DOI:10.12158/j.2096-3203.2021.02.015

基于并行化 *K*-means 的综合能源服务客户识别

沈子垚，袁晓玲
（河海大学能源与电气学院，江苏 南京 211100）

摘 要：随着电力体制改革的不断深入以及大数据技术的发展，传统的供电公司和综合能源服务企业急需改善现有的粗放型营销模式，实现不同用户需求的快速响应。针对综合能源服务潜在客户的精准识别问题，文中通过对综合能源服务潜在客户的标签进行分析，基于 Spark 内存计算平台提出了一种改进的并行化 *K*-means 聚类算法。首先，对聚类过程中初始聚类中心的选取和样本影响因素的权值进行改进；其次，基于优化后的权值对客户数据集进行聚类分析，对综合能源服务潜在客户进行识别；最后，采集综合能源服务企业的近期交易数据，在多节点的物理机上进行实验与分析。结果表明改进后的聚类算法更准确。在执行效率上，并发度高的算法执行效率优于单线程的算法具有较好的并行能力。

关键词：潜在客户识别；大数据；Spark 框架；*K*-means 聚类；并行计算
中图分类号：TM73 **文献标志码**：A **文章编号**：2096-3203(2021)02-0107-07

0 引言

随着综合能源服务的不断推广和互联网技术的高速发展，客户的档案信息与交易数据激增。传统的供电公司与综合能源服务企业积累了海量的营销数据。通过对营销数据的挖掘分析，能够获取客户的行为信息和状态数据，识别现有客户的特征与交易模式，预测综合能源服务需求，提高企业决策的前瞻性[1—3]。综合能源服务企业如想在激烈的竞争中保持优势，需要做好各类客户的识别与服务拓展工作。如企业能采取有效措施精准识别潜在客户，就能以较小的成本发展潜在客户，针对客户需求制定综合能源服务策略[4—8]，提高投入产出比。文献[9]构建了数据仓库以整合数据资源并提取属性特征，通过信息匹配实现对潜在客户的识别。

传统的聚类算法在处理海量数据时，存在计算复杂度高和计算能力不足等问题。文献[10]优化了 *K*-means 聚类算法初始聚类中心的选取，并选用 MapReduce 并行编程方法，提高了传统聚类方法的计算效率。文献[11]将用户的用电行为数据按行保存于 Hadoop 分布式文件系统(Hadoop distributed file system，HDFS)，将用电行为数据集划分为不同切片产生子数据集，利用 MapReduce 计算模块对各切片数据进行读取。Hadoop 支持 TB 级别的数据和流式数据访问，但实时性较差，不适合大量小文件存储[12—13]。

收稿日期：2020-09-05；修回日期：2020-10-19
基金项目：国家电网有限公司科技项目“能源互联网环境下的多源互联配电网及多样化用电方式的需求策略系统研究”

在数据挖掘中，需处理大量数据，应构建简单有效的模型。数据挖掘是指通过对海量杂乱无章的数据进行挖掘，找到其中蕴含的规律和有价值的信息。数据挖掘的主要步骤包括：确定数据挖掘的需求，采集相关数据并预处理，采用合适的数据挖掘算法构建识别模型，对识别结果进行分析评估。

为进一步减少数据聚类迭代过程中的冗余计算，提高聚类算法的效率和准确性，文中针对综合能源服务潜在客户识别问题，研究基于 Spark[14—17] 的 *K*-means 聚类算法，优化了初始点的选取和聚类时影响因素的权值选取。文中通过并行计算提高了数据的处理速度，依据实验结果分析了算法的准确率和并行计算性能。

1 潜在客户识别数据库设计

潜在客户识别数据库的设计包括分析潜在客户的各项标签[18]，利用数据仓库技术对数据资源进行整合。通过潜在客户数据库完成客户的精准识别工作，可构建完成综合能源服务潜在客户识别模型。

1.1 客户识别数据仓库

客户识别数据仓库的建立基于现有的营销业务系统及外部数据获取渠道[19]，需要提取客户信息形成客户数据集。数据预处理操作包括缺失数据的补全，重复数据的删除，数据泛化等工作。基于预处理之后的数据集，可以建立综合能源服务潜在客户数据仓库。数据仓库的构建基于收入贡献、成本占用、成长性、信誉度、忠诚度等标签，潜在客户特征分析指标如表 1 所示。

表 1 潜在客户特征分析指标表

Table 1 Characteristic analysis index for potential customers

指标名称	字段名称	字段类型
收入贡献	income	decimal(10,2)
成本占用	cost	decimal(10,2)
成长性	growth	bigint(20)
信誉度	credibility	bigint(20)
忠诚度	loyalty	bigint(20)

