中国科学 : 数学　2020 年　第 50 卷　第 9 期 : 1387 ∼ 1404
SCIENTIA SINICA Mathematica
论　文

k- 均值问题的理论与算法综述

献给越民义教授 100 华诞

张冬梅[1], 李敏[2*], 徐大川[3], 张真宁[3]

1. 山东建筑大学计算机科学与技术学院, 济南 250101;
2. 山东师范大学数学与统计学院, 济南 250014;
3. 北京工业大学数学学院, 北京 100124
E-mail: zhangdongmei@sdjzu.edu.cn, liminemily@sdnu.edu.cn, xudc@bjut.edu.cn, zhangzhenning@bjut.edu.cn

收稿日期: 2019-11-16; 接受日期: 2020-02-26; 网络出版日期: 2020-09-09; * 通信作者
国家自然科学基金 (批准号: 11531014 和 11871081)、山东省高校科研计划 (批准号: J17KA171)、山东省自然科学基金 (批准号: ZR2019MA032) 和北京市教委科技项目 (批准号: KM201810005006) 资助项目

摘要　k- 均值问题是理论计算机科学和组合优化领域的经典问题之一. 相应的 Lloyd 算法是数据挖掘的十大经典算法之一, 在各种领域被广泛研究和应用, 特别是在图像处理和特征工程方面. 随着数据多样性和数据量的爆炸性增长, 在实际应用中遇到的 k- 均值聚类问题更加复杂多样, 产生了各种亟需解决的具有挑战性的研究课题. k- 均值问题在理论上是 NP- 难的. 本文介绍经典 k- 均值问题及其变形的基于局部搜索、线性规划舍入、原始对偶、对偶拟合和 Lagrange 松弛等技术的有效算法. 首先介绍经典 k- 均值问题的近似算法、加倍度量空间中的有效多项式时间近似方案及满足稳定性实例的多项式可解性, 然后介绍 k- 均值问题的若干重要变形, 包括 k- 中位、球面 k- 均值、鲁棒 k- 均值、带约束的 k- 均值和隐私保护 k- 均值等问题, 最后列出 k- 均值领域中的若干公开问题.

关键词　k- 均值　近似算法　线性规划

MSC (2010) 主题分类　90C27, 68W25

1　引言

在机器学习中, 无监督学习 (unsupervised learning) 的目的是通过对无标记样本的学习来揭示数据的内在性质和规律. 无监督学习中研究最多、应用最广的是聚类 (clustering) (参见文献 [1]). 给定若干对象组成的集合, 聚类就是将这个集合分成多个聚簇 (cluster), 每个聚簇由相似的对象组成, 不同聚簇中的对象差异较大. 与分类不同, 聚类对要划分的对象的类别是未知的. 互联网技术的深入应用带来了数据多样性和数据量的爆炸增长, 获得各种类型有标签的数据变得非常困难, 有时预先获得标签

英文引用格式: Zhang D M, Li M, Xu D C, et al. A survey on theory and algorithms for k-means problems (in Chinese). Sci Sin Math, 2020, 50: 1387–1404, doi: 10.1360/SSM-2019-0280

都是困难的, 因此, 数据聚类技术愈来愈受到重视. 聚类算法在许多数据驱动的应用领域都是非常核心的算法.

k- 均值 (k-means) 聚类是一种重要的数据聚类技术. 经典的 k- 均值问题可描述为: 给定 n 个元素的观测集, 其中每个观测点都是 d 维实向量, 目标是选取 k ($\leqslant n$) 个点作为聚类中心, 把 n 个观测点划分到 k 个集合 (每个集合对应一个聚类中心) 中, 使得所有观测点到对应的聚类中心的距离的平方和最小. 容易证明, 对于任意给定集合, 最优的聚类中心是该集合中所有观测点的均值点. 在 k- 均值聚类中, 针对各种实际问题, 可能会对距离给出不同的定义, 对聚类中心采取不同的选取方式, 或采用不同的优化目标函数, 这样就引出了与 k- 均值相关的各种各样的变形. 与综述文献 [2, 3] 相比, 本文囊括了最近两年的结果, 增加了一些新的变形的介绍, 并系统总结了该领域的研究技巧.

k- 均值问题在不同领域里被分别提出 (参见文献 [4–7]), 是理论计算机科学和机器学习领域的研究热点 (参见文献 [8–11]). Lloyd 算法是求解 k- 均值问题的一种简单有效的算法, 在聚类分析及相关领域具有广泛的应用, 特别是在图像处理和特征工程方面有典型的应用. Lloyd 算法在图像处理方面常用于进行图像分割和图像压缩. 而在特征工程方面, k- 均值聚类更是广泛应用于特征选择 (feature selection) 和特征抽取 (feature extraction) 等方面. 社交网络和大数据等带来的各种新的应用环境, 对 k- 均值聚类带来了新的挑战, 产生了各种亟需解决的具有挑战性的研究课题, 需要我们研究求解 k- 均值问题及其变形的各种算法.

本文结构如下: 第 2 节介绍经典 k- 均值问题的研究进展, 第 3 节介绍 k- 均值问题的重要变形, 第 4 节总结待研究的问题.

2 k- 均值问题

设计近似算法是求解 NP- 难问题的方法之一, 近似算法能对问题给出具有性能保证的近似解. 关于 k- 均值问题近似算法的研究主要分为两方面: 一是得到一般 k- 均值问题的近似结果, 二是寻找特殊 k- 均值问题的多项式时间近似方案 (polynomial-time approximation scheme, PTAS). 为了平衡解的质量与可行性, 本文引入双准则近似. k- 均值的双准则 (β, α)- 近似算法表示算法得到的解把观测点分成 βk 类, 目标值不超过最优值的 α 倍. 在分析 k- 均值问题的算法性能时, 除了与最优值 OPT 相比之外, 有时还引入附加误差项: 算法输出解的质量不超过 $a \cdot \mathrm{OPT} + b$, 其中 b 称为附加近似误差 (additive approximation error), a 称为乘法近似误差 (multiplicative approximation error), 在不引起混淆的情形下, 仍称 a 为近似比.

给定观测集 $\mathcal{X}$, 用 $\mathrm{OPT}_k(\mathcal{X})$ 表示相应 k- 均值问题的最优值. 如果 $\mathrm{OPT}_k(\mathcal{X})/\mathrm{OPT}_{k-1}(\mathcal{X}) \leqslant \delta^2$, 则称 $\mathcal{X}$ 关于 k- 均值问题是 δ- 分离的, δ 称为分离比值. 为了叙述方便, 引入一些记号: L 表示问题输入的字节长度; S_0 通常是算法的初始可行解, $\mathrm{cost}(S_0)$ 表示 S_0 的费用; $\mathrm{Var}(\mathcal{X}) := \mathrm{OPT}_1(\mathcal{X})$; $\tilde{O}(\cdot)$ 较之于 $O(\cdot)$ 隐藏了 log 的多项式量级因子.

2.1 问题描述

给定 n 个元素的观测集 $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ 和整数 k, k- 均值问题的目标是选取中心点集合 $\mathcal{C} = \{c_1, c_2, \ldots, c_k\} \subseteq \mathbb{R}^d$, 使得下面的函数值达到最小:

$$\sum_{i=1}^{n} \min_{j \in \{1,2,\ldots,k\}} \|x_i - c_j\|^2.$$

对于任意集合 $U \subseteq \mathcal{X}$ 和点 $c \in \mathbb{R}^d$, 可以证明 (参见文献 [12])

$$\sum_{x\in U} \|x-c\|^2 = \sum_{x\in U} \|x-\mathrm{cen}(U)\|^2 + |U|\cdot\|c-\mathrm{cen}(U)\|^2,$$

其中

$$\mathrm{cen}(U) := \frac{\sum_{x\in U} x}{|U|},$$

称为集合 U 的质心点. 根据上述性质, k- 均值问题的目标也可以这样描述: 将观测集 $\mathcal{X}$ 划分为 k 个部分 $\{X_1, X_2, \ldots, X_k\}$, 使得下面的函数值达到最小:

$$\sum_{j=1}^{k}\sum_{x\in X_j} \|x-\mathrm{cen}(X_j)\|^2.$$

2.2 近似难度

经典的 k- 均值问题是 NP- 难问题, 即使在 d 为常数或者 k 为常数的情形下, 该问题依然是 NP- 难的. 表 1 列出了 k- 均值问题在不同情形下的近似难度, 其中 UGC 表示唯一博弈猜想 (unique games conjecture). Cohen-Addad 等 [21] 研究了 k- 均值问题运行时间的下界. 他们证明了在指数时间假设 (exponential-time hypothesis, ETH) 下, 对任意可计算函数 f, 不存在 $f(k)n^{o(k)}$ 时间算法求解 k- 均值问题.

2.3 Lloyd 算法的理论研究

求解 k- 均值问题最常用的算法是 Lloyd 算法 [5]. 由于该算法的广泛应用, 通常又被称为 k- 均值算法. 为了避免与 k- 均值问题的其他算法混淆, 本文采用 Lloyd 算法的称呼. 该算法是一种迭代优化算法, 交替进行分配 (将观测点分配到离其最近的中心点) 和更新 (计算新的聚类的中心点) 两个步骤, 直到算法收敛为止.

