文章编号:1672-3961(2011)04-0125-08

k-means 聚类问题的改进近似算法

王守强[1],朱大铭[2],史士英[1]
(1. 山东交通学院信息工程系,山东 济南 250023;2. 山东大学计算机科学与技术学院,山东 济南 250101)

摘要:研究了 Ostrovsky 给出 k-means 问题的$(1+\varepsilon)$-近似算法,针对算法中取样参数小的以及枚举数量大的不足,证明了可以选择一个更大的取样参数减小取样点集,基于随机算法,提出新的枚举策略减少枚举数量。本文分析了算法的成功概率。改进算法的期望时间复杂度为 $O(2^{O(k\alpha^2/\varepsilon)}dn)$,其中 d、n 分别为问题实例的空间维数和输入点个数,α 是小于1的分隔系数。算法的成功概率为 $\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\sqrt{\alpha}))$。与 Ostrovsky 给出的算法相比,算法的运算效率得到很大的提高。

关键词:算法;聚类;概率;质心点
中图分类号:TP301.6 **文献标志码**:A

Improved approximation algorithm for the k-means clustering problem

WANG Shou-qiang[1], ZHU Da-ming[2], SHI Shi-ying[1]
(1. Department of Information Engineering, Shandong Jiaotong University, Jinan 250023, China;
2. School of Computer Science and Technology, Shandong University, Jinan 250100, China)

Abstract: The $(1+\varepsilon)$-randomized approximation algorithm proposed by Ostrovsky was investigated in depth, and an improved algorithm was proposed. The sample parameter was enlarged to reduce the size of sample set. Based on the randomized algorithm, a new method was proposed to decrease the enumerating number. Also, the successful probability of the improved algorithm was analyzed. The running time of the improved algorithm was $O(2^{O(\frac{k\alpha^2}{\varepsilon})}nd)$, where d, n denote the dimension and the number of the input points respectively, and α represents the separated coefficient that is lower than 1. The probability was $\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\sqrt{\alpha}))$. Compared to the original algorithm, the improved algorithm runs more efficiency.

Key words: algorithm; clustering; probability; centroid

0 引言

给定 d 维空间中的点集 P,k-means 聚类问题要求选取 k 个中心点,使 P 中的点与其距离最近的中心点的距离平方和最小。形式化描述为:

实例:点集 $P\in\mathbf{R}^d$,正整数 $k\in Z^+$。

目标:寻找点集 $C=\{c_1,c_2,\cdots,c_k\}$,最小化 $\sum_{p\in P}[d(p,C)]^2$,其中:$d(p,C)=\min_{c_i\in C}\{d(p,c_i)\}$。

k-means 问题的教科书算法为 Lloyd 等给出的启发式方法[1-5],该方法首先从给定点集中随机取出 k 个点作为初始中心点,然后进行收敛计算直到局部收敛为止。Lloyd 算法简单而容易实现,但运行结果依赖于初始值,算法无法保证一个确切的求

收稿日期:2011-06-18
基金项目:山东省高校科技计划资助项目(J08LI69);山东交通学院自然科学基金资助项目(Z201024);山东交通学院博士启动基金项目
作者简介:王守强(1967-),男,山东济南人,副教授,博士,主要研究方向为智能计算,计算机网络等. E-mail: wangshouqiang888@163.com

解近似度。由于局部搜索技术在求解 k-median 问题算法上表现出较好的结果[6-7]，Kanungo 与 Song 分别探讨了利用局部搜索技术求解 k-means 问题算法。Kanungo 等给出 k-means 问题 $9+\varepsilon$ 近似度局部搜索算法[8]，但算法需要利用空间结构划分求到一个候选中心点集，候选中心点集求解十分复杂，算法显得不够实用。Song 等进一步证明，如果将给定实例点集 P 作为中心点的候选点集，通过对候选中心点集局部搜索，可使算法的近似度达到 $O(1)$[9]。但 $O(1)$隐含的常数值较大，这是由于算法在迭代过程中，当每次交换 1 个中心点时，其值可达 162；而当执行多于 1 个中心点的交换时，该值为 50。1994 年，M. Inaba 等给出求解 2-means 问题的$(1+\varepsilon)$-近似算法[10]。2000 年，Matousek 等利用点的空间结构划分给出一个时间复杂度为 $O(n\varepsilon^{-2k^2d}\log^k n)$的$(1+\varepsilon)$-近似算法[11]。2004 年，Har-Peled 利用核心点集技术给出了一个时间复杂度为 $O(n+k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1} n\log^k(1/\varepsilon))$的$(1+\varepsilon)$-近似算法[12]。近几年，国内外部分学者研究利用随机算法求解 k-means 聚类问题$(1+\varepsilon)$-近似算法[13-16]。Amit Kumar 基于随机取样，给出近似性能比为$(1+\varepsilon)$的算法[13]，算法时间复杂度为 $O(2^{(k/\varepsilon)^{O(1)}}dn)$。设 $\Delta_k^2(P)$、$\Delta_{k-1}^2(P)$分别表示针对点集 P 求解 k 个中心和 $k-1$ 个中心作为最优解的解值，Ostrovsky 基于初始中心点选择，对于满足条件 $\Delta_k^2(P)/\Delta_{k-1}^2(P)\leqslant\alpha^2(0<\alpha^2\ll 1)$ 的 k-means 聚类子问题将 Kumar 所给出算法时间复杂度由 $O(2^{(k/\varepsilon)^{O(1)}}dn)$改进为 $O(2^{O(k(1+\alpha^2)/\varepsilon)}dn)$[16]。

本文主要工作为：(1) 放大了选取样本参数值 β；(2)改进了从样本点集中枚举部分样本点，用以计算质心点的方法；(3)原算法描述中未能给出计算$(1+\varepsilon)$-近似解的成功概率表达式，我们分析了改进算法的成功概率。改进算法的期望时间复杂度为 $O(2^{O(k\alpha^2/\varepsilon)}dn)$，成功概率为

$$\left(\frac{1}{2}(1-\mathrm{e}^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\sqrt{\alpha}))。$$

1 符号标记与基本结论

给定 k-means 问题实例：点集 $P\in\mathbf{R}^d$ 和正整数 $k\in Z^+$，记 $\Delta_k^2(P)$为该实例最优解的解值，设 $C=\{c_1,c_2,\cdots,c_k\}$为该实例任意 k 个中心点的集合，则记：$\Delta^2(P,C)=\sum_{x_i\in P}\min_{c\in C}\|x_i-c\|^2$。