对于大量数据,如无法补全缺失数据的记录和属性,应删除该部分信息。对于缺失少量非关键数据的记录,可根据部分未缺失数据的均值、众数进行填充。综合能源服务潜在客户识别的数据来源众多,可能存在大量冗余的数据。因此,可通过优化数据库范式结构以删除重复数据,对粒度较小的标签进行泛化处理。数据的泛化是指用高层次的概念取代低层次的概念,能够进一步明确数据属性的取值差异,且减少数据的计算量。

1.2 客户价值评价

综合能源服务潜在客户的价值可以用当前价值与潜在价值进行综合评估,具体的客户价值评估体系如图 1 所示。

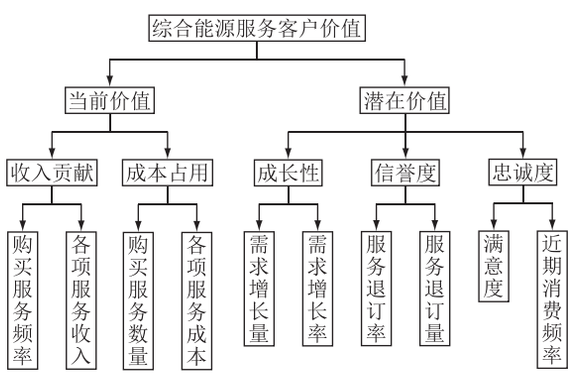

图 1 客户价值评估指标体系

Fig.1 Evaluation index system of customers

基于当前价值与潜在价值,可综合评估综合能源服务潜在客户。综合能源服务潜在客户的当前价值可以用收入贡献和成本占用两个指标衡量。收入贡献可以用购买服务频率和各项服务收入两个细分指标衡量。成本占用可以进一步细分为购买服务数量和各项服务成本。综合能源潜在客户的潜在价值可以用成长性、信誉度、忠诚度等指标衡量。客户的成长性包括需求增长量和需求增长率;信誉包括服务的退订率和服务的退订量;忠诚度包括满意度和近期消费频率。

1.3 指标映射数据库

由于评估潜在客户指标的量纲不同,需要将不同指标值进行无量纲化处理。文中通过构造综合能源服务潜在客户的指标映射数据库,定量计算客户之间的差异度。

若综合能源服务客户的当前价值为主导因素,则映射区间可以超出标准区间限制;而对于非主导因素的潜在价值,映射区间可以限制在一定范围内。

数据库映射关系 f 表示为:

$$f:(x_1,x_2,\cdots,x_n)\rightarrow(y_1,y_2,\cdots,y_n) \quad (1)$$

式中:$(x_1,x_2,\cdots,x_n)$ 为评价客户价值的各个指标;$(y_1,y_2,\cdots,y_n)$ 为评价指标通过数据库映射,投影到映射区间的值。部分映射区间的取值如表 2 所示。

表 2 部分映射数据库取值

Table 2 Database value of partial mapping

评价指标	映射前取值	映射后取值
购买频率	0~0.2	4
购买频率	0.2~0.4	8
⋮	⋮	⋮
购买频率	0.8~1.0	16
退订频率	0~0.2	8
退订频率	0.2~0.4	6
⋮	⋮	⋮
退订频率	0.8~1.0	0

在综合能源服务发展的过程中,如综合能源服务企业对潜在客户的评估标准改变,那么指标映射数据库的映射取值可能会发生变化。

2 *K*-means 并行聚类识别潜在客户

聚类分析通过反复分区,将数据进行归类,使得同类的对象之间能够彼此联系。聚类算法能够在没有客户类别标识的前提下对客户进行分类,最大化不同类别客户的差异。聚类对象根据最大化同一簇中的相似性,最小化不同簇之间相似性原则进行划分[20]。

综合能源服务企业可以修改客户价值的映射数据库,通过聚类效果探索出适合自身的映射关系。在与综合能源客户交易的过程中,用户的价值评价体系和企业的偏好可能会发生改变,综合能源服务企业可以根据偏好改变聚类过程中映射区间的取值。

文中采用 Spark 平台对大数据进行并行化处理。Spark 是为大数据处理专门设计的快速通用的计算引擎,该框架多任务之间的数据基于内存进行通信,消除了冗余的 Hadoop 分布式文件系统读写,并针对 Java 虚拟机(Java virtual machine,JVM)进行了优化。因此,Spark 更加适用于实时处理等数据