Lloyd 算法的优点在于算法简单, 实际应用中运行时间很快. Duda 等 [22] 指出, 实际应用中 Lloyd 算法的迭代数通常远比观测点数少. 但是从理论上来看, Lloyd 算法最坏时间复杂度是指数量级的, 算法输出解的质量可以任意差. 很多关于 Lloyd 算法的理论研究集中在两方面:

(1) 从理论上解释 Lloyd 算法实际运行时间快的原因;

(2) 修改 Lloyd 算法使得输出解的质量有理论保证.

表 1 k- 均值问题的近似难度

文献	d	k	近似难度
Inaba 等 [13]	常数	常数	P 问题, $O(n^{dk+1})$ 时间内精确求解
Aloise 等 [14]、Dasgupta [15] 和 Drineas 等 [16]	任意	2	NP- 难
Mahajan 等 [17]	2	任意	NP- 难
Awasthi 等 [18]	$\Omega(\log n)$	任意	APX- 难
Lee 等 [19]	任意	任意	假设 $P \neq$ NP, 近似比下界 1.0013
Cohen-Addad 和 Karthik [20]	任意	任意	假设 UGC, 近似比下界 1.07

2.3.1 Lloyd 算法的时间复杂度

Lloyd 算法的最坏时间复杂度已经研究得比较清楚. Dasgupta [23] 考虑 $d=1$ 的情形, 证明了当 $k=2$ 时存在实例使得 Lloyd 算法需要运行 $\Omega(n)$ 步; 当 $k<5$ 时, Lloyd 算法最多 $O(n)$ 步终止. Har-Peled 和 Sadri [24] 仍然考虑 $d=1$ 的情形, 证明了对于任意 k, Lloyd 算法最多 $O(n\Delta^2)$ 步终止, 其中 Δ 是观测集中两点之间最大距离与最小距离的比值. Arthur 和 Vassilvitskii [25] 构造例子说明 Lloyd 算法最坏时间复杂度是 $2^{\Omega(\sqrt{n})}$, 进一步说明即使初始中心点在观测点集中随机选取, Lloyd 算法的运行时间高概率是超多项式的 (superpolynomial); 作为对实际数值表现的初步解释, 他们研究了数据点从 $\Omega(n/\log n)$ 维标准正态分布独立选取的情形, 这时 Lloyd 算法以高概率多项式时间终止. Vattani [26] 构造 2 维实例说明 Lloyd 算法最坏时间复杂度是 $2^{\Omega(n)}$.

为了缩小 Lloyd 算法实际表现和最坏时间复杂度理论分析的间隙, 人们开始研究 Lloyd 算法的平滑时间复杂度. Arthur 和 Vassilvitskii [27] 给出了 Lloyd 算法的第一个平滑分析. 他们证明了下面的结论: 如果任意观测集中的每个点独立地被均值为 0、标准差为 σ 的正态分布扰动, 那么对扰动后的观测集应用 Lloyd 算法得到的平均运行时间是关于 n^k、d 和 D/σ 的多项式, 其中 D 是扰动后的观测集的直径. Manthey 和 Röglin [28] 改进了上面的分析, 得到了两个平均运行时间的估计: 第一个是关于 $n^{\sqrt{k}}$ 和 $1/\sigma$ 的多项式, 第二个是 $k^{kd}\cdot\mathrm{poly}(n,1/\sigma)$. Arthur 等 [29] 首次给出了 Lloyd 算法的多项式平滑时间复杂度, 得到的平均运行时间是关于 n、k、d 和 $1/\sigma$ 的多项式 $O(k^{34}d^8\sigma^{-6}n^{34}\log^4 n)$.

2.3.2 Lloyd 算法的初始化方法

在 Lloyd 算法中, 初始的 k 个聚类中心是从观测点中任意 (或者随机) 选取的, 输出结果的好坏依赖于初始解的选取. 人们开始研究如何修改初始解, 在此基础上再运行 Lloyd 算法. 这类为 Lloyd 算法选取初始解的方法被称为 Lloyd 算法的初始化方法. 本文重点介绍随机初始化方法, 特别是 k- 均值++ 及其变形 (参见表 2).

Ostrovsky 等 [30,31] 引入了观测集 δ- 分离的概念, 在 δ 充分小时给出了具有常数近似比的随机初

表 2 基于 Lloyd 的初始化算法

文献	限制条件	近似比	运行时间
Ostrovsky 等 [30,31]	$\mathcal{X}$ 是 δ- 分离的, δ 充分小	$(1-\delta^2)/(1-37\delta^2)$	$O(ndk+dk^3)$
Arthur 和 Vassilvitskii [32]	–	$8(\log k+2)$	$O(ndk)$
Aggarwal 等 [33]	–	$(\lceil 16(1+1/\sqrt{k})\rceil, 20)$	$O(ndk)$
	–	$(O(1/\epsilon\cdot\log(1/\epsilon)), 4+\epsilon)$	$O(ndk/\epsilon\cdot\log(1/\epsilon))$
Ailon 等 [34]	–	$(3\log k, 64)$	$O(ndk\log k)$
Wei [35]	–	$(\beta, 8(1+1.618/(\beta-1)))$	$O(\beta ndk)$
Jaiswal 等 [36]	k 为常数	$1+\epsilon$	$\tilde{O}(nd\cdot 2^{\tilde{O}(k^2/\epsilon)})$
Jaiswal 等 [37]	k 为常数	$1+\epsilon$	$\tilde{O}(nd\cdot 2^{\tilde{O}(k/\epsilon)})$
Lattanzi 和 Sohler [10]	–	$O(1)$	$O(ndk^2\log\log k)$
Choo 等 [38]	–	$O(1)$	$O(ndk\log k)$
Bachem 等 [39]	观测集满足假设	$O(\log k)$	$O(k^3 d\log^2 n\log k)$
Bachem 等 [40]	–	$8(\log_2 k+2)$, 附加近似误差: $\epsilon\mathrm{Var}(\mathcal{X})$	$O(nd+(1/\epsilon)k^2 d\log(k/\epsilon))$
	观测集满足较弱假设	$8(\log_2 k+3)$	$O(nd+k^3 d\log k)$

始化方法. Arthur 和 Vassilvitskii [32] 独立地提出了另外一种随机初始化方法—k- 均值++. 在 k- 均值++ 中, 迭代 k 轮选取 k 个中心点, 每轮按照概率选一个点作为新的中心点, 选点的概率与该点产生的费用 (该点到当前最近中心点的距离的平方) 成正比. 上述选点方法称为 D^2 抽样, 也称为自适应抽样.

Ailon 等 [34] 提出了双准则近似的初始化方法: k- 均值 $\sharp$, 并设计了 k- 均值问题的流算法. k- 均值 $\sharp$ 借鉴了 k- 均值++ 的思想, 仍然进行 k 轮迭代, 但是每轮中同时独立选取 $3\log k$ 个中心点. Ackermann 等 [41] 根据 k- 均值++ 设计了另外的流算法. Aggarwal 等 [33] 独立地提出了类似于 k- 均值 $\sharp$ 的双准则近似初始化方法. D^2 抽样的思想也启发了 Jaiswal 等 [36] 和 Jaiswal 等 [37] 在 k 为常数时给出了 PTAS. Wei [35] 将 k- 均值++ 的迭代次数从 k 增大到 βk (这里 $\beta > 1$), 得到改进的双准则算法.

Bachem 等 [39] 采用 Markov 链 Monte Carlo 抽样技术近似 k- 均值++ 从而加速算法, 时间复杂度关于 n 是次线性的, 该算法记为 K-MC^2. Bachem 等 [40] 修改了建议分布 (proposal distribution) 得到 AFK-MC^2, 其中 AF (assumption-free), 表示不需要对观测集做任何假设就可以得到带附加近似误差的理论估计. Lattanzi 和 Sohler [10] 指出运行 k- 均值++ 后, 再运行 $O(k\log\log k)$ 步局部搜索, 可以得到具有常数近似比的输出. Choo 等 [38] 改进了上面的分析, 论证了后续的局部搜索只需要运行 ϵk 步.

k- 均值++ 是 k 轮的串行算法, 每轮都要扫描整个观测集来计算最短距离, 因而处理大数据时产生困难. Bahmani 等 [42] 提出了并行的初始化方法: k- 均值‖. 记 ψ 为 $\mathcal{X}$ 的量化误差 (quantization error), 即在 $\mathcal{X}$ 中随机选一点为中心时对应的 1- 均值问题的聚类目标值. k- 均值‖ 由两个阶段组成: 阶段 I 包括 $O(\log\psi)$ 轮迭代, 通过引入过抽样因子 l, 每轮并行抽样 $O(l)$ 个点; 阶段 II 对抽样的 $O(l\log\psi)$ 个点利用 k- 均值++ 进行聚类得到最终的 k 个中心点. 他们证明了阶段 I 产生的解 (注意不是可行解) 对应的目标值可以用乘法/附加近似误差来估计, 阶段 II 产生的可行解近似比为 $O(\log k)$. Bachem 等 [43] (限定 $l \geqslant k$) 和 Rozhoň [44] (限定 $l = k$) 分别给出了改进的理论分析; 其改进主要针对阶段 I, 阶段 II 的近似比保持不变. 表 3 给出了上面 3 种分析的对比, 其中

$$\alpha = \exp(-(1-\mathrm{e}^{-l/(2k)})) \approx \mathrm{e}^{-l/(2k)}.$$

2.4 固定参数 d 或 k

参数 d 或 k 固定时, 学者们主要采用了以下 5 类技巧, 给出了一系列 PTAS 结果 (参见表 4).

