计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):215−219 [doi: 10.15888/j.cnki.csa.006564]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BIRCH 聚类的物流配送设施选址算法①

李捷承[1,2], 陶耀东[2], 孙　咏[2], 高　岑[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘　要: 物流配送设施的选址对于物流成本、在途时间影响巨大. 其特点包括: 配送设施选址和配送路线交互影响、多层级选址、配送设施存件数量均衡性等. 本文通过分析物流配送设施选址的特点设计了一个基于 BIRCH 聚类的物流配送设施选址算法, 融合了 BIRCH 聚类算法和基于 Dijkstra 距离的重心法, 为物流配送设施选址提供了更好的方案, 大幅节约长期运营成本.

关键词: 选址问题; 容量限制; 多层级聚类; BIRCH 聚类

引用格式: 李捷承,陶耀东,孙咏,高岑.基于 BIRCH 聚类的物流配送设施选址算法.计算机系统应用,2018,27(9):215–219. http://www.c-s-a.org.cn/1003-3254/6564.html

Location Algorithm of Logistics Distribution Facilities Based on BIRCH Clustering

LI Jie-Cheng[1,2], TAO Yao-Dong[2], SUN Yong[2], GAO Cen[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: The location of logistics distribution facilities has a great impact on logistics costs and deliver time. Its features include: the interaction between the location of delivery facilities and the delivery route planning, the multi-level location, the balance of shipment quantity, etc. Through the analysis of the characteristics of logistics distribution facilities location, a BIRCH-based logistics distribution facility location algorithm, a combination of BIRCH clustering algorithm and Dijkstra-based gravity center method, is designed to provide a better location and save long-term operating costs.

Key words: location problem; capacity limitation; multi-level clustering; BIRCH clustering

物流业近年来发展迅速, 对促进第三产业、带动第一第二产业发展作用明显. 物流配送设施是快件集散的关键节点, 在整个物流系统中起着承上启下的作用, 配送设施的选址对物流成本影响巨大, 一个好的配送设施选址决策可以使得快件在汇集、中转、分发、配送的过程达到最少的费用和时间[1]. 因而对物流配送设施选址进行研究具有较大的经济意义和现实意义. 同时随着电子商务的兴起, 面向普通消费者的快递物流业成为了新的爆发点, 本文即是对面向普通消费者的物流业务的配送设施选址问题进行研究.

针对一般设施的选址研究已有许多, 主要有以下 4 种类型: 1) 基于专家咨询的层次分析法 (AHP) 及模糊综合评价法, 该类方法主观判断占主导地位, 决策会受到专家的经验、领域知识等因素的限制; 2) 将选址问题当作整数或混合整数规划进行求解, 但设施选址是 NP-Hard 问题, 一旦数据规模较大, 就会运算时间过长, 难以求解出最优解; 3) 重心法[2], 将运输成本作为唯一的选址决策因素, 在单设施选址问题中实用性强, 缺点是考虑因素单一, 重心点可能位于已有建筑物、河流等无法建造配送中心的地方; 4) 聚类方法, 应用基于划分的聚类算法如 K-means 算法等对客户进行聚类, 每个类别的中心即所求选址方案, 但这些聚类算法

① 收稿时间: 2018-02-12; 修改时间: 2018-03-07; 采用时间: 2018-03-21; csa 在线出版时间: 2018-08-16

也有各自的缺点, K-means 算法必须指定划分数目, 容易陷入局部最小值, 且对孤立点敏感.

虽然目前已有许多针对选址问题的求解算法, 但物流配送设施的选址与传统的单/多设施选址有所不同, 传统选址问题的研究大多未考虑到物流配送设施选址的实际特点, 在具体实施中效果不佳, 所以仍然有较大的优化空间.

1 物流配送设施选址的特点

通过与物流企业沟通, 并深入分析了配送中心对物流配送流程的影响因素, 总结发现物流配送设施的选址有以下三个特点:

(1) 配送设施选址与配送线路的规划是相互影响的两个 NP-Hard 问题[3]. 配送设施与各需求点之间的距离不是简单的欧氏距离, 考虑到配送车辆是进行巡回配送的, 不同的配送范围划分、不同线路选择都对最终结果有影响. 而传统选址问题一般假设所选中心到需求点之间的距离是欧氏距离, 同时假设配送设施与需求点之间是点对点服务, 这两个假设与实际情况不符, 使得选址的结果不尽如人意.

(2) 物流配送设施的选址是一个多层级的选址问题. 物流业历经多年发展, 大多采用多级分拣配送体系: 大区中转仓库→一级区域分拣中心→二级区域分拣中心→末端快递站 (→快递柜). 还要注意的是快件的集散过程中不会出现跨层运输.

