计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):251−255 [doi: 10.15888/j.cnki.csa.007029]

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

基于改进的 K-means 聚类的多区域物流中心选址算法①

鲁玲岚, 秦江涛

(上海理工大学 管理学院, 上海 200093)

摘　要: 针对当前多区域物流中心选址需建立配送中心个数不定、位置、覆盖范围不明的问题, 本文提出了一种改进的 k-means 聚类算法, 以城市经济引力模型为基础, 将城市运输距离与居民消费能力的指标相结合, 重新定义对象之间相似性度量的距离因子. 并将密度思想引入 k-means 算法, 提出类内差分均值的概念确定最优聚类数. 实现分区后, 分别在这些区域中利用重心法对配送中心进行最终的确定. 最后实例分析了在西部地区 37 个城市创建物流配送中心的选址过程, 并通过和传统的 k-means 聚类的选址结果对比, 说明改进后的算法不仅可以节省配送时间, 而且大大降低了运输成本, 有很好的经济利用价值.

关键词: 多区域配送中心选址; k-means 聚类; 城市经济引力模型; 重心法; 西北物流

引用格式: 鲁玲岚,秦江涛.基于改进的 K-means 聚类的多区域物流中心选址算法.计算机系统应用,2019,28(8):251–255. http://www.c-s-a.org.cn/1003-3254/7029.html

Multi-Regional Logistics Distribution Center Location Method Based on Improved K-means Algorithm

LU Ling-Lan, QIN Jiang-Tao

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Focusing on the issues that the number, location, and coverage of multi-regional logistics centers of distribution centers are unknown, an improved k-means clustering algorithm is proposed. Based on the urban economic gravity model, this algorithm combines the urban transportation distance with the indicators of household consumption capacity, redefines the distance factor of the similarity measure between objects. The idea of density is introduced into the k-means algorithm, and the concept of intra-class difference mean is raised to determine the optimal number of clusters. After the partition is implemented, the centroid method is used to determine the final distribution center in these areas. Finally, in case study, we analyze the location process of constructing logistics distribution centers in 37 cities in the western region, and compares them with the traditional k-means clustering results. The comparing result shows that the improved algorithm not only saves the delivery time, but also greatly reduces the transportation cost and has sound economic value.

Key words: multi-regional distribution center location; k-means algorithm; urban economic gravity model; center of gravity method; northwest logistics

区域配送中心配送规模较大, 用户较多, 是物流中重要节点. 其合理的物流规划不仅关系到是否能达到资源利用的效益最大化, 满足客户的物流需求, 保证物流时效性, 也决定着投入物流成本的高低. 多区域物流配送中心选址需确定配送中心最佳数目、位置及覆盖范围. 聚类算法将一整个区域划分, 以此来确定配送中

① 收稿时间: 2019-01-29; 修改时间: 2019-02-26; 采用时间: 2019-03-14; csa 在线出版时间: 2019-08-08

心数目以及其覆盖范围, 非常适用于多配送中心的选址问题. 诸如 K-means 的聚类算法通常用于多区域配送中心的选址, 但它们的初始聚类中心是随机的, 易陷入局部极小解, 且 K 值不同导致聚类效果各异[1], 并且有些城市地处偏僻, 常被作为噪音数据排除, 最主要的是由于山川、河流的阻碍, 两地之间的欧式距离并不能用来衡量实际的运输距离.

1 目前主要 K-means 聚类划分方法综述

K-means 聚类算法技术较为成熟, 操作便捷, 经常被用于选址中对区域的划分上. 考虑到该聚类方法对区域划分时存在的劣势, 许多学者对该聚类方法进行了改进. 朱培芬结合密度的思想, 在可选范围中优先考虑边缘点, 具有良好的全局收敛性[2], 但并未科学衡量距离因子, 也未曾考虑物流需求等重要影响因子. 谷炜提出两阶段 K-means 聚类算法, 在传统 K-means 聚类后, 通过不断迭代来检验是否满足配送时间最少的收敛条件, 最终确定聚类结果[3], 避免了算法陷入局部最优, 但不断的迭代大大增加了算法的运算时间. 于晓寒考虑到河流、公路等地理障碍, 以“障碍距离”为差异度量标准, 站点工作量为约束, 提出基于障碍距离的约束聚类算法[4], 但距离并不是简单的空间距离, 虽其在抽象意义上是同质的, 但诸如城市这样实质的经济主体, 距离明显包含了更多的非空间因素.