(1) 降维 (dimension reduction). 由于观测点所在 Euclid 空间的维数可能非常高, 如何降维是聚类中的重要问题. Johnson 和 Lindenstrauss [58] 提出了著名的 Johnson-Lindenstrauss 引理: 任给 Euclid 空间 $\mathbb{R}^d$ 中的 n 个点, 可以映射到 $\mathbb{R}^t$, 距离偏差不超过 $1+\epsilon$, 这里 $t = O(\log n/\epsilon^2)$, 且该映射可以在 $O(nd\log n/\epsilon^2)$ 时间内构造出来. Frankl 和 Maehara [59] 给出了简化的证明. Linial 等 [60] 推广了上面的结果.

表 3 k- 均值‖ 阶段 I 指标

文献	轮数 t	乘法近似误差	附加近似误差
Bahmani 等 [42]	$O(\log\psi)$	$16/(1-\alpha)$	$((1+\alpha)/2)^t\psi$
Bachem 等 [43]	$O(\log(\mathrm{Var}(\mathcal{X})))$	26	$2(k/(\mathrm{e}l))^t\mathrm{Var}(\mathcal{X})$
Rozhoň [44]	$O(\log(\mathrm{Var}(\mathcal{X})/\mathrm{OPT})/\log\log(\mathrm{Var}(\mathcal{X})/\mathrm{OPT}))$	20	0

表 4 k- 均值问题的 PTAS

文献	d	k	运行时间
Matoušek [45]	常数	2	$O(n\log n\cdot\epsilon^{-2d}\log(1/\epsilon)+n\epsilon^{-(4d-2)}\log(1/\epsilon))$
	常数	常数	$O(n\log^k n\cdot\epsilon^{-2k^2d})$
Bădoiu 等 [46]	任意	常数	$O(2^{(k/\epsilon)^{O(1)}}\mathrm{poly}(d)n\log^k n)$
De La Vega 等 [47]	任意	常数	$O(2^{(k^3/\epsilon^8)(\log(k/\epsilon))\log k}dn\log^k n)$
Har-Peled 和 Mazumdar [48]	任意	常数	$O(n+k^{k+2}\epsilon^{-(2d+1)k}\log^{k+1}n\log^k(1/\epsilon))$
Har-Peled 和 Kushal [49]	任意	常数	$O(n+\mathrm{poly}(k,\log n,1/\epsilon)+k^3\epsilon^{-(d+1)}(k^3\epsilon^{-(2d+1)}\log(1/\epsilon))^{k+1})$
Kumar 等 [50, 51]	任意	常数	$O(2^{(k/\epsilon)^{O(1)}}dn)$
Chen [52]	任意	常数	$O(ndk+2^{(k/\epsilon)^{O(1)}}d^2n^\sigma)$, 注: $\sigma>0$ 是任一常数
Ostrovsky 等 [31]	任意	常数	$O(2^{O(k(1+\delta^2)/\epsilon)}dn)$, 注: 要求 $\mathcal{X}$ 是 δ- 分离的且 δ 充分小
Feldman 等 [53]	任意	常数	$O(ndk+d\cdot\mathrm{poly}(k/\epsilon)+2^{\tilde{O}(k/\epsilon)})$
Friggstad 等 [54]	常数	任意	$O((k/\epsilon)^{d^{O(d)}\cdot\epsilon^{-O(d/\epsilon)}}\log(\mathrm{cost}(S_0)/\mathrm{OPT}))$
Cohen-Addad 等 [55]	常数	任意	$O(n^{(1/\epsilon)^{O(d)}})$
Cohen-Addad [56]	常数	任意	$nk(\log n)^{(d/\epsilon)^{O(d)}}$
Cohen-Addad 等 [57]	常数	任意	$((1/\epsilon)^{1/\epsilon})^{2^{O(d^2)}}n\log^5 n+2^{O(d)}n\log^9 n$

(2) 核心集 (coreset). 许多学者研究利用核心集的概念设计 k- 均值问题的快速算法 [46]. 核心集是观测集的规模较小的子集 (可能每个点带权重), 在该子集上的聚类问题可以很好地近似在整个观测集上的聚类问题. 一般来讲, 核心集越小, 问题越容易被近似, 也意味着人们可以更有效地概括观测集. 最小核心集的基数是聚类问题的基本组合性质. 针对 k- 均值问题, Har-Peled 和 Mazumdar [48] 找到了基数为 $O(k\epsilon^{-d}\log n)$ 的核心集; Har-Peled 和 Kushal [49] 将核心集的基数改进为 $O(k^3\epsilon^{-(d+1)})$ (与 n 无关); Chen [52] 采用随机抽样技术构造核心集, 对于事先给定的参数 $\lambda\in(0,1)$, 找到了基数为

$$O\left(k\epsilon^{-2}\log n\left(kd\log\frac{1}{\epsilon}+k\log k+k\log\log n+\log\frac{1}{\lambda}\right)\right)$$

的核心集.

(3) 近似质心集 (approximate centroid set). Matoušek [45] 提出了近似质心集的概念: 如果聚类中心限定在集合 $\mathcal{C}$ 中选取, 相应的聚类费用不超过 $(1+\epsilon)$OPT, 则称 $\mathcal{C}$ 为 ϵ- 近似质心集. 他证明了 ϵ- 近似质心集 $\mathcal{C}$ 可以在 $O(n\log n+n\epsilon^{-d}\log(1/\epsilon))$ 时间内得到, 并且 $|\mathcal{C}|=O(n\epsilon^{-d}\log(1/\epsilon))$. 根据上述结论, 可以将原问题转化为离散型 k- 均值问题. De La Vega 等 [47] 利用枚举方式构造了近似质心集. Kumar 等 [50, 51] 采用随机抽样技术构造近似质心集. Feldman 等 [53] 结合核心集和近似质心集的特点引入了弱核心集.

(4) 局部搜索 (local search). 当 d 为常数时, 许多学者采用局部搜索算法, 分析时利用了全局解和局部解的随机划分得到 PTAS [54–56].

(5) 动态规划 (dynamic program). 当 d 为常数时, 目前最好的 PTAS 由 Cohen-Addad 等 [57] 利用随机层次分解和动态规划技术得到.

2.5 任意参数 d 和 k

Jain 和 Vazirani [61] 采用观测集本身作为近似质心集, 这时近似比损失为 2; 利用 Lagrange 松弛将 k- 均值问题转化为开设费用相同的设施选址问题, 再应用原始对偶算法得到 Lagrange 乘子保持

(Lagrangean multiplier preserving, LMP) 近似; 最后通过对 Lagrange 乘子的二分法得到两个解, 借助于两点舍入 (bi-point rounding) 技巧, 得到 k- 均值问题的第一个常数比近似算法.

结合第 2.4 小节的 Johnson-Lindenstrauss 引理和 Matoušek [45] 关于近似质心集的构造, 可以在多项式时间 $O(nd\log n/\epsilon^2+n^{O(1/\epsilon^2)\log(1/\epsilon)}\log(1/\epsilon))$ 中得到基数为多项式量级 $O(n^{O(1/\epsilon^2)\log(1/\epsilon)}\log(1/\epsilon))$ 的近似质心集, 近似比损失为 $1+\epsilon$. 限定在近似质心集中选取中心点, 得到离散型 k- 均值问题. 接下来的改进均是基于上述近似质心集. Kanungo 等 [12] 将 Arya 等 [62] 针对 k- 中位问题的局部搜索算法巧妙地应用到 k- 均值问题上, 得到 $(9+\epsilon)$- 近似算法. Ahmadian 等 [63] 沿用了 Jain 和 Vazirani [61] 的框架, 在两个地方进行了实质性改进: (1) 利用 k- 均值问题的几何结构得到 LMP 6.357- 近似; (2) 通过多项式次枚举 Lagrange 乘子, 设计对偶上升算法构造可行解, 保持了中心点个数的某种连续性. 最终得到的 $6.357+\epsilon$ 是目前为止最好的近似比. 在表 5 进行了总结, 其中

$$\tilde{n}=n^{O(1/\epsilon^2)\log(1/\epsilon)}\log\frac{1}{\epsilon}.$$

Cohen-Addad 等 [64] 综合利用核心集、枚举和次模优化技术, 得到了运行时间为 $f(k,\epsilon)n^{O(1)}$、近似比为 $1+8/\mathrm{e}+\epsilon$ 的 FPT (fixed-parameter tractability) 算法. 他们同时指出在 Gap-ETH (gap-exponential time hypothesis) 下, 上述结果是紧的, 即存在函数 $g:\mathbb{R}^+\to\mathbb{R}^+$ 使得任意 $(1+8/\mathrm{e}-\epsilon)$- 近似算法的运行时间至少为 $n^{k^{g(\epsilon)}}$.

2.6 双准则近似

除了在第 2.3.2 小节介绍的 Aggarwal 等 [33] 给出的双准则近似之外, 还有更多的双准则结果, 我们在此介绍具有代表性的 3 个结果. Makarychev 等 [65] 基于线性规划和局部搜索技巧, 提出了 $(\beta,\alpha(\beta))$ 双标准近似算法, 其中 $\alpha(\beta)$ 是关于 β 的单调减函数, 上界是 $9+\epsilon$, 其中 $\alpha(2)<2.59$, $\alpha(3)<1.4$, 我们可以把该结果理解为 Kanungo 等 [12] $(9+\epsilon)$- 近似算法的推广. Hsu 和 Telgarsky [66] 采用贪婪技巧给出了 k- 均值的 $(O(\log(1/\epsilon),1+\epsilon)$ 双标准近似算法, 运行时间为 $O(dk\log(1/\epsilon)n^{1+\lceil 1/\epsilon\rceil})$. 当维数固定时, Bandyapadhyay 和 Varadarajan [67] 利用局部搜索技巧得到 k- 均值问题的 $(1+\epsilon,1+\epsilon)$ 双标准近似算法.