定义 1.1 给定点集 P，其质心点定义为：$c(P)=(\sum_{p\in P}p)/|P|$。设 $C'=\{c'_1,c'_2,\cdots,c'_k\}$为 d 维空间中的点集，$0<\varepsilon<1$，如果 C'满足 $\Delta^2(P,C')\leqslant(1+\varepsilon)\Delta_k^2(P)$，则称 C'为 P 的 ε 近似质心点集。

定义 1.2 给定点集 P 和正整数 k，若存在 $0<\alpha<\frac{1}{2}$，满足 $\Delta_k^2(P)\leqslant\alpha^2\Delta_{k-1}^2(P)$，则称点集 P 满足 α 可分割性。

定义 1.3 给定点集 P，$X=\{x_1,x_2,\cdots,x_k\}$，$R(x_i)=\{p\in P\mid d(p,x_i)\leqslant d(p,x_j)(j\neq i)\}$，则称 $R(x_1),\cdots,R(x_k)$为 P 关于 X 的一个 Voronoi 划分，称 $R(x_i)$为 P 的一个 Voronoi 划分子集。

引理 1.1[13] 给定点集 P，P 的质心点为关于点集 P 的 1-means 最优解。

Inaba 等[1]指出在欧氏空间中，只需从 P 中随机选取部分点，取样点集的质心点则以较大的概率成为 P 的 ε 近似质心点。该结论详细描述为：

引理 1.2[10] 给定点集 P，从 P 中均匀地随机选取部分点，设 S 为取样点的集合，$m=|S|$，$c(S)$为 S 的质心点，则存在 $\delta(0<\delta<1)$，使下述不等式成立的概率至少为 $1-\delta$。

$$\Delta^2(P,c(S))<\left(1+\frac{1}{\delta m}\right)\Delta_1^2(P)。$$

根据该引理，如果从集合 P 中随机取样 $m=\frac{2}{\varepsilon}$ 个点，选择 $\delta=1/2$，则取样点集的质心点满足 P 的 1-means 聚类$(1+\varepsilon)$-近似解的概率至少为 1/2。由此我们给出如下定义：

定义 1.4 给定点集 P，$\varepsilon\in(0,1)$，设 $L\subseteq P$。若 L 满足：当$|P|<\frac{2}{\varepsilon}$时 $L=P$；当$|P|\geqslant\frac{2}{\varepsilon}$时$|L|=\frac{2}{\varepsilon}$，则称 L 为 P 的一个标志点集。

显然，对于一个给定点集 P，P 的标志点集的个数 $C_{|P|}^{\frac{2}{\varepsilon}}$。

2 改进前的$(1+\varepsilon)$-近似算法

给定点集 P，设 P 所对应的 k-means 聚类最优划分子集为 $P_1,P_2,\cdots,P_k$，$c_1,c_2,\cdots,c_k$ 分别为 $P_1,\cdots,P_k$ 的质心点。文献[16]关于 k-means 的$(1+\varepsilon)$-近似算法主要想法为：计算 k 个初始点，对每个初始点，确定一个子集 R_i，R_i 以较高的概率满足：$P_i\subseteq R_i$。从 R_i 中随机选取部分点，基于该取样点集，枚举少量样本点，以枚举出的少量样本点集的质心点作为 P_i 的一个中心点，引理 1.2 保证这样得到的中心点以较大的概率满足 P_i 的$(1+\varepsilon)$-近似解。

2.1 初始点的选取

随机选取 $O(k)$ 个点，采用反向贪心策略将

$O(k)$个点减少为 k 个点,以这 k 个点作为初始点。在 Chrobak 等给出的 k-median 问题反向贪心算法[17]中使用过同样策略。

2.1.1　$O(k)$个点的选取

设 $\rho=\sqrt{\alpha}, N=2k/(1-5\rho)+2\ln(2/\rho)/(1-5\rho)^2$,算法非均匀地从 P 中随机选取 $N=O(k)$ 个点,得到的取样点集记为 S。选取过程如下:首先从 P 中随机选取两个点 $S=\{x_1,x_2\}$,遵循两点距离越大,被选取概率越大的原则。再依次随机选取点 $x_3,\cdots,x_N$,选取第 $i(i>2)$ 个点时,遵循一个点与已选择的点距离平方和越大,则该点被选取概率越大的原则。因此算法规定第一次选择两个点$\{x_1,x_2\}$的概率表达式为:$\dfrac{\|x_1-x_2\|^2}{\sum_{\{x,y\}\subseteq P}\|x-y\|^2}$。设 $d(x,S)=\min\limits_{c_j\in S}\{d(x,c_j)\}$,规定选择第 i 个点 x_i 的概率表达式为:$\dfrac{d(x_i,S)^2}{\sum_{x\in P}d(x,S)^2}$。算法可描述为:

算法 1:$O(k)$个点选取算法

输入:点集 P

(1) 从 P 中随机选取两个点 x_1,x_2,选择概率为

$$\frac{\|x_1-x_2\|^2}{\sum_{\{x,y\}\subseteq P}\|x-y\|^2},S=\{x_1,x_2\}$$

(2) While 选取点数$\leqslant N$

(3)　从 P 中随机选取一点 $x_i(i\geqslant 3)$,选取概率为$\dfrac{d(x_i,S)^2}{\sum_{x\in P}d(x,S)^2},S=S\cup\{x_i\}$。

(4) End While

(5) 返回选取点$\{x_1,\cdots,x_N\}$。

结论 2.1[16]　给定点集 P,如果 P 满足 α 可分割性,记$\rho=\sqrt{\alpha}, r_i^2=\dfrac{\Delta_1^2(P_i)}{|P_i|}$,那么算法 1 至少以 $1-O(\rho)$的概率满足:针对每个最优划分子集的质心点 c_i,S 中存在相应的某个取样点 $x_j\in S$,满足 $\|x_j-c_i\|\leqslant r_i/\sqrt{\rho^3}$,算法的时间复杂度为 $O(nkd)$。

2.1.2　k 个点的选择

设算法 1 得到的点集为 $S=\{x_1,\cdots,x_N\}$,基于点集 P,计算 S 的一个 Voronoi 划分。设 $R(x)(x\in S)$为 P 的一个 Voronoi 划分子集,$R(x)$的质心点记为 $c(R(x))$。计算每个划分子集 $R(x)$的质心点,得到一个集合 $\hat{S}=\{\hat{x}=c(R(x))\mid x\in S\}$。对每个 $\hat{x}\in\hat{S}$,赋权 $w(\hat{x})=|R(x)|$。