本文针对选址的实际情况, 以城市经济引力模型为基础, 考虑到距离的“非空间因素”, 重新定义了对象间度量距离的因子. 同时把密度的思想引入 K-means 算法以解决 K 值不确性问题, 并提出类内差分均值的概念确定最优聚类数. 借助改进的 K-means 聚类算法完成西北地区物流网络的划分, 实现分区后, 分别在这些区域中利用重心法对配送中心进行最终的确定. 并通过和传统 K-means 聚类结果对比, 来验证改进后方法更优.

2 改进的 K-means 聚类算法和重心法

2.1 改进 K-means 聚类对区域划分

(1) 将密度思想引入 K-means 算法

基于传统 K-means 算法易受噪声和孤立点影响的事实, 本文将密度的思想引入 K-means 算法来确定 k 个聚类中心. 不仅可以避免噪声数据干扰, 而且可以有效降低算法时间复杂度. 考虑到 M 维空间的 n 维数据点$X_i(i=0,1,2,\cdots,n-1)$ 其基本思路如下: 计算每个数据点X_i的密度. $j=0,1,2\cdots n-1$. r_a为正数, 定义为该点的领域半径, 取$r_a=\frac{1}{2}\max\{\|X_i-X_k\|\}$, 当$i$取一个值时,$k=0,1,2,\cdots,n-1$, X_i的密度指标记为D_i具体计算公式如下:

$$D_i=\sum_{j=0}^{n-1}\exp\left[-\frac{\|X_i-X_j\|^2}{(0.5r_a)^2}\right] \tag{1}$$

根据式 (1) 获取所有样本点的密度指数, 按照从大到小的顺序排列, 选取前 k 个数据点作为聚类中心.

(2) 基于经济引力模型对距离的定义

在度量两个城市之间相似度的过程中, 把距离的概念仅定义为两个城市间的空间距离是不可取的, 为了更科学的衡量距离且更加有效的运用于实际情况, 本文以居民可支配收入作为衡量城市间的经济引力因子, 交通运输时间作为衡量城市间的距离因子. 相应的公式如下:

$$D(X_i,X_j)=\frac{{D_{ij}}^2}{(M_iM_j)^u} \tag{2}$$

式 (2) 中M_iM_j分别为城市i和城市 j 的居民人均可支配收入指标. 居民人均可支配收入是衡量居民消费能力的重要指标, 在消费性支出中包括食品烟酒、衣着、居住、生活用品以及其他用品和服务. 随着科技的进步, 出现了越来越多样化的消费渠道. 这里主要考虑线上和线下消费. 在物流网络建设不完善的地区, 居民线上消费遇到较大限制, 消费被迫转到线下, 这些地区物流需求远远不及实际的消费需求, 用居民的可支配收入来衡量物流需求更为现实. D_{ij}为城市 i 和城市 j 的交通距离, 本文以车辆在两座城市之间行驶一趟的交通时间来衡量. u 作为调节物流网络划分时受地区居民消费水平影响的程度. 一般来说, u 越大, 则说明在区域物流网络划分时, 各地区的居民消费水平的吸引力占主导地位, 城市间的距离因素被较大程度的弱化; 反之 u 越小, 则认为各地区的居民消费水平影响较弱, 城市间的距离因素主导地位越强; u 为 0 时, 距离等价于欧式距离.

(3) k 值的确定

聚类的目的是使同一类的样本点间相似度高, 而不同类间相似度低. 为精确测量聚类结果, 进一步确定 k 值. 本文采用计算类内差分均值的评估方法. 计算样本点与各自聚类中心的距离之和的平均值. 值越小, 总

的聚类距离越小, 类间相似度越高, 聚类效果越好; 反之, 值越大, 聚类距离越大,类间相似度越低, 聚类效果越差. 用内类差均值的方法衡量最佳聚类个数, 以达到修建最少的配送中心实现最大区域的覆盖.

2.2 重心法

重心法是一种简单可行的选址方法, 通常用于解决连续点的单个配送中心选址问题, 其唯一的决策依据是运输成本. 应用时, 它对候选位置没有任何限制, 在已知各个备选地点的位置、物流需求量、各个备选地点的直线距离的前提下, 运用重心法可以很好地确定配送中心的位置. 其运算方便, 计算速度快, 通过几次迭代就能计算出理论位置. 但此种方法只适用于确定单一配送中心, 对于确定多个配送中心的选址问题, 此法并不适用[5].