2.7 度量空间 k- 均值问题

经典 k- 均值问题的 $(9+\epsilon)$- 和 $(6.357+\epsilon)$- 近似算法可以推广到一般度量空间 k- 均值问题的 $(25+\epsilon)$- 和 $(9+\epsilon)$- 近似算法. 度量空间的加倍维数 (doubling dimension) 是满足下面条件的最小 τ: 对任意半径为 $2r$ 的球都可以用不超过 2^τ 个半径为 r 的球覆盖. 加倍度量空间 (doubling metric) 是加倍维数为常数的度量空间 (参见文献 [68]). 下面介绍的 3 个结果均针对固定加倍维数 d 的度量空间 (包括固定维数的 Euclid 空间) k- 均值. Friggstad 等 [54] 证明了经典的局部搜索算法是运行时间为

表 5 k- 均值问题的近似算法

文献	研究技巧	近似比	运行时间
Jain 和 Vazirani [61]	Lagrange 松弛 + 原始对偶 + 双点舍入	108	$O((L+\log n)n^2\log n)$
Kanungo 等 [12]	局部搜索	$9+\epsilon$	$O(nd\log n/\epsilon^2+\tilde{n}^{O(1/\epsilon)}\log(\mathrm{cost}(S_0)/\mathrm{OPT})/\epsilon)$
Ahmadian 等 [63]	Lagrange 松弛 + 原始对偶 + 枚举舍入	$6.357+\epsilon$	$O(nd\log n/\epsilon^2+\tilde{n}^{O(\epsilon^{-5})})$

$O(n^{(d/\epsilon)^{O(d)}})$ 的 PTAS, Cohen-Addad 等 [57] 得到了运行时间为

$$((1/\epsilon)^{1/\epsilon})^{2^{O(d^2)}} n\log^5 n + 2^{O(d)} n\log^9 n$$

的 PTAS. 给定 k- 均值问题的实例, 如果距离延伸不超过 α 倍时, 该实例具有不变的唯一最优解, 称该实例是 α- 稳定的. 对于加倍度量空间 k- 均值的稳定实例, Friggstad 等 [9] 证明了多交换 (multi-swap) 的局部搜索算法可以多项式时间找到最优解.

3 k- 均值问题的重要变形

与许多经典的优化问题一样, k- 均值也有诸多相关的变形. 如果定义不同的距离或目标函数, 或选取不同的聚类中心, 就引出了与 k- 均值相关的各种各样的变形. 例如, 当距离的定义是一类 Bregman 散度函数时, 此问题称为 Bregman 散度 k- 均值 [69].

3.1 k- 中位问题

k- 中位 (median) 问题是与 k- 均值问题紧密联系的一个经典问题, 在该问题中, 所有观测点在一般的度量空间中, 距离满足三角不等式, 要从给定的设施集合中选取 k 个中心点, 使得每个观测点到最近的选取中心点的距离之和最小. Arya 等 [62] 给出了基于局部搜索技术的 $(3+\epsilon)$- 近似算法, 目前最好近似比为 $2.675+\epsilon$ [70]. 有序 k- 中位问题是 k- 中位问题和设施选址问题的推广. 在该问题中, 给定有限度量空间 (V, dist) 中的 n 个点、惩罚权重序列 $\lambda_1 \geqslant \cdots \geqslant \lambda_n \geqslant 0$ 和整数 k, 对于任意设施集合 $F \subset V$, 每个顶点 $v \in V$ 对应距离 $\mathrm{dist}(v, F)$, 根据这些距离顺序乘以相应的惩罚权重再求和得到该问题的费用函数. 对于有序 k- 中位问题, Aouad 和 Segev [71] 利用局部搜索技术, 得到了 $O(\log n)$- 近似, Byrka 等 [72] 利用线性规划舍入技术得到 $(38+\epsilon)$- 近似, Chakrabarty 和 Swamy [73] 利用原始对偶方法得到 $(18+\epsilon)$- 近似.

3.2 球面 k- 均值问题

真实世界的数据有相当一部分以自然语言文本的形式存在. 在社交网络数据中, 文本是最主要的载体. 以文本数据创建的向量空间模型具有两大特点: (1) 文档向量的维数非常高; (2) 文档向量非常稀疏. 当这些向量的方向比其模长更重要或者作用更大时, 我们可以假设文档向量被正规化 (具有单位模长), 从而它们可以被看成是高维单位球面上的点. 所以在文本分析中, 文本观测点间的相似性采用余弦相似度来度量更为合适, 在这种距离定义下的 k- 均值问题, 也称为球面 k- 均值问题 (spherical k-means problem) [74]. 与一般的 k- 均值问题不同的是, 球面 k- 均值问题要求聚类点必须在单位球上.

给定 n 个元素的观测集 $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ 和整数 k, 球面 k- 均值问题的目标是选取中心点集合 $\mathcal{C} = \{c_1, c_2, \ldots, c_k\} \subseteq \mathbb{R}^d$, 使得下面的函数达到最小:

$$\sum_{i=1}^{n} \min_{j\in\{1,2,\ldots,k\}} (1-\cos(x_i, c_j)).$$

注意到

$$1-\cos(x_i, c_j) = \frac{1}{2}\left\| \frac{x_i}{\|x_i\|} - \frac{c_j}{\|c_j\|} \right\|^2,$$

球面 k- 均值问题可以等价描述为: 给定 n 个元素的观测集 $\mathcal{X}=\{x_1,x_2,\dots,x_n\}\subseteq S^d$ 和整数 k, 其中 S^d 是 $\mathbb{R}^d$ 中的单位球面, 目标是选取中心点集合 $\mathcal{C}=\{c_1,c_2,\dots,c_k\}\subseteq S^d$, 使得下面的函数达到最小:

$$\frac{1}{2}\sum_{i=1}^{n}\min_{j\in\{1,2,\dots,k\}}\|x_i-c_j\|^2.$$

对于任意集合 $U\subseteq\mathcal{X}\subseteq S^d$ 和点 $c\in S^d$, 可以证明 (参见文献 [75])

$$\sum_{x\in U}\|x-c\|^2=\sum_{x\in U}\|x-\operatorname{scen}(U)\|^2+\left\|\sum_{x\in U}x\right\|\cdot\|c-\operatorname{scen}(U)\|^2,$$

其中

$$\operatorname{scen}(U):=\frac{\sum_{x\in U}x}{\|\sum_{x\in U}x\|},$$

称为集合 U 的球面质心点. 根据上述性质, 球面 k- 均值问题的目标也可以这样描述: 将观测集 $\mathcal{X}\subseteq S^d$ 划分为 k 个部分 $\{X_1,X_2,\dots,X_k\}$, 使得下面的函数达到最小:

$$\frac{1}{2}\sum_{j=1}^{k}\sum_{x\in X_j}\|x-\operatorname{scen}(X_j)\|^2.$$

球面 k- 均值问题有类似于 k- 均值问题的 Lloyd 算法和对应的初始化方法 (参见文献 [75]). Li 等 [76] 研究了该问题的初始化算法并给出理论分析. Zhang 等 [77] 利用局部搜索技术给出了球面 k- 均值问题的 $(2(4+\sqrt{7})+\epsilon)$- 近似.

3.3 鲁棒 k- 均值问题

当观测集中的数据出现缺失或失真时, 需要更稳定的聚类技术来处理这类问题, 鲁棒 (robust) k- 均值问题研究的就是对带噪声的观测点的聚类 [78]. Cai 等 [79] 在集成了大规模数据异质表示的基础上提出了一种鲁棒的大规模数据的多视角聚类方法. 为了极小化目标而对观测点作出取舍, 即并不将所有的观测点都进行聚类, 舍弃的观测点称其为异常点.

近期有大量的工作从不同角度、用不同方法研究了带异常点 k- 均值问题. Chawla 和 Gionis [80] 推广了 Lloyd 算法, 并将其应用到了带有异常点的 k- 均值问题中. Hautamäki 等 [81] 提出了异常点移除聚类算法, 该算法先将数据以 k- 均值方式聚类, 再移除远离聚类中心的那些点. Ott 等 [82] 将聚类和异常点检测用整数规划来描述, 通过 Lagrange 松弛, 利用次梯度方法求得异常点和划分. Rujeerapaiboon 等 [83] 将问题用混合的整数规划来描述, 利用半定规划和线性规划松弛求解, 通过设计一个确定的舍入规则得到该问题的可行解. Malkomes 等 [84] 对带有异常点的大数据聚类问题提出了一个快速的分布式算法. Ben-David 和 Haghtalab [85] 则是改造一些聚类算法, 使其具有鲁棒性. Deb 和 Dey [86] 将 k- 均值聚类方法和树状图相结合达到聚类和异常点检测的目的. Gan 和 Ng [87] 将所有异常点聚集为一簇, 进而 k- 均值问题可以理解为 $(k+1)$- 均值问题, 推广了 k- 均值算法. Gupta 等 [88] 在可以违反异常点数量限制的条件下, 基于局部搜索技术给出了一个双标准的 $O(1)$- 近似算法. Friggstad 等 [89] 利用局部搜索提出了双准则 PTAS: 聚类中心有 $k(1+\epsilon)$ 个, 针对 Euclid 空间和加倍度量空间近似比为 $1+\epsilon$, 针对一般度量空间近似比为 $25+\epsilon$. Krishnaswamy 等 [90] 给出了基于迭代线性规划舍入技术的 $(53.002+\epsilon)$- 近似算法, 这是该问题的第一个常数近似比算法. Krishnaswamy 等 [90] 的算法思想如下: 由于带异常点 k- 均值问题的自然线性规划松弛的整数间隙无界, 他们先把线性规划松弛的解舍入为

费用损失很少的几乎整数解, 在该解中至多有两个分数开设的中心; 由此可知, 线性规划整数间隙来自于几乎整数解和完全整数解的间隙. 采用预处理程序, 他们把几乎整数解转化为完全整数解, 仅损失了近似比中的常数因子; 进一步采用稀疏化技巧, 上述转化导致的额外损失可以减少到任意 $\epsilon > 0$.