挖掘工作,在大规模的数据计算上优于传统的 Map-Reduce 编程模式[21]。基于 Spark 的潜在客户识别的并行聚类模型如图 2 所示。

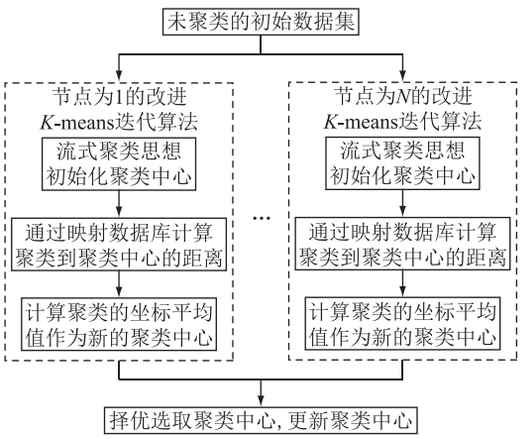

图 2 并行聚类模型

Fig.2 Parallel clustering model

对于传统的 *K*-means 算法,首先利用流式聚类思想优化选取初始聚类中心点,再通过映射数据库衡量不同标签对聚类算法的影响,计算聚类中心。*K*-means 并行聚类算法运行于 Spark 平台,并行聚类运算完成后更新聚类中心。

2.1 初始化聚类中心的优化

文中提出一种基于流式动态聚类思想的单遍权重 *K* 均值聚类方法(single-pass-weighted *K*-means, SWP *K*-means)。首先,在原数据集上通过随机抽样构造出 s 个大小为 n 的数据子集 $\boldsymbol{X}$。算法最初将每个数据子集 $\boldsymbol{X}$ 中的样本权重设置为 1。然后,计算第一组权值为 1 的样本 *K*-means 的聚类中心 $\boldsymbol{y}$,得到最小的聚类误差平方和 $D(\boldsymbol{y})$:

$$D(\boldsymbol{y})=\min_{1\leqslant k\leqslant n-1}\sum_{i=1}^{n}\omega_i\left(y_i-y_{i,k}\right)^2 \tag{2}$$

式中:w_i为样本第 i 个指标的权值;y_i为该样本第 i 个指标的数据库映射值;$y_{i,k}$为第 k 个聚类中心第 i 个指标的数据库映射值。

迭代剩余的 $\boldsymbol{X}$ 的数据子集,每一次迭代运用 *K*-means 聚类算法在一组更大的数据集上进行聚类划分,数据集由上一次迭代的聚类中心 y_{t-1}和本次的样本子集 $\boldsymbol{X}_{t-1}$组成。第 t 次迭代有 $t+n-1$ 个实体进行聚类,重复迭代 s 次 *K*-means 算法,直至选择出共 k 个聚类中心。该方法基于上一次初始化的聚类中心加速收敛,代替了传统算法的阈值收敛方法,大大降低了聚类算法的迭代次数,在进行海量数据的聚类分析时更具有优势。

2.2 客户样本隶属的中心点

定义基于 Spark 平台的聚类过程中的距离为欧几里德距离,对于两点 $\boldsymbol{y}'=[y'_1\ y'_2\ \cdots\ y'_n]$ 和 $\boldsymbol{y}''=[y''_1\ y''_2\ \cdots\ y''_n]$ 之间的欧几里德距离计算公式为:

$$d(\boldsymbol{y}',\boldsymbol{y}'')=\sqrt{\sum_{i=1}^{n}\left(y'_i-y''_i\right)^2}=\sqrt{\sum_{i=1}^{n}\left(y'_i\right)^2+\sum_{i=1}^{n}\left(y''_i\right)^2-2\sum_{i=1}^{n}\left(y'_iy''_i\right)} \tag{3}$$

对于公式(3),$\sum_{i=1}^{n}\left(y'_i\right)^2+\sum_{i=1}^{n}\left(y''_i\right)^2$ 部分可以提前计算,$\sum_{i=1}^{n}\left(y'_iy''_i\right)$ 需要实时计算。Spark 平台实现了一种快速距离算法,该方法假设两点 $\boldsymbol{y}'=[y'_1\ y'_2\ \cdots\ y'_n]$ 和 $\boldsymbol{y}''=[y''_1\ y''_2\ \cdots\ y''_n]$ 之间的快速距离 $d_{\text{quick}}(\boldsymbol{y}',\boldsymbol{y}'')$ 为:

$$d_{\text{quick}}(\boldsymbol{y}',\boldsymbol{y}'')=\left|\sqrt{\sum_{i=1}^{n}\left(y'_i\right)^2}-\sqrt{\sum_{i=1}^{n}\left(y''_i\right)^2}\right| \tag{4}$$