3 多区域配送中心选址模型

针对多区域配送中心选址问题, 本文采用目前比较主流的两阶段模型:第一阶段把所有的需求点划分为若干个配送区域, 第二步阶段选取相应区域中最佳配送中心[6–8]. 借助该模型的选址步骤, 本文首先用改进的 k-means 聚类对区域进行划分, 确定需建立配送中心的个数以及其配送范围. 然后运用重心法在划分的区域中选取合适的配送中心位置.

3.1 基于改进的 K-means 聚类对区域划分

本文运用 MyEclipse Professional 2014 软件, 采用 JAVA 语言, 在 Win10 64 位系统环境下运行.其中按照样本大小和聚类个数 k 值的关系, $k \leqslant \sqrt{n}$[9], 这里 n 为样本的大小. 具体操作步骤如下:

1) 存储数据, 对数据进行归一化处理.

2) 按照式 (1) 计算城市密度, 按密度从大到小的顺序排列, 取排序前 m 个城市作为聚类中心候选点.

3) 取前 k 个候选点作为聚类中心, 按照式 (2) 计算当 k 取不同值时的距离因子, 这里 u=0.1, 当 u=0.1 时, 最能科学均衡经济引力与距离之间的影响力. 按照最小距离原则分配$n-k$个城市样本. 得到$\sqrt{n}$种聚类结果.

4) 计算$\sqrt{n}$种聚类结果的类内差分均值, 选取最优优 k 值, 确定最佳聚类方案.

5) 结束, 输出最优聚类结果.

3.2 重心法对单一区域配送中心选址

在上一阶段过程中确定了配送中心的个数以及其覆盖范围的问题, 但聚类过程中选取的聚类中心仅仅考虑到了城市交通的密度, 而未考虑其他因素, 不足以认为是最合适的配送中心地址. 需结合单一配送选址模型在各个区域中对聚类中心进行修正, 从而确定单一区域配送中心的最佳位置, 本文选取重心法来进行此阶段工作.

4 评估模型建立

区域物流中心选址模型是带有复杂约束的非线性规划模型, 在构建评估模型时先明确如下假设: (1) 中心容量总是可以满足所覆盖需求点的需求量, 并由该中心供应的所有单位的需求量确定. (2) 一个需求点仅由一个物流配送中心供应. (3) 从物流中心到需求点通过零担物流的方式进行配送, 且不考虑装载的问题, 仅以最大车载量进行运算. (4) 运输费用仅由运价、实际运输路程决定, 不考虑装卸等人工成本. (5) 配送中心每辆车每天仅往需求点配送一趟. 基于以上 5 条假设建立衡量选址是否有效的模型. 目标函数是各个配送中心到需求点的运输费用之和最小, 目标函数为:

$$F_{MIN}=\sum_{i=0}^{k-1}\sum_{j\in N_i} m_{ij}w_{ij}J \tag{3}$$

式 (3) 中 i 表示配送中心, 其取值为$0,1,2,\cdots,k-1$, k 表示配送中心个数. j 为配送中心所供应需求点个数, N_i表示第 i 个配送中心需求点集合. m_{ij}表示配送中心 i 到需求点 j 的计费里程, 以 km 为单位. w_{ij}表示配送中心 i 到需求点 j 运输货物的重量, 以 kg 为单位. J 表示从配送中心 i 到需求点 j 的单位运价, 以元/kg*km 为单位.

5 实例分析

5.1 数据来源与处理

西北地区地域广阔, 自然资源丰富, 但地区天气恶劣, 城市之间道路险阻较多, 交通十分不便, 物资运输上的不便是西北地区电子商务的发展主要阻碍. 本文对西北物流网络进行研究, 由于新疆地区地域物流管理体制较混乱, 机构多元化, 多采用外包的物流模式, 这里主要讨论除新疆之外的 37 个地级市的物流区域配送中心选址问题. 本次共采集 37 个城市间交通时间数据共 703 条记录. 选取甘肃、青海、陕西、宁夏 2016 年的统计年鉴中选取四省的城镇居民可支配收

入. 并对已有数据进行了归一化处理, 来消除数据量纲的影响.

为验证方法的正确性和实用性, 抽象实验数据为: 零担运输的机动车辆车身均为 5~7 米,车载量 20 000 kg, 平均行驶速度 68 km/h. 考虑到西北地区一体化趋势日益明显, 可忽略地域的运价费用差异, 这里单位运价均定 1 元/kg*km. 通过城镇居民可支配收入折算成每日的需求量以及配送所需车辆如表 1 所示.

