北方部分城市电网典型用电负荷曲线

图5 分布式协同聚类典型用户负荷曲线
Fig. 5 Typical user load curves of distributed collaborative clustering

以电网6月数据结果为例,如图5所示,可以看出用户用电模式丰富多变,I型用户负荷水平普遍很高,一般有两个高峰用电时段,集中在08:00—12:00和14:00—18:00时段,晚间仍有高负荷水平,属于高负荷部分迎峰用电;II型曲线负荷水平比其他3类都低且较为均匀,用电量基本不超过0.8 MW,高峰用电通常在06:00—09:00、12:00—14:00和19:00—23:00这3个时段,属于低负荷迎峰用电;III型用户日间08:00—19:00时段的负荷水平较大,午间和晚间有负荷小峰值,夜间负荷水平急速下降,是典型的日间高负荷用电;IV型用户06:00—11:00和14:00—19:00时段负荷水平较高,夜间用电趋近于0,峰谷形态与II型互补,属于低负荷部分迎峰用电。

采用本文所提的分布式聚类算法,可以清楚地将电网负荷用户划分成4类负荷用户类型,为电网后期运行、规划打下基础,也证明了该算法的有效性和可行性。

6 结语

本文针对地理分布式背景下的电力数据,构建了考虑特征迁移的分布式聚类模型框架,提出了一种基于参数共识利用局部信息得到全局聚类模型的分布式协同聚类算法。算法针对单地理节点采用PCA-负荷指标获得加权组合特征,考虑密度峰值信息确定类簇数目,通过参数共识利用局部模型参数使得每一个电力数据中心获得包含全局信息的全局聚类模型。针对新加入的数据中心,采用TCA算法进行迁移学习,实现在线构建分布式聚类模型。通过算法对比试验表明,本文所提的分布式协同聚类算法能够在地理分布式背景下借助少量脱敏数据传输,实现用户隐私保护,有效降低数据时延,同时能保留区域特征,并利用整体区域特征实现加速收敛快速构建全局聚类模型,获得很好的负荷用户分类效果,帮助分析电力用户负荷特性。

然而,算法中的共识步骤需要节点之间多次交互迭代,传输成本问题仍有待进一步解决。此外,多个地理节点聚类数目要求一致的条件稍显强硬,对方法的应用上增加了很大的限制。在今后的研究工作中,一方面需要改善参数共识算法,减少节点之间交互,另一方面也需增强算法的灵活性,针对不同地理节点不同聚类数目也可构建和迁移模型。