快速距离算法的优势明显,可以提前计算样本向量的 2 范数,极大地降低计算量。易知 $d_{\text{quick}}(\boldsymbol{y}',\boldsymbol{y}'')\leqslant d(\boldsymbol{y}',\boldsymbol{y}'')$,对于同一个样本,当聚类中心 p_i的快速距离大于聚类中心 p_j的欧几里德距离时,聚类中心 p_i的欧几里德距离必大于聚类中心 p_j的欧几里德距离。此时,该样本所在簇的聚类中心应为 p_j。倘若聚类中心 p_i的快速距离小于聚类中心 p_j的欧几里德距离,快速距离算法失效,需要重新计算样本与聚类中心 p_i的欧几里德距离。

2.3 聚类效果评估

聚类效果评估采用集合内误差平方和(within set sum of squared error, WSSSE)W_{SSSE},W_{SSSE}为所有数据点到距离该点最近的聚类中心的平方和:

$$W_{\text{SSSE}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{i,j}-y_{\text{close},j}\right)^2 \tag{5}$$

式中:m 为样本总个数;n 为指标投影向量的维数;$y_{i,j}$为 i 个样本的第 j 个指标投影到映射区间的值;$y_{\text{close},j}$为第 i 个样本最近的聚类中心的第 j 个指标投影到映射区间的值。

易知随着聚类个数 K 的增大,W_{SSSE}减少。当聚类个数 $K=m$ 时,$W_{\text{SSSE}}=0$。一般来说,最优的 K 取值是 K-W_{SSSE}曲线的拐点位置。在拐点处,K 值的增加能最大程度地优化聚类效果。

3 案例分析

文中选取 2017—2019 年常州市供电局综合能源服务相关的部分负荷数据以组成客户识别数据仓库。此外,客户数据集加入实地客户集中调研及获取外部数据渠道提取的客户信息。将数据集的收入贡献、成本占用、成长性、信誉度、忠诚度等标

签进行泛化处理，转化为客户的当前价值和潜在价值作为输入。对潜在客户案例及不同的数据挖掘算法进行对比分析，得出最优的挖掘算法。最终输出综合能源潜在客户及客户类型，针对性地对各类客户推广综合能源服务。

3.1 模型构建

客户的当前价值能帮助综合能源服务企业评估客户的购买力，且基于客户的潜在价值可衡量客户在后续交易中带来的利润。文中结合综合能源客户的历史数据，采用基于数据挖掘中的 *K*-means 聚类方法进行定量分析，利用矩阵分类法建立综合能源潜在客户的二维细分模型，如图 3 所示。

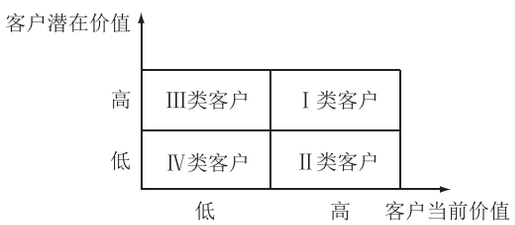

图 3 基于当前价值与潜在价值的客户分类

Fig.3 Customer classification based on current value and potential value

Ⅰ类用户的当前价值较高，且该类客户较为稳定，综合能源服务企业与该类用户合作可以获取较大的利润。Ⅱ类用户的当前价值同样较高，但综合能源服务企业无法满足客户的综合能源服务需求，导致后续交易过程中用户的潜在价值较低，需要投入一定资源激活，避免该类用户转向竞争对手。Ⅲ类用户的当前价值较低，但具有较大的发展潜力，同样属于综合能源服务企业的发展对象。Ⅳ类客户的当前价值和潜在价值都较低，该类客户购买力有限，对于综合能源服务需求较少，属于综合能源服务中的“劣质客户”，不属于综合能源服务的交易对象。

3.2 结果分析

3.2.1 聚类效果分析

文中通过综合评价指标 F 评估算法的性能：

$$\begin{cases} F = \dfrac{2PR}{P+R} \\ P = \dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \\ R = \dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \end{cases} \tag{6}$$

式中：T_{P}为正确识别潜在用户的数量；P 为被分为潜在用户的类别中实际为潜在用户的比例；R 为潜在用户被正确识别的比例，用于衡量覆盖面；F_{P}为将非潜在用户识别为潜在用户的数量；F_{N}为将潜在用户识别为非潜在用户的数量。