带惩罚的 k- 均值问题则是给每个异常点设置惩罚费用, 通过在目标函数中增加惩罚项, 自动过滤掉异常点. 该方法由 Charikar 等 [91] 针对 k- 中位和设施选址问题引入. Tseng [92] 研究该类问题时需要假定每个点的惩罚值是相同的, 我们称其为一致惩罚的 k- 均值问题. 对于带惩罚的 k- 均值问题, Zhang 等 [93] 利用局部搜索技术给出了第一个常数近似比算法, 其近似比为 $25+\epsilon$. Feng 等 [94] 利用原始对偶技巧将上述近似比改进为 $19.849+\epsilon$. Alimi 等 [95] 研究了更一般的带惩罚的 k- 均值问题, 给出了 $O(\log^{1.5} n \log\log n)$- 近似算法. Li 等 [96] 利用初始化算法得到了一个关于 $\log k$ 和惩罚函数比值有关的近似算法. Ji 等 [97] 提出了带惩罚的球面 k- 均值问题, 并根据初始化算法设计了其近似算法.

3.4 带约束的 k- 均值问题

在各种应用背景下的实际问题可能会对 k- 均值聚类有不同的特殊要求, 例如, 为了避免得到的局部解中某些类中含的观测点过少, 或者不希望分类的总数过多, 往往会在 k- 均值算法中再加上各种各样的约束, 此类变形称为带约束的 k- 均值问题 [98].

Ding 和 Xu [99] 通过推广 Kumar 等 [51] 的方法 (从无约束 k- 均值到约束 k- 均值), 利用均匀采样和单纯形引理几何技巧, 给出了运行时间为 $O(nd\cdot(\log n)^k\cdot 2^{\mathrm{poly}(k/\epsilon)})$ 的算法, 产生规模为 $O((\log n)^k\cdot 2^{\mathrm{poly}(k/\epsilon)})$ 的 k- 元组候选集合, 其中一个 k- 元组为 $(1+\epsilon)$- 近似解. Bhattacharya 等 [100] 利用 D^2- 采样技巧, 给出了运行时间为 $O(knd\cdot(2123\mathrm{e}k/\epsilon^3)^{64k/\epsilon}\cdot 2^k)$ 的算法, 产生规模为 $O((2123\mathrm{e}k/\epsilon^3)^{64k/\epsilon}\cdot 2^k)$ 的 k- 元组候选集合, 其中一个 k- 元组为 $(1+\epsilon)$- 近似解. Feng 等 [101] 给出了运行时间为 $O(nd\cdot(1891\mathrm{e}k/\epsilon^2)^{8k/\epsilon})$ 的算法, 产生规模为 $O(n(1891\mathrm{e}k/\epsilon^2)^{8k/\epsilon})$ 的 k- 元组候选集合, 其中一个 k- 元组为 $(1+\epsilon)$- 近似解. 针对带约束的 2- 均值问题, Feng 和 Fu [102] 给出了运行时间为 $O(dn+d(1/\epsilon)^{O(1/\epsilon)}\log n)$ 的算法, 产生规模为 $O((1/\epsilon)^{O(1/\epsilon)}\log n)$ 的 2- 元组候选集合, 其中一个 2- 元组为 $(1+\epsilon)$- 近似解; 利用该技巧可以将现有针对带约束的 k- 均值问题的运行时间为 $C(k,n,d,\epsilon)$ 的 PTAS 转化为运行时间为 $C(k,n,d,\epsilon)/k^{\Omega(1/\epsilon)}$ 的 PTAS.

与带容量约束的 k- 均值问题相关的带容量约束的 k- 中位问题和 k- 设施选址问题已经有若干结果. Arya 等 [62] 给出了带相同容量约束的 k- 中位问题的局部搜索双准则近似算法; Han 等 [103] 给出了带相同容量约束的 k- 设施选址问题的局部搜索双准则近似算法; Adamczyk 等 [104] 研究了带容量约束的 k- 中位问题, 给出了运行时间为 $2^{O(k\log k)}n^{O(1)}$、近似比为 $7+\epsilon$ 的 FPT 算法. 由此引发了下面 3 篇关于带容量约束的 k- 均值问题的研究结果. Xu 等 [105] 给出了运行时间为 $2^{O(k\log k)}n^{O(1)}$、近似比为 $69+\epsilon$ 的 FPT 算法. 结合 Kumar 等 [51] 的技巧, Cohen-Addad 和 Li [106] 给出了运行时间为 $(k/\epsilon)^{k(1/\epsilon)^{O(1)}}n^{O(1)}$ (与 d 无关) 的 PTAS. Cohen-Addad [8] 给出了运行时间为 $n^{((2/\epsilon)^2\log n)^{O(d)}}$ (与 k 无关) 的 PTAS.

带下界约束的 k- 均值问题尚没有近似算法, 我们简要介绍与之相关的带下界约束的设施选址 (lower bounded facility location) 问题的若干结果. 对于下界一致的情形, Svitkina [107] 通过归约到容量约束的设施选址问题, 给出了一致下界约束的设施选址问题的第一个常数近似比为 488 的算法; Ahmadian 和 Swamy [108] 通过归约到有更特殊结构的容量约束设施选址问题, 将上面的近似比改进到 82.6. 对于一般情形 (下界不要求一致), Li [109] 利用更复杂的归约, 给出了非一致下界约束的设施选址

问题的第一个常数近似比为 4,000 的算法.

3.5 隐私保护 k- 均值问题

由于用户隐私保护 (privacy preserving) 的意识和需求日益增长, 人们提出了隐私保护 k- 均值问题. 差分隐私 (differential privacy) 技术可以处理该类问题, 模型分为两类: 中心模型 (centralized model) 和本地模型 (local model). 中心模型和本地模型也分别称为非交互式模型和交互式模型. 在中心模型中, 假设有一个信托机构 (trusted curator) 可以收集所有用户信息并进行分析, 分析结果隐藏了任一单个用户的信息 (但是这些信息对于信托机构是公开的), 从而保护隐私. 在本地模型中, 有 n 个用户和一个服务器, 每个用户 i 的数据 $x_i \in \mathbb{R}^d$ 是私有信息. 用户不会将真实数据发送给服务器, 而是将自己本地的数据随机处理后发送有噪声的数据给服务器. 服务器汇总所有带噪声的数据, 计算相应的 k- 均值目标函数. 带噪声的数据可以保护隐私, 同时噪声对于数据的整体分布几乎没有影响. 称用户输入数据 $S = (x_1, \ldots, x_n)$ 为分布式数据库, 这些数据不是存储在一个位置, 每个 x_i 为用户 i 本地所有. 在实际应用中, 大公司通常采用本地模型来保护用户隐私, 这时用户隐私数据不能被公司服务器清晰地采集到.

在隐私保护 k- 均值问题的近似算法分析中需要引进刻画输入数据半径的误差项. 假设所有用户的数据都来自于 d 维单位球. 目前隐私保护 k- 均值中心模型的最好近似算法由 Kaplan 和 Stemmer[110] 给出, 算法得到的解不超过 $O(1) \cdot \mathrm{OPT} + \mathrm{poly}(\log(n), k, d)$. 隐私保护 k- 均值本地模型的最好近似算法由 Stemmer[11] 给出, 通过 $O(1)$ 轮用户与服务器之间的交互 (interaction), 算法得到的解不超过

$$O(1) \cdot \mathrm{OPT} + \tilde{O}(n^{1/2+a} \cdot k \cdot \max\{\sqrt{d}, \sqrt{k}\}),$$

其中 $a > 0$ 是任意小的常数.

3.6 泛函 k- 均值问题

随着数据采集技术的发展, 从气象学、医学、经济学、金融学、化学计量学和生物学等不同领域获取的数据可能都是函数性数据, 即动态数据. 例如, 特定时间段内某一区域的温度就是一种函数性数据. 当 "观测点" 是函数性数据时, 泛函 k- 均值问题 (functional k-means problem) 便成了重点研究对象, 这是一种非参数聚类方法. Meng 等[111] 结合函数的特点, 为了更深程度地度量函数样本之间的相似性, 将梯度信息引入 "距离" 中, 得到类似于一般 k- 均值问题的质心引理, 并将 Lloyd 算法成功应用到该问题中. Li 等[112] 将初始化算法应用到该问题中, 得到 $O(\log k)$- 近似算法. 更多关于泛函 k- 均值问题的研究可参见文献 [113, 114].

