# 基于局部密度的最小生成树聚类算法及其在电力大数据的应用

靳文星[1],王电钢[2],张哲敏[1]
(1. 上海电力大学计算机科学与技术学院,上海 200090;
2. 国网四川省电力公司信息通信公司,四川 成都 610041)

**摘 要**:电力大数据主要来源于电力生产和电能使用的发电、输电、变电、配电、用电和调度各个环节,,如何运用这些数据提高电力管理工作的智能化水平已经成为相关电力环节十分重要的研究课题之一。但现有电力大数据中用到的聚类方法却不能发现任意形状的数据集聚类类别(即类簇),这影响了电力大数据在应用中的计算精度与计算时长。因此提出了一种新的算法,即使用局部密度峰值和基于共享邻点的距离,更好地结合了密度与距离的关系,表示出数据之间的差异。使用局部密度峰值并用基于共享邻点的距离来构造最小生成树,然后重复切割最长的边,直到找到给定数量的簇。在电力大数据应用上的实验结果表明,该算法在具有良好的效果。

**关键词**:最小生成树;聚类;局部密度峰值;基于共享邻点的距离
**中图分类号**:TM 769 **文献标志码**:A **文章编号**:1003 - 6954(2021)04 - 0016 - 04
DOI:10.16527/j.issn.1003 - 6954.20210404

# Minimum Spanning Tree Clustering Based on Local Density and Its Application to Power Big Data

Jin Wenxing[1], Wang Diangang[2], Zhang Zhemin[1]
(1. School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China;
2. State Grid Sichuan Information and Communication Company, Chengdu 610041, Sichuan, China)

**Abstract**: Power big data mainly comes from all aspects of power generation, transmission, transformation, distribution, power consumption and dispatching of power production and energy use. How to use these data to improve the intelligent level of power management has become one of the most important research topics of the related power links. However, the existing clustering methods used in power big data can not find clusters of arbitrary shape, which affects the calculation accuracy and calculation time in the application to power big data. A new algorithm is proposed, which uses the local density peak and the distance based on shared neighbor points to better combine the relationship between density and distance and express the differences between data. The minimum spanning tree (MST) is constructed by using the local density peak and the distance based on the shared neighbor, and then the longest edge is cut repeatedly until a given fnumber of clusters is ound. The experimental results show that the proposed algorithm has a good effect in the application to power big data.

**Key words**: minimum spanning tree; clustering; local density peaks; shared neighbor-based distance

## 0 引 言

近些年,针对电力大数据收集和存储中数据量大、数据收集不精准的问题,先后提出并采用了 K - means、K - medoids[1] 和一些改进之后的K - means 算法,但是这些算法的使用都必须初始化聚类中心。为了避免初始化聚类中心,在算法领域中的 AP 算法[2]将所有数据点都视为潜在的中心。K - AP[3]是 AP 算法的改进,它在消息传递过程中引入约束,利用 K 簇产生的直接结果,然而,由于每个点总是分配到最近的中心,导致这些算法不能发现任意形状的聚类(即类簇)。还有一种快速搜索发现密度峰值[4](density peak,DP)的聚类算法,选择局部密度

**基金项目**:国家电网有限公司总部科技项目(输变电设备物联网边缘智能关键技术研究及应用 52199920002W)

最大的点作为聚类中心,将其余点作为密度最大的近邻分配到同一个类别中。假设每个类簇都有收缩的密度核,大致保留了类簇的形状,并提出了一种基于密度核的聚类算法,称为 Dcore[5]。基于密度的聚类算法 DBSCAN[6]将聚类定义为由稀疏区域分隔的稠密区域。它的关键思想是,设定集群的每个核心点,在每个核心点周围给定半径内必须包含有参数设定数量的点(如参数设定为 30,则若一点给定半径范围内有超过 30 点,即认定此点为核心点)。Dcore 和 DBSCAN 可以有效地识别具有任意形状的数据集,但是它们必须设置许多参数。