在不同的映射权值下，聚类结果不同。客户当前价值与潜在价值的映射权值为(0.67，1.33)时，聚类结果如图 4 所示。

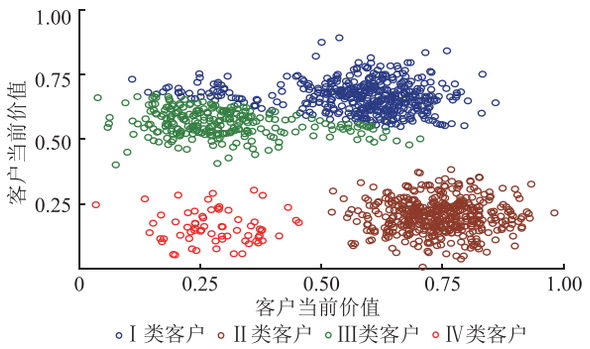

图 4 *K*-means 聚类结果

Fig.4 *K*-means clustering results

Ⅰ类客户 449 个，正确识别 416 个；Ⅱ类客户 448 个，正确识别 448 个；Ⅲ类客户 262 个，正确识别 229 个；Ⅳ类客户 67 个，正确识别 67 个。综合评价指标 F 为 92.6%。

映射权值表示综合能源服务企业对该类价值的重视程度。以客户潜在价值为主导时，客户的潜在价值映射区间大于当前价值的映射区间，Ⅰ类用户与Ⅲ类用户的潜在价值都较高，主要差异为客户当前价值。当前价值的映射区间较小时区分度不明显，此时聚类模型无法很好地区分Ⅰ类用户与Ⅲ类用户。

客户当前价值与潜在价值的映射权值为(1，1)时，基于权值的 *K*-means 算法聚类结果如图 5 所示。

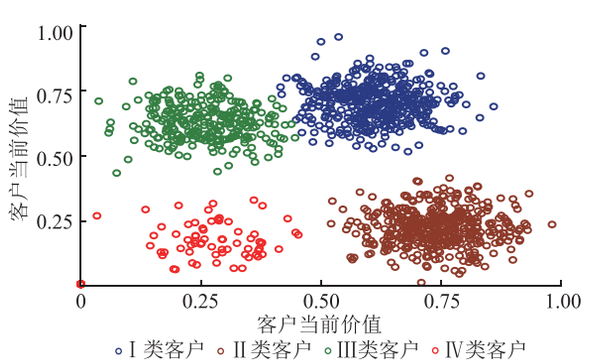

图 5 基于权值的 *K*-means 聚类结果

Fig.5 Weighted *K*-means clustering results

Ⅰ类客户 431 个，正确识别 423 个；Ⅱ类客户 448 个，正确识别 448 个；Ⅲ类客户 280 个，正确识别 272 个；Ⅳ类客户 67 个，正确识别 67 个。综合评价指标 F 为 98.6%。

设映射权值为(1，1)，SPW *K*-means 算法聚类

结果如图 6 所示。

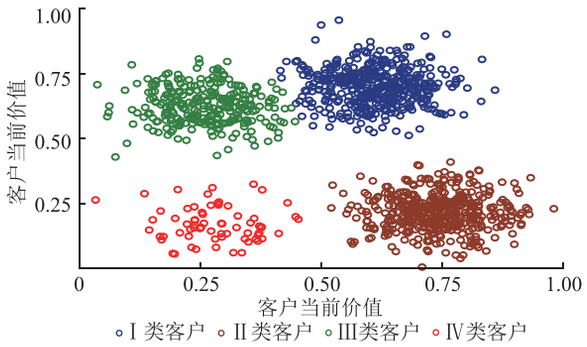

图 6　SPW *K*-means 聚类结果

Fig.6　SPW *K*-means clustering results

Ⅰ类客户 431 个,正确识别 426 个;Ⅱ类客户 448 个,正确识别 448 个;Ⅲ类客户 280 个,正确识别 275 个;Ⅳ类客户 67 个,正确识别 67 个。综合评价指标 F 为 98.7%。

客户的当前价值与潜在价值的映射区间相同时,客户的当前价值与潜在价值重要程度相近,聚类结果更符合矩阵分类法建立的二维细分模型。优化特征向量的权值将改善基于权值的 *K*-means 算法与 SPW *K*-means 算法的聚类性能。在权值相同且测试数据较少的情况下,初始中心点的选取对最终的聚类结果影响不大,这表明基于权值的 *K*-means 算法具有良好的稳定性。