附录见本刊网络版(<http://www.aeps-info.com/aeps/ch/index.aspx>),扫英文摘要后二维码可以阅读网络全文。

参考文献

- [1] 薛禹胜, 赖业宁. 大能源思维与大数据思维的融合(一)大数据与电力大数据[J]. 电力系统自动化, 2016, 40(1): 1-8.
XUE Yusheng, LAI Yening. Integration of macro energy thinking and big data thinking part one big data and power big data [J]. Automation of Electric Power Systems, 2016, 40(1): 1-8.
- [2] 孙宏斌, 郭庆来, 潘昭光. 能源互联网: 理念、架构与前沿展望[J]. 电力系统自动化, 2015, 39(19): 1-8.
SUN Hongbin, GUO Qinglai, PAN Zhaoguang. Energy Internet: concept, architecture and frontier outlook [J]. Automation of Electric Power Systems, 2015, 39(19): 1-8.
- [3] JIBRAN M, NASIR H A, QURESHI F A, et al. A demand response-based solution to overloading in underdeveloped distribution networks [J]. IEEE Transactions on Smart Grid, 2021, 12(5): 4059-4067.
- [4] BAEK K, LEE E, KIM J. Resident behavior detection model for environment responsive demand response [J]. IEEE Transactions on Smart Grid, 2021, 12(5): 3980-3989.
- [5] 葛磊蛟, 赵康, 孙永辉, 等. 基于孪生网络和长短期记忆网络结合的配电网短期负荷预测[J]. 电力系统自动化, 2021, 45(23): 41-50.
GE Leijiao, ZHAO Kang, SUN Yonghui, et al. Short-term load forecasting of distribution network based on combination of Siamese network and long short-term memory network [J]. Automation of Electric Power Systems, 2021, 45(23): 41-50.
- [6] WEN H L, GU J, MA J H, et al. Probabilistic load forecasting via neural basis expansion model based prediction intervals [J]. IEEE Transactions on Smart Grid, 2021, 12(4): 3648-3660.
- [7] ALAVI S A, MEHRAN K, VAHIDINASAB V, et al. Forecast-based consensus control for DC microgrids using distributed long short-term memory deep learning models [J]. IEEE Transactions on Smart Grid, 2021, 12(5): 3718-3730.
- [8] WANG C Y, PANG K Y, SHAHIDEPOUR M, et al. MILP-based fault diagnosis model in active power distribution networks [J]. IEEE Transactions on Smart Grid, 2021, 12(5): 3847-3857.
- [9] 王浩, 杨东升, 周博文, 等. 基于并联卷积神经网络的多端直流输电线路故障诊断[J]. 电力系统自动化, 2020, 44(12): 84-92.
WANG Hao, YANG Dongsheng, ZHOU Bowen, et al. Fault diagnosis of multi-terminal HVDC transmission line based on parallel convolutional neural network [J]. Automation of Electric Power Systems, 2020, 44(12): 84-92.
- [10] 万磊, 陈成, 黄文杰, 等. 基于BRB和LSTM网络的电力大数据用电异常检测方法[J]. 电力建设, 2021, 42(8): 38-45.
WAN Lei, CHEN Cheng, HUANG Wenjie, et al. Power abnormality detection method based on power big data applying BRB and LSTM network [J]. Electric Power Construction, 2021, 42(8): 38-45.
- [11] 肖勇, 郑楷洪, 余忠忠, 等. 基于三次指数平滑模型与DBSCAN聚类的电量数据异常检测[J]. 电网技术, 2020, 44(3): 1099-1104.
XIAO Yong, ZHENG Kaihong, YU Zhongzhong, et al. Power data anomaly detection based on holt-winters model and DBSCAN clustering [J]. Power System Technology, 2020, 44(3): 1099-1104.
- [12] ARAVENA I, CHAPIN S J, PONCE C. Decentralized failure-tolerant optimization of electric vehicle charging [J]. IEEE Transactions on Smart Grid, 2021, 12(5): 4068-4078.
- [13] XU M D, ZHANG G G, LI W F, et al. Coordinated control strategy for grid-connected integrated energy system of wind, photovoltaic and hydrogen [C]// 2019 IEEE 8th International Conference on Advanced Power System Automation and Protection, October 21-24, 2019, Xi'an, China: 1120-1124.
- [14] KHAZAEI J, NGUYEN D H. Multi-agent consensus design for heterogeneous energy storage devices with droop control in smart grids [J]. IEEE Transactions on Smart Grid, 2019, 10(2): 1395-1404.
- [15] 吴庆曦, 彭晖, 王瑾, 等. 电网调控集群分布式实时数据库的设计与关键技术[J]. 电力系统自动化, 2017, 41(22): 89-95.
WU Qingxi, PENG Hui, WANG Jin, et al. Design and key technologies of cluster based distributed real-time database for power dispatching and control system [J]. Automation of Electric Power Systems, 2017, 41(22): 89-95.
- [16] HU Z M, LI B C, LUO J. Time- and cost- efficient task scheduling across geo-distributed data centers [J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(3): 705-718.
- [17] PAL S, SIKDAR B, CHOW J H. Classification and detection of PMU data manipulation attacks using transmission line parameters [J]. IEEE Transactions on Smart Grid, 2018, 9(5): 5057-5066.
- [18] CHEN L, LU R X, CAO Z F, et al. MuDA: Multifunctional data aggregation in privacy-preserving smart grid communications [J]. Peer-to-Peer Networking and Applications, 2015, 8(5): 777-792.
- [19] 张少敏, 赵乙桥, 王保义. 智能电网下保护用户隐私的无证书环签名方案[J]. 电力系统自动化, 2018, 42(3): 118-123.
ZHANG Shaomin, ZHAO Yiqiao, WANG Baoyi. Certificateless ring signature scheme for preserving user privacy in smart grid [J]. Automation of Electric Power Systems, 2018, 42(3): 118-123.
- [20] GOPE P, SIKDAR B. Privacy-aware authenticated key agreement scheme for secure smart grid communication [J]. IEEE Transactions on Smart Grid, 2019, 10(4): 3953-3962.
- [21] NOORWALI A, RAO R, SHAMI A. Modeling and delay analysis of wide area network in smart grid communications [C]// 2016 IEEE Smart Energy Grid Engineering (SEGE), August 21-24, 2016, Oshawa, Canada: 347-352.
- [22] VULIMIRI A, CURINO C, GODFREY B, et al. Global analytics in the face of bandwidth and regulatory constraints [C]// Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation, May 1, 2015, Vancouver, Canada: 323-336.
- [23] CANO I, WEIMER M, MAHAJAN D, et al. Towards geo-distributed machine learning [EB/OL]. [2020-01-09]. <https://arxiv.org/abs/1603.09035v1>.
- [24] 史佳琪, 谭涛, 郭经, 等. 基于深度结构多任务学习的园区型综合能源系统多元负荷预测[J]. 电网技术, 2018, 42(3): 698-707.
SHI Jiaqi, TAN Tao, GUO Jing, et al. Multi-task learning based on deep architecture for various types of load forecasting in regional energy system integration [J]. Power System Technology, 2018, 42(3): 698-707.
- [25] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344: 1492-1496.