下面给出利用反向贪心算法由 $\hat{S}$ 计算 k 个中心点的实现过程。开始,$\hat{C}=\hat{S}$,首先基于 $\hat{C}$ 计算 $\hat{S}$ 的 Voronoi 划分。对于每个 $\hat{x}\in\hat{C}$,记 $R(\hat{x})$为 $\hat{S}$ 的一个 Voronoi 划分子集,该划分的赋权代价定义为:$T=\sum_{\hat{x}\in\hat{C}}\sum_{y\in R(\hat{x})}w(y)\|y-\hat{x}\|^2$。对于每个 $\hat{x}\in\hat{C}$,以点集 $\hat{C}-\{\hat{x}\}$重新对 $\hat{S}$ 作 Voronoi 划分。设 $R_{-\hat{x}}(z)$为以 $z\in\hat{C}-\{\hat{x}\}$ 为中心点的划分子集,划分的赋权代价为 $T_{\hat{x}}=\sum_{z\in\hat{C}\setminus\{\hat{x}\}}\sum_{y\in R_{-\hat{x}}(z)}w(y)\|y-z\|^2$。选择 $T_{\hat{x}}-T$ 最小的 $\hat{x}$,从 $\hat{C}$ 中删除 $\hat{x}$。重复该过程,直到剩余 k 个中心点为止。算法描述为:

算法 2:求解 k 个初始点

输入:点集 $\hat{S}$,$\hat{S}$ 上相应点的权集合

(1) $\hat{C}=\hat{S}$

(2) While $|\hat{C}|>k$

(3)　计算以 $\hat{C}$ 为中心点服务于 $\hat{S}$ 中所有点的赋权代价 T

(4)　对每个 $\hat{x}\in\hat{C}$,分别计算以 $\hat{C}-\{\hat{x}\}$ 为中心点服务于 $\hat{S}$ 中所有点的赋权代价 $T_{\hat{x}}$

(5)　选择使 T_y-T 最小的 $y\in\hat{C}$,$\hat{C}=\hat{C}-\{y\}$

(6)　对于每个 $\hat{x}\in\hat{C}$,置 $R(\hat{x})=R_{-y}(\hat{x})$,$\hat{C}=\hat{C}\setminus\{\hat{x}\}\cup\{c(R(\hat{x}))\}$。

(7) End While

(8) 返回选取 N 个点

上述算法第(5)步删除 y 后的中心点集为 $\hat{C}=\hat{C}-\{y\}$。算法第(6)步实现可描述为:重新赋值其关于 $\hat{S}$ 的 Voronoi 划分子集 $R_{-y}(\hat{x})$;计算每个新划分子集的质心点,将其加入到集合 $\hat{C}$ 中,同时删除相应的中心点 $\hat{x}$,即 $\hat{C}=\hat{C}\setminus\{\hat{x}\}\cup\{c(R(\hat{x}))\}$。

设 $C=\{c_1,\cdots,c_k\}$ 为 k-means 实例最优解的中心点集,记 $D_i=\min_{j\neq i}\|c_j-c_i\|$,其中 $c_i,c_j\in C$。调用算法 2 所求得的 k 个初始点设为$\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$,文[16]给出如下结论:

结论 2.2[16]　给定点集 P,P 满足 α 可分割性,记$\rho=\sqrt{\alpha}$。如果对于每个 c_i,存在相应的某个取样点 $x_i\in S$,满足 $\|x_i-c_i\|\leqslant r_i/\sqrt{\rho^3}$,则算法 2 得到的 $\hat{C}$ 中必存在一个点 $\hat{c}_i$,满足 $\|c_i-\hat{c}_i\|\leqslant\dfrac{D_i}{10}$,算法的时间复杂度为 $O(k^3d)$。

2.2　k-means 聚类的$(1+\varepsilon)$-近似算法

给定点集 P,记算法 1 和算法 2 所求得的初始中心点集为 $\hat{C}=\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$,结论 2.1 与 2.2 确保 $\hat{C}$ 与 C 会以较大的概率接近到距离足够小,即每个 $\hat{c}_i$ 与 c_i 的距离不超过 $D_i/10$ 的概率至少为 $1-O(\rho)$。记$\rho_1=\dfrac{36\alpha^2}{1-\alpha^2}$,$\hat{d}_i=\min_{j\neq i}\|\hat{c}_j-\hat{c}_i\|$,对于每一个点 $x\in P$,设 $\hat{c}(x)\in\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$ 为满足 $\|x-\hat{c}(x)\|=\min_j\|x-\hat{c}_j\|$ 的初始点。分别定义点集

B_i、R_i 和 P_i^{cor} 如下：

$$B_i=\left\{x\mid x\in P\wedge \|x-\hat{c}_i\|\leqslant\frac{\hat{d}_i}{3}\right\};$$

$$R_i=\{x\mid x\in P\wedge \|x-\hat{c}_i\|\leqslant\|x-\hat{c}(x)\|+\frac{\|\hat{c}_i-\hat{c}(x)\|}{4}\};$$

$$P_i^{\mathrm{cor}}=\left\{x\mid x\in P_i\wedge \|x-c_i\|^2\leqslant\frac{r_i^2}{\rho_1}\right\}。$$

显然 B_i、R_i 均可根据 P 和 $\hat{C}$ 计算得到。文[16]给出点集 B_i、R_i、P_i^{cor} 的包含关系如下：

结论 2.3[16] 设 P 的最优划分子集为 $P_1,P_2,\cdots,P_k$，其中 $P_i(i=1,2,\cdots k)$ 的质心点为 c_i。若 $\|c_i-\hat{c}_i\|\leqslant\frac{D_i}{10}$，则下述3个式子均成立：

$$P_i^{\mathrm{cor}}\subseteq B_i\subseteq P_i\subseteq R;_i$$

$$|P_i|\geqslant\beta|R_i|\left(\beta=\frac{1}{1+144\alpha^2}\right);$$

$$|P_i^{\mathrm{cor}}|\geqslant(1-\rho_1)|P_i|。$$

文献[16]仍然利用 Kumar 等给出 $1+\varepsilon$ 近似算法的思想，求到 k-means 实例的 $(1+\varepsilon)$-近似解。先从每个集合 R_i 中随机选取 $\frac{4}{\beta\varepsilon}$ 个点，设取样点集为 S_i $(1\leqslant i\leqslant k)$。每个 S_i 含有 $C_{\frac{4}{\beta\varepsilon}}^{\frac{2}{\beta\varepsilon}}$ 个标志点集。设 S_i 中所有标志点集的质心点集为 $c\mathrm{tr}(S_i)$。从 $c\mathrm{tr}(S_1)$，$\cdots$，$c\mathrm{tr}(S_k)$ 中各选择一个点，得到的 k 个点即为 P 的一个 k-means 可行解。分别在 $c\mathrm{tr}(S_1)$，$\cdots$，$c\mathrm{tr}(S_k)$ 中枚举所有可能的 k 个中心点，选择目标函数最小的 k 个中心点，即为算法的最终解。该算法的时间复杂度为：$O(2^{\frac{4k}{\beta\varepsilon}}dn)$，进一步可简化为 $O(2^{O(k(1+\alpha^2)/\varepsilon)}dn)$[16]。因 $\alpha^2\ll1$，所以文献[16]将该算法时间复杂度简略为 $O(2^{O(k/\varepsilon)}dn)$。