为验证进一步不同算法的聚类效果,对原有的数据集进行扩容,不同扩容倍率下聚类算法的误差平方和如图 7 所示。

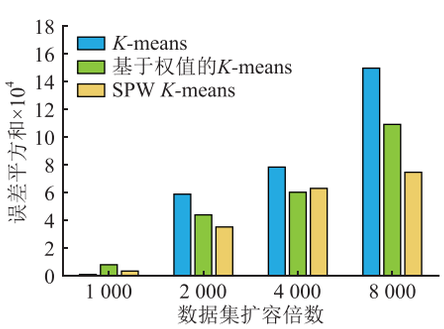

图 7　不同扩容倍数下的误差平方和

Fig.7　MSE with different expansion times

由图 7 可知,随着数据集扩容倍数的增长,SPW *K*-means 的聚类误差增加趋势小于 *K*-means 算法和基于权值的 *K*-means 算法。在数据集扩大的情况下,初始聚类中心点选取的优劣程度将决定最终的聚类性能。

3.2.2　性能分析

为了检验算法的执行效率,对 20 000 个测试数据进行算法的时间复杂度分析,并行聚类部分执行时间,如图 8 所示。

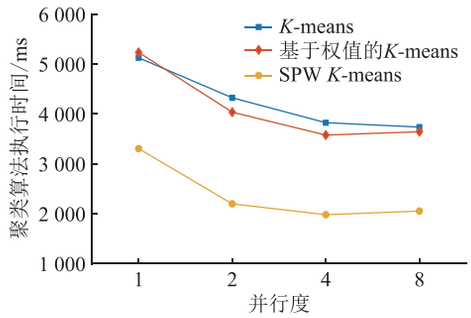

图 8　不同并行度下的算法执行时间

Fig.8　Algorithm execution time with different parallelism

随着算法并行度的增加,聚类模型执行的时间优化幅度减少。这是因为增加算法的并行度能充分利用空闲线程,提高运行效率。同时,并行度为 4 时的算法运行时间小于并行度为 8 的时间,这是由于随着并行度的提高,运行节点之间的数据传输会消耗资源。

采用加速比 S_{peedup} 和扩展比 E 测试并行 *K*-means 算法的并行化性能,加速比 S_{peedup} 和扩展比 E 的公式为:

$$S_{peedup}=\frac{T_s}{T_p} \tag{7}$$

$$E=\frac{S_{peedup}}{p} \tag{8}$$

式中:T_s 为单节点进行运算消耗的时间;T_p 为 p 个节点进行运算所消耗的时间。

通过聚类算法的串行执行时间与并行执行时间的比率来判断并行效果,不同聚类算法的加速比如图 9 所示。

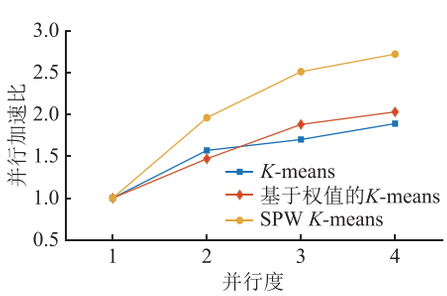

图 9　不同并行度下聚类算法的加速比

Fig.9　Speedup of clustering algorithm with different parallelism

在不同并行度下,文中提出的 SPW *K*-means 算法加速比优于其他聚类算法。添加节点后,处理速度变快,但加速比未能符合线性增长。

文中通过对数据集的规模进行扩展,比较 SPW *K*-means 算法在不同节点数量下的扩展比,如表 3 所示。

表 3 SPW *K*-means 聚类算法的扩展比

Table 3 The expansion ratio of SPW *K*-means clustering algorithm

数据集扩展倍数	扩展比		
	1	2	4
1 000	1	0.78	0.43
2 000	1	0.73	0.47
3 000	1	0.88	0.53

Spark 框架更加适合大量数据的处理，当数据量足够大时，集群并行化能够有效提高聚类算法的速度，数据集越大，并行效果越明显。这是因为数据量增加，节点更容易发挥它的计算能力，节点利用率提高。而随着节点的增加，扩展比未能线性增加。这是因为在集群上运行时，平台启动，任务调动与数据通信等因素会影响聚类算法的运行。

4 结语

文中针对综合能源服务潜在客户的精准识别问题，整合了近期的客户信息，建立客户识别数据库；基于流式动态聚类的思想优化初始聚类中心的选取，分析了客户的不同标签并用标签的映射权值来衡量对价值的影响；基于分布式内存计算框架 Spark 进行并行化聚类，达到了综合能源服务潜在客户精准识别的目的。