针对电力大数据中无法高效识别具有任意形状数据集的问题,提出了基于最小生成树(minimum spanning tree,MST)和局部密度峰值(local density peak,LDP)的聚类算法,称为 LDP – MST,它在发现复杂数据时,不仅计算效率高,而且可以与其他先进的聚类方法相媲美。在 LDP – MST 中,首先找到局部密度峰值,将剩余的点分配到相应的局部密度峰值;然后,定义一个新的基于共享邻点的局部密度峰值之间的距离,并利用新的距离在局部密度峰值上构造最小生成树;最后通过不断地去除最长边,得到了最终的聚类。

## 1　基于局部密度峰值和共享邻点的 MST 聚类

现有的基于 MST 的聚类算法,在整个数据集上构造 MST 的时候,因为只利用树中包含的边缘信息对数据集进行划分,导致数据的计算量很大,而且容易受到噪声点的影响。基于此问题,提出了一种基于局部密度峰值的最小生成树聚类算法(以图 1 所示的一个数集为例)。首先,选取相邻区域中局部密度最大的点作为局部密度峰值,并将其余点分配到相应的局部密度峰值附近,如图 1(a)所示;然后,定义一个新的局部密度峰值之间的距离分类(它考虑了欧几里得距离和邻点信息),利用局部密度峰值和距离构建 MST,如图 1(b)所示。在此之后,根据新的距离不断地去除最长的边,并进行距离连线,直到得到期望的簇数。图 1(c)中链接不同簇之间的边是需要从 MST 中更正的边,最后得到如图 1(d)所示的聚类结果。整个算法过程由于只在局部密度峰值上构造 MST,减少了噪声点的干扰,大大提高了算法的效率。

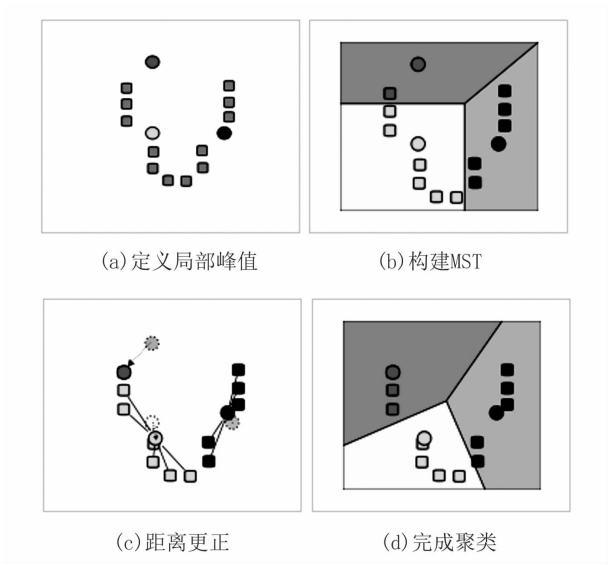



**图 1　LDP – MST 的主要思想**

### 1.1　局部密度

为了找到局部密度峰值,首先定义点的局部密度。因为稠密区域的点与其近邻点的距离总和通常小于稀疏区域的点与近邻点的距离之和,在稠密区域,nb 值较大;在稀疏区域,nb 值较小,所以,点 $p$ 的局部密度与 $\mathrm{nb}(p)$ 的值成正比,与点 $p$ 和其相邻点之间的距离成反比。利用这一特性,计算局部密度 $\rho(p)$:

$$\rho(p)=\frac{\mathrm{nb}(p)}{\sum_{q\in \mathrm{NNK}(p)} d(p,q)} \tag{1}$$

式中:$\mathrm{nb}(p)$ 为到达自然特征值时的 $p$ 的反向近邻数;$\mathrm{NNK}(p)$ 为 $p$ 的反向 $k$ 近邻;$d(p,q)$ 为 $p$ 和 $q$ 之间的距离。