3.7 软聚类 k- 均值问题

以上介绍的聚类分析都是一种硬性划分, 每个观测点都被严格地划分到某一个聚类中, 即任意两个聚类的交集是空集, 具有非此即彼的特点. 然而在实际问题中, 大多数研究对象并不具有这样严格的界限, 即它们可以同时属于多个聚类, 具有亦此亦彼的特点, 这是一种软性聚类. 例如, 要在城市中选择一些位置建立几座超市用于服务周边市民, 市民到超市的最短距离是硬聚类问题考察的主要因素, 实际情况是每个市民不一定只到最近超市购物, 而很大可能是根据所选物品的性价比去多家超市. 因此实际中的实例更适合用软性划分. 模糊 C- 均值问题 (fuzzy C-means problem) 就是基于这种理念提出的 (参见文献 [115]). 该问题中聚类的定义 (界限) 是模糊的, 每个观测点到每个簇都存在一个隶属

度 (在 $[0,1]$ 区间里面取值), 要求每个观测点到所有簇的隶属度的和为 1. 在 k- 均值问题中, 簇是确定的并以质心为中心, 显然, 每个观测点到每个簇的隶属度是 0 或 1.

给定 n 个元素的观测集 $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}\subseteq\mathbb{R}^d$、正整数 k 和 m, 模糊 C- 均值问题的目标是选取中心点集合 $\mathcal{C}=\{c_1,c_2,\ldots,c_k\}\subseteq\mathbb{R}^d$, $\mu_{ij}\in[0,1]$ $(i=1,2,\ldots,n;\ j=1,2,\ldots,k)$ 满足

$$\sum_{j=1}^{k}\mu_{ij}=1,\quad i=1,2,\ldots,n,$$

使得下面的函数值达到最小:

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^{m}\|x_i-c_j\|^2,$$

其中 m 称为模糊参数. 容易看出, 当 $m=1$ 时, 最优解一定是对观测集的划分, 对于任意 i, 都有 $\{\mu_{ij}\}_{j=1,\ldots,k}$ 中恰好一个取 1, 此时的模糊 C- 均值问题退化为 k- 均值问题. 对于模糊 C- 均值问题的可行解 $(\mathcal{C},\{\mu_{ij}\})$, 可以证明下面的性质:

(1) 给定 $\mathcal{C}=\{c_1,c_2,\ldots,c_k\}$, 使得目标函数达到最小的 $\{\mu_{ij}\}$ 为

$$\mu_{ij}=\frac{1}{\sum_{l=1}^{k}\left(\frac{\|x_i-c_j\|}{\|x_i-c_l\|}\right)^{\frac{2}{m-1}}},\quad i=1,2,\ldots,n,\quad j=1,2,\ldots,k;$$

(2) 给定 $\{\mu_{ij}\}$, 使得目标函数达到最小的 $\mathcal{C}=\{c_1,c_2,\ldots,c_k\}$ 为

$$c_j=\frac{\sum_{i=1}^{n}\mu_{ij}^{m}x_i}{\sum_{i=1}^{n}\mu_{ij}^{m}},\quad j=1,2,\ldots,k.$$

Bezdek 等 [116] 最早将 Lloyd 算法用来求解模糊 C- 均值问题. Stetco 等 [117] 将 k- 均值++ 算法应用到模糊 C- 均值问题中, 并给出有效的数值运算. 更多关于模糊 C- 均值问题的研究和应用可以参见文献 [118, 119].

3.8 其他变形

k- 均值问题的其他变形还有很多, 如数据流模型下的聚类 [120–124]、分布式 k- 均值问题 [125, 126] 和在线 k- 均值问题 [127–129]. 由于篇幅所限, 本文不再赘述.

4 讨论

根据第 2 和 3 节的研究现状, 总结出下面一些值得进一步开展深入研究的问题.

(1) 与设施选址问题的近似间隙 $[1.463, 1.488]$ 相比, 经典 k- 均值问题的近似间隙 $[1.07, 6.357]$ 比较大, 还有很大的改进空间, 说明我们对于该问题的理解还不够深刻. 一个可能的改进上界的思路是, 巧妙结合 Lagrange 松弛和对偶拟合技术, 设计连续化的对偶拟合算法进行舍入 (参见文献 [63]).

(2) 关于度量空间 k- 均值问题, 最近的研究成果集中在两个方向: 一是加倍度量空间的 PTAS, 二是稳定实例的快速算法. 对于一般度量空间 k- 均值问题的近似算法研究非常少, 可以考虑借鉴 k- 中位问题的研究技巧, 考虑无损的伪近似方法 (参见文献 [130]).

(3) 关于球面 k- 均值问题, 针对该问题设计的近似算法尚不多见, 寻找更好的近似球面质心集是迫切需要研究的问题. 如果能够得到任意近似的球面质心集, 那么结合文献 [77] 中的分析马上可以得到改进的近似算法.

(4) 关于鲁棒 k- 均值问题, 带惩罚和带异常点的 k- 均值是两种最重要的变形, 虽然都有常数近似比, 但是结果都非常少, 近似间隙仍然很大. 特别是目前针对带异常点的 k- 均值问题的最好近似算法利用了迭代线性规划舍入技巧 (参见文献 [90]), 算法实现起来比较复杂, 是否能得到简单的近似算法?

(5) 关于隐私保护 k- 均值问题, 中心模型研究结果比较多, 对于本地模型研究结果比较少, 其他聚类问题也存在隐私保护的需求. 更多复杂环境下的隐私保护 k- 均值问题值得深入研究.

(6) 关于带约束的 k- 均值问题, 当 k 固定时, 有若干 PTAS 研究成果. 对于带容量约束的 k- 均值问题, 目前给出的结果多数为 FPT 算法, 可以考虑研究 k 和 d 任意的双准则算法. 下界约束的 k- 均值问题, 目前尚无任何结果.

(7) 针对大数据环境下的 k- 均值问题, 深入研究相应的流算法、并行算法和分布式算法等.

参考文献

1 Jain A K, Murty M N, Flynn P J. Data clustering: A review. ACM Comput Surv, 1999, 31: 264–323
2 徐大川, 许宜诚, 张冬梅. k- 平均问题及其变形的算法综述. 运筹学学报, 2017, 21: 101–109
3 徐大川, 许宜诚, 张冬梅. k- 均值算法的初始化方法综述. 运筹学学报, 2018, 22: 31–40
4 Ball G H, Hall D J. ISODATA, a novel method of data analysis and pattern classification. Technical Reports NTIS AD 699616. Stanford: Stanford Research Institute, 1965
5 Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory, 1982, 28: 129–137
6 MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. California: University of California Press, 1967, 281–297
7 Steinhaus H. Sur la division des corp materiels en parties. Bull Acad Polon Sci, 1956, 4: 801–804
8 Cohen-Addad V. Approximation schemes for capacitated clustering in doubling metrics. In: Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2020, 2241–2259
9 Friggstad Z, Khodamoradi K, Salavatipour M R. Exact algorithms and lower bounds for stable instances of Euclidean k-means. In: Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2019, 2958–2972
10 Lattanzi S, Sohler C. A better k-means++ algorithm via local search. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach: Proceedings of Machine Learning Research, 2019, 3662–3671
11 Stemmer U. Locally private k-means clustering. In: Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2020, 548–559
12 Kanungo T, Mount D M, Netanyahu N S, et al. A local search approximation algorithm for k-means clustering. In: Proceedings of the 18th Annual ACM Symposium on Computational Geometry, Barcelona. New York: Association for Computing Machinery, 2002, 10–18
13 Inaba M, Katoh N, Imai H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In: Proceedings of the Tenth Annual Symposium on Computational Geometry. New York: Association for Computing Machinery, 1994, 332–339
14 Aloise D, Deshpande A, Hansen P, et al. NP-hardness of Euclidean sum-of-squares clustering. Mach Learn, 2009, 75: 245–248
15 Dasgupta S. The hardness of k-means clustering. Technical Report CS2008–0916. San Diego: University of California, 2008
16 Drineas P, Frieze A, Kannan R, et al. Clustering large graphs via the singular value decomposition. Mach Learn, 2004, 56: 9–33
17 Mahajan M, Nimbhorkar P, Varadarajan K. The planar k-means problem is NP-hard. In: Proceedings of the International Workshop on Algorithms and Computation. Berlin-Heidelberg: Springer, 2009, 274–285
18 Awasthi P, Charikar M, Krishnaswamy R, et al. The hardness of approximation of Euclidean k-means. In: Proceedings of the 31st International Symposium on Computational Geometry. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2015, 754–767

19 Lee E, Schmidt M, Wright J. Improved and simplified inapproximability for k-means. Inform Process Lett, 2017, 120: 40–43

20 Cohen-Addad V, Karthik C S. Inapproximability of clustering in L_p metrics. In: Proceedings of the 60th Annual IEEE Symposium on Foundations of Computer Science. Washington: IEEE Computer Society, 2019, 519–539

21 Cohen-Addad V, De Mesmay A, Rotenberg E, et al. The bane of low-dimensionality clustering. In: Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms. New Orleans: SIAM, 2018, 441–456

22 Duda R O, Hart P E, Stork D G. Pattern Classification. New York: John Wiley & Sons, 2012

23 Dasgupta S. How fast is k-means? In: Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop. Berlin: Springer, 2003, 735–735