文中将改进后的 *K*-means 聚类算法在集群上并行运行，比较聚类的准确率。通过比较不同并行度下的算法执行时间、加速比、并行度，验证并行化计算的高效性。结果表明基于 Spark 平台的改进 *K*-means 算法不仅能够有效利用闲置 CPU 内核的运算能力缩短训练建模时间，且能根据各类标签对客户价值的优化调节映射权值提高分类的精度。针对海量数据集，调节聚类算法的并行度可以减少算法执行时间，这说明该算法具有良好的扩展性。但基于 Spark 框架的 *K*-means 聚类算法本身具有局限性，未来如何在不同场景下对聚类算法进行并行化设计有待进一步实践。

参考文献：

[1] 容志. 大数据背景下公共服务需求精准识别机制创新[J]. 上海行政学院学报，2019，20(4)：44-53.
RONG Zhi. Innovation of accurate recognition mechanism of public service demand in the background of big data[J]. The Journal of Shanghai Administration Institute，2019，20(4)：44-53.

[2] 薛少华，李宁，周星明，等. 考虑综合需求响应的综合能源系统优化运行[J]. 电力需求侧管理，2020，22(5)：7-12.
XUE Shaohua，LI Ning，ZHOU Xingming，et al. Optimal operation of integrated energy system considering integrated demand response[J]. Power Demand Side Management，2020，22(5)：7-12.

[3] 荀挺，雷胜华，丁晓辰，等. 区域综合能源系统的多目标最优潮流算法研究[J]. 智慧电力，2019，47(9)：19-28.
XUN Ting，LEI Shenghua，DING Xiaochen，et al. Multi-objective optimal power flow algorithms for integrated community energy systems[J]. Smart Power，2019，47(9)：19-28.

[4] 尹鹏，张剑，董兵，等. 基于大数据的电力优质客户识别及市场营销服务策略分析[C]//中国电力科学研究院. 2018 智能电网新技术发展与应用研讨会论文集. 北京：《计算机工程与应用》编辑部，2018：13-15.
YIN Peng，ZHANG Jian，DONG Bing，et al. Power quality customer identification and marketing service strategy analysis based on big data[C]//China Electric Power Research Institute. Proceedings of 2018 smart grid new technology development and application seminar. Beijing：Editorial Department of computer engineering and application，2018：13-15.

[5] 李夫宝，李明轩. 园区综合能源服务各利益方诉求和耦合关系研究[J]. 电力需求侧管理，2020，22(4)：62-65.
LI Fubao，LI Mingxuan. Research on the demands and coupling relations among the stakeholders of park level integrated energy services[J]. Power Demand Side Management，2020，22(4)：62-65.

[6] 贾楚蕴，李华强，高红均. 基于合同能源管理的园区能耗优化及多主体利益分配研究[J]. 智慧电力，2020，48(10)：30-36，98.
JIA Chuyun，LI Huaqiang，GAO Hongjun. Research on energy optimization and benefit sharing among owners in industrial park based on energy management contract[J]. Smart Power，2020，48(10)：30-36，98.

[7] 崔杨，闫石，王铮，等. 多主体利益制衡的综合能源系统日前-实时出清方法[J]. 电力系统自动化，2020，44(24)：68-76.
CUI Yang，YAN Shi，WANG Zheng，et al. Day-ahead and real-time clearing method of integrated energy system considering interest balance between multiple entities[J]. Automation of Electric Power Systems，2020，44(24)：68-76.

[8] 杨挺，赵黎媛，刘亚闯，等. 基于深度强化学习的综合能源系统动态经济调度[J]. 电力系统自动化，2021，45(5)：39-47.
YANG Ting，ZHAO Liyuan，LIU Yachuang，et al. Dynamic economic dispatch for integrated energy system based on deep reinforcement learning[J]. Automation of Electric Power Systems，2021，45(5)：39-47.

[9] 陈宁，孙晓阳，龚德鹏. 基于商业智能的铁路货运客户精准识别方案[J]. 综合运输，2018，40(7)：103-109.
CHEN Ning，SUN Xiaoyang，GONG Depeng. Study on accurate identification of railway freight customers based on business intelligence[J]. China Transportation Review，2018，40(7)：103-109.

[10] 贾金伟，吴旭鹏，李启本，等. 基于并行计算的大数据挖掘在电网中的应用[J]. 电力与能源，2017，38(6)：724-729.
JIA Jinwei，WU Xupeng，LI Qiben，et al. Application of big

data mining in power grid based on parallel computing[J]. Power & Energy,2017,38(6):724-729.

[11] 王永才. 基于 Hadoop 平台的用电行为数据特征挖掘方法[J]. 自动化与仪器仪表,2020(11):227-230.
WANG Yongcai. Data mining method of power consumption behavior based on Hadoop platform[J]. Automation & Instrumentation,2020(11):227-230.