- [26] 陈君,黄立峰,高成英,等.基于Gabor小波加权组合特征的性别识别[J].计算机辅助设计与图形学学报,2015,27(9):1767-1774.
CHEN Jun, HUANG Lifeng, GAO Chengying, et al. Gender recognition based on weighted combination of Gabor wavelets [J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(9): 1767-1774.
- [27] LIU P Y, YANG C Y, WU J, et al. Hybrid Features based K-means Clustering Algorithm for use in Electricity Customer Load Pattern Analysis[C]// 37th Chinese Control Conference (CCC), July 25-27, 2018, Wuhan, China.
- [28] ABDESSAMEUD A, TAYEBI A. Distributed consensus algorithms for a class of high-order multi-agent systems on directed graphs[J]. IEEE Transactions on Automatic Control, 2018, 63(10): 3464-3470.
- [29] PANG Y P, HU G Q. Randomized gradient-free distributed optimization methods for a multiagent system with unknown cost function[J]. IEEE Transactions on Automatic Control, 2020, 65(1): 333-340.
- [30] LIN P, WANG Y H, QI H S, et al. Distributed consensus-based K-means algorithm in switching multi-agent networks[J]. Journal of Systems Science and Complexity, 2018, 31(5): 1128-1145.
- [31] POPOV N V, KLIONSKIY D M. Research of dimension reduction algorithms of feature space in data analysis task[C]// 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering. January 27-30, 2020, St. Petersburg and Moscow, Russia: 458-462.
- [32] MATASCI G, VOLPI M, KANEVSKI M, et al. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(7): 3550-3564.
- [33] CHIDAMBARAM B, RAVICHANDRAN M, SESHADRI A, et al. Computational heat transfer analysis and genetic algorithm-artificial neural network-genetic algorithm-based multiobjective optimization of rectangular perforated plate fins [J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2017, 7(2): 208-216.
- [34] Irish Social Science Data Archive. SN: 0013-00. Commission for energy regulation (CER) smart metering project [EB/OL]. [2020-01-09]. <https://www.ucd.ie/issda/data/commissionforenergyregulationcer>.
- [35] CHEN J X, LI Y L, SUN P, et al. An improved distributed clustering algorithm based on density [C]// 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), November 1-3, 2015, Tianjin, China: 133-136.
- 刘家丞(1996—),男,硕士研究生,主要研究方向:分布式机器学习、电力大数据。E-mail: xjsk1016tansuo@stu.xjtu.edu.cn
- 吴江(1980—),男,通信作者,博士,副教授,博士生导师,主要研究方向:电力系统优化调度、分布式机器学习算法。E-mail: jiangwu@xjtu.edu.cn
- 刘鹏远(1993—),男,硕士,主要研究方向:电力大数据、智能电网。E-mail: lpy1001@stu.xjtu.edu.cn