3 改进的 k-means 聚类 $(1+\varepsilon)$-近似算法

3.1 参数 β 的改进与成功概率计算

调用算法1和算法2所求的 k 个初始点仍记为 $\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$，针对每个初始点 $\hat{c}_i$，计算两个子集 R_i 和 B_i。结论2.3表明，若 $\|c_i-\hat{c}_i\|\leqslant\frac{D_i}{10}$ 成立，则 $P_i\subseteq R_i$，且 $|P_i|\geqslant\beta|R_i|$，其中 $\beta=\frac{1}{1+144\alpha^2}$。下面证明可得到一个更大的数值 β，满足 $|P_i|\geqslant\beta|R_i|$。

定理 3.1 如果 $\|c_i-\hat{c}_i\|\leqslant\frac{D_i}{10}(1\leqslant i\leqslant k)$ 成立，则 $|P_i|\geqslant\beta'|R_i|$，其中 $\beta'=\frac{49}{49+3\,600\alpha^2}$。

证明 考虑最优划分子集 $P_1,\cdots,P_k$，假设存在某个 P_i，满足 $|P_i|<\beta'|R_i|$，记 $a_j=\frac{|R_i\cap P_j|}{|R_i|}(1\leqslant j\leqslant k)$，则：

$$\frac{a_i}{1-a_i}<\frac{\beta'}{1-\beta'}。\tag{1}$$

针对每个子集 $P_j(j\neq i)$，从 P_i 中选取 $\frac{a_j}{1-a_i}|P_i|$ 个不同的点，记为 $P_{ij}(1\leqslant j\leqslant k,j\neq i)$，将这些点重新分配到 P_j 中，即 $P_j=P_j\cup P_{ij}$。由于 $\sum_{j\neq i}\frac{a_j}{1-a_i}|P_i|=|P_i|$，因此经过重新分配后，$P_i$ 中的所有点被重新分配到其它子集 $P_j(j\neq i)$ 中，点集 P 被重新划分为 $k-1$ 个子集，下面分析这由 $k-1$ 个子集所确定的值的上界。

根据三角不等式，对于 P_i 中的任一点 x，$\|x-c_j\|\leqslant\|x-c_i\|+\|c_i-c_j\|$，由此进一步得到 $\|x-c_j\|^2\leqslant2(\|x-c_i\|^2+\|c_i-c_j\|^2)$。将 P_i 中的所有点分配到其他子集 $P_j(j\neq i)$ 后，新划分的 $(k-1)$ 个子聚类的值为：

$$\Delta=\sum_{j\neq i}\sum_{x\in P_{ij}}\|x-c_j\|^2+\sum_{j\neq i}\Delta_1^2(P_j)\leqslant$$

$$\sum_{j\neq i}\sum_{x\in P_{ij}}2(\|x-c_i\|^2+\|c_i-c_j\|^2)+\sum_{j\neq i}\Delta_1^2(P_j)=$$

$$2\Delta_1^2(P_i)+2\frac{|P_i|}{1-a_i}\sum_{j\neq i}a_j\|c_i-c_j\|^2+\sum_{j\neq i}\Delta_1^2(P_j)=$$

$$2\Delta_1^2(P_i)+2\frac{a_i|R_i|}{1-a_i}\sum_{j\neq i}a_j\|c_i-c_j\|^2+\sum_{j\neq i}\Delta_1^2(P_j)\leqslant$$

$$2\Delta_1^2(P_i)+2\frac{\beta'}{1-\beta'}\sum_{j\neq i}a_j|R_i|\cdot\|c_i-c_j\|^2+$$

$$\sum_{j\neq i}\Delta_1^2(P_j)\leqslant$$

$$2\Delta_1^2(P_i)+2\frac{\beta'}{1-\beta'}\sum_{j\neq i}\sum_{y\in R_i\cap P_j}\|c_i-c_j\|^2+$$

$$\sum_{j\neq i}\Delta_1^2(P_j)。\tag{2}$$

记 $\hat{c}(x)\in\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$ 为满足条件 $\|x-\hat{c}(x)\|=\min_j\|x-\hat{c}_j\|$ 的初始点，对于任一点 $y\in R_i\cap P_j$，由于 $y\in R_i$，根据 R_i 定义得：

$$\|y-\hat{c}_i\|\leqslant\|y-\hat{c}(y)\|+\|\hat{c}_i-\hat{c}(y)\|/4。\tag{3}$$

根据 $\hat{c}(y)$ 定义可得：

$$\|y-\hat{c}(y)\|\leqslant\|y-\hat{c}_i\|,$$

所以：

$$\|\hat{c}_i-\hat{c}(y)\|\leqslant\|y-\hat{c}_i\|+\|y-\hat{c}(y)\|\leqslant2\|y-\hat{c}_i\|。\tag{4}$$

(4)代入(3)可得：

$$\| y-\hat{c}_i \| \leqslant 2\| y-\hat{c}(y) \| \leqslant 2\| y-\hat{c}_j \|, \quad (5)$$

又由于:

$$\begin{aligned}\| y-c_i \| &\leqslant \| y-\hat{c}_i \| + \| \hat{c}_i-c_i \| \leqslant \\ &2\| y-\hat{c}_j \| + \| \hat{c}_i-c_i \| \leqslant \\ &2(\| y-c_j \| + \| \hat{c}_j-c_j \|) + \| \hat{c}_i-c_i \| \leqslant \\ &2\| y-c_j \| + \frac{D_j}{5} + \frac{D_i}{10}。\end{aligned} \quad (6)$$