(3) 多个配送设施之间所负责的快件数量的均衡性. 物流业比较看重快件的流通速度, 若使大多数快件集中在一个分拣中心, 容易造成快件积压, 延长了所有快件的流通时间, 极大影响配送效率和客户体验, 所以在考虑配送设施与下层节点的运输路径成本的同时, 还要考虑多个配送设施之间所负责的快件数量的均衡性. 这一点在末端快递站的选址上尤为重要.

综上所述, 物流配送设施的选址是一个多层级均衡选址问题.

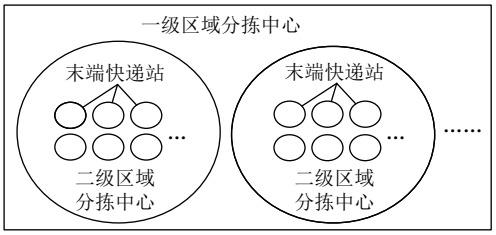

图 1 多级分拣配送体系

对特点 (1) 的一个可行的解决办法: 先进行配送范围的划分, 再决策配送设施的位置, 避免同时决策配送设施的位置和快件的配送线路这两个 NP-Hard 问题. 因为每天的快递订单数据会持续变动, 而且快递配送线路要根据配送设施位置和当天具体的快递订单数据来进行计算, 所以同时决策配送设施的位置和未来快件的配送线路不是一个好的方法. 较好地解决办法是对配送范围进行划分, 再根据配送范围和历史订单数据决策配送设施的位置, 最后根据配送设施位置和当天具体的快递订单数据规划配送线路.

对特点 (2) 的解决: 既然物流配送设施的选址是一个多层级的选址问题, 易想到可以使用层次聚类来对数据进行逐层聚类, 但配送设施选址还有一个特点: 快件的集散过程中不会出现跨层运输, 也即第 N+1 层的数据信息对于第 N 层的配送设施选址是有用的, 但对第 N–1 层的配送设施选址就没用了. 所以直接套用常见的层次聚类算法并不适用, 要对层次聚类算法进行改进使之更符合多层级的选址问题的应用场景.

对特点 (3) 的解决: 这个特点使得我们使用的算法要具有控制每个簇大小的能力.

本文针对以上物流配送设施选址问题的三个特点设计了一个基于 BIRCH 聚类的物流配送设施选址算法, 融合了 BIRCH 聚类算法和基于 Dijkstra 距离的重心法, 较好地解决了物流配送设施选址问题.

2 BIRCH 聚类算法

BIRCH 算法[4]是一种层次聚类算法, 采用自底向上的策略进行聚类, 并通过迭代重定位改进结果. BIRCH 算法只需要单遍扫描数据集就能进行有效聚类, 最小化数据集的输入输出, 尤其适用于对大数据集的处理. BIRCH 算法有两个核心概念: 聚类特征 CF 和聚类特征树 CF-Tree.

定义 1. 给定一个包含 N 个 d 维数据点的簇: $\{x_i\}(x_i = 1,2,\cdots,N)$, 则该簇的聚类特征 CF 是一个三元组: $CF = (N, LS, SS)$, 其中, N 是簇中数据点的个数, LS 是 N 个数据点的线性和[5], 即 $\sum_{i=1}^{n} x_i$, SS 是 N 个数据点的平方和, 即 $\sum_{i=1}^{n} x_i^2$.

定义 2. 一棵聚类特征树是一棵平衡树, 存储了层次聚类的簇的特征. 其每个结点代表一个簇, 且对其每个子结点 (/子簇) 都包含一个 CF 条目. CF 树的形态由三个参数决定: 非叶结点分支因子 B、叶结点分支因

子 L 和子簇最大半径阈值 T. 分支因子 B 限定每个非叶子结点的最大孩子个数; 分支因子 L 限定每个叶子结点的最大子簇数; 最大半径阈值 T 限定了子簇的最大半径, 保证簇的紧凑程度[5]. 如图 2 所示.