(编辑 孔丽蓓)

Geographically Distributed Collaborative Clustering Method for Load Pattern Analysis

LIU Jiacheng¹, WU Jiang¹, LIU Pengyuan^{1,2}, XU Zhanbo¹, LI Xiaopeng¹, GUAN Xiaohong¹

(1. Ministry of Education Key Lab for Intelligent Networks and Network Security (Xi'an Jiaotong University), Xi'an 710049, China; 2. Xi'an Power Supply Company of State Grid Shaanxi Electric Power Company, Xi'an 710048, China)

Abstract: Under the framework of Energy Internet, data transmission delay to users and regulatory requirements of power grid company promote the establishment of power data centers all over the country, therefore the power data naturally presents geographic distribution, in this regard the clustering algorithm of user load characteristics in geographically distributed scenarios is studied. Firstly, in the face of the characteristic analysis of user load in geographic nodes, based on the principal component analysis (PCA)-load index feature weighted combination algorithm, a K -means algorithm considering density peak information is proposed, which provides the support for information consensus among geographical nodes. Secondly, according to the geographically distributed network perception structure of user needs, the framework of distributed clustering model considering feature migration is constructed, and the distributed K -means algorithm is proposed to get the global clustering model realizing on the premise that only the public information is transmitted between nodes, from the local information of nodes through parameter consensus. Finally, the model and algorithm are verified by user load data from some cities in Ireland and Northern China. The results show that distributed K -means can use global information and consider differences in different regions to better identify typical power consumption curves, and the algorithm has better transferability.

This work is supported by National Key R&D Program of China (No. 2016YFB0901900) and National Natural Science Foundation of China (No. 61803297).

Key words: distributed clustering; load characteristics analysis; parameter consensus; clustering by fast search and find of density peaks (CFSFDP); transfer learning



附录 A

表 A1 负荷特征指标
Table A1 Load characteristic indicators

负荷特征指标	表达式
负荷峰值	$L_1 = \max \{d(t); t \in (0, T_f)\}$
平均负荷	$L_2 = \frac{1}{T_f} \int_0^{T_f} d(t) dt = \frac{\sum_{j=1}^{N_T} d_j \Delta t_j}{\sum_{j=1}^{N_T} \Delta t_j}$
均方负荷	$L_3 = \sqrt{\frac{\int_0^{T_f} d^2(t) dt}{T_f}} = \sqrt{\frac{\sum_{j=1}^{N_T} d_j^2 \Delta t_j}{\sum_{j=1}^{N_T} \Delta t_j}}$
负荷率	$L_4 = \frac{1}{T_f d_{\max}^2} \int_0^{T_f} d(t) dt = \frac{\sum_{j=1}^{N_T} d_j \Delta t_j}{d_{\max} \sum_{j=1}^{N_T} \Delta t_j}$
负荷偏度	$L_5 = \frac{\sum_{j=1}^{N_T} (d_j - \bar{d})^3}{(N_T - 1) \sigma_x^3}$
负荷散度	$L_6 = \frac{\sum_{j=1}^{N_T} (d_j - \bar{d})^4}{(N_T - 1) \sigma_x^4}$