根据三角不等式得:

$$\| y-c_i \| \geqslant \| c_i-c_j \| - \| y-c_j \|。 \quad (7)$$

根据 D_i, D_j 定义得:

$$D_i \leqslant \| c_i-c_j \|, D_j \leqslant \| c_i-c_j \|。 \quad (8)$$

(7)、(8)代入(6)得:

$$\frac{7}{10}\| c_i-c_j \| \leqslant 3\| y-c_j \|。$$

进一步可得:

$$\| c_i-c_j \| \leqslant \frac{30}{7}\| y-c_j \|。 \quad (9)$$

将(9)代入(2),得到:

$$\begin{aligned}\Delta &\leqslant 2\Delta_1^2(P_i) + \frac{1\,800\beta'}{49(1-\beta')}\sum_{j\neq i}\sum_{y\in R_i\cap P_j}\| y-c_j \|^2 + \\ &\sum_{j\neq i}\Delta_1^2(P_j) \leqslant \\ &2\Delta_1^2(P_i) + \frac{1\,800\beta'}{49(1-\beta')}\sum_{j\neq i}\Delta_1^2(P_j) + \sum_{j\neq i}\Delta_1^2(P_j) \leqslant \\ &2\Delta_1^2(P_i) + \left(1+\frac{1\,800\beta'}{49(1-\beta')}\right)\sum_{j\neq i}\Delta_1^2(P_j) \leqslant \\ &\max\left\{2, 1+\frac{1\,800\beta'}{49(1-\beta')}\right\}(\Delta_1^2(P_i) + \sum_{j\neq i}\Delta_1^2(P_j)) = \\ &\max\left\{2, 1+\frac{1\,800\beta'}{49(1-\beta')}\right\}\Delta_k^2(P) = \\ &\max\left\{2, 1+\frac{1}{2\alpha^2}\right\}\Delta_k^2(P)。\end{aligned} \quad (10)$$

由于 $\alpha^2 \ll \frac{1}{2}$ 及实例点集 P 满足 α 可分割性条件,即 $\Delta_k^2(P) \leqslant \alpha^2\Delta_{k-1}^2(P)$,将其代入(10)可得 Δ 的上界至多为:

$$\begin{aligned}&\max\left\{2, 1+\frac{1}{2\alpha^2}\right\}\alpha^2\Delta_{k-1}^2(P) = \\ &\max\left\{2\alpha^2, \frac{1}{2}+\alpha^2\right\}\Delta_{k-1}^2(P) < \Delta_{k-1}^2(P)。\end{aligned}$$

另一方面由于 $\Delta_{k-1}^2(P)$ 代表实例点集 P 被分成 $(k-1)$ 个聚类所对应的最优解值,因此上述新划分的 $(k-1)$ 个聚类的下界显然大于或等于 $\Delta_{k-1}^2(P)$,二者显然相矛盾,因此结论成立。

当从集合 R_i 中随机选取 $\frac{4}{\beta\varepsilon}$ 个点后,仍设 S_i 为取样点的集合。证明 $|S_i\cap P_i| \geqslant \frac{2}{\varepsilon}$ 成立的概率足够大。(注:文献[8,12]仅说明存在常数概率,但未能给出概率的具体表达式)

定理 3.2　给定点集 P_i 和 R_i, $P_i \subseteq R_i$ 并且满足 $|P_i| \geqslant \beta|R_i|$,其中 $0<\beta\leqslant 1$。如果从 R_i 中随机均匀选择 $\frac{4}{\beta\varepsilon}$ 个样本点,记 S_i 为样本点集,则 $|S_i\cap P_i| \geqslant \frac{2}{\varepsilon}$ 成立的概率至少为 $1-e^{(-\frac{1}{2\varepsilon})}$。

证明　设 $S_i=\{x_1, x_2, \cdots, x_m\}$, $m=\frac{4}{\beta\varepsilon}$。引入随机变量 X_i,如果点 $x_i \in P_i$ 则 $X_i=1$,否则 $X_i=0$。以 X 表示 S_i 中属于 P_i 的点数,则 $X=X_1+X_2+\cdots+X_m$,下面讨论 X 的期望值 $E(X)$。

由条件 $|P_i| \geqslant \beta|R_i|$ 知,S_i 中每个点 x_i 属于 P_i 的概率至少为 β,由此可得 X 的期望值满足:$E(X) \geqslant \beta m$。记 $0<\lambda\leqslant 1$,根据 Chernoff-Bounds 不等式[18],可得:

$$\Pr(X \geqslant \lambda\beta m) \geqslant 1-e^{\left(-\frac{(1-\lambda)^2}{2}\beta m\right)}, \quad (11)$$

由于 $m=\frac{4}{\beta\varepsilon}$,选择 $\lambda=\frac{1}{2}$,代入(11)得:

$$\Pr\left(X \geqslant \frac{2}{\varepsilon}\right) \geqslant 1-e^{(-\frac{1}{2\varepsilon})}。$$

记 $\mathrm{ctr}(S_i)$ 为 $S_i\left(|S_i|=\frac{4}{\beta\varepsilon}\right)$ 中所有标志点集的质心点的集合,如果 S_i 中存在一标志点集 L 属于 P_i,计算 L 的质心点 $c' \in \mathrm{ctr}(S_i)$,根据引理 2.2,c' 满足 P_i 的 $(1+\varepsilon)$-近似解的概率大于等于 1/2,即:$\Delta(P_i, c') \leqslant (1+\varepsilon)\Delta_1(P_i)$ 成立的概率大于等于 1/2。所以文献[16]通过枚举 $\mathrm{ctr}(S_i)$ 中的点,求出 P_i 的 $(1+\varepsilon)$-近似解。显然,$|S_i|$ 越小,$|\mathrm{ctr}(S_i)|$ 越小。由于 $|S_i|=\frac{4}{\beta\varepsilon}$,为减小 S_i,可选择更大的 β 值。为此,我们将 R_i 点中的点分为 3 部分:

(1) $x \in B_i$;

(2) $x \in R_i-B_i$,且 $x \notin R_j (j\neq i)$;

(3) $x \in R_i-B_i$,且 $x \in R_j (j\neq i)$。

对于情况(1),由结论 3.3 知 $x \in P_i$。证明如下定理:

定理 3.3　P 中任一点 x,如果 $x \in R_i-B_i$ 并且 $x \notin R_j (j\neq i)$,则 $x \in P_i$。

证明　假设存在 $j\neq i$,满足 $x \in P_j$,由结论 3.3 知 $P_j \subseteq R_j$,由此可推得 $x \in R_j$,这与已知条件 $x \notin R_j$ 相矛盾,所以 $x \in P_i$。

设 $\hat{B}_i$ 为 R_i 中所有满足(1)和(2)的点集,则 $\hat{B}_i=(R_i-\cup_{j\neq i}R_j)$。当 $\| c_i-\hat{c}_i \| \leqslant \frac{D_i}{10}$ 成立时,由结