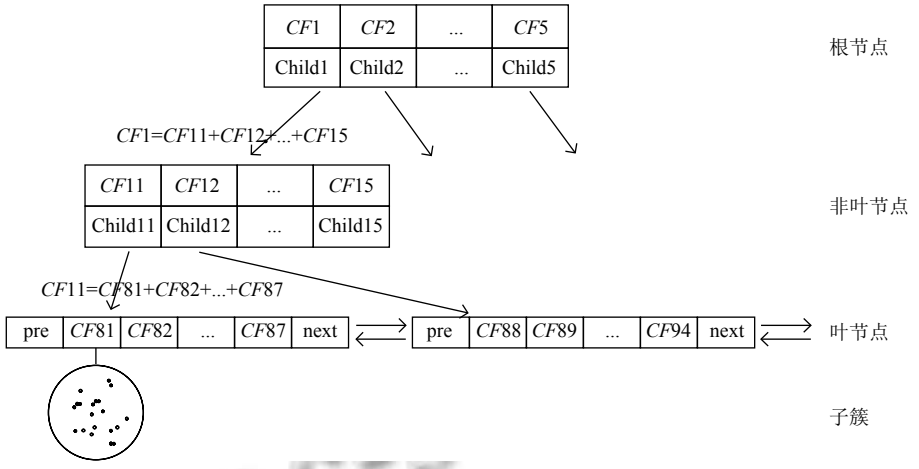

图 2 一棵 CF 树的例子 (B=5, L=7)

聚类特征 CF 概括了簇的基本信息, 并且是高度压缩的, 其中 LS 反映了聚类的中心,

$$x_0 = LS/N \tag{1}$$

x_0 为簇的中心, 用于计算簇 (/点) 与簇 (/点) 之间的距离; SS 反映了簇内对象的平均距离 (簇半径),

$$R = \sqrt{\sum_{i=1}^{N}(x_i - x_0)^2/N} = \sqrt{SS/N - LS^2/N^2} \tag{2}$$

而且 CF 满足线性可加性, 即

$$CF_1 + CF_2 = (N_1 + N_2, LS_1 + LS_2, SS_1 + SS_2) \tag{3}$$

这是一个很好的性质, 使得 CF-Tree 中结点的 CF 条目 (N,LS,SS) 值等于这个 CF 条目所指向的子结点的所有 CF 条目之和, 使得更新 CF-Tree 的时候效率很高.

进行 BIRCH 聚类的过程就是用所有的样本构建一颗聚类特征树的过程, 每个结点就是一个聚类的簇. BIRCH 算法十分高效, 处理大数据集时对内存的需求也不高; 但如果数据集的簇不是类似于超球体, 或者说非凸的, 则聚类效果不好[4]; 同时样本的读入顺序会导致不合理的 CF 结点分裂, 影响聚类效果.

结合本文所探讨的多层级均衡选址问题, 可以发现: 1) BIRCH 算法实现了层次化的聚类, 但并没有给出每个聚类的中心点; 2) BIRCH 算法根据参数 B、L、T 控制结点的分裂, 限制了每个聚类的子类个数, 而我们需要的是对每个聚类大小的控制以实现同层聚类之间的大小均衡. 所以 BIRCH 算法实现了对物流选址问题特点 (2) 的解决, 并具有加以修改以实现特点 (3) 的潜质. 于是本文基于 BIRCH 算法并结合基于 Dijkstra 距离的重心法, 设计了基于 BIRCH 聚类的物流配送设施选址算法.

3 基于 BIRCH 聚类的物流配送设施选址算法

该算法先使用带容量限制的 BIRCH 算法划分配送范围, 再使用基于 Dijkstra 距离的重心法在各个配送范围内进行单配送中心选址, 两步交替执行完成从底层到顶层的全部选址决策. 本算法避免了同时决策配送设施的位置和快件的配送线路; 从底层到顶层逐层递进决策; 在划分配送范围时保证了划分的均衡性, 对物流配送设施选址的三个特点都有较好的处理.

3.1 划分配送范围-带容量限制的 BIRCH 算法

假设目前要对第 K 层配送中心进行选址决策, 首先进行配送范围的划分. 输入数据为下一层的配送目的地和配送权重, 即已求得的第 K+1 层的需求点 (/配送中心) 及其需求量. 决策目标是将第 K+1 层的需求点集合划分成 m 个子集, 使得各个子集在地理位置上足够内聚且互不相交, 同时要满足各个子集所包含的需求量大致相当, 即满足均衡性.

带容量限制的 BIRCH 算法是在 BIRCH 算法的基础上修改而来. BIRCH 算法根据非叶结点分支因子

B、叶结点分支因子 L 对结点大小进行限制, 限制了每个聚类的子类个数[6]. 而我们需要控制每个聚类的大小以实现同层聚类之间的大小均衡, 所以作以下修改:

(1) 对聚类特征 CF 增加一个属性: 容量 W, 表明该聚类的容量大小, $CF=(N,LS,SS,W)$, 易知添加该属性后 CF 依然满足线性可加性.

$$CF_1+CF_2=(N_1+N_2,LS_1+LS_2,SS_1+SS_2,W_1+W_2) \tag{4}$$

(2) 树结构限定为三层: 根结点、中间结点、叶结点, 因为最终目标是 m-均衡划分, 使用三层即可.