如图 2 中给出了每个局部密度峰值的邻域(图中粗线表示),其中包括其成员和一些额外的最近邻域,在图中用不同点间的连线表示。共享邻点的数量和密度越大,表示它们之间的距离越小。

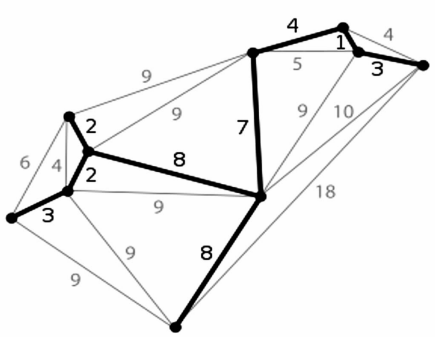
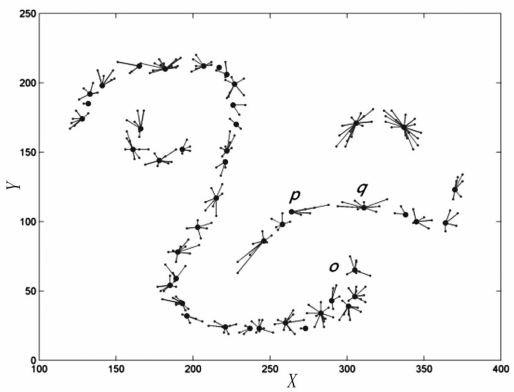



**图 2　LDP 的邻点和共享邻点**

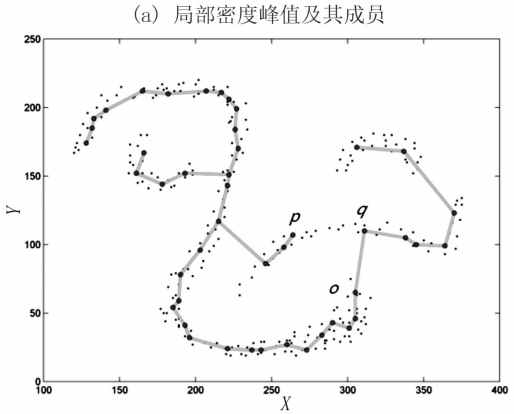
### 1.2 基于共享邻点的局部密度峰值之间的距离

由于欧几里得距离不能很好地对复杂数据进行恰当度量,且由于大多时候都测量不到图形点位置的先验信息,导致不能直接得到准确的测量距离。基于局部密度峰值的共享邻域,采用了一个新的距离,即基于共享邻点的局部密度峰值之间的距离。

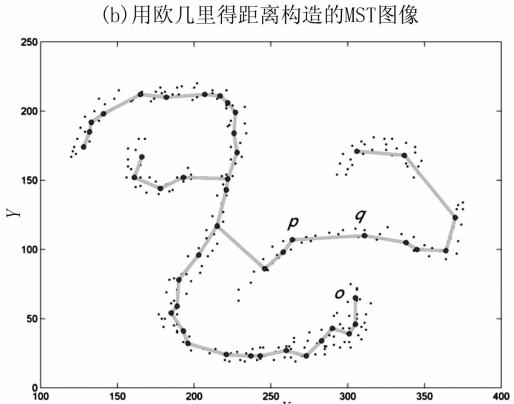
由于数据集中局部密度峰值分布不均匀,欧氏距离不适用于测量局部密度峰值之间的差异。所以使用基于邻域的共享距离利用局部密度峰值之间的邻域信息,缩短被稠密区域紧密相连的局部密度峰值之间的距离的方法更恰当地表示了局部密度峰值之间的差异。