附录 B

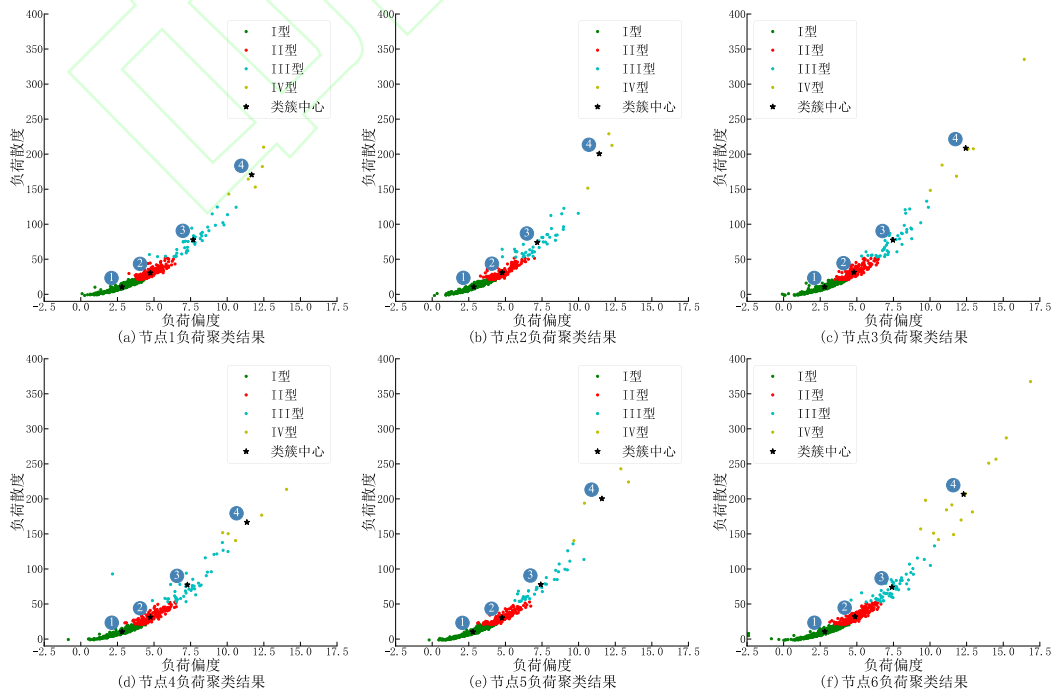


图 B1 爱尔兰电网各地理节点分布式聚类结果

Fig. B1 Distributed clustering results of geographical nodes of CER

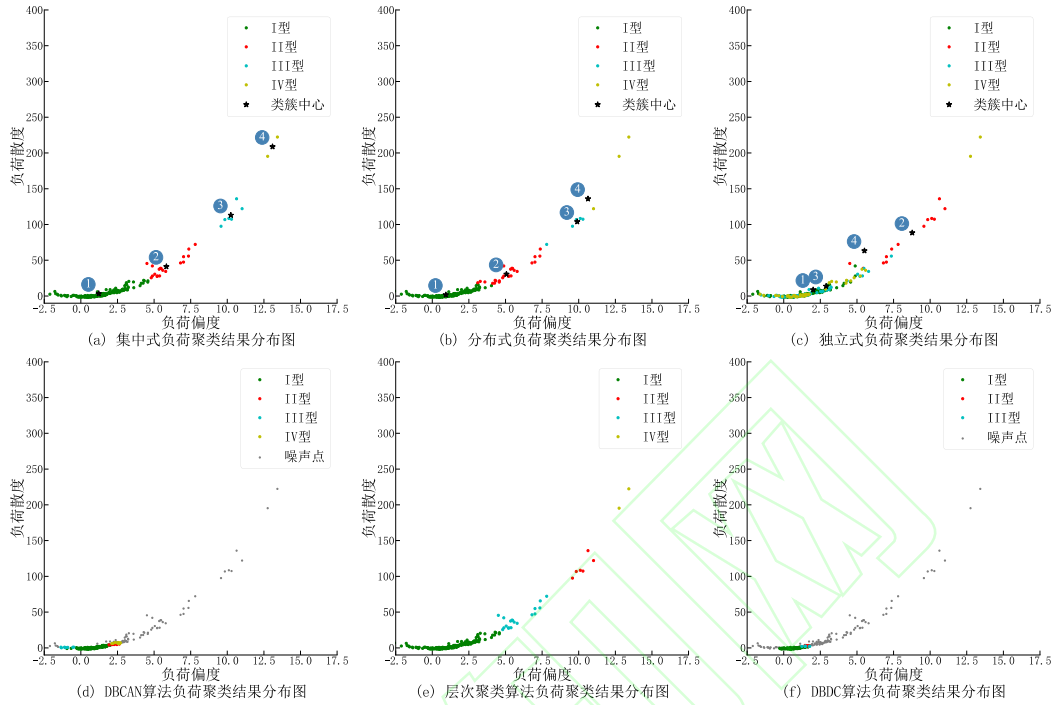


图 B2 中国北方部分城市电网负荷聚类结果
Fig. B2 Clustering results of state grid of some cities in northern China