24 Har-Peled S, Sadri B. How fast is the k-means method? Algorithmica, 2005, 41: 185–202

25 Arthur D, Vassilvitskii S. How slow is the k-means method? In: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry. New York: Association for Computing Machinery, 2006, 144–153

26 Vattani A. k-means requires exponentially many iterations even in the plane. Discrete Comput Geom, 2011, 45: 596–616

27 Arthur D, Vassilvitskii S. Worst-case and smoothed analysis of the ICP algorithm, with an application to the k-means method. SIAM J Comput, 2009, 39: 766–782

28 Manthey B, Röglin H. Improved smoothed analysis of the k-means method. In: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2009, 461–470

29 Arthur D, Manthey B, Röglin H. Smoothed analysis of the k-means method. J ACM, 2011, 58: 1–31

30 Ostrovsky R, Rabani Y, Schulman L J, et al. The effectiveness of Lloyd-type methods for the k-means problem. In: Proceedings of the 7th Annual IEEE Symposium on Foundations of Computer Science. Berkeley: IEEE, 2006, 165–174

31 Ostrovsky R, Rabani Y, Schulman L J, et al. The effectiveness of Lloyd-type methods for the k-means problem. J ACM, 2013, 59: 1–22

32 Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2007, 1027–1035

33 Aggarwal A, Deshpande A, Kannan R. Adaptive sampling for k-means clustering. In: Proceedings of the 12th International Workshop on Approximation Algorithms for Combinatorial Optimization and 13th International Workshop on Randomization and Approximation Techniques in Computer Science. Berlin: Springer, 2009, 15–28

34 Ailon N, Jaiswal R, Monteleoni C. Streaming k-means approximation. In: Proceedings of the Neural Information Processing Systems. Vancouver: Curran Associates, 2009, 10–18

35 Wei D. A constant-factor bi-criteria approximation guarantee for k-means++. In: Proceedings of the Neural Information Processing Systems. Vancouver: Curran Associates, 2016, 604–612

36 Jaiswal R, Kumar A, Sen S. A simple D^2-sampling based PTAS for k-means and other clustering problems. Algorithmica, 2014, 70: 22–46

37 Jaiswal R, Kumar M, Yadav P. Improved analysis of D^2-sampling based PTAS for k-means and other clustering problems. Inform Process Lett, 2015, 115: 100–103

38 Choo D, Grunau C, Portmann J, et al. K-means++: Few more steps yield constant approximation. arXiv:2002.07784, 2020

39 Bachem O, Lucic M, Hassani S H, et al. Approximate k-means++ in sublinear time. In: Proceedings of the Association for the Advancement of Artificial Intelligence. Toronto: AAAI Press, 2016, 1459–1467

40 Bachem O, Lucic M, Hassani S H, et al. Fast and provably good seedings for k-means. In: Proceedings of the Neural Information Processing Systems. Barcelona: dblp.org, 2016, 55–63

41 Ackermann M R, Märtens M, Raupach C, et al. StreamKM++: A clustering algorithm for data streams. J Exp Algorithmics, 2012, 17: 2.1–2.30

42 Bahmani B, Moseley B, Vattani A, et al. Scalable k-means++. In: Proceedings of the 38th International Conference on Very Large Data Bases. New York: Association for Computing Machinery, 2012, 622–633

43 Bachem O, Lucic M, Krause A. Distributed and provably good seedings for k-means in constant rounds. In: Proceedings of the 34th International Conference on Machine Learning. Sydney: Proceedings of Machine Learning Research, 2017, 292–300

44 Rozhoň V. Simple and sharp analysis of k-means‖. arXiv:2003.02518, 2020

45 Matoušek J. On approximate geometric k-clustering. Discrete Comput Geom, 2000, 24: 61–84

46 Bădoiu, M, Har-Peled S, Indyk P. Approximate clustering via core-sets. In: Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2002, 250–257

47 De La Vega W F, Karpinski M, Kenyon C, et al. Approximation schemes for clustering problems. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2003, 50–58

48 Har-Peled S, Mazumdar S. On coresets for k-means and k-median clustering. In: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2004, 291–300

49 Har-Peled S, Kushal A. Smaller coresets for k-median and k-means clustering. In: Proceedings of the Twenty-First Annual Symposium on Computational Geometry. New York: Association for Computing Machinery, 2005, 126–134

50 Kumar A, Sabharwal Y, Sen S. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In: Proceedings of the Annual IEEE Symposium on Foundations of Computer Science. Washington: IEEE Computer Society, 2004, 454–462

51 Kumar A, Sabharwal Y, Sen S. Linear-time approximation schemes for clustering problems in any dimensions. J ACM, 2010, 57: 1–32

52 Chen K. On k-median clustering in high dimensions. In: Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms. New York: Association for Computing Machinery, 2006, 1177–1185

53 Feldman D, Monemizadeh M, Sohler C. A PTAS for k-means clustering based on weak coresets. In: Proceedings of the Twenty-Third Annual Symposium on Computational Geometry. New York: Association for Computing Machinery, 2007, 11–18

54 Friggstad Z, Rezapour M, Salavatipour M R. Local search yields a PTAS for k-means in doubling metrics. SIAM J Comput, 2019, 48: 452–480

55 Cohen-Addad V, Klein P N, Mathieu C. Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. SIAM J Comput, 2019, 48: 644–667

56 Cohen-Addad V. A fast approximation scheme for low-dimensional k-means. In: Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2018, 430–440

57 Cohen-Addad V, Feldmann A E, Saulpic D. Near-linear time approximation schemes for clustering in doubling metrics. In: Proceedings of the 60th Annual IEEE Symposium on Foundations of Computer Science. Baltimore: IEEE, 2019, 540–559

58 Johnson W B, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. Contemp Math, 1984, 26: 189–206

59 Frankl P, Maehara H. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. J Combin Theory Ser B, 1988, 44: 355–362

60 Linial N, London E, Rabinovich Y. The geometry of graphs and some of its algorithmic applications. Combinatorica, 1995, 15: 215–245

61 Jain K, Vazirani V V. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. J ACM, 2001, 48: 274–296

62 Arya V, Garg N, Khandekar R, et al. Local search heuristics for k-median and facility location problems. SIAM J Comput, 2004, 33: 544–562

63 Ahmadian S, Norouzi-Fard A, Svensson O, et al. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. In: Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science. Washington: IEEE Computer Society, 2017, 61–72

64 Cohen-Addad V, Gupta A, Kumar A, et al. Tight FPT approximations for k-median and k-means. In: Proceedings of ICALP, 2019, 42:1–42:14

65 Makarychev K, Makarychev Y, Sviridenko M, et al. A bi-criteria approximation algorithm for k-means. In: Proceedings of the Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Berlin: Springe, 2016, 14:1–14:20

66 Hsu D, Telgarsky M. Greedy bi-criteria approximations for k-medians and k-means. arXiv:1607.06203, 2016

67 Bandyapadhyay S, Varadarajan K. On variants of k-means clustering. In: Proceedings of the 32nd International Symposium on Computational Geometry. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2016, 1–15

68 Talwar K. Bypassing the embedding: Algorithms for low dimensional metrics. In: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2004, 281–290

69 Banerjee A, Merugu S, Dhillon I S, et al. Clustering with Bregman divergences. J Mach Learn Res, 2005, 6: 1705–1749

70 Byrka J, Pensyl T, Rybicki B, et al. An improved approximation for k-median, and positive correlation in budgeted optimization. ACM Trans Algorithms, 2017, 13: 1–31

71 Aouad A, Segev D. The ordered k-median problem: Surrogate models and approximation algorithms. Math Program, 2019, 177: 55–83

72 Byrka J, Sornat K, Spoerhase J. Constant-factor approximation for ordered k-median. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. New York: Association for Computing Machinery, 2018, 620–631

73 Chakrabarty D, Swamy C. Interpolating between k-median and k-center: Approximation algorithms for ordered k-median. In: Proceedings of the 45th International Colloquium on Automata, Languages, and Programming. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2018, 29:1–29:14

74 Buchta C, Kober M, Feinerer I, et al. Spherical k-means clustering. J Statist Softw, 2012, 50: 1–22

75 Endo Y, Miyamoto S. Spherical k-means++ clustering. In: Proceedings of the 12th International Conference on Modeling Decisions for Artificial Intelligence. Berlin: Springer, 2015, 103–114

76 Li M, Xu D, Zhang D, et al. The seeding algorithms for spherical k-means clustering. J Global Optim, 2020, 76: 695–708

77 Zhang D, Cheng Y, Li M, et al. Local search approximation algorithms for the spherical k-means problem. In: Proceedings of Algorithmic Aspects in Information and Management. Cham: Springer, 2019, 341–351

78 Georgogiannis A. Robust k-means: A theoretical revisit. In: Proceedings of the Neural Information Processing Systems. Barcelona: dblp.org, 2016, 2883–2891

79 Cai X, Nie F, Huang H. Multi-view k-means clustering on big data. In: Proceedings of the International Joint Conference on Artificial Intelligence. Toronto: AAAI Press, 2013, 2598–2604

80 Chawla S, Gionis A. K-means--: A unified approach to clustering and outlier detection. In: Proceedings of the 13th SIAM International Conference on Data Mining. Philadelphia: SIAM, 2013, 189–197

81 Hautamäki V, Cherednichenko S, Kärkkäinen I, et al. Improving k-means by outlier removal. In: Proceedings of the 14th Scandinavian Conference on Image Analysis. Berlin: Springer, 2005, 978–987