[12] 赵卫中,马慧芳,傅燕翔,等. 基于云计算平台 Hadoop 的并行 K-means 聚类算法设计研究[J]. 计算机科学,2011,38(10):166-168,176.
ZHAO Weizhong, MA Huifang, FU Yanxiang, et al. Research on parallel K-means algorithm design based on Hadoop platform[J]. Computer Science,2011,38(10):166-168,176.

[13] 江小平,李成华,向文,等. K-means 聚类算法的 Map Reduce 并行化实现[J]. 华中科技大学学报(自然科学版),2011,39(S1):120-124.
JIANG Xiaoping, LI Chenghua, XIANG Wen, et al. Parallel implementing K-means clustering algorithm using Map Reduce programming mode[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2011, 39(S1):120-124.

[14] 赵玉明,舒红平,魏培阳,等. 基于 Spark 的聚类算法优化与实现[J]. 现代电子技术,2020,43(8):52-55,59.
ZHAO Yuming, SHU Hongping, WEI Peiyang, et al. Optimization and implementation of clustering algorithm based on Spark[J]. Modern Electronics Technique,2020,43(8):52-55,59.

[15] 余胜辉,李玲娟. 基于 Spark 的层次聚类算法的并行化研究[J]. 计算机技术与发展,2020,30(6):19-22.
YU Shenghui, LI Lingjuan. Research on parallelization of hierarchical clustering algorithm based on Spark[J]. Computer Technology and Development,2020,30(6):19-22.

[16] 胡俊,胡贤德,程家兴. 基于 Spark 的大数据混合计算模型[J]. 计算机系统应用,2015,24(4):214-218.
HU Jun, HU Xiande, CHENG Jiaxing. Big data hybrid computing mode based on Spark[J]. Computer Systems & Applications,2015,24(4):214-218.

[17] SHEN J F, WANG HH. Fusion effect of SVM in Spark architecture for speech data mining in cluster structure[J]. International Journal of Speech Technology,2020,23(3):481-488.

[18] 孔颖. 基于数据挖掘的分类算法在潜在客户识别中的应用[J]. 计算机时代,2008(9):31-32.
KONG Ying. Application of classification algorithm based on data mining in potential customer identification[J]. Computer Era,2008(9):31-32.

[19] CELUCH K, WALZ A M. The role of active identification in driving retail customer feedback[J]. Services Marketing Quarterly,2020,41(2):163-181.

[20] 傅世权. 大数据时代下数据挖掘技术在电力企业中的应用探讨[J]. 信息记录材料,2019,20(11):128-129.
FU Shiquan. Discussion on the application of data mining technology in power enterprises in the era of big data[J]. Information Recording Materials,2019,20(11):128-129.

[21] 陶婧. 基于 Spark 的分布式大数据并行化聚类方法研究[J]. 湖北第二师范学院学报,2019,36(8):49-53.
TAO Jing. Parallel clustering method of distributed big data based on Spark[J]. Journal of Hubei University of Education, 2019,36(8):49-53.

作者简介:


沈子垚

沈子垚(1995),男,硕士在读,研究方向为泛在电力物联网与综合能源服务(E-mail:ashassnow@126.com);

袁晓玲(1971),女,博士,副教授,研究方向为新能源并网及其控制、泛在电力物联网与综合能源服务。

Implementation of integrated energy service for customer identification based on parallel *K*-means clustering

SHEN Ziyao, YUAN Xiaoling

(College of Energy and Electrical Engineering, Hohai University, Nanjing 211100, China)

Abstract: With the deepening reform of electric power enterprise and the development of big data technology, traditional power supply companies and integrated energy service enterprises have to change the present extensive marketing mode for offering rapid response to consumers' requirement. In order to improve the accurate identification of potential customers in integrated energy services, this paper marks the tags of potential customers, and proposes an improved parallel *K*-means clustering algorithm based on spark memory computing platform. Firstly, the selection of initial cluster center and the evaluation of sample influencing factors are improved. Secondly, based on the optimized weight of factors, cluster analysis is carried out on the data setting to identify the potential customers of integrated energy services. Finally, the recent transaction data of integrated energy service enterprises are collected, and the experimental results are carried out on a multi-node physical machine. The results show that the accuracy of improved *K*-means clustering model is boosted. In terms of executive effectiveness, the algorithm with high concurrency has better parallel ability than that with single thread.

Keywords: customer identification; big data; spark framework; *K*-means clustering; parallel computation

(编辑 李栋)