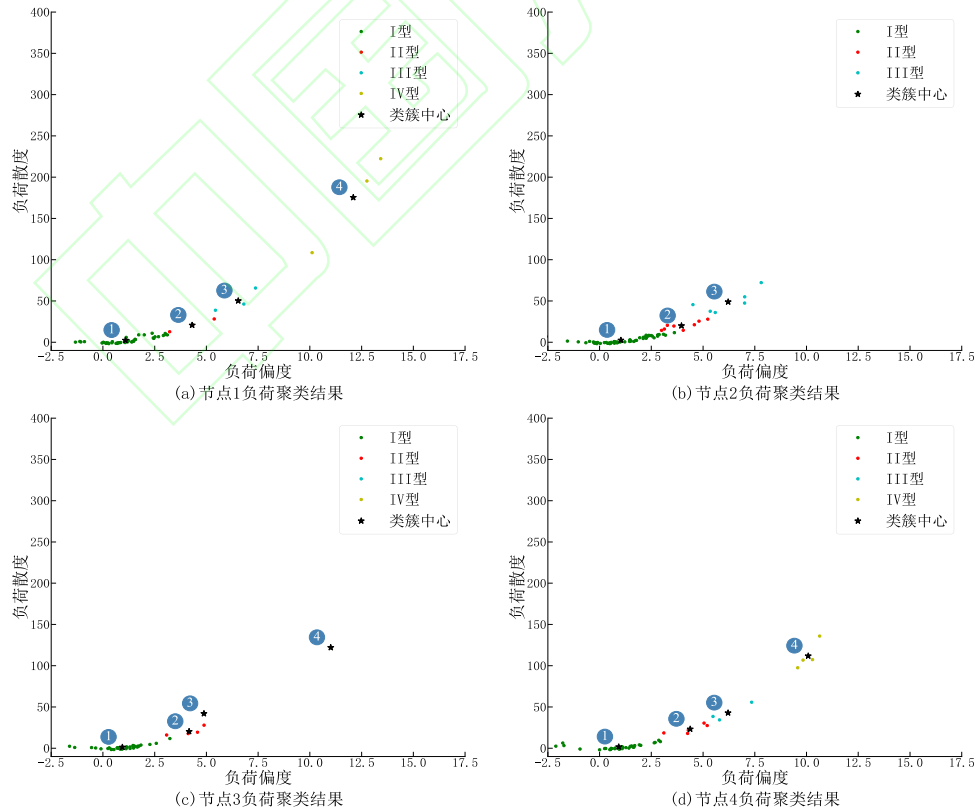


图 B3 中国北方部分城市电网各地理节点分布式聚类结果
Fig. B3 Distributed clustering results of geographical nodes of some cities in northern China

表 B1 不同算法性能指标及计算时长
Table B1 Indicator values and calculation time of different clustering algorithms

算法	爱尔兰电网					中国北方部分城市电网				
	SC	DBI	CH	DVI	计算时长/s	SC	DBI	CH	DVI	计算时长/s
分布式聚类	0.645	0.501	12281.00	0.241	6.782	0.797	0.374	1127.77	0.447	0.039
集中式聚类	0.632	0.512	12338.86	0.235	14.602	0.775	0.381	1103.14	0.441	0.092
独立式聚类	-0.041	37.868	347.85	0.005	6.131	0.069	14.349	81.62	0.051	0.023
DBSCAN	0.011	28.743	413.23	0.012	3.012	0.005	43.743	401.23	0.082	0.012
层次聚类	0.547	0.448	11138.11	0.224	5.361	0.711	0.311	985.43	0.314	0.022
DBDC	0.445	0.980	8145.23	0.002	12.772	-0.015	1.24	8.63	0.042	0.075

表 B2 集中式聚类和分布式聚类数据传输比较
Table B2 Comparison of data transmission between centralized clustering and distributed clustering