论2.3及定理3.3可得 $B_i \subseteq \hat{B}_i \subseteq P_i \subseteq R_i$。设 $\beta_i = \frac{|\hat{B}_i|}{|R_i|}$,因 $\hat{B}_i \subseteq P_i \subseteq R_i$,所以 $|P_i| \geqslant \beta_i |R_i|$。选择 $\beta_i = \max\{\beta_i, \beta'\}$ $\left(\text{其中}\ \beta' = \frac{49}{49+3\,600\alpha^2}\right)$,再从集合 R_i 中随机选取 $\frac{4}{\beta_i \varepsilon}$ 个取样点,这些点记为 S_i。由定理3.2,S_i 中含有标志点集 L 属于 P_i 的概率仍然不小于 $1 - e^{\left(-\frac{1}{2\varepsilon}\right)}$。

定理 3.4 S_i 中取样点数小于文献[16]从 R_i 中选择的取样点数。

证明 由于文献[16]在 R_i 中选择的取样点数为 $\frac{4}{\beta\varepsilon}$,其中 $\beta = \frac{1}{1+144\alpha^2}$。显然 $|S_i| \leqslant \frac{4}{\beta'\varepsilon} < \frac{4}{\beta\varepsilon}$,所以 S_i 的取样点数要小于文[16]的取样点数。

3.2 改进算法的实现

记 $S_i\left(|S_i| = \frac{4}{\beta_i\varepsilon}\right)$ 为 R_i 的取样点集。根据定理3.2知,S_i 中存在一标志点集 L 属于 P_i 的概率至少为 $(1 - e^{\left(-\frac{1}{2\varepsilon}\right)})$,而由引理1.2知,$L$ 的质心点满足 P_i 的 $(1+\varepsilon)$-近似解的概率大于等于 $\frac{1}{2}$。为求出该质心点,文献[16]通过枚举 S_i 中所有标志点集的方法求到最好的质心点。下面我们利用集合 B_i、P_i、R_i 的关系 $B_i \subseteq P_i \subseteq R_i$ 来减少从 S_i 中枚举标志点集的个数,从而达到提高算法运算效率目的。

S_i 中的点可分为两部分:一部分属于 $\hat{B}_i$,另外一部分属于 $R_i - \hat{B}_i$ 中。设 $S_{i,1} = S_i \cap \hat{B}_i$, $S_{i,2} = S_i \cap (R_i - \hat{B}_i)$。根据定理3.3及 $\hat{B}_i$ 的构造可知:$S_{i,1} \subseteq P_i$。只需从 $S_{i,2}$ 中选择 $\max\left\{\frac{2}{\varepsilon} - |S_{i,1}|, 0\right\}$ 个点与 $S_{i,1}$ 合并即可凑足 $\frac{2}{\varepsilon}$ 个点成为 S_i 的一个标志点集,以该标志点集的质心点作为 P_i 的中心点。针对每个子集 $P_i (1 \leqslant i \leqslant k)$,利用该策略,通过枚举,计算所有可能标志点集的质心点求到一个关于 P_i 的候选中心点的集合 $C(P_i)$。从 $C(P_1), \cdots, C(P_k)$ 枚举所有可能的 k 个中心点,取值最小的 k 个中心点作为算法的最终解。算法的实现描述如下:

算法 3 求解 k-means 的 $(1+\varepsilon)$-近似解

输入:点集 P, 正整数 k (1) 调用算法3-1及算法3-2求出 k 个初始点 $\{\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_k\}$。

(2) For $i = 1$ to k

(3) 根据初始点 $\hat{c}_i$,点集 P,计算点集 B_i, R_i。

(4) End for

(5) For $i = 1$ to k

(6) 计算 $\hat{B}_i = (R_i - \cup_{j \neq i} R_j)$,

取 $\beta_i = \max\left\{\frac{|\hat{B}_i|}{|R_i|}, \frac{49}{49+3\,600\alpha^2}\right\}$。

(7) 从 R_i 中随机均匀选取 $\frac{4}{\beta_i\varepsilon}$ 个点的集合 S_i

(8) $S_{i,1} = S_i \cap \hat{B}_i$, $S_{i,2} = S_i \cap (R_i - \hat{B}_i)$。

(9) End for

(10) 用函数 Cost = Irred-K-Means($\varnothing$, 0, ε)

(11) 返回 Cost 的值。

根据定理3.2,若 $P_i \subseteq R_i$ 并且 $|P_i| \geqslant \beta_i |R_i|$ 成立,第(7)步所求 S_i 中至少存在一个标志点集属于 P_i 的概率大于等于 $1 - e^{\left(-\frac{1}{2\varepsilon}\right)}$。算法第(8)步将 S_i 划分为 $S_{i,1}$ 和 $S_{i,2}$。算法第(10)步调用函数 Irred-K-Means(·)计算出 P 的 k-means 解。

函数 Irred-K-Means(·)的实现采用递归的方法:从 $i = 1$ 开始,枚举 $S_{i,2}$ 中所有大小为 ℓ $\left(\ell = \max\left\{\frac{2}{\varepsilon} - |S_{i,1}|, 0\right\}\right)$ 的子集。对每个子集 Y,计算 $S_{i,1} \cup Y$ 的质心点,以该质心点作为 P_i 的中心点。同时对于满足 $S_{i,2} \cap S_{j,2} \neq \varnothing$ $(j = i+1, \cdots, k)$ 的样本集合 $S_{j,2}$,执行 $S_{j,2} = S_{j,2} - Y$,该操作目的在于当枚举 $S_{j,2} (j = i+1, \cdots, k)$ 时,可进一步减少从 $S_{j,2}$ 中所枚举子集的个数,从而提高算法效率。该算法描述如下:

算法 4 函数 Irred-K-Means(C, m, ε)

输入:已求中心点集 C,C 的大小 m 及参数 ε

(1) If $m = k$ then

(2) 以 C 作为中心点,计算解值 S。

(3) if $S <$ Mincost then Mincost = S,保存 C 和 Mincost

(4) For $i = |C| + 1$ to k

(5) Repeat

(6) 从 $S_{i,2}$ 中任取 ℓ ($\ell = \max\{\frac{2}{\varepsilon} - |S_{i,1}|, 0\}$) 个点的子集 Y; $S_{j,2} = S_{j,2} - Y (j = i+1, \cdots, k)$

(7) 计算 $S_{i,1} \cup Y$ 的质心点 c'_i, $C = C \cup \{c'_i\}$

(8) Irred-K-Means($i+1, C, \varepsilon$)