表 1 城镇每日物流需求量以及所需配送车辆量化表

城市	需求量 (kg)	所需车辆 (辆)	城市	需求量 (kg)	所需车辆 (辆)	城市	需求量 (kg)	所需车辆 (辆)
酒泉	30 072	2	渭南	27 485	2	德令哈	27 720	2
嘉峪关	33 540	2	商洛	25 468	2	玉树	27 978	2
石嘴山	25 970	2	安康	35 510	2	张掖	21 502	2
临夏	17 912	1	庆阳	25 300	2	金昌	32 073	2
同仁	26 567	2	榆林	29 781	2	西宁	27 539	2
咸阳	31 662	2	延安	30 693	2	海晏	26 828	2
西安	35 630	2	固原	22 716	2	共和	26 218	2
铜川	27 549	2	吴忠	23 351	2	玛沁	28 133	2
陇南	20 504	2	中卫	23 276	2	海东	25 492	2
银川	30 477	2	兰州	29 661	2	武威	23 612	2
宝鸡	20 816	2	白银	25 313	2	合作	21 327	2
天水	22 648	2	定西	20 815	2	平凉	23 446	2
汉中	25 595	2						

5.2 改进方法选址求解

(1) 确定聚类中心候选点, 本文取前 10 位城市以及相应密度, 从结果看出, 嘉峪关、酒泉、西安等作为当前交通的枢纽中心, 是聚类中心的首选. 在区域划分中把密度思想引入是合理有效的. 结果如表 2 如下.

表 2 前 10 城市密度

密度排名	城市	城市密度
1	嘉峪关	1.018 653 556 307 322 5
2	酒泉	1.018 248 425 610 582 2
3	西安	1.014 244 608 966 195 4
4	咸阳	1.014 077 791 757 837 4
5	西宁	1.000 362 721 585 526 9
6	海东	1.000 257 033 444 363 5
7	渭南	1.000 014 151 695 080 7
8	铜川市	1.000 008 098 251 050 9
9	银川	1.000 000 005 215 422 5
10	石嘴山	1.000 000 002 596 885 4

(2) 通过类内差分均值评估, 绘制曲线图观察最适 k 值. 从折线图很直观看出, 当 k=6 时, 曲线出现拐点, 达到最小值, 当 k=7,8 时类内差均值小范围增加, 证明当建立 6 个配送中心时物流网络已经达到了较好的覆盖效果. 随着配送中心的增多, 覆盖效果反而降低. 即当 37 个城市被分为 6 类时, 类间的相似度越高, 聚类结果越好. 如图 1 所示.

确定最佳聚类个数后, 根据改进的 K-means 算法聚类, 当 k=6 时, 聚类结果如表 3 所示.

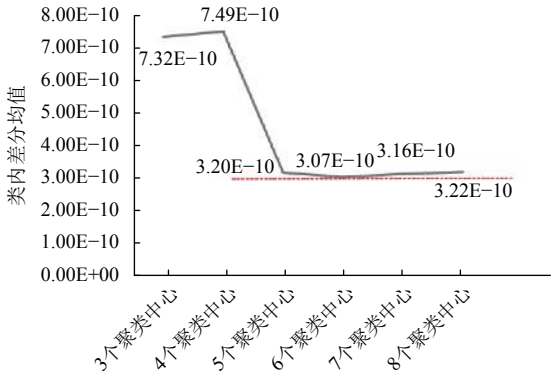

图 1 3~8 个聚类的类内差分均值

表 3 k=6 时的城市分区结果

区域	城市聚类中心	覆盖城市
区域一	酒泉市	张掖
区域二	咸阳市	宝鸡、固原、汉中、天水、陇南
区域三	嘉峪关市	金昌
区域四	西安市	铜川市、渭南、商洛、安康、平凉、庆阳、榆林、延安
区域五	西宁市	共和县、德令哈、海晏县、玛沁县、玉树
区域六	海东市	吴忠、中卫、武威、兰州、白银、定西、合作、临夏、石嘴山、银川、同仁

(3) 把每个区域中城市作为物流服务需求点,人均可支配收入作为需求量, 用重心法修正的配送中心结果如下, 区域一: 张掖, 区域二: 宝鸡, 区域三: 嘉峪关, 区域四: 铜川市, 区域五: 共和县, 区域六: 白银.配送中心地址及其辐射区域如图 2 所示.