算法	爱尔兰电网		中国北方部分城市电网	
	传输量/MB	传输时间/s	传输量/MB	传输时间/s
分布式聚类	0.58	0.05	0.24	0.01
集中式聚类	530	327	95	57

附录 C

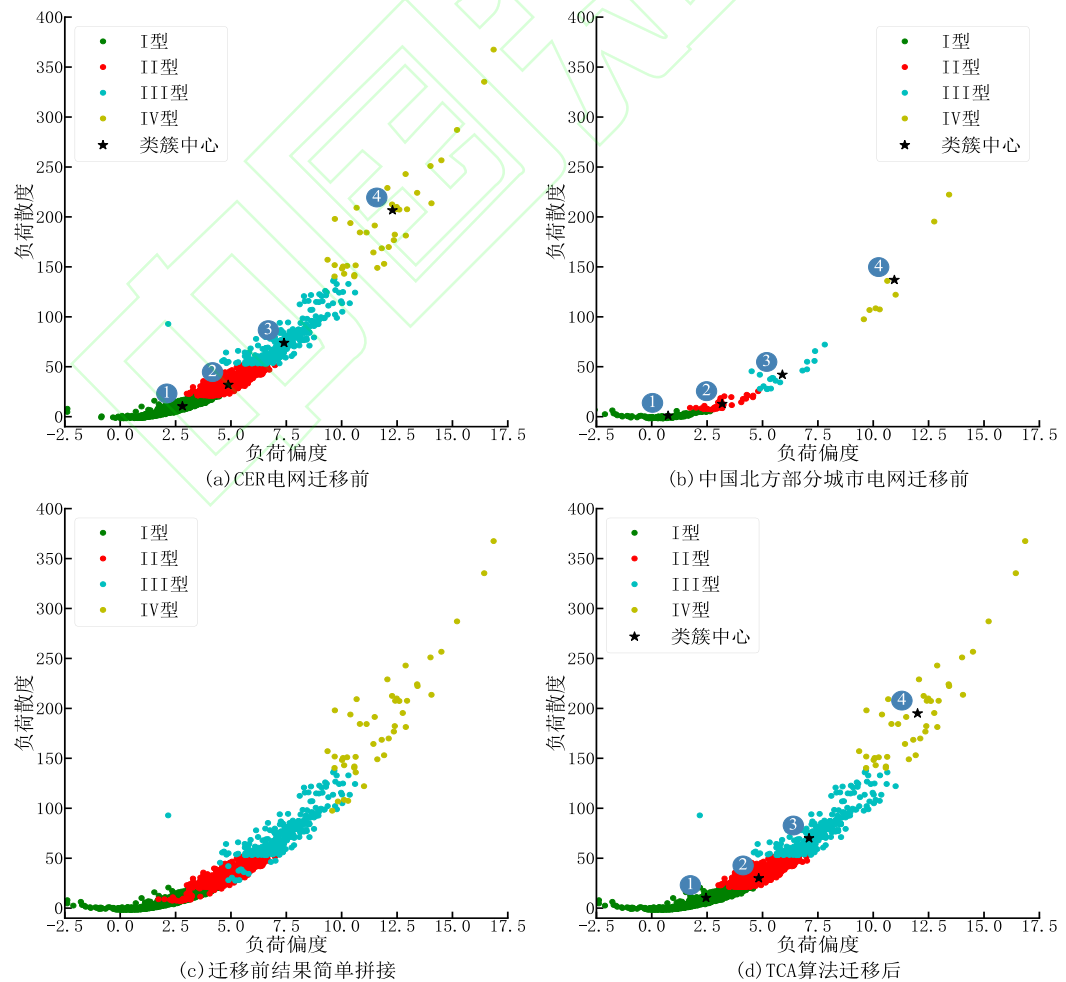


图 C1 分布式聚类模型特征迁移结果
Fig. C1 Feature migration results of distributed clustering model

表 C1 分布式聚类模型特征迁移前后指标评价
Table C1 Indicator values of distributed clustering model before and after feature migration

指标	特征迁移前	特征迁移后
MMD	0.112	0.003
SC	0.637	0.632
DBI	0.573	0.611
CH	12013.271	11431.213