(3) 不使用非叶结点分支因子 B 控制非叶结点的分裂, 改用中间结点容量阈值 H, 当中间结点的容量达到 H, 就进行结点的分裂, 分裂规则: 选择该中间结点距离最远的两个 CF 条目 (根据聚类中心 x_0 计算) 作为新中间结点的初始叶结点, 其余叶结点 CF 条目依次合并到较近的新中间结点.

(4) 依然使用叶结点分支因子 L 和子簇最大半径阈值 T 来控制叶结点和子簇的大小, 防止叶结点过大使中间结点分裂失败.

具体算法步骤如算法 1.

算法 1. 带容量限制的 BIRCH 算法构建聚类特征树过程

1) 读入新样本 x, 根据其坐标位置在所有子簇中寻找距离 x 最近的子簇 D(根据公式 (1)), 将 x 加入到 D 中, 如果 x 加入后, 子簇 D 的簇半径 (根据公式 (2)) 小于阈值 T, 转入 4). 否则转入 2);

2) 分裂子簇: 将子簇 D 中最远的一个样本点独立为一个新的子簇, 并归入这个叶结点, 如果当前叶结点的子簇个数小于阈值 L, 转入 4). 否则转入 3);

3) 分裂叶结点: 选择叶结点中距离最远的两个 CF 条目独立为两个新叶结点的第一个 CF 条目, 将其余 CF 条目按照距离远近归入对应的叶结点;

4) 更新从子簇向上到根的路径上所有的 CF 四元组 (N,LS,SS,W), 检查中间结点的容量 W 是否超过容量阈值 H, 超过则按 3) 所述分裂规则分裂中间结点;

5) 若还有未处理样本, 转 1), 否则检查中间结点, 将容量小于 $H/2$ 的中间结点所含样本点重新处理, 使所有中间结点的容量大于 $H/2$ 小于 H.

此时所有中间结点容量大于 $H/2$ 小于 H, 全部中间结点即为第 K+1 层需求点集合的 m 个均衡划分子集, 也即所求的 m 个配送范围. 下面要针对每一个配送范围求解其配送中心.

3.2 配送范围内选择配送中心-基于 Dijkstra 距离的重心法

对第 K 层进行选址决策, 在上一小节已经将第 K+1 层的需求点集合划分成了 m 个容量大致均衡的不交子集, 第二步即是为每个子集选择一个位置作为该配送范围的配送中心. 这个子问题即是连续区域单设施选址问题, 采用常用的基于 Dijkstra 距离的重心法进行决策即可.

重心法是连续区域单设施选址问题的最常用的一种方法, 假设各个需求点的位置和需求量已知, 优化目标是使运输总成本最小[2].

$$\begin{aligned}&\min 运输总成本 =\\&\min\sum 中心与各需求点之间的运输距离\times 需求量\end{aligned} \tag{5}$$

可求得由各需求点构成的超多面体的几何重心即为单设施选址问题的最优解.

$$\bar{x}=\frac{\sum_{i=1}^{N} 运输距离\times 需求量\times x_i}{\sum_{i=1}^{N} 运输距离\times 需求量} \tag{6}$$

重心法考虑因素单一, 将运输成本作为唯一的决策因素, 未考虑城市交通状况和地产价格等; 以中心与需求点之间的直线距离作为运输距离, 与事实不符; 只能用于单设施选址. 优点则是计算简单.

基于 Dijkstra 距离的重心法采用 Dijkstra 距离作为运输距离[7]. Dijkstra 距离是使用 Dijkstra 算法求得的结点间最短距离, 符合快件配送时巡回配送的特点, 比使用直线距离作为运输距离的模拟效果要好许多. 因此这里使用基于 Dijkstra 距离的重心法作为划分了配送范围后, 在配送范围内进行单配送中心选址的方法, 具体算法步骤如算法 2.

算法 2. 基于 Dijkstra 距离的重心法

1) 对要求解配送中心的子集 D 进行初始化: 相邻结点之间的边的权重为综合考虑了运输距离、交通条件的“虚拟距离”;

2) 计算任两个结点之间 Dijkstra 距离, 即使用 Dijkstra 算法迭代求得的最短距离;

3) 使用 Dijkstra 距离作为运输距离代入公式 (6), 所求得的重心点即为该配送范围的配送中心.