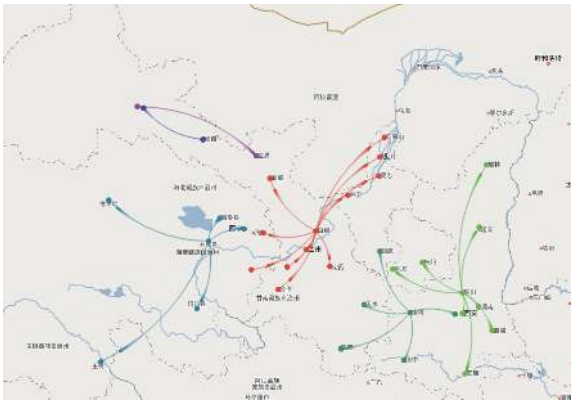

图 2 西北四省配送中心地址及其覆盖城市图

5.3 结果评估

本文参照传统的 K-means 算法对配送中心选址结果进行对比. 在传统的 K-means 算法中把城市之间的直线距离作为相似性度量的因子, 以每个城市之间的经纬度计算城市距离, 设置 k=6, 阈值为 0.000 001, 共迭代 50 次, 得到区域划分后, 再用重心法选出各个区域的配送中心. 以式 (3) 计算运输成本, 实验结果如表 4 所示.

表 4 配送中心选址效果对比

	传统 K-means	改进 K-means
总运输成本 (元)	350 640 840	309 014 420
总运输时间 (小时)	257.90	227.28

从配送中心选址效果对比中可以看出, 对比传统的 K-means 聚类算法, 改进后的算法每天可节约运输成本 41 626 420 元, 配送时间可节省 30.62 小时. 在保证配送时效的同时也节约了运输成本, 在一定程度上说明改进后的算法较传统的算法在实际中能创造更好的经济效益.

6 结论

通过本文分析, 得出以下结论: 在进行区域划分时, 考虑当地的实际情况越来越受到学者们的重视, 本文以城市运输距离与居民消费能力的指标相结合, 重新定义了对象间的相似性度量的距离因子, 区域划分后后, 利用重心法对聚类中心进行修正, 科学的选出了每个区域内的配送中心, 更符合西北地区发展的实际情况. 为了验证改进方法的有效性, 本文对比传统 K-means 聚类对区域进行划分后选址结果, 实验可知: 较传统的 K-means 聚类, 改进后的算法不仅节省了配送时间, 而且大大降低了运输成本, 具有很好的实际运用价值, 为西北物流建设提供了参考. 但本文考虑的是西北地区这一宏观的区域性概念, 还有地处偏僻的部落并未考虑在内, 因此划分的区域未能全部覆盖, 想要在西北地区完成物流的全部覆盖及其布局, 还需要一些发展的契机, 有待以后学者们去探索.

参考文献

1 Kanungo T, Mount DM, Netanyahu NS, *et al*. An efficient K-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881–892. [doi: 10.1109/TPAMI.2002.1017616]

2 朱培芬, 汉吉庆, 杨华龙, 等. 基于改进 K-means 算法的烟草配送区域划分. 物流工程与管理, 2009, 31(6): 84–85. [doi: 10.3969/j.issn.1674-4993.2009.06.034]

3 谷炜, 张群, 胡睿. 基于改进 K-means 聚类的物流配送区域划分方法研究. 中国管理信息化, 2010, 13(24): 60–63. [doi: 10.3969/j.issn.1673-0194.2010.24.028]

4 于晓寒, 王东. 基于带约束 K-means 聚类的城市快递配送区域划分. 哈尔滨商业大学学报 (自然科学版), 2016, 32(5): 631–634.

5 沈默, 戴冰洁. 物流配送中心重心法选址解析——以苏宁为例. 物流技术, 2015, 34(4): 192–194. [doi: 10.3969/j.issn.1005-152X.2015.04.058]

6 叶浔宇. 基于聚类和重心法的区域配送中心选址应用研究. 中国市场, 2009, (23): 83–85. [doi: 10.3969/j.issn.1005-6432.2009.23.026]

7 胡贤满, 张燕, 李珍萍. 带车辆路线安排的多配送中心选址问题的求解——基于 SPSS 和遗传算法. 物流技术, 2010, 29(1): 83–86. [doi: 10.3969/j.issn.1005-152X.2010.01.028]

8 孔继利, 顾苧, 孙欣, 等. 系统聚类和重心法在多节点配送中心选址中的研究. 物流技术, 2010, 29(3): 83–85.

9 杨善林, 李永森, 胡笑旋, 等. K-means 算法中的 k 值优化问题研究. 系统工程理论与实践, 2006, 26(2): 97–101. [doi: 10.3321/j.issn:1000-6788.2006.02.013]