(a) 局部密度峰值及其成员




(b)用欧几里得距离构造的MST图像




(c) 用基于共享邻点的距离构造的MST图像

**图3 各个方法距离的区别**

以图3所示的数据集为例,图3(a)为局部密度峰值及其邻域点,图3(b)为用欧几里得法构造的局部密度峰值的MST图像,图3(c)为基于共享邻点的距离构造的MST图像。局部密度峰值 $p$ 和 $q$ 在同一簇,$q$ 和 $o$ 在不同簇,但是 $p$ 和 $q$ 之间的欧氏距离大于 $q$ 和 $o$ 之间的欧氏距离,所以用欧氏距离构造的MST会出现错误。但是,基于共享邻点的距离构建的MST正确地保留了原始数据集的结构。

### 1.3 算法流程

首先,使用局部密度峰值和基于共享邻点的距离来构建MST;然后,重复切割最长的边(边的长度是采用基于共享邻点距离的),并保证切割该边导致的两个簇的大小都大于松散估计的最小点数,直到找到给定数量的簇为止。对局部密度峰值进行聚类后,将每个剩余点分配到与对应的局部密度峰值所属的相同类簇中。LDP-MST算法主要包括以下步骤:1)搜索局部密度峰值;2)计算局部密度峰值之间基于共享邻点的距离;3)采用基于MST的聚类算法对局部密度峰值进行聚类。

## 2 LDP-MST在电力大数据中的应用

如今,智能电网建设速度不断加快,与之而来的是大量的数据,这些数据主要来源于电网的发、输、配、用四大环节。聚类分析可以从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中的人们事先不知道但又具有潜在价值的信息。其中,最具有显著效果的聚类分析就是对用户用电行为的聚类和异常检测。用户用电行为聚类基于用户用电行为模式对相似性用户进行划分类别,而异常检测主要是指检测电力偷窃、电能表错误、计费错误等非技术损失造成的异常用电情况。

LDP-MST算法在电力大数据领域具有良好的应用前景,尤其体现在异常值检测中。异常值检测的目标是将不属于任何簇的样本点与正常点进行区别,从数据的角度来说,就是找出样本点数量较小的簇。故使用LDP-MST算法将样本点较少的簇提取出来,就可以得到异常样本。为验证算法的实用性,以某网站3个月的访问量和网络流量为基础,使用LDD-MST算法检测了其中的异常值。

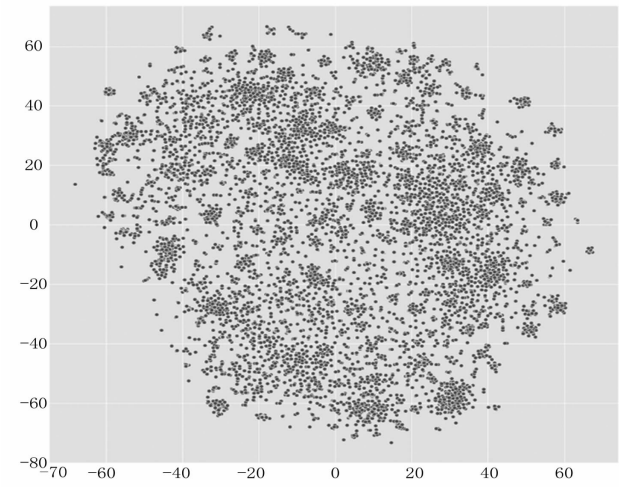



**图 4　LDP – MST 算法聚类结果**

在进行聚类之前,先对数据进行了预处理,即用缺失点外的其他值的均值代替该属性的缺失值。最终得到 LDP – MST 算法聚类结果如图 4 所示。由于只通过聚类法不容易用肉眼判别聚类结果,所以要对数据进行归一化处理。这里采取的归一化的方式为

$$\overline{P} = \frac{p_i - p_{\min}}{p_{\max} - p_{\min}} \tag{2}$$

图 5 为归一化处理后的数据。由图可以看出,在 3 月之初以及 4 月中后期有一些数据的网络流量与正常用户访问次数差距较大,明显偏离了正常数值。将这些异常值输出,并经聚类分析和异常值判定后,得到如表 1 所示的异常值分布。可发现所提算法将数据集中的异常值全部检测出来,说明 LDP – MST 算法对异常值检测具有比较良好的效果。