附录 D

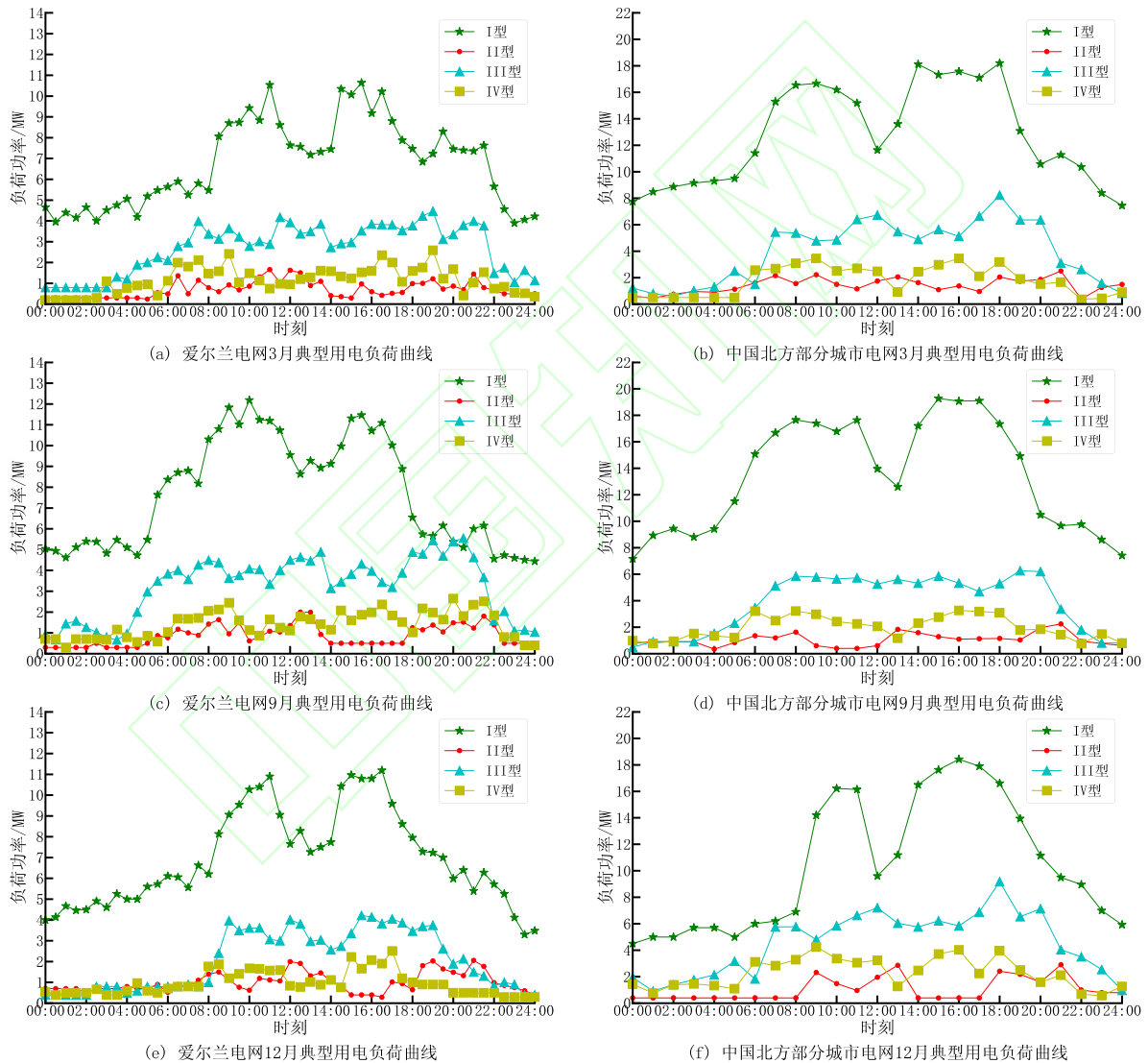


图 D1 不同月份电网典型用户负荷曲线
Fig. D1 Typical user load curve of power grid in different months