第 K+1 层全部 m 个划分的重心点即为第 K 层的选址决策, 接下来将之作为第 K 层的需求点, 需求量为对应配送范围的总需求量, 继续进行“划分配送范围-配送范围内进行单配送中心选址”这两步即可得到第 K–1 层的选址决策, 如此循环可得从底层到顶层的全部选址决策.

4 实验验证

为了验证本算法的有效性, 选取了一次典型的案

例: 国内某三线城市的物流配送设施选址规划, 利用 2017 年二、三季度的调研数据进行一座二级区域分拣中心、若干末端快递站及快递柜的选址决策.

根据分拣中心和末端快递站标配的设备及人员, 可以确定分拣中心日处理订单数最多 12 000 件, 末端快递站日处理订单数最多 700 件. 本案例中需要决策两级配送设施, 在决策末端快递站时, 确定参数 H=700, L=10, T=500; 决策二级区域分拣中心时, 确定参数 H=12 000, L=10, T=50 000. 最终得到 19 座末端快递站和 1 座二级区域分拣中心的选址决策. 另外, 在决策末端快递站时, 对聚类后的子簇进行筛选, 容量足够大的子簇说明样本点足够密集, 可以选作快递柜, 在此选容量大于 30 的子簇建设快递柜, 共 530 处.

作为对照, 使用 k=19 和 k=1 的 K-means 算法同样得到 19 座末端快递站和 1 座二级区域分拣中心的选址决策. 并通过以下三个指标比较两种算法的优劣.

(1) 平均送货距离 $d=\sum d(x_i,x_0)/$总件数, 其中 x_i 为第 i 件快递的配送目的地, x_0 为负责配送第 i 件快递的配送站. 该指标用来衡量选址结果的内聚性.

(2) 快递站日均负载均衡度 $\sigma^2=\sum(x_i-\mu)^2/N$, 即所有快递站日均负载快件数的方差, μ 为所有快递站日均负载的均值. 该指标用来衡量选址结果的均衡性.

(3) 超出负载的快递站数目.

实验结果见表 1.

表 1 算法效果对比

	本文算法	K-means
d(米)	1073	1659
σ^2(百件^2)	0.64	30.3
超出负载的快递站数目 (个)	0	5

通过上述指标可以看出, K-means 算法的选址结果均衡性较差, 导致在快件密集的城区快递站不够, 超出负载的快递站数目较多, 使得平均送货距离指标也比较差.

另外从功能性上来说 K-means 算法无法实现不确定个数的快递柜选址, 而本文算法可以完成决策. 综上, 本文设计的基于 BIRCH 聚类的物流配送设施选址算法较好的解决了物流配送设施的选址问题.

5 结语

本文通过分析物流配送设施选址的特点, 通过改进 BIRCH 聚类算法并结合基于 Dijkstra 距离的重心法设计了有针对性的物流配送设施选址算法, 较好地解决了物流配送设施的选址问题[8,9]. 同时, 该算法对如通讯基站、银行营业网点、垃圾站等需要考虑到负载均衡的设施选址问题以及部门机构选址等多层级选址问题有一定参考性.

参考文献

1 关菲, 张强. 模糊多目标物流配送中心选址模型及其求解算法. 中国管理科学, 2013, 21(S1): 57–62.

2 程珩, 牟瑞芳. 物流配送中心选址的重心法探讨. 交通运输工程与信息学报, 2013, 11(1): 91–95. [doi: 10.3969/j.issn.1672-4747.2013.01.017]

3 石兆. 物流配送选址—运输路径优化问题研究[博士学位论文]. 长沙: 中南大学, 2014.

4 Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. ACM SIGMOD Conference, 1999, 25(2): 103–114. [doi: 10.1145/233269.233324]

5 韦相. 基于密度的改进 BIRCH 聚类算法. 计算机工程与应用, 2013, 49(10): 201–205. [doi: 10.3778/j.issn.1002-8331.1112-0567]

6 Karatas M, Yakıcı E. An iterative solution approach to a multi-objective facility location problem. Applied Soft Computing, 2018, 62: 272–287. [doi: 10.1016/j.asoc.2017.10.035]

7 武方方. 基于大数据的物流配送中心选址优化研究[硕士学位论文]. 合肥: 合肥工业大学, 2015.

8 Karatas M, Nasuh R, Hakan T. A Comparison of P-median and maximal coverage location models with Q–coverage requirement. Procedia Engineering, 2016, 149: 169–176. [doi: 10.1016/j.proeng.2016.06.652]

9 王鹏飞. 基于聚类算法的快递服务网点布局研究[硕士学位论文]. 成都: 成都理工大学, 2016.