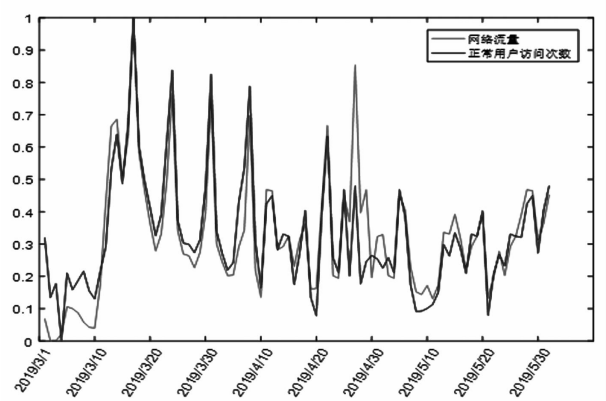



**图 5　归一化处理后的数据分布**

**表 1　异常值数据分布**

| 异常值标号 | 日期 |
|---|---|
| 1 | 3 月 1 日 |
| 3 | 3 月 3 日 |
| 58 | 4 月 26 日 |
| 59 | 4 月 27 日 |
| 60 | 4 月 28 日 |

## 3　结　语

上面提出了一种新的聚类算法 LDP – MST,其核心思想是选择局部密度峰值来构建 MST,避免了噪声点的干扰,减少了基于 MST 的聚类算法的运行时间。电力综合数据集的实验表明,该聚类算法能较好地识别数据集中的复杂模式,且比现有的聚类算法更有效。在进行电力大数据的异常检测时,算法在短时间内有效地检测出了异常结果。今后,将继续完善本算法的缺点以及将这一基于聚类算法的异常检测方法应用到电力系统的更多方面。

**参考文献**

[1]　王千, 王成, 冯振元,等. K – means 聚类算法研究综述[J]. 电子设计工程, 2012,20(7):21 – 24.

[2]　周涛, 陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用, 2012,48(12):100 – 111.

[3]　Zhang X, Wang W, Nrvg K, et al. K – AP: Generating Specified K Clusters by Efficient Affinity Propagation [C]// ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 2010.

[4]　陈春涛. 快速搜索与密度峰值发现算法的研究与应用[D]. 上海:华东师范大学,2019.

[5]　Yeuang Chen, Shengyu Tang, Lida Zhou, et al. Decentralized Clustering by Finding Loose and Distributed Density Cores[J]. Information Sciences, 2016:433 – 434.

[6]　荣秋生, 颜君彪, 郭国强. 基于 DBSCAN 聚类算法的研究与实现[J]. 计算机应用, 2004,24(4):47 – 48.

[7]　陈恒飞. Chameleon 聚类算法研究[D]. 西安:西安理工大学,2017.

[8]　崔凤山,刘博,贾凯,等. 用电采集系统电力大数据应用探究——电力数据看居民房屋空置率[J]. 农电管理,2021(2):28 – 29.

[9]　曹成,陶继群,郑湃. 基于 Kudu 的电力辅助设备实时监控业务解决方案[J]. 科技创新与应用,2021(8):130 – 134.

[10]　刘博,钱勇,沈阿美. 大数据技术视域下电力配电网智能运维管控系统研究[J]. 工业加热,2021,50(1):46 – 48.

[11]　葛一统,向锋铭,余桂华,等. 大数据背景下的电力营销信息化建设研究[J]. 华电技术,2021,43(1):76 – 82.