82 Ott L, Pang L, Ramos F T, et al. On integrated clustering and outlier detection. In: Proceedings of the Neural Information Processing Systems. Montreal: dblp.org, 2014, 1359–1367

83 Rujeerapaiboon N, Schindler K, Kuhn D, et al. Size matters: Cardinality-constrained clustering and outlier detection via conic optimization. SIAM J Optim, 2019, 29: 1211–1239

84 Malkomes G, Kusner M J, Chen W, et al. Fast distributed k-center clustering with outliers on massive data. In: Proceedings of the Neural Information Processing Systems. Montreal: dblp.org, 2015, 1063–1071

85 Ben-David S, Haghtalab N. Clustering in the presence of background noise. In: Proceedings of the 30th International Conference on Machine Learning. Beijing: JMLR, 2014, 280–288

86 Deb A B, Dey L. Outlier detection and removal algorithm in k-means and hierarchical clustering. World J Comput Appl Technol, 2017, 5: 24–29

87 Gan G, Ng M K P. k-means clustering with outlier removal. Pattern Recognition Lett, 2017, 90: 8–14

88 Gupta S, Kumar R, Lu K, et al. Local search methods for k-means with outliers. In: Proceedings of the 43rd International Conference on Very Large Databases. Munich: CEUR-WS.org, 2017, 757–768

89 Friggstad Z, Khodamoradi K, Rezapour M, et al. Approximation schemes for clustering with outliers. ACM Trans Algorithms, 2019, 15: 1–26

90 Krishnaswamy R, Li S, Sandeep S. Constant approximation for k-median and k-means with outliers via iterative rounding. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. New York: Association for Computing Machinery, 2018, 646–659

91 Charikar M, Khuller S, Mount D M, et al. Algorithms for facility location problems with outliers. In: Proceedings of the 12th Annual Symposium on Discrete Algorithms. New York: Association for Computing Machinery, 2001, 642–651

92 Tseng G C. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. Bioinformatics, 2007, 23: 2247–2255

93 Zhang D, Hao C, Wu C, et al. Local search approximation algorithms for the k-means problem with penalties. J Comb Optim, 2019, 37: 439–453

94 Feng Q, Zhang Z, Shi F, et al. An improved approximation algorithm for the k-means problem with penalties. In: Frontiers in Algorithmics. FAW 2019. Lecture Notes in Computer Science, vol. 11458. Cham: Springer, 2019,

170–181
95 Alimi M, Daneshgar A, Foroughmand-Araabi M H. An $O(\log^{1.5} n \log\log n)$ approximation algorithm for mean isoperimetry and robust k-means. arXiv:1807.05125, 2018
96 Li M, Xu D, Yue J, et al. The seeding algorithm for k-means problem with penalties. J Comb Optim, 2020, 39: 15–32
97 Ji S, Xu D, Guo L, et al. The seeding algorithm for spherical k-means clustering with penalties. In: Proceedings of Algorithmic Aspects in Information and Management. Berlin: Springer, 2019, 149–158
98 Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown: Morgan Kaufmann, 2001, 577–584
99 Ding H, Xu J. A unified framework for clustering constrained data without locality property. In: Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2015, 1471–1490
100 Bhattacharya A, Jaiswal R, Kumar A. Faster algorithms for the constrained k-means problem. Theory Comput Syst, 2018, 62: 93–115
101 Feng Q, Hu J, Huang N, et al. Improved PTAS for the constrained k-means problem. J Comb Optim, 2019, 37: 1091–1110
102 Feng Q, Fu B. Speeding up constrained k-means through 2-means. arXiv:1808.04062, 2018
103 Han L, Xu D, Du D, et al. A local search approximation algorithm for the uniform capacitated k-facility location problem. J Comb Optim, 2018, 35: 409–423
104 Adamczyk M, Byrka J, Marcinkowski J, et al. Constant-factor FPT approximation for capacitated k-median. In: Proceedings of the 27th Annual European Symposium on Algorithms. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2019, 1–14
105 Xu Y, Möhring R H, Xu D, et al. A constant FPT approximation algorithm for hard-capacitated k-means. Optim Eng, 2020, 21: 709–722
106 Cohen-Addad V, Li J. On the fixed-parameter tractability of capacitated clustering. In: Proceedings of the 46th International Colloquium on Automata, Languages, and Programming. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2019, 41:1–41:14
107 Svitkina Z. Lower-bounded facility location. ACM Trans Algorithms, 2010, 6: 1–16
108 Ahmadian S, Swamy C. Improved approximation guarantees for lower-bounded facility location. In: Proceedings of the 10th International Workshop on Approximation and Online Algorithms. Berlin: Springer, 2012, 257–271
109 Li S. On facility location with general lower bounds. In: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2019, 2279–2290
110 Kaplan H, Stemmer U. Differentially private k-means with constant multiplicative error. In: Proceedings of the 32nd International Conference on Neural Information Processing. Berlin: Springer, 2018, 5436–5446
111 Meng Y, Liang J, Cao F, et al. A new distance with derivative information for functional k-means clustering algorithm. Inform Sci, 2018, 463: 166–185
112 Li M, Wang Y, Xu D, et al. The seeding algorithm for functional k-means problem. In: Proceedings of Computing and Combinatorics. Lecture Notes in Computer Science, vol. 11653. Cham: Springer, 2019, 387–396
113 Gamasaee R, Zarandi M H F. A new Dirichlet process for mining dynamic patterns in functional data. Inform Sci, 2017, 405: 55–80
114 Park J, Ahn J. Clustering multivariate functional data with phase variation. Biometrics, 2017, 73: 324–333
115 Bezdek J. Pattern Recognition with Fuzzy Objective Function Algorithms. Norwell: Kluwer Academic Publishers, 1981
116 Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci, 1984, 10: 191–203
117 Stetco A, Zeng X J, Keane J. Fuzzy C-means++: Fuzzy C-means with effective seeding initialization. Expert Syst Appl, 2015, 42: 7541–7548
118 Gafar A, Tahyudin I. Comparison between k-means and fuzzy C-means clustering in network traffic activities. In: Proceedings of the Eleventh International Conference on Management Science and Engineering Management. Lecture Notes on Multidisciplinary Industrial Engineering. Cham: Springer, 2018, 300–310
119 Soomro S, Munir A, Choi K N. Fuzzy c-means clustering based active contour model driven by edge scaled region information. Expert Syst Appl, 2019, 120: 387–396
120 Braverman V, Feldman D, Lang H, et al. Streaming coreset constructions for M-estimators. In: Proceedings of the 22nd International Workshop on Approximation Algorithms for Combinatorial Optimization and 23rd International

Workshop on Randomization and Approximation Techniques in Computer Science. Munich: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2019, 62:1–62:15
121 Braverman V, Lang H, Levin K, et al. Clustering problems on sliding windows. In: Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2016, 1374–1390
122 Braverman V, Meyerson A, Ostrovsky R, et al. Streaming k-means on well-clusterable data. In: Proceedings of the 2011 Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM, 2011, 26–40
123 Goyal D, Jaiswal R, Kumar A. Streaming PTAS for constrained k-means. arXiv:1909.07511, 2019
124 Shindler M, Wong A, Meyerson A W. Fast and accurate k-means for large datasets. In: Proceedings of the Neural Information Processing Systems. Granada: dblp.org, 2011, 2375–2383
125 Ding H, Liu Y, Huang L, et al. K-means clustering with distributed dimensions. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York: JMLR, 2016, 1339–1348
126 Guha S, Li Y, Zhang Q. Distributed partial clustering. ACM Trans Parallel Comput, 2019, 6: 1–20
127 Bhaskara A, Ruwanpathirana A K. Robust algorithms for online k-means clustering. Proc Mach Learn Res, 2020, 117: 1–26
128 Cohen-Addad V, Guedj B, Kanade V, et al. Online k-means clustering. arXiv:1909.06861, 2019
129 Liberty E, Sriharsha R, Sviridenko M. An algorithm for online k-means clustering. In: Proceedings of the Meeting on Algorithm Engineering and Experiments. Philadelphia: SIAM, 2016, 81–89
130 Li S, Svensson O. Approximating k-median via pseudo-approximation. SIAM J Comput, 2016, 45: 530–547

A survey on theory and algorithms for k-means problems

Dongmei Zhang, Min Li, Dachuan Xu & Zhenning Zhang

Abstract The k-means problem is one of the classical problems in theoretical computer science and combinatorial optimization. Meanwhile, the corresponding Lloyd algorithm is one of the ten classical algorithms in data mining. It has been studied in various fields and has a lot of applications, especially on image processing and feature engineering. With the explosive growth of data diversity and quantity, the k-means clustering in practical applications are more complex and diversified. A variety of challenging research topics have emerged that need to be solved urgently. The k-means problem is theoretically NP-hard. In this paper, we introduce effective algorithms based on local search, linear programming rounding, primal-dual, dual-fitting, Lagrange relaxation and other techniques for the classical k-means problem and its variants. We begin with the review of improving approximation algorithms for the classical k-means problem. Then we introduce effective polynomial-time approximation scheme in the doubling metric space and polynomial solvability of stable instances. We further survey several important variants of k-means problems including k-median, spherical k-means, robust k-means, constrained k-means, privacy preserving k-means, etc. Finally, we discuss some open problems for k-means problems.

Keywords k-means, approximation algorithm, linear programming

MSC(2010) 90C27, 68W25

doi: 10.1360/SSM-2019-0280