(9) 恢复 $S_{j,2} (j = i+1, \cdots, k)$, $C = C - \{c'_i\}$

(10) Until 穷举完 $S_{i,2}$ 中所有子集 Y

(11) End for

(12) 返回 MinCost 的解值。

3.3 算法分析

由定理3.3,若 $P_i \subseteq R_i$ 并且 $|P_i| \geqslant \beta_i |R_i|$ 成立,则 R_i 中取样子集 S_i 至少存在一标志点集属于 P_i 的概率大于等于 $1 - e^{\left(-\frac{1}{2\varepsilon}\right)}$。我们先讨论算法3-1能够

求到 *k*-means 问题实例$(1+\varepsilon)$-近似解的概率。

定理 3.5　给定点集 P,P 满足 α 可分割性,记 $\rho=\sqrt{\alpha}$,算法 3 求到 P 的$(1+\varepsilon)$-近似解的概率不小于$\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\rho))$。

证明　设 P 的最优划分子集为 $P_1,P_2,\cdots,P_k$,记 $c_1,c_2,\cdots,c_k$ 为 $P_1,P_2,\cdots,P_k$ 的质心点。根据结论 2.1 及结论 2.2,算法 2 求到 k 个初始点$\{\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_k\}$,满足条件$\|c_i-\hat{c}_i\|\leqslant D_i/10$ 成立的概率至少为 $1-O(\rho)$,其中 $D_i=\min_{j\neq i}\|c_j-c_i\|$。

针对每个初始点 $\hat{c}_i$,当$\|c_i-\hat{c}_i\|\leqslant D_i/10$ 成立时,由结论 2.3 得 $P_i\subseteq R_i$。算法 3 从每个子集 R_i 内随机选取一个取样点集 $S_i\left(|S_i|=\frac{4}{\beta_i\varepsilon}\right)$。根据定理 3.2,$S_i$ 中包含属于 P_i 一个标志点集的概率大于等于 $1-e^{(-\frac{1}{2\varepsilon})}$。针对每个 P_i,函数 Irred-K-Means(·)通过枚举用来找出 S_i 中属于 P_i 的标志点集,以该标志点集的质心点作为 P_i 的中心点,由引理 1.2,标志点集的质心点满足 P_i 的$(1+\varepsilon)$-近似解的概率不小于$\frac{1}{2}$。所以,当条件$\|c_i-\hat{c}_i\|\leqslant D_i/10$ 成立时,算法 3-1 求到 P_i 的$(1+\varepsilon)$-近似解的概率至少为$\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})$。

因此,当 k 个初始点均满足$\|c_i-\hat{c}_i\|\leqslant D_i/10$ 时,算法 3 求到 *k*-means 实例的$(1+\varepsilon)$-近似解的概率至少为$\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k$。考虑到求 k 个初始点的概率,可得出算法得到的 k 个中心点是给定 *k*-means 实例$(1+\varepsilon)$-近似解的概率至少为$\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\rho))$。

下面再分析算法的时间复杂性。

根据结论 2.1、2.2,算法 3 第 1 步求解 k 个初始点时间复杂度为 $O(nkd+k^3d)$。算法 3 中 2-9 步的时间复杂度为 $O(nkd)$。由于 S_i 是从 R_i 中随机选择的点集,S_i 中分别属于 $\hat{B}_i$ 及 $R_i-\hat{B}_i$ 的点的个数也是随机的,所以我们只能讨论算法 Irred-*K*-Means(·)的平均时间复杂度。

记 $\hat{\beta}_i=\frac{|\hat{B}_i|}{|R_i|}$,针对 S_i 中的每一个点 x,显然该点属于 $\hat{B}_i$ 中一个点的概率为 $\hat{\beta}_i$。设随机变量 Y_i,Z_i 分别代表 S_i 中属于 $\hat{B}_i$ 以及 $R_i-\hat{B}_i$ 的取样点数,则:

$$E(Y_i)=\hat{\beta}_i\times\frac{4}{\beta_i\varepsilon}。$$

当 $\hat{\beta}_i\geqslant\beta$ 时$\left(\beta=\frac{49}{49+3\,600\alpha^2}\right)$时,$\beta_i=\max\{\hat{\beta}_i,\beta\}=\hat{\beta}_i$。所以 $E(Y_i)=\frac{4}{\varepsilon}$。

由此得到:

$$E(Z_i)=\frac{4}{\beta_i\varepsilon}-E(Y_i)=\frac{4}{\beta_i\varepsilon}-\frac{4}{\varepsilon}=\frac{4}{\varepsilon}\left(\frac{1}{\beta_i}-1\right)\leqslant\frac{4}{\varepsilon}\left(\frac{1}{\beta}-1\right)=\frac{4\times3600\alpha^2}{49\varepsilon}。\quad(12)$$

当 $\hat{\beta}_i<\beta$ 时,$\beta_i=\max\{\hat{\beta}_i,\beta\}=\beta$。由结论 3.3 知,$P_i^{\mathrm{cor}}\subseteq B_i$ 并且$|P_i^{\mathrm{cor}}|\geqslant(1-\rho_1)|P_i|$(其中 $\rho_1=\frac{36\alpha^2}{1-\alpha^2}$)。根据 $\hat{B}_i$ 的定义可得,$|\hat{B}_i|\geqslant|B_i|\geqslant|P_i^{\mathrm{cor}}|$,又由于$|P_i|\geqslant\beta|R_i|$(定理 4.1),因此:

$$\hat{\beta}_i=\frac{|\hat{B}_i|}{|R_i|}\geqslant\frac{|P_i^{\mathrm{cor}}|}{|R_i|}\geqslant(1-\rho_1)\beta|R_i|。$$

所以:

$$E(Z_i)=\frac{4}{\beta_i\varepsilon}-E(Y_i)=\frac{4}{\beta\varepsilon}-\hat{\beta}_i\times\frac{4}{\beta\varepsilon}\leqslant\frac{4}{\beta\varepsilon}-(1-\rho_1)\times\beta\times\frac{4}{\beta\varepsilon}=\frac{4}{\varepsilon}\left(\frac{1}{\beta}-(1-\rho_1)\right)=\frac{4}{\varepsilon}\times\left(\frac{3600\alpha^2}{49}+\rho_1\right)。\quad(13)$$

由于 $\rho_1=\frac{36\alpha^2}{1-\alpha^2}$ 以及 $\alpha^2\ll1$,根据(12)、(13)得,S_i 中枚举 ℓ($\ell=\max\{\frac{2}{\varepsilon}-|S_{i,1}|,0\}$)个点的子集的期望个数至多为 $2^{O(\alpha^2/\varepsilon)}$),由此可得当枚举 k 个中心点时,枚举子集个数期望值至多为 $2^{O(k\alpha^2/\varepsilon)}$,所以函数 Irred-*K*-Means(·)的时间复杂度期望值为:$O(2^{O(k\alpha^2)/\varepsilon}dn)$。

综上分析,由下述定理:

定理 3.6　给定点集 P,如果 P 满足 α 可分割性,记 $\rho=\sqrt{\alpha}$,本文算法至少以 γ 概率求出 *k*-means 问题实例的$(1+\varepsilon)$-近似解,其中 $\gamma=\left(\frac{1}{2}(1-e^{(-\frac{1}{2\varepsilon})})\right)^k(1-O(\rho))$。算法的时间复杂度期望值为 $O(2^{O(k\alpha^2)/\varepsilon}dn)$。

与文[16]算法相比,该算法的时间复杂度期望值要优于文[16]的算法时间复杂度 $O(2^{O(k/\varepsilon)}dn)$。

4　结束语

本文探讨了满足 α 可分割性的 *k*-means 聚类问题的$(1+\varepsilon)$-近似算法,改进了文[16]算法的时间复杂度。放大了文[16]的取样参数 β 值,并改进了样本点的选取方法。改进算法基本思想:针对每个

最优子集 P_i，计算两个子集 B_i 和 R_i，使之满足 $B_i \subseteq P_i \subseteq R_i$。从每个子集 R_i 中随机选取部分点。将取样点划分为两部分：一部分点是能够确定属于最优子集 P_i 中的点；另外一部分是不能确定是否属于 P_i 中的点。本文首先从取样点中找出满足第一部分点，仅在第二部分点中枚举得到标志点集，有效地减少枚举点的组合数目，从而降低算法的时间复杂度。最后，本文分析算法的时间复杂度以及求到$(1+\varepsilon)$解的成功概率。基于该算法，还有几个问题值得探讨。(1)该算法在计算质心点时只是减少了各点之间的组合数目，显然时间效率较低。能否不利用枚举策略，而是通过其它策略从每个最优子集中找出满足条件的一定数量的点？(2)该算法能否进一步提高成功概率？(3)该算法是针对满足 α 可分割性的一类问题，对于一般问题如何求解？因此，如何减少组合数目，提高每次成功概率以及是否适用于一般 k-means 聚类问题求解是需要进一步研究的课题。

参考文献：

[1] PENA J M, LOZANO J A, ARRANAGA P L. An empirical comparison of four initialization methods for the k-means algorithm [J]. Pattern Recognition Lett, 1999 (20):1027-1040.

[2] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k-means clustering algorithm: Analysis and implementation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(7):881-892.

[3] 钱伟宁,周傲英. 从多角度分析现有聚类算法[J]. 软件学报, 2002, 13(8):1382-1394.
QIAN Weining, ZHOU Aoying. Analyzing popular clustering algorithm from different viewpoints[J]. Journal of Software, 2002, 13(8):1382-1394.

[4] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):48-61.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithm research[J]. Journal of Software, 2008, 19(1):48-61.

[5] 张莉,周伟达. 核聚类算法[J]. 计算机学报,2002,25(6):587-590.
ZHAN Li, ZHOU Weida. Kernel clustering algorithm [J]. Journal of Computer, 2002, 25(6):587-590.

[6] ARYA V, GARG N, KHANDEKAR R, et al. Local search heuristics for k-Median and facility location problems[C]//Proceedings of the 33rd Annual ACM Symp on Theory of Computing. New York, USA: the ACM Press, 2001:21-29.

[7] 潘锐,朱大铭,马绍汉. k-median 近似计算复杂度与局部搜索近似算法分析[J]. 软件学报,2005,16(3):392-399.
PAN Rui, ZHU Daming, MA Shaohan. Approximated computational hardness and local search approximated algorithm analysis for k-median problem[J]. Journal of Software, 2005, 16(3):392-399.

[8] KANUNGO T, MOUNT D M, NETANYAHU N, et al. A local search approximation algorithm for k-means clustering[J]. Computational Geometry, 2004(28): 89-112.

[9] SONG M, RAJASEKARAN S. Fast k-means algorithms with constant approximation[C]//Proceedings of the 16th Annual International Symposium on Algorithms and Computation. Sanya, China: SPRINGER-VERLAG, 2005: 1029-1038.

[10] INABA M, KAOTH N, IMAI H. Application of weighted Voronoi diagrams and randomization to variance-based k-clustering (extended abstract) [C]//Proceeding of the tenth annual symposium on Computational Geometry. Stony Brook, New York: the ACM Press, 1994: 332-339.

[11] MATOUSEK J. On approximate geometric k-clustering [J]. Discrete and Computational Geometry, 2000, 24 (1):61-84.

[12] HAR-PELED S, MAZUMDAR S. Coresets for k-means and k-median clustering and their applications[C]//Proceedings of the 36th Annual Symposium on Theory of Computing. Chicago, USA: ACM Press, 2004: 291-300.

[13] KUMAR A, SABHARWAL Y, SEN S. A sample linear time $(1+\varepsilon)$ algorithm for k-means clustering in any dimensions[C]// Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science. Rome, Italy: IEEE Press, 2004:454-462.

[14] 王守强,朱大铭. 基于最小聚类求解 k-means 问题算法[J]. 通信学报,2010,31(7):46-52.
WANG Shouqiang, ZHU Daming. Algorithm for the k-means clustering based on minimum cluster size [J]. Journal of Communications, 2010, 31(7):46-52.

[15] 王守强,朱大铭. 基于最小聚类划分的 k-means 聚类 $(1+\varepsilon)$近似算法[J]. 计算机研究与发展,2008,45 (z1):26-30.
WANG Shouqiang, ZHU Daming. The $(1+\varepsilon)$ approximate algorithm for k-means based on the minimum size of sub-cluster[J]. Journal of Computer Research and Development, 2008, 45(z1):26-30.

[16] OSTROVSKY R, RABANI Y, SCHULMAN L J. The effectiveness of Lloyd-type methods for the k-means problem[C]// Proceedings of 47th Annual IEEE Symposium on the Foundations of Computer Science. Berkeley, CA: IEEE Press, 2006:165-176.

[17] CHROBAK M, KENYON C, YOUNG N. The reverse greedy algorithm for the metric k-median problem[J]. Information Processing Letters, 2006, 97(2):68-72.

[18] MOTWANI R, RAGHAVAN P. Randomized algorithms [M]. London: Cambridge University Press, 1995:90-95.

（编辑：陈斌）