(下转第 49 页)

桌面放置的小型化、精简的安全终端设备经网络接入终端管控系统,获取相应的虚拟机资源和交互画面并对其进行操作,使用体验与使用物理工作站无差别。

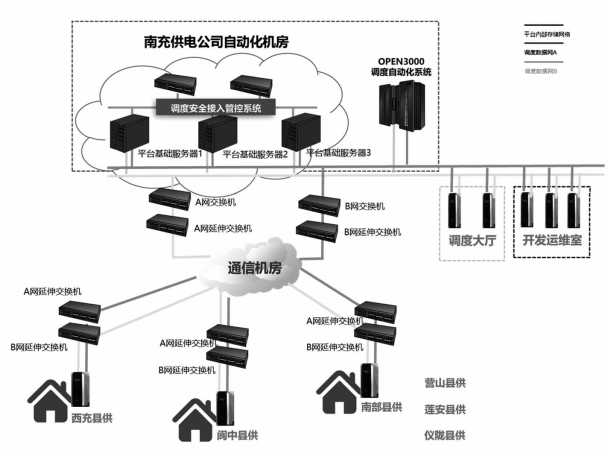



**图 5　示范应用拓扑**

特别是在西充(距地调 40 km)和阆中(距地调 80 km)两个县供电公司远程部署的场景下,终端管控系统优化了自身传输协议,对于网络带宽的需求降至 10 M 以下,满足窄带宽工作要求,相较传统物理工作站 30 M 以上的带宽需求,明显降低了对于网络带宽的消耗。同时对于所穿透的纵向加密设备,也减轻了其加密负载。数据实时刷新、场站图形和模型加载、历史库数据查询速度较传统工作站提升 60% 以上。

## 8　结　语

所提出的调控终端安全管控系统极大提高了日常对工作站的管理和维护效率,结合终端管控系统的安全审计、认证管理、授权管理等运维安全管理功能,可对调控终端进行统一管理和集中调配,实现了对调度员和维护人员使用调控终端的安全审计、风险防范,加强了调度工作站使用的机密性和规范性,满足等保要求。

项目试运行以后运行良好、稳定、可靠,达到了项目预期效果,可以推广应用。

### 参考文献

[1] 殷自力, 钱静, 陈宇星,等. 基于 D5000 平台的调配一体技术方案[J]. 电力系统自动化, 2016, 40(18): 162 – 167.

[2] 李敏子. 电力调度自动化中的一体化技术[J]. 电子技术与软件工程, 2018(10):123 – 123.

[3] 邱丽娜. 电力调度自动化系统中一体化技术的应用[J]. 现代信息科技, 2017,1(6):28 – 29.

[4] 余时强,张为华. GPU 虚拟化相关技术研究综述[J]. 计算机系统应用,2017,26(12):25 – 31.

[5] 仝伯兵,杨昕吉,谢振平,等. GPU 虚拟化技术及应用研究[J]. 软件导刊,2015,14(6):153 – 156.

[6] 陈钢,吴百锋. 面向 OpenCL 模型的 GPU 性能优化[J]. 计算机辅助设计与图形学学报,2011(4):571 – 581.

[7] 卢风顺,宋君强,银福康,等. CPU/GPU 协同并行计算研究综述[J]. 计算机科学,2011,38(3):5 – 9.

**作者简介:**

王先强(1979),男,高级工程师,从事电网及二次系统管理工作;

张　睿(1982),男,硕士,高级工程师,从事电网及二次系统管理工作;

张　华(1985),男,硕士,高级工程师,从事调度自动化管理工作。

**(收稿日期:2020 – 01 – 17)**

(上接第 19 页)

[12] 卜云,高传海,李文芳,等. 大数据架构下电力系统风险评估[J]. 电网与清洁能源,2021,37(1):77 – 83.

[13] 田园,原野. 基于改进 K – means 算法的电力大数据系统研究[J]. 电子设计工程,2021,29(2):76 – 80.

**作者简介:**

靳文星(1996),男,在读硕士,研究方向为电力信息化及自动化;

王电钢(1973),男,教授级高级工程师,研究方向为电力信息化;

张哲敏(1997),男,在读硕士,研究方向为电力信息化及自动化。

**(收稿日期:2021 – 03 – 01)**