采样与搜索：一种适用于高维学习增强型 k-中位数聚类的有效算法

Kangke Cheng[1], **Shihong Song**[2],**Guanlin Mo**[1],**Hu Ding**[1*]

[1]中国科学技术大学，中国合肥 [2]爱丁堡大学信息学院，英国爱丁堡 ke314159@mail.ustc.edu.cn, S.Song-29@sms.ed.ac.uk, moguanlin@mail.ustc.edu.cn, huding@ustc.edu.cn

arXiv:2603.10721v1 [cs.DS] 11 Mar 2026

摘要

本文研究学习增强型 k-中位数聚类问题，旨在通过使用错误率预测器 $\alpha \in [0,1)$ 对点集进行预处理，提升传统聚类算法的性能。此预处理步骤会在聚类前为各点分配潜在标签。我们针对该问题提出了一种基于简单而有效的采样方法的算法，其时间复杂度显著优于现有算法。此外，我们还缓解了算法对欧几里得空间维度的指数级依赖。最后，我们通过实验将所提方法与若干最先进的学习增强型 k-中位数聚类方法进行比较。实验结果表明，我们提出的方法在实践中能够显著降低计算复杂度，同时实现更低的聚类代价。

代码— https://github.com/KangkeCheng/Learning-Augmented-k-Median-Sample-and-Search

1 引言

作为无监督学习的核心主题， k-中位数 和k-均值 聚类广泛应用于众多领域，如生物信息学（Kiselev 等，2017）、计算机视觉（Caron 等，2020）和社交网络（Ghaffari、Mosavi 和 Shamshirband，2021）。这些基于中心的聚类问题的主要目标是将一组未标记的数据点划分为多个簇，使同一簇中的数据点彼此相似（按照某种度量），而不同簇中的数据点则存在显著差异。

k-means问题旨在寻找 k 个中心，使每个点到其最近中心的平方欧几里得距离之和最小。形式上，目标是最小化$\sum_{x\in X} \min_{c\in C} \|x-c\|_2^2$，其中 X 是输入数据集，C 是 k 个中心的集合。尽管广受欢迎， k-means已知对离群值和噪声较为敏感，因为平方距离目标函数会以二次方式放大极端值的影响。相比之下， k-中位数问题最小化欧几里得距离之和： $\sum_{x\in X} \min_{c\in C} \|x-c\|_2$，从而对离群值和重尾分布具有更强的稳健性。因此， k-中位数聚类在许多实际应用中更为可取，尤其是在数据存在噪声时。因此，我们的工作聚焦于 k-中位数设定，旨在保留其稳健性优势，同时通过所提出的框架应对算法挑战。

学习增强算法。算法设计领域的一项核心挑战，是在维持可靠近似比的同时降低算法时间复杂度。海量数据的激增和机器学习的发展，为许多经典算法问题带来了获取有价值先验知识的机会。为克服传统最坏情况分析往往过于悲观的界，理论计算机科学界提出了学习增强算法（Hsu 等，2018；Antoniadis 等，2020；Dinitz 等，2021；Lykouris 和 Vassilvitskii，2021；Mitzenmacher 和 Vassilvitskii，2022）——这一新范式属于“超越最坏情况分析”（Roughgarden，2021）的范畴。其核心思想是设计能够利用辅助信息（通常来自机器学习模型）来提升性能的算法。

对于学习增强的 k-均值和 k-中位数聚类，Gamlath 等人（2022）研究了噪声标签，在平衡对抗噪声下实现了 $(1+O(\epsilon))$ 近似，并在随机噪声模型下实现了 $O(1)$ 近似。这里，近似比是衡量解质量的指标，定义为算法所得解的成本与最优解成本之比。Ergun 等人（2022）提出了一种学习增强框架，其中数据点附加了预测标签，并通过错误率 $\alpha \in [0,1)$ 进行量化。他们针对 k-均值问题提出了一种随机化算法，在对 α 和簇大小施加特定约束的条件下，以 $O(nd\log n)$ 的时间实现了 $(1+20\alpha)$ 近似，其中 n 表示待聚类的点数， d 表示空间维度。他们还针对 k-中位数问题提出了一种算法，该算法在满足 $\alpha = \tilde{O}\left(\frac{1}{k}\right)$ 的条件下，实现了 $\tilde{O}((k\alpha)^{1/4})$ 近似。

Nguyen、Chaturvedi 和 Nguyen（2023）进一步改进了这些算法。他们的 k-均值算法直接在预测的簇中按维度估计局部最优中心，实现了 $(1+O(\alpha))$ 近似保证，耗时

*通讯作者 版权所有 © 2026，美国人工智能促进协会（www.aaai.org）。保留所有权利。

方法	近似比	标签误差范围	时间复杂度
Ergun 等人（2022）	$1+O((k\alpha)^{1/4})$	$\frac{\tilde{O}}{\binom{1}{k}}$	$O(nd$ 对数$^3\, n+$ 多项式$(k,$ 对数 $n))$
Nguyen、Chaturvedi 和 Nguyen（2023年）	$1+\frac{7\alpha+10\alpha^2-10\alpha^3}{(1-\alpha)(1-2\alpha)}$	$[0,1/2)$	$O(\frac{1}{1-2\alpha}nd\log^3 n\log^2\frac{k}{\delta})$
Huang 等人（2025）	$1+\frac{(6+\epsilon)\alpha-4\alpha^2}{(1-\alpha)(1-2\alpha)}$	$[0,1/2)$	$O(nd\log(kd)\log(n\Delta)\cdot(\frac{\sqrt{d}}{\alpha\epsilon})^{O(d)})$
采样与搜索（我们的方法）	$1+\frac{(6+\epsilon)\alpha-4\alpha^2}{(1-\alpha)(1-2\alpha)}$	$[0,1/2)$	$O(2^{O(1/(\alpha\varepsilon)^4)}nd$对数 $\frac{k}{\delta})$

表 1：我们提出的采样与搜索算法与最先进方法的比较。这里，$\epsilon>0$ 和 $\delta\in(0,1)$是控制近似精度和成功概率的参数。Δ 表示给定点集的长宽比。

$O(nd\log n)$ 时间，其中 $\alpha\in[0,1/2)$。他们的 k-中位数算法通过采用多次随机采样和剪枝技术，显著提升了近似保证。

最近，Huang 等人（2025）将逐维估计方法扩展到学习增强型k-均值聚类中，通过采样避免排序，从而降低时间复杂度，同时保持相近的近似比。他们还提出了一种 k-中位数算法，可实现 $1+\frac{(6+\epsilon)\alpha-4\alpha^2}{(1-\alpha)(1-2\alpha)}$近似，据我们所知，这代表了学习增强型 k-中位数聚类在近似比方面最先进的水平。然而，在他们的工作中，k-均值与 k-中位数之间的结构差异造成了根本性的算法差距。对于 k-均值，均值中心有一个闭式解，可以在各维度上独立计算。相比之下，k-中位数的中心没有闭式表达式，也无法按维度分解，因此即使有点标签预测，计算起来仍然困难得多。因此，他们的 k-中位数方法需要在原始高维空间中采用穷举网格划分和搜索过程，从而引入对 d的指数级依赖，这通常被认为在实践中不可接受，尤其是在高维场景中。因此，一个关键的开放问题是：能否设计一种算法，在达到最先进近似比的同时，克服对维度的指数级依赖 d?

我们的核心思想与主要贡献。我们的关键洞见是：对于每个预测簇，正确标记的子集的真实中位数位于由一个小随机样本张成的低维子空间附近。这使得我们能够利用低维网格高效地离散化搜索空间，从而降低计算成本。主要的技术挑战在于，预测簇可能存在噪声——部分点被错误分类，且预测簇中心可能远离真实中心。为应对这一障碍，我们设计了一种新颖的采样与搜索框架，能够有效地选出合适的候选簇中心。其关键思想是在上述低维网格中采用贪心搜索策略，从而巧妙地避免显式区分正确标记的点与错误标记的点。

我们的贡献总结如下：我们提出了

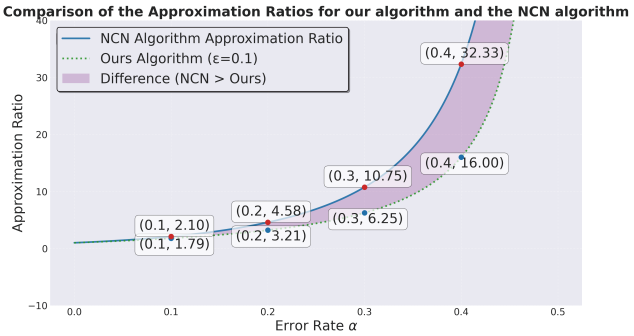


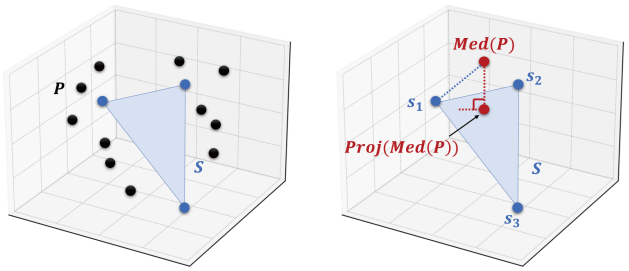
图1：我们的算法（集合 $\epsilon=0.1$）与 NCN 算法在错误率 α变化方面的近似比比较。该图表明，在错误率 $\alpha\in[0,1/2$的所有取值上，我们的算法（绿色虚线）始终取得低于 NCN 算法（Nguyen、Chaturvedi 和 Nguyen（2023））（蓝色实线）的近似比。紫色阴影区域凸显了这一性能差距，且该差距随着α 的增大而更加显著。

一种简单而有效的学习增强型 k-中位数聚类算法。其时间复杂度为**线性于** n **和**d，避免了对维度的指数级依赖 d。与此同时，我们的算法针对 $\alpha<\frac{1}{2}$实现了 $1+\frac{(6+\epsilon)\alpha-4\alpha^2}{(1-\alpha)(1-2\alpha)}$的近似比，达到当前最先进水平。此外，我们在高维数据集上进行了一系列实验，结果表明，与以往方法相比，我们的方法实现了最高达 $10\times$ 的加速，同时保持了较高的聚类质量。图 1 比较了其算法（记为 NCN）与我们的方法的近似比。表 1 对我们的方法与现有方法的结果进行了详细比较。

1.1 预备知识

记号。令 X 表示 $\mathbb{R}^d$中的 n 个点组成的输入集合。对于任意两个点 $p,\ q\in\mathbb{R}^d$，它们之间的欧几里得距离为$\|p-q\|_2$。

给定任意点集 C，点 p 到 C 中最近点的距离记为 $\text{dist}(p,\ C)=\min_{c\in C}\|p-c\|_2$。特别地，当 C 是点集 X给定的中心集时，相应的成本记为 $\text{Cost}(X,\ C)$,



(a) (b)

图2：(a) 简要示意了子空间的生成方式。我们从原始点集 P（黑点）中采样一个子集 S（蓝点），S 构成了一个子空间 span(S)。(b) 表明，span(S) 包含 Med(P) 的一个投影，记为 Proj(Med(P))，该投影与 Med(P) 很接近。此外，S 包含一个点（例如 s_1），它与 Med(P) 的距离有界。

其定义为 Cost(X，C) = $\sum_{x\in X}$dist(p，C)。稍微滥用一下记号，我们也用 Cost(P，c) 表示点集 P到单个点 c的距离之和。

我们将给定实例 X 的最优 k 簇记为 $\{X_1^*,\ldots,X_k^*\}$，集合 $C^* = \{c_1^*,\ldots,c_k^*\}$ 包含它们对应的最优中心。

对于任意点集 P，我们用 Med(P) 表示其**中位点**，即

$$\mathtt{Med}(P) = \arg\min_{q\in\mathbb{R}^d} \sum_{p\in P} \|p-q\|_2. \tag{1}$$

因此，对于每个 $1 \le i \le k$，$c_i^* = \mathtt{Med}(X_i^*)$。

定义 1.1 (**学习增强型 k-中位数聚类**)。假设存在一个预测器，可以输出一个针对 X 的带标签划分 $\{\tilde{X}_1,\tilde{X}_2,\ldots,\tilde{X}_k\}$，其参数为标签错误率 $\alpha\in[0,1)$，并满足：

$$|\tilde{X}_i \cap X_i^*| \geqslant (1-\alpha)\max(|\tilde{X}_i|,|X_i^*|)$$

其中，$|\cdot|$ 表示集合中的点数。学习增强型 k-中位数聚类的目标是利用这样的部分正确结果，计算出一个中心集 $C\subset\mathbb{R}^d$，使 成本(P,C) 最小。

我们还介绍欧几里得空间中几何中位点的两个重要命题，它们对于后续证明至关重要。

命题 *1.1*。（*Badoiu*、*Har-Peled* 和 *Indyk*，*2002*）设P是$\mathbb{R}^d$中的一个点集。给定两个参数$1>\varepsilon>0$和$\gamma>1$，我们从 P 中随机抽取一个大小为 $\frac{\gamma}{\varepsilon^3}\log\frac{1}{\varepsilon}$ 的随机样本 S。那么，以至少 $1-1/\gamma$的概率，以下两个事件会发生：*(i)* 仿射张成*(S)*中包含一个点，其到 *Med(P)*的距离不超过$\epsilon\cdot\frac{\mathtt{Cost}(P,\mathtt{Med}(P))}{|P|}$，其中 *span(S)*表示由 S中的点张成的子空间。*(ii)* 集合S中包含一个点，其到 *Med(P)*的距离不超过$2\times\frac{\mathtt{Cost}(P,\mathtt{Med}(P))}{|P|}$。

命题 1.1 表明，从 P 中抽取足够大的随机样本，其张成的子空间必定包含 Med(P)的良好近似，这使我们能够通过用网格划分该子空间来构造候选中心集，从而近似最优中心。图2展示了如何根据样本生成子空间。命题 1.2 用于估计各簇的平均代价，进而指导网格单元边长的设计。

命题 1.2。（*Kumar*、*Sabharwal* 和 *Sen*，*2010*）设 P 是 $\mathbb{R}^d$ 中的一个点集。给定参数 $\zeta\in(0,1/12)$，我们从 P 中随机抽取一个点 p_0 和一个大小为 $1/\zeta$ 的集合 S。定义值 $v=$成本(S,p_0)。那么，以概率 $>(1-\zeta^2)_2^{1/\zeta+1}$，有：

$$\frac{v\zeta^3}{2} \leqslant \frac{\mathtt{Cost}(P,\mathrm{Med}(P))}{|P|} \leqslant \frac{v}{\zeta}.$$

错误率上界 α。我们假设标签错误率 α 的上界为 $1/2$。正如 Nguyen、Chaturvedi 和 Nguyen（2023）所讨论的，当 α 达到$1/2$时，预测簇与最优簇之间的关系可能会彻底失效。

1.2 其他相关工作

k-中位数算法。由于 k-中位数问题具有 NP-困难性（Cohen-Addad 和 C.S.Karthik 2019），过去半个世纪以来，人们对其近似算法进行了广泛研究。尽管已有多种多项式时间近似方案（PTAS）算法被提出，但它们的运行时间会随维度 d 或簇的数量 k 呈指数增长（Arora、Raghavan 和 Rao 1998；Cohen-Addad、Feldmann 和 Saulpic 2021；Kumar、Sabharwal 和 Sen 2010），因此在许多场景中并不实用。其他算法包括 Arya 等人（2001）提出的通过局部搜索实现的 $(3+\epsilon)$-近似算法，以及 Charikar 和 Li（2012）提出的使用 LP 舍入方法实现的 3。25-近似算法。

k-均值算法。同样，对于 k-中位数问题，现有的 k-均值问题多项式时间近似方案在 d 或 k 方面表现出指数级依赖（Cohen-Addad、Feldmann 和 Saulpic 2021年；Kumar、Sabharwal 和 Sen 2010）。与此同时，广泛使用的 k-均值++算法的时间复杂度为 $O(ndk)$，并达到 $O(\log k)$ 的近似比（Arthur 和 Vassilvitskii 2007 年）；对于良好分离的数据，其近似比为 $O(1)$（Ostrovsky 等人 2013）。

2 我们的算法与理论分析

本节中，我们提出一种用于学习增强型 k-中位数聚类的“采样与搜索”算法。我们的主要思路是通过均匀采样从预测标签中提取信息，然后利用中位点的性质（基于命题 1.1）在低维子空间中构造候选中心集，最后采用贪心搜索方法从候选中心集中找出所需的解。这种采样与搜索策略避免了在原始空间中进行搜索，而原始空间可能

远大于根据命题 1.1 得出的子空间，因此能够大幅降低总体计算复杂度。第 2.1 节将介绍具体算法及主要理论结果，即定理 2.1。随后，我们在第 2.2 节给出定理 2.1 的证明。

2.1 我们提出的算法与主要定理

我们在算法 1 中给出了采样与搜索算法。总体而言，该算法包含三个主要阶段：

1.**基于采样的子空间构造：** 对于每个预测簇，我们采样一个较小的点子集，形成一个**“基”**，以捕捉一个与最优簇中心距离有界的“邻近点”。请注意，一旦找到这样的基，我们只需在这个小基所张成的低维子空间内搜索，即可近似该簇的最优中心。这使得搜索空间的大小仅取决于 ϵ，而不取决于维度 d。

2. **基于网格的候选生成：** 在生成 k个合适的子空间（每个子空间都足够接近相应的最优中心）之后，我们在每个子空间中构造网格结构，以生成 k 个较小的中心点候选集。这样便无需在原始高维空间中进行搜索。

3. **贪心中心选择：** 我们采用成本最小化的贪心选择流程，从候选集中选出最佳中心。

下面给出我们算法的主要理论结果。

定理 2.1。设 $1 > \epsilon > 0$ 和 $1 > \delta > 0$为两个参数。给定一个如定义 *1.1* 所述的实例 X，若假设错误率为 $\alpha < \frac{1}{2}$，则算法 *1* 可以输出一个近似比为 $\left(1 + \frac{6\alpha - 4\alpha^2 + \epsilon\alpha}{1-\alpha}\right)$ 的解，且该结果以 $1 - \delta$ 的概率成立。时间复杂度为 $O(2^{O(1/(\alpha\varepsilon)^4)} nd \log \frac{k}{\delta})$。

2.2 定理 2.1 的证明

我们将证明分为三个主要步骤：首先，我们证明对于每个预测簇，以高概率构造出的候选中心集至少包含一个点，该点接近预测簇中正确标记子集的真实中位数。这一点在引理 2.2 中进行了形式化表述；该引理基于我们的采样设计，利用了命题 1.1 和命题 1.2 中的几何性质。其次，我们分析从候选集中选出的中心的成本。在引理 2.3 中，我们证明，尽管标签存在噪声，该中心仍能使聚类代价接近最优值；为此，我们仔细界定了误分类点带来的额外成本，并分析了贪心选择的最优性。

最后，我们汇总所有簇的界，得到总聚类成本，并分析算法的候选集大小和运行时间。

算法 1：用于学习增强型 k-中位数的采样与搜索 1：输入：一个 k-中位数实例，由有限点集 $X \subset \mathbb{R}^d$ 和整数 k 组成；X 的预测划分 $\{X_1, \ldots, X_k\}$，其错误率为 $\alpha \in [0, 1/2)$；精度参数 $\epsilon \in$ (0, 1) 和失败概率 $\delta \in$ (0, 1)。2：输出：中心集 C，满足 $|C| = k$ 3：初始化 $C \leftarrow \emptyset, \zeta \leftarrow \frac{1}{13}$ 4：对于 $i \leftarrow 1$ 到 k，执行 5：初始化候选集 $C_i \leftarrow \emptyset$ 6：对于 $j \leftarrow 1$ 到 $\left\lceil \frac{\log(\delta/k)}{\log(0.975)} \right\rceil$，执行 7：采样：首先，随机采样一个点 $y_i^j \in X_i$，然后从 X_i 中均匀随机采样两个分离集：一个大小为 $\lceil \frac{2}{(1-\alpha)\zeta} \rceil$ 的集合 $Q_i^j \subseteq \tilde{X}_i$，以及一个大小为 $\lceil \frac{4\log(1/\delta)}{(\frac{\alpha\epsilon}{2})(\frac{1-\alpha}{2})^3} \rceil$ 的集合 $R_i^j \subseteq \tilde{X}_i$ 8：对于每个大小为 $1/\zeta$ 的子集 $Q \subseteq Q_i^j$，执行 9：$v \leftarrow \text{Cost}(Q, y_i^j)$ 10：$a \leftarrow \frac{v\zeta^3}{2}$ 和 $b \leftarrow \frac{v}{\zeta}$ 11：对于每个整数 $l \in \{\lfloor \log_2 a \rfloor, \ldots, \lceil \log_2 b \rceil\}$，执行 12：$t \leftarrow 2^l$ 13：运行算法 2：$S \leftarrow \text{CSC}(R_i^j, t, \alpha, \epsilon)$ 14：$C_i \leftarrow C_i \cup S$ 15：结束循环 16：结束循环 17：结束循环 18：对于每个 $c \in C_i$，将 $N_i(c)$ 定义为 X_i 中距离 c 最近的 $\lceil (1 - \alpha)|X_i| \rceil$ 个点组成的集合 19：贪心选择：为第 i 个簇中心找到最佳候选项，$\hat{c}_i \leftarrow \arg\min_{c \in C_i} \sum_{x \in N_i(c)} \|x - c\|_2$ 20：$C \leftarrow C \cup \{\hat{c}_i\}$ 21：结束循环 22：返回 C

算法 2：候选集构造（CSC）

1：输入：点集 $R \subset \mathbb{R}^d$；近似平均成本 $t > 0$；参数 α 和 ϵ。 2：输出：候选中心集 $S \subset \mathbb{R}^d$。 3：初始化 $S \leftarrow \emptyset, \theta \leftarrow \frac{\alpha\epsilon t}{4|R|}$ 4：对于每个点 $r \in R$，执行 5：初始化候选集 $S_r \leftarrow \emptyset$ 6：在 span(R) 上构造网格 G_r，以 r 为中心，边长为 θ 7：定义球 $B(r, 2t) \leftarrow \{x \in \mathbb{R}^d \mid \|x - r\|_2 \le 2t\}$ 8：$S_r \leftarrow G_r \cap B(r, 2t)$ 9：$S \leftarrow S \cup S_r$ 10：结束循环 11：返回 S

~为方便起见，我们将 X_i 与 X_i^*~的交集记为 T_i，即 $T_i = X_i \cap X_i^*$。

~**引理 2.2.** 对于预测簇 X_i，以$1 - \frac{\delta}{k}$ 的概率，存在一个点 $q \in \tilde{C}_i$ 满足：

$$||q - \texttt{Med}(T_i)||_2 \leqslant \frac{\alpha\epsilon \times \texttt{Cost}(T_i, \texttt{Med}(T_i))}{|T_i|}. \quad (2)$$

证明。首先，在学习增强设置下，我们有$|T_i| \geqslant (1-\alpha)\max(|\tilde{X}_i|, |X_i^*|) > \frac{1}{2}|\tilde{X}_i|$。由于我们在算法第一阶段从 $\tilde{X}_i$中均匀采样一个点 y_i，并从 $\tilde{X}_i$ 中均匀采样一个大小为^2 $(1-\alpha)\zeta$ 的集合 Q_i^j，利用马尔可夫不等式可推得，以下两个事件以至少 $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$的概率同时发生：

$$y_i \in T_i, |T_i \cap Q_i^j| \geqslant \frac{1}{\zeta}.$$

现在假设这两个事件都发生了。存在一个大小为 $\frac{1}{\zeta}$的子集$Q \subseteq (T_i \cap Q_i^j)$。根据命题 1.2，如果我们设定 $p_0 = y_i^j$，$S = Q$，$P = T_i$，对于 $v =$成本$(Q,\ y_i^j)$，我们有

$$\frac{v\zeta^3}{2} \leqslant \frac{\texttt{Cost}(T_i, \texttt{Med}(T_i))}{|T_i|} \leqslant \frac{v}{\zeta} \quad (3)$$

其概率至少为$(1-\zeta^2)_2^{1/\zeta+1}$。在我们的算法的第二阶段，我们遍历 Q_i^j 的所有大小为 $\frac{1}{\zeta}$ 的子集，因此，区间$\lfloor \log \frac{v\zeta^3}{2}, \log \frac{v}{\zeta} \rceil$ 内存在一个整数 l，使得 $t = 2^l$ 满足

$$t/2 \leqslant \frac{\texttt{Cost}(T_i, \texttt{Med}(T_i))}{|T_i|} \leqslant t. \quad (4)$$

类似地，在算法的第一阶段，我们从 $\tilde{X}_i$ 中均匀采样一个大小为$\frac{1}{1}\log(1/\frac{\alpha\epsilon}{2})(1-\alpha)(\frac{\alpha\epsilon}{2})^3$的集合 R_i^j；利用马尔可夫不等式可得，$|T_i \cap R_i^j| \geqslant {}^2\log(1/\frac{\alpha\epsilon}{2})(\frac{\alpha\epsilon}{2})^3$的概率至少为 $1/2$。因此，根据命题 1.1，以至少 $1/2$ 的概率，以下两个事件会发生：

1. 仿射张成空间(R_i^j)中存在一个点，其到 Med(T_i)的距离为 $\leqslant \alpha\epsilon\texttt{Cost}(T_i,\texttt{Med}(T_i))$ $2|T_i|$；2. R_i^j 中存在一个点，其到中心 Med(T_i)的距离为 $\leqslant 2 \times \frac{\texttt{Cost}(T_i,}{\texttt{Med}(T_i))}|T_i|$。

因此，仿射张成空间(R_i^j)中存在一个点 o，使得

$$||o - \texttt{Med}(T_i)||_2 \leqslant \frac{\alpha\epsilon \times \texttt{Cost}(T_i, \texttt{Med}(T_i))}{2|T_i|}. \quad (5)$$

因此，在算法 2 的网格构造下，必定存在一个点 $q \in S$ 满足

$$||q - o||_2 \leqslant \frac{\alpha\varepsilon t}{4} \leqslant \frac{\alpha\varepsilon \times \texttt{Cost}(T_i, \texttt{Med}(T_i))}{2|T_i|}, \quad (6)$$

其中第二个不等式来自不等式 (4)。结合不等式 (5) 和不等式 (6)，由三角不等式可得

$$||q - \texttt{Med}(T_i)||_2 \leqslant \frac{\alpha\epsilon \times \texttt{Cost}(T_i, \texttt{Med}(T_i))}{|T_i|}.$$

现在，我们计算单次试验中所有事件均成功的概率。综合成功概率为 $(1-\zeta^2)_{32}^{1/\zeta+1}$. 我们有 $(1-\zeta^2)_{32}^{1/\zeta+1} > e^{\zeta}_{1-\zeta} \geq 0.025$ 当 $\zeta \in (0, 1/12)$。这里，第一个不等式直接应用了 $\ln(1-x) > -\frac{x}{1-x}$，第二个不等式则利用了该函数单调递减这一事实。因此，每次试验的失败概率小于 0。975。由于我们执行$\lceil \frac{\log(\delta/k)}{\log(0.975)} \rceil$次运行，因此总体成功概率大于 $1 - 0.975^{\frac{\log(\delta/k)}{\log(0.975)}} = 1 - \frac{\delta}{k}$. □

现在我们来评估选定中心所产生的聚类代价。引理 2.3 在此分析中起着核心作用——它量化了由于噪声标签和抽样方差，选定中心可能偏离真实簇中心的程度，以及这一误差如何转化为总体聚类代价。

~**引理 2.3**。对于每个预测簇 X_i，有：

$$\texttt{Cost}(X_i^*, \hat{c}_i) \leqslant \left(1 + \frac{6\alpha - 4\alpha^2 + \alpha\epsilon}{(1-\alpha)(1-2\alpha)}\right) \texttt{Cost}(X_i^*, c_i^*).$$

证明。分析的一个关键环节，是将选定中心$\hat{c}_i$与真实中心 c_i^* 以及正确标记点的中位数 Med(T_i)联系起来。我们首先将成本拆分为两部分，

$$\texttt{Cost}(X_i^*, \hat{c}_i) = \texttt{Cost}(T_i, \hat{c}_i) + \texttt{Cost}(X_i^* \backslash T_i, \hat{c}_i) \quad (7)$$

我们先关注“Cost$(T_i,\ \hat{c}_i)$”这一项。为证明该等式，我们先通过将集合T_i 和 $N_i(\hat{c}_i)$拆分为三个互不相交的划分，计算将 T_i 中的点分配给$\hat{c}_i$ 所产生的额外成本：$S_1 = T_i \cap N_i(\hat{c}_i)$ $S_2 = T_i \setminus N_i(\hat{c}_i)$ $S_3 = N_i(\hat{c}_i) \setminus T_i$. 这意味着 $T_i = S_1 \cup S_2$ 和 $N_i(\hat{c}_i) = S_1 \cup S_3$。因此，Cost$(T_i,\ \hat{c}_i)$ − Cost$(T_i,\ Med(T_i))$ 可以写作

$$\begin{aligned}
&\texttt{Cost}(T_i, \hat{c}_i) - \texttt{Cost}(T_i, \texttt{Med}(T_i)) \\
&= [\texttt{Cost}(S_1, \hat{c}_i) + \texttt{Cost}(S_2, \hat{c}_i)] \\
&\quad - [\texttt{Cost}(S_1, \texttt{Med}(T_i)) + \texttt{Cost}(S_2, \texttt{Med}(T_i))] \\
&= [\texttt{Cost}(S_1, \hat{c}_i) + \texttt{Cost}(S_3, \hat{c}_i)] \\
&\quad - [\texttt{Cost}(S_1, \texttt{Med}(T_i)) + \texttt{Cost}(S_3, \texttt{Med}(T_i))] \\
&\quad - [\texttt{Cost}(S_3, \hat{c}_i) - \texttt{Cost}(S_3, \texttt{Med}(T_i))] \\
&\quad + [\texttt{Cost}(S_2, \hat{c}_i) - \texttt{Cost}(S_2, \texttt{Med}(T_i))] \\
&= [\texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i) - \texttt{Cost}(N_i(\hat{c}_i), \texttt{Med}(T_i))] \\
&\quad - [\texttt{Cost}(N_i(\hat{c}_i) \setminus T_i, \hat{c}_i) - \texttt{Cost}(N_i(\hat{c}_i) \setminus T_i, \texttt{Med}(T_i))] \\
&\quad + [\texttt{Cost}(T_i \setminus N_i(\hat{c}_i), \hat{c}_i) - \texttt{Cost}(T_i \setminus N_i(\hat{c}_i), \texttt{Med}(T_i))]
\end{aligned} \quad (8)$$

~我们还有 $|T_i \setminus N_i(\hat{c}_i))| \leqslant \alpha|X_i|$ 和 $|N_i(\hat{c}_i)) \setminus T_i| \leqslant \alpha|X_i|$。因此，我们可以求得 Cost$(T_i \setminus$ 的一个上界

$N_i(\hat{c}_i)$，$\hat{c}_i$)–根据三角不等式，Cost($T_i \setminus N_i(\hat{c}_i)$，Med($T_i$)) 的上界为

$$\begin{aligned}&\text{成本}(T_i \setminus N_i(\hat{c}_i),\hat{c}_i) \ -\text{成本}(T_i \setminus N_i(\hat{c}_i),\text{Med}(T_i))\\ &\leqslant |T_i \setminus N_i| \times ||\text{Med}(T_i) - \hat{c}_i||_2 \leqslant \alpha|\tilde{X}_i|||\text{Med}(T_i) - \hat{c}_i||_2.\end{aligned} \tag{9}$$

我们可以通过三角不等式得到不等式$\|\text{Med}(T_i) - \hat{c}_i\|_2 \leqslant \frac{(2+\alpha\epsilon)\text{Cost}(X_i^*, c_i^*)}{(1-2\alpha)|\tilde{X}_i|}$（详细推导见论文完整版。）。然后我们有不等式(9) $\leqslant \frac{(2\alpha+\alpha^2\epsilon)\text{Cost}(X_i^*, c_i^*)}{1-2\alpha}$ 类似地，我们得到

$$\begin{aligned}&\text{Cost}(N_i(\hat{c}_i) \setminus T_i, \hat{c}_i) - \text{Cost}(N_i(\hat{c}_i) \setminus T_i, \text{Med}(T_i))\\ &\leqslant \frac{(2\alpha + \alpha^2\epsilon)\text{Cost}(X_i^*, c_i^*)}{1-2\alpha}.\end{aligned} \tag{10}$$

现在我们为成本($N_i(\hat{c}_i),\hat{c}_i$) –成本($N_i(\hat{c}_i)$,Med(T_i)). 找到一个上界。由于我们的贪心选择，我们有成本($N_i(\hat{c}_i)$, $\hat{c}_i$)$\leqslant$ 成本 ($N_i(q)$ q)$\leqslant$成本 ($N_i(\hat{c}_i)$ q)。因此

$$\begin{aligned}&\text{Cost}(N_i(\hat{c}_i), \hat{c}_i) - \text{Cost}(N_i(\hat{c}_i), \text{Med}(T_i))\\ &\leqslant \text{Cost}(N_i(\hat{c}_i), q) - \text{Cost}(N_i(\hat{c}_i), \text{Med}(T_i))\\ &\leqslant |(1-\alpha)|\tilde{X}_i|| \times ||q - \text{Med}(T_i)||_2.\end{aligned} \tag{11}$$

应用引理 2.2，我们可以得到不等式 (11) $\leqslant\alpha\epsilon \times$ 成本 (T_i，Med(T_i))。将不等式 (9)、不等式 (10) 和不等式 (11) 合并，可得到方程 (7) 左侧的以下界

$$\begin{aligned}&\text{Cost}(T_i, \hat{c}_i) - \text{Cost}(T_i, c_i^*)\\ &\leqslant \text{Cost}(T_i, \hat{c}_i) - \text{Cost}(T_i, \text{Med}(T_i))\\ &\leqslant \frac{(4\alpha + \epsilon\alpha)\text{Cost}(X_i^*, c_i^*)}{1-2\alpha}.\end{aligned} \tag{12}$$

接下来，我们考察方程 (7) 中的第二项“成本($X_i^* \setminus T_i$，$\hat{c}_i$)”。根据三角不等式

$$\begin{aligned}\text{Cost}(X_i^* \setminus T_i, \hat{c}_i) &\leqslant \text{Cost}(X_i^*/T_i, c_i^*)\\ &+ |X_i^* \setminus T_i| \times ||\hat{c}_i - c_i^*||_2,\end{aligned} \tag{13}$$

随后，我们对 $|X_i^* \setminus T_i|$ 进行界定。由定义 1.1 可知，$(1-\alpha)|X_i^*| \leqslant |T_i|$，$|T_i| \leqslant |X_i|$，因此，我们可以将 $|X_i^* \setminus T_i|$ 界定为

$$|X_i^* \setminus T_i| = |X_i^*| - |T_i| \leqslant \alpha|X_i^*| \leqslant \frac{\alpha|\tilde{X}_i|}{1-\alpha}. \tag{14}$$

结合 (13)、(14) 和不等式 $\|\hat{c}_i - c_i^*\|_2 \leqslant \frac{(2+\alpha\epsilon)\text{Cost}(X_i^*, c_i^*)}{(1-2\alpha)|\tilde{X}_i|}$，我们在论文完整版中得到

$$\begin{aligned}&\text{Cost}(X_i^* \setminus T_i, \hat{c}_i) - \text{Cost}(X_i^* \setminus T_i, c_i^*)\\ &\leqslant \frac{\alpha(2 + \alpha\epsilon)\text{Cost}(X_i^*, c_i^*)}{(1-\alpha)(1-2\alpha)}.\end{aligned} \tag{15}$$

现在，我们推导最终的近似保证。结合不等式 (12) 和不等式 (15)，可得

$$\begin{aligned}\text{Cost}(X_i^*, \hat{c}_i) &= \text{Cost}(T_i, \hat{c}_i) + \text{Cost}(X_i^*/T_i, \hat{c}_i)\\ &\leqslant \left(1 + \frac{6\alpha - 4\alpha^2 + \epsilon\alpha}{(1-\alpha)(1-2\alpha)}\right)\text{Cost}(X_i^*, c_i^*).\end{aligned}$$

□

现在，我们正式证明定理 2.1，确立其近似比和运行时间复杂度。

定理*2.1*的证明. 我们的第一步是计算该算法的近似比。在每个簇中，根据引理 2.3，我们得到

$$\text{Cost}(X_i^*, \hat{c}_i) \leqslant \left(1 + \frac{6\alpha - 4\alpha^2 + \epsilon\alpha}{(1-\alpha)(1-2\alpha)}\right)\text{Cost}(X_i^*, c_i^*).$$

因此，对于整个实例，我们有

$$\begin{aligned}&\sum_{i\in[k]} \text{Cost}(X_i^*, \{\widehat{c}_j\}_{j=1}^k)\\ &\leqslant \left(1 + \frac{6\alpha - 4\alpha^2 + \epsilon\alpha}{(1-\alpha)(1-2\alpha)}\right)\sum_{i\in[k]} \text{Cost}(X_i^*, c_i^*).\end{aligned}$$

我们现在评估该算法的时间复杂度。这包括分析通过采样和网格离散化生成的候选中心集的规模，以及评估所有候选中心所产生的成本。

首先，我们计算候选中心集的规模。我们最终构造的候选中心集的规模为

$$\begin{aligned}&O\left(|R|\left(\frac{1}{(\alpha\epsilon)^4}\log\frac{1}{(\alpha\epsilon)}\right)^{O(|R|)}\log\frac{k}{\delta}\right)\\ &= O\left(2^{O(\frac{1}{(\alpha\epsilon)^3}\log^2\frac{1}{(\alpha\epsilon)})}\log\frac{k}{\delta}\right) \leqslant O\left(2^{O(\frac{1}{(\alpha\epsilon)^4})}\log\frac{k}{\delta}\right).\end{aligned}$$

对于候选中心集中的每个候选点，计算其成本所需的~时间为 $|X_i|d$。因此，该算法的总体时间复杂度为

$$\sum_{i\in[k]} 2^{O(\frac{1}{(\alpha\epsilon)^4})}|\tilde{X}_i|d\log\frac{k}{\delta} = 2^{O(\frac{1}{(\alpha\epsilon)^4})}nd\log\frac{k}{\delta}.$$

接下来，我们分析该算法的成功概率。正如我们在引理 2.2 中所得到的，该算法在每个簇中的成功概率为 $1-\frac{\delta}{k}$，因此，根据并集界，该算法的总~体成功概率 $\geqslant 1 - k \times \frac{\delta}{k} = 1 - \delta$ □。

3 实验

我们在真实世界数据集上评估了我们的算法。实验在一台配备 Intel(R) Xeon(R) Gold 6154 CPU 和 1024GB 随机存取存储器的服务器上进行。对于所有实验，我们报告了 10 次独立运行的平均聚类代价及其标准差。

数据集。遵循 Nguyen、Chaturvedi 和 Nguyen（2023）、Ergun 等人（2022）以及 Huang 等人（2025）的研究，我们使用一系列错误率 α，在 CIFAR-10 (n =50，000，d = 3，072)（Krizhevsky 和 Hinton 2009）、PHY (n = 10，000，d = 50)（KDD 2004）和 MNIST (n =1，797，$d = 64$)（Deng 2012）数据集上评估了我们的算法。此外，我们还在另一个高维数据集 Fashion-MNIST ($n = 60000$，$d = 784$)（Xiao、Rasul 和 Vollgraf 2017）上评估了算法的性能

预测器生成与误差模拟 为了评估我们的算法，我们首先使用以 KMedoids++ 初始化的 Lloyd 算法，为每个数据集计算真实划分（记作 KMed++）。随后，我们以错误率 α 随机选取每个真实簇中的 α 比例的点，并将它们重新分配到随机选取的簇中，从而生成受损划分（记作预测器）。为确保公平比较，在任意给定错误率 α 下，每种算法都在完全相同的一组受损标签上进行测试。

算法在我们的实验中，我们评估了所提出的采样与搜索算法。我们将其性能与其他最先进的学习增强方法进行比较，包括 Ergun 等人（2022）提出的算法（记为 EFS+）、Nguyen、Chaturvedi 和 Nguyen（2023）提出的方法（记为 NCN），以及 Huang 等人（2025）的最新工作（记为 HFH+）。正如 Nguyen、Chaturvedi 和 Nguyen（2023）所指出的，实践中通常无法获知真实错误率 α，因此需要搜索其最优值。为确保比较公平，我们对所有评估的算法采用统一的超参数调优策略。具体而言，我们遍历 α 的 10 个候选值，这些值从区间 $[0.01, 0.5]$ 内均匀间隔的点中选取。对于每种方法，我们选择使所得 k-中位数聚类代价最小的 α，以生成最终输出。为了根据真实标签评估最终的聚类质量，我们还在论文完整版中报告了调整兰德指数（*ARI*）和归一化互信息（*NMI*）。

结果我们在表2和表3中，将我们的算法与几种基线方法进行了比较评估。表2详细展示了在 Fashion-MNIST（$n = 60000$，$d = 784$）上的性能，其中 $k = 10$ 固定，取值涵盖一系列 α。表3展示了在 PHY（$n = 10,000$，$d = 50$）上的结果，其中 $\alpha = 0.2$ 固定，并采用了 k的多种取值。两组结果均表明，与所有竞争方法相比，我们的算法明显更快，同时通常能够获得更好的近似质量。关于其他数据集的补充实验以及更详细的结果展示，请参阅补充材料。在这些数据集上，我们的算法在运行时间和成本方面也展现出显著优势。

4 结论与未来工作

本文研究学习增强的 k-中位数聚类问题。我们首先介绍一种基于简单而有效的采样方法的算法，然后从理论上研究其质量保证，最后开展一系列实验，与其他学习增强的 k-中位数算法进行比较。理论和实验结果均表明，与现有方法相比，我们的方法效率更高，且达到了最先进的近似比。在此基础上，未来还有若干机会可以从理论和实践两个角度进一步改进我们的方法。例如，能否通过减轻或消除对 ϵ的指数级依赖，进一步降低算法的时间复杂度？能否在不牺牲时间复杂度显著增加？此外，能否针对流式模型设计一种学习增强聚类算法，以更有效地处理大规模数据?

条件		Cost		时间（秒）	
α	方法	Avg.	标准差	Avg.	标准差
0	KMed++	8.4054e+07	-	-	-
0.1	预测器	8.4259e+07	-	-	-
	EFS+	8.4050e+07	115.17	270.47	12.97
	HFH+	8.4049e+07	834.92	749.72	18.47
	NCN	8.4050e+07	181.80	272.22	4.37
	Ours	**8.4048e+07**	933.64	**47.37**	0.78
0.2	预测器	8.4935e+07	-	-	-
	EFS+	8.4057e+07	287.83	283.13	24.07
	HFH+	8.4053e+07	1598.91	751.66	25.42
	NCN	8.4057e+07	309.52	282.97	13.78
	Ours	**8.4052e+07**	961.03	**47.96**	3.33
0.3	预测器	8.6223e+07	-	-	-
	EFS+	8.4076e+07	467.75	282.57	8.66
	HFH+	8.4065e+07	3527.00	751.13	22.67
	NCN	8.4077e+07	695.49	299.50	22.54
	Ours	**8.4062e+07**	3848.20	**45.38**	1.33
0.4	预测器	8.8209e+07	-	-	-
	EFS+	8.4109e+07	631.67	297.27	13.66
	HFH+	8.4101e+07	11512.25	758.16	29.89
	NCN	8.4111e+07	1206.20	302.45	25.71
	Ours	**8.4100e+07**	12684.42	**45.29**	2.18
0.5	预测器	9.0897e+07	-	-	-
	EFS+	8.4150e+07	1320.47	304.67	11.82
	HFH+	8.4148e+07	12671.01	751.08	26.53
	NCN	8.4152e+07	2503.70	305.95	10.98
	Ours	**8.4145e+07**	17385.62	**47.87**	2.04

表2：在 Fashion-MNIST数据集上，当 $k = 10$和 α变化时的性能比较。

条件		Cost		时间（秒）	
k	方法	平均值	标准差	Avg.	标准差
10	KMed++	2.0224e+08	-	-	-
	预测器	2.0427e+08	-	-	-
	EFS+	2.0204e+08	4444.80	362.31	52.30
	HFH+	2.0147e+08	105661.96	42.33	1.91
	NCN	2.0163e+08	109131.95	160.15	14.33
	Ours	**2.0134e+08**	82812.18	**20.72**	0.59
30	KMed++	8.4404e+07	-	-	-
	预测器	8.5018e+07	-	-	-
	EFS+	8.4490e+07	721.59	294.21	54.13
	HFH+	8.4404e+07	4372.98	42.26	2.05
	NCN	8.4480e+07	14266.80	221.81	49.30
	Ours	**8.4404e+07**	3043.38	**27.08**	3.26
50	KMed++	6.2758e+07	-	-	-
	预测器	6.3111e+07	-	-	-
	EFS+	6.2796e+07	503.13	285.51	22.48
	HFH+	6.2755e+07	1072.71	44.69	1.34
	NCN	6.2791e+07	5662.71	208.89	29.26
	Ours	**6.275456e+07**	677.03	**36.87**	0.73

表3：在 PHY数据集上对固定$\alpha = 0.2$和变化的 k进行性能比较。

5 致谢

本研究部分得到了国家自然科学基金委员会资助，项目编号为62432016和62272432；国家重点研发计划资助，项目编号为2021YFA1000900；以及安徽省自然科学基金资助，项目编号为2208085MF163。作者还要感谢匿名审稿人提出的宝贵意见和建议。

参考文献

Antoniadis, A.；Gouleakis, T.；Kleer, P.；以及 Kolev, P. 2020年。基于机器学习建议的秘书与在线匹配问题。载于第*34*届神经信息处理系统国际会议论文集，NIPS '20。美国纽约州雷德胡克：Curran Associates Inc. ISBN9781713829546。

Arora, S.；Raghavan, P.；以及 Rao, S. 1998年。欧几里得 k-中位数及相关问题的近似方案。载于第三十届 *ACM* 计算理论研讨会年会论文集，STOC '98，106–113。美国纽约州纽约市：计算机协会。ISBN 0897919629。

Arthur, D.；以及 Vassilvitskii, S. 2007年。k-均值++：精细初始化的优势。载于第十八届 *ACM-SIAM* 离散算法研讨会年会论文集，SODA '07，1027–1035。美国：工业与应用数学学会。ISBN 9780898716245。

Arya, V.；Garg, N.；Khandekar, R.；Meyerson, A.；Muna- gala, K.；以及 Pandit, V. 2001年。用于 k-中位数和设施选址问题的局部搜索启发式算法。载于第三十三届 *ACM* 计算理论研讨会年会论文集，STOC '01，21–29。美国纽约州纽约市：计算机协会。ISBN 1581133499。

Badoiu, M.；Har-Peled, S.；以及 Indyk, P. 2002年。通过核心集进行近似聚类。载于第三十四届 *ACM* 计算理论研讨会年会论文集，STOC '02，250–257。美国纽约州纽约市：计算机协会。ISBN 1581134959。

Caron, M.；Misra, I.；Mairal, J.；Goyal, P.；Bojanowski, P.；以及 Joulin, A. 2020年。通过在视图之间交换分配进行视觉特征的无监督学习。载于神经信息处理系统进展，第33卷，9912–9924。

Charikar, M.；和 Li, S. 2012年。针对 k-中位数问题的相关 LP 舍入方法。载于第 *39* 届国际自动机、语言与程序设计研讨会论文集——第 *I* 卷，ICALP'12，194–205。柏林、海德堡：Springer-Verlag。ISBN 9783642315930。

Cohen-Addad, V.；和 C.S.Karthik。2019。Lp 度量中的聚类不可近似性。载于*2019 IEEE* 第 *60* 届计算机科学基础年会（*FOCS*），519–539。

Cohen-Addad, V.；Feldmann, A. E.；以及 Saulpic, D. 2021年。倍增度量空间中的近线性时间聚类近似方案。《*ACM* 杂志》，68(6)。

Deng, L. 2012年。用于机器学习研究的 MNIST 手写数字图像数据库 [网络精选]。《*IEEE* 信号处理杂志》，29(6)：141–142。Dinitz, M.；Im, S.；Lavastida, T.；Moseley, B.；以及 Vassil- vitskii, S. 2021年。通过学习得到的对偶变量实现更快的匹配。在第 *35* 届神经信息处理系统国际会议论文集，NIPS '21。美国纽约州雷德胡克：Curran Associates Inc. ISBN 9781713845393。Ergun, J. C.；Feng, Z.；Silwal, S.；Woodruff, D.；以及 Zhou, S. 2022年。学习增强的 k-means 聚类。在学习表征国际会议。

Gamlath, B.；Lattanzi, S.；Norouzi-Fard, A.；和 Svensson, O. 2022。基于噪声标签的近似簇恢复。载于 Loh, P.-L.；和 Raginsky, M. 编，第三十五届学习理论会议论文集，第 178 卷《机器学习研究论文集》，1463–1509。PMLR。

Ghaffari, M.；Mosavi, A.；与 Shamshirband, S. 2021. 面向机器人检测的社交网络用户行为聚类与高维表示。见*2021*年万维网会议配套论文集，19–22。

Hsu, C.-Y.；Indyk, P.；Katabi, D.；与 Vakilian, A. 2018. 基于学习的频率估计算法。见 学习表征国际会议。

Huang, J.；Feng, Q.；Huang, Z.；Zhang, Z.；Xu, J.；与 Wang, J. 2025. 学习增强k均值问题的新算法。见第十三届学习表征国际会议。

Inaba, M.；Katoh, N.；与 Imai, H. 1994. 加权沃罗诺伊图与随机化在基于方差的k-聚类中的应用：（扩展摘要）。见第十届计算几何年度研讨会论文集，SCG '94，332–339。美国纽约州纽约市：美国计算机协会。ISBN0897916484。

KDD. 2004. KDD杯. https://osmot.cs.cornell.edu/kddcup/index.html.

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K. N.; Reik, W.; Barahona, M.; Green, A. R.; 和 Hemberg, M. 2017. SC3：单细胞RNA测序数据的一致性聚类。《自然·方法》，14(5)：483–486。

Krizhevsky, A.; 和 Hinton, G. 2009. 从微型图像中学习多层特征。技术报告0，多伦多大学，多伦多，安大略。

Kumar, A.; Sabharwal, Y.; 和 Sen, S. 2010. 任意维度聚类问题的线性时间近似方案。《*ACM*期刊》，57(2)。

Lykouris, T.; and Vassilvitskii, S. 2021. 基于机器学习建议的竞争性缓存。《*J. ACM*》，68(4)。

Mitzenmacher, M.; and Vassilvitskii, S. 2022. 带预测的算法。《*ACM*通讯》，65(7)：33–35。Nguyen, T. D.; Chaturvedi, A.; and Nguyen, H. 2023年。改进的学习增强算法：用于 k-均值和 k-中位数聚类。发表于第十一届国际学习表征会议。

Ostrovsky, R.; Rabani, Y.; Schulman, L. J.; and Swamy, C. 2013. Lloyd 型方法对 k-均值问题的有效性。《*J. ACM*》，59(6)。
Roughgarden，T. 2021。超越算法的最坏情况分析。剑桥大学出版社。
Xiao，H.；Rasul，K.；以及 Vollgraf，R. 2017。Fashion-MNIST：用于评测机器学习算法的新型图像数据集。*arXiv*，abs/1708.07747。

一个缺失的证明：k-中位数

命题 A.1。 两个距离 $||\hat{c}_i - \texttt{Med}(T_i)||_2$ 和 $||\hat{c}_i - c_i^*||_2$ 均不大于 $\frac{(2+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*)}{(1-2\alpha)|\tilde{X}_i|}$。

证明。首先，由于采用了贪心搜索方法，我们可以轻松建立 $\texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i)$ 与 $\texttt{Cost}(T_i, \texttt{Med}(T_i))$ 之间的不等式关系，得到

$$\texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i) \leqslant \texttt{Cost}(N_i(q), q).$$

由于 $N_i(q)$ 是从 $\hat{c}_i$ 中选出的、大小为 $(1-\alpha)|X_i|$ 的最近点集合，因此我们也有

$$\texttt{Cost}(N_i(q), q) \leqslant \texttt{Cost}(T_i, q).$$

我们有

$$\texttt{Cost}(T_i, q) \leqslant (1+\alpha\epsilon)\texttt{Cost}(T_i, \texttt{Med}(T_i)).$$

由上述不等式可得

$$\texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i) \leqslant (1+\alpha\epsilon)\texttt{Cost}(T_i, \texttt{Med}(T_i)). \quad (16)$$

根据学习增强的定义，$|T_i| \geqslant (1-\alpha)n_i$，并且由于 $|N_i(\hat{c}_i)| = (1-\alpha)n_i$ and $N_i(\hat{c}_i)$ $T_i \subseteq X_i$，我们可以推导出

$$\begin{aligned}|N_i(\hat{c}_i) \cap T_i| &\geqslant |N_i(\hat{c}_i)| - |\tilde{X}_i \backslash T_i| \geqslant (1-\alpha-\alpha)|\tilde{X}_i| \\ &= (1-2\alpha)|\tilde{X}_i|. \quad (17)\end{aligned}$$

然后，根据三角不等式和不等式 (17)，可得

$$\begin{aligned}&\sum_{p \in N_i(\hat{c}_i) \cap T_i} ||p - \hat{c}_i||_2 + ||p - \texttt{Med}(T_i)||_2 \quad (18)\\ &\geqslant |N_i(\hat{c}_i) \cap T_i| \times ||\hat{c}_i - T_i||_2 \\ &\geqslant (1-2\alpha)|\tilde{X}_i| \times ||\hat{c}_i - \texttt{Med}(T_i)||_2. \quad (19)\end{aligned}$$

此外，由不等式 (16) 可得

$$\begin{aligned}\sum_{p \in N_i(\hat{c}_i) \cap T_i} ||p - \hat{c}_i||_2 &\leqslant \texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i) \\ &\leqslant (1+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*) \quad (20)\end{aligned}$$

$$\sum_{p \in N_i(\hat{c}_i) \cap T_i} ||p - \texttt{Med}(T_i)||_2 \leqslant \texttt{Cost}(X_i^*, c_i^*) \quad (21)$$

综合不等式（18）、（20）和（21），我们得到

$$(2+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*) \geqslant (1-2\alpha)|\tilde{X}_i|||\hat{c}_i - \texttt{Med}(T_i)||_2, \quad (22)$$

这直接意味着

$$||\hat{c}_i - \texttt{Med}(T_i)||_2 \leqslant \frac{(2+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*)}{(1-2\alpha)|\tilde{X}_i|}. \quad (23)$$

至此，我们完成了第一个不等式的证明。

第二个不等式的证明与之类似，如同不等式 (18)，我们有

$$\begin{aligned}&\sum_{p \in N_i(\hat{c}_i) \cap X_i^*} ||p - \hat{c}_i||_2 + ||p - \texttt{Med}(X_i^*)||_2 \\ &\geqslant |N_i(\hat{c}_i) \cap X_i^*| \times ||\hat{c}_i - \texttt{Med}(X_i^*)||_2 \\ &\geqslant (1-2\alpha)|X_i| \times ||\hat{c}_i - \texttt{Med}(X_i^*)||_2, \quad (24)\end{aligned}$$

第二个不等式来自 $T_i \subseteq X_i^*$，因此 $|N_i(\hat{c}_i) \cap X_i^*| \geqslant |N_i(\hat{c}_i) \cap T_i^*|$. 因为

$$\begin{aligned}\sum_{p \in N_i(\hat{c}_i) \cap X_i^*} ||p - \hat{c}_i||_2 &\leqslant \texttt{Cost}(N_i(\hat{c}_i), \hat{c}_i) \\ &\leqslant (1+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*)\end{aligned}$$

$$\sum_{p \in N_i(\hat{c}_i) \cap X_i^*} ||p - \texttt{Med}(X_i^*)||_2 \leqslant \texttt{Cost}(X_i^*, c_i^*),$$

我们有

$$||\hat{c}_i - c_i^*||_2 \leqslant \frac{(2+\alpha\epsilon)\texttt{Cost}(X_i^*, c_i^*)}{(1-2\alpha)|\tilde{X}_i|}.$$

这就是我们要证明的第二个不等式。 □

B k-均值算法

记号。 对于 k-均值，给定任意点集 C，从点 p 到其在 C 中的最近点的距离记为 $\text{dist}_2(p, C)$。特别地，当 C 是点集 X 的给定中心点集时，相应的成本记为 $\text{Cost}_2(X, C)$，定义为 $\text{Cost}_2(X, C) = \sum_{x \in X} \text{dist}_2(x, C)$。 对于任意点集 P，我们用 $\text{Cen}(P)$ 表示其**均值点**，即，

$$\texttt{Cen}(P) = \arg\min_{q \in \mathbb{R}^d} \sum_{p \in P} \|p - q\|_2^2.$$

在本节中，我们将采样与搜索算法扩展到学习增强的 k-均值问题。我们的方法仍分为三个阶段，但在两个方面不同于算法 1：我们修改了采样点的数量以及构建候选中心集的方法。具体而言，我们首先采样一个常数规模的数据点集合，然后以关于样本规模呈指数的时间构造候选中心集，最后以关于数据集规模呈线性的时间识别局部最优中心。

B.1 我们提出的算法与主要定理该算法的详细实现在算法 3 中描述。表 4 对学习增强的 k-均值算法的结果进行了详细比较。下面我们给出本算法的主要理论结果。

定理 B.1. 算法 *3* 是一种用于学习增强 k-均值聚类的算法。给定数据集 $X \in \mathbb{R}^{n \times d}$ 以及错误率为 $\alpha < \frac{1}{2}$ 的划分 $(X_1, \ldots, X_k)$，该算法输出近似比为 $1 + \frac{\alpha}{1-\alpha} + \frac{4\alpha+\alpha\epsilon}{1-2\alpha}$ 的解，并以常数概率在 $O(2^{O(1/\varepsilon)} nd \log k)$ 时间复杂度内完成。

表 4：学习增强型 k-均值算法的比较结果

方法与参考文献	近似比	标签误差范围	时间复杂度
Ergun 等人（2022）	$1+20\alpha$	$[\frac{10\log m}{\sqrt{m}}\ 1/7]$	$O(nd\log\ n)$
Nguyen、Chaturvedi 和 Nguyen（2023）	$1+\frac{\alpha}{1-\alpha}+\frac{4\alpha}{(1-2\alpha)(1\ -\alpha)}$	$[0,1/2)$	$O(nd\log\ n)$
Huang 等人（2025）	$1+\frac{\alpha}{1-\alpha}+\frac{4\alpha+\alpha\epsilon}{(1-2\alpha)(1-\alpha)}$	$[0,1/2)$	$O(\epsilon^{-1/2}nd\log(kd))$
Huang 等人（2025）	$1+\frac{\alpha}{1-\alpha}+\frac{12\alpha-18\alpha^2}{(1-3\alpha-\epsilon)(1-2\alpha-\epsilon)}$	$(0,\ 1/3-\epsilon)$	$O(nd)\ +\tilde{O}(\epsilon^{-5}kd)$
采样与搜索（我们的方法）	$1+\frac{\alpha}{1-\alpha}+\frac{4\alpha+\alpha\epsilon}{(1-2\alpha)(1-\alpha)}$	$[0,1/2)$	$O(2^{O(1/\epsilon)}nd$对数 $k)$

算法 3：用于学习增强k-均值的采样与搜索 **1：输入：一个 k-均值实例（X,k,d）、一组错误率为 α 的划分（$X_1\dots X_k$），以及参数 $\epsilon\in(0,1)$ 2：输出：一组具有 $|C|=k$ 的中心集合 $C\in\mathbb{R}^d$。3：** $C\leftarrow\{\}$ **4：对于 $i\in[k]$ 执行 5：** $C_i\leftarrow\{\}$ **6：对于 $j=1$ 到 $\lceil^{\log(\delta/k)}\log\ _{0.75}\rceil$ 执行 7：从 $\tilde{X}_i$ 中随机且独立地采样一个大小为 $\lceil\ \ ^4(1-\alpha)\epsilon\rceil$ 的集合 R_i^j 8：对于 R_i^j 的每一个 $\lceil\ \ ^1(1-\alpha)\epsilon\rceil$ 子集 R 执行 9：** $C_i=C_i\cup$**Cen$_2$(R) 10：结束循环 11：结束循环 12：对于每个 $c\in C_i$，将 $N_i(c)$ 定义为 X_i 中最接近 c 的 $\lceil(1-\alpha)m_i\rceil$ 个点的集合。13：** $\hat{c}_i=$**arg min**$_{c\in C'}$**成本$_2$($N_i(c)$ c)** ^ ^**14：** $C=C\cup$**Cen($N_i(\hat{c}_i)$) 15：结束循环 ^16：返回** C

B.2 定理 B.1 的证明

我们首先介绍 k 均值聚类领域中两个广为人知且应用广泛的结果。

命题 B.2。设 $X\subseteq\mathbb{R}^d$ 是一组 n 点，且 $c\in\mathbb{R}^d$。那么：

$$\text{Cost}_2(X,c)=\text{Cost}_2(X,\text{Cen}_2(X))+n\cdot\|c-\text{Cen}_2(X)\|_2^2.$$

命题 B.3。对于集合 $X\subseteq\mathbb{R}^d$ 的任意划分 $X_1\cup X_2$，其中 X 的大小为 n，若 $|X_1|\geq(1-\lambda)n$，则：

$$\|\texttt{Cen}_2(X),\texttt{Cen}_2(X_1)\|_2^2\leq\frac{\lambda}{(1-\lambda)n}\texttt{Cost}_2(X,\texttt{Cen}_2(X)).$$

我们还介绍了一个关于欧几里得空间中几何均值点的重要命题，这对后续证明至关重要。

命题 B.4。（*Inaba*、*Katoh* 和 *Imai*（*1994*））设 S 是从点集 P 中均匀随机且独立采样 m 个点得到的一组 m 点。那么，对于任意 $\delta>0$，

$$||\text{Cen}_2(P)-\text{Cen}_2(T_i)||_2\leqslant\frac{\epsilon\texttt{Cost}_2(T_i,c(T_i))}{|T_i|}.$$

以至少 $1-\delta$ 的概率成立。

命题 B.4 表明，如果从 P 中随机采样足够多的点，那么采样点的质心以较高概率接近 Med(P)，这使我们能够直接利用采样点的质心构造候选中心集。

我们将证明分为三个主要步骤：首先，我们证明，对于每个预测簇，以高概率，构造出的候选中心集至少包含一个点，该点接近预测簇中正确标记子集的真实中位数。引理 B.5 对此进行了形式化，并利用了命题 B.4 中关于我们采样设计的几何性质。其次，我们分析从候选集中选出的中心的成本。在引理 B.7 中，我们通过仔细界定误分类点带来的额外成本以及贪心选择的最优性，证明尽管存在噪声标签，该中心仍能使聚类代价接近最优值。最后，我们汇总所有簇的界，得到总聚类代价，并分析候选集的大小和算法的运行时间。

引理 B.5。 对于每个预测簇 X_i，以概率 $1-\frac{\delta}{k}$，存在一个点 $q \in C'$ 满足：

$$||q - \texttt{Cen}_2(T_i)||_2 \leqslant \frac{\epsilon \texttt{Cost}_2(T_i, c(T_i))}{|T_i|}.$$

证明。 首先，在学习增强设置下，我们有

$$|T_i| \geqslant (1-\alpha)\max(|\tilde{X}_i|, |\tilde{X}_i|).$$

然后，由于我们从 $\tilde{X}_i$ 中采样了一个大小为 $\frac{4}{(1-\alpha)\epsilon}$ 的点集 Q_i，利用马尔可夫不等式可得，

$$|T_i \cap R_i^j| \geqslant \frac{2}{\epsilon}.$$

其概率至少为 $\frac{1}{2}$。由此可知，存在一个子集 R 满足

$$R \subset T_i, |R| = \frac{2}{\epsilon}.$$

根据命题 B.4，令 $P = T_i, S = R, \delta = \frac{1}{2}, m = \frac{2}{\epsilon}$，我们有

$$||q - \texttt{Cen}_2(T_i)||_2 \leqslant \frac{\epsilon \texttt{Cost}_2(T_i, c(T_i))}{|T_i|}.$$

其概率至少为 $\frac{1}{2}$。现在，我们计算单次试验中所有事件均成功的概率。综合成功概率为 $\frac{1}{4}$。由于我们运行 $\lceil \frac{\log(\delta/k)}{\log 0.75} \rceil$ 次，因此总体成功概率大于

$$1 - 0.75^{\frac{\log(\delta/k)}{\log 0.75}} = 1 - \frac{\delta}{k}.$$

□

现在，我们来评估选定中心产生的聚类代价。引理 B.7 在此分析中起着核心作用，它量化了由于噪声标签和抽样方差，选定中心可能偏离真实簇中心的程度，以及这种误差如何转化为总体聚类代价。我们首先证明以下论断。

断言 B.6 c_i 与 $\text{Cen}_2(T_i)$ 之间的距离满足以下不等式：

$$||\hat{c}_i - \text{Cen}_2(T_i)||_2^2 \leq \frac{4\alpha + \epsilon\alpha}{1-2\alpha} \frac{\text{Cost}_2(T_i, \text{Cen}_2(T_i))}{|T_i|}.$$

证明. 首先，根据我们的贪心搜索过程，有

$$\texttt{Cost}_2(N_i(\hat{c}_i), \hat{c}_i) \leqslant \texttt{Cost}_2(N_i(q), q). \tag{25}$$

由于 $N_i(q)$ 是距离 $\hat{c}_i$ 最近的 $(1-\alpha)|X_i|$ 个点的集合，因此我们也有

$$\texttt{Cost}_2(N_i(q), q) \leqslant \texttt{Cost}_2(T_i, q). \tag{26}$$

由引理 B.5 可直接得出

$$\texttt{Cost}_2(T_i, q) \leqslant (1+\epsilon)\texttt{Cost}_2(T_i, \texttt{Cen}_2(T_i)). \tag{27}$$

由于 $|T_i| \geqslant (1-\alpha)n_i$，由上文给出的不等式 (25)、(26) 和 (27) 可得

$$\frac{\texttt{Cost}_2(N_i(\hat{c}_i), \hat{c}_i)}{(1-\alpha)n_i} \leqslant (1+\epsilon)\frac{\texttt{Cost}_2(T_i, \texttt{Cen}_2(T_i))}{|T_i|}.$$

因此，我们有

$$\begin{aligned}&\texttt{Cost}_2(N_i(\hat{c}_i), \hat{c}_i)\\ &\leqslant (1+\epsilon)\frac{(1-\alpha)n_i}{|T_i|}\texttt{Cost}_2(T_i, \texttt{Cen}_2(T_i))),\end{aligned} \tag{28}$$

根据命题 B.3，若设 $C = N_i(\hat{c}_i)$ $C_1 = N_i(\hat{c}_i) \cap T_i$，众所周知

$$\frac{|C_1|}{C} \geqslant \frac{|N_i(\hat{c}_i)| - |X_i \backslash T_i|}{|N_i(\hat{c}_i)|} \geqslant \frac{1-2\alpha}{1-\alpha} = 1 - \frac{\alpha}{1-\alpha},$$

则可得

$$\begin{aligned}||\hat{c}_i - \texttt{Cen}_2(N_i(\hat{c}_i) \cap T_i)||_2^2 &\leqslant \frac{\frac{\alpha}{1-\alpha}}{1-\frac{\alpha}{1-\alpha}} \frac{\texttt{Cost}_2(N_i(\hat{c}_i), \hat{c}_i)}{|N_i(\hat{c}_i)|}\\ &= \frac{\alpha \texttt{Cost}_2(N_i(\hat{c}_i), \hat{c}_i)}{(1-2\alpha)(1-\alpha)n_i}.\end{aligned} \tag{29}$$

类似地，如果我们设定 $C = T_i$，$C_1 = N_i(\hat{c}_i) \cap T_i$，我们也可以界定 $\texttt{Cen}_2(N_i(\hat{c}_i) \cap T_i)$ 与 $\texttt{Cen}_2(T_i)$ 之间的距离，如下所示

$$||\texttt{Cen}_2(T_i) - \texttt{Cen}_2(N_i(\hat{c}_i) \cap T_i)||_2^2 \leqslant \frac{\alpha \text{Cost}(T_i, \texttt{Cen}_2(T_i)}{(1-2\alpha)|T_i|}, \tag{30}$$

根据不等式 (28)、(29) 和 (30)，我们能够界定 $\hat{c}_i$ 与 $\texttt{Cen}_2(T_i)$ 之间的距离

$$\begin{aligned}||\hat{c}_i - \texttt{Cen}_2(T_i)||_2^2 &\leqslant 2||\hat{c}_i - \texttt{Cen}_2(N_i(\hat{c}_i) \cap T_i)||_2^2\\ &\quad + 2||\texttt{Cen}_2(N_i(\hat{c}_i) \cap T_i) - \texttt{Cen}_2(T_i)||_2^2\\ &\leqslant 2\frac{\alpha \text{Cost}_2(N_i(\hat{c}_i), \hat{c}_i)}{(1-2\alpha)(1-\alpha)n_i}\\ &\quad + 2\frac{\alpha \text{Cost}_2(T_i, \texttt{Cen}_2(T_i)}{(1-2\alpha)|T_i|}\\ &= \frac{(4\alpha + 2\epsilon\alpha)\text{Cost}_2(T_i, \texttt{Cen}_2(T_i))}{(1-2\alpha)|T_i|}.\end{aligned}$$

□

现在我们来评估选定中心所产生的聚类代价。引理 B.7 在此分析中起着核心作用——它量化了由于噪声标签和抽样方差，选定中心可能偏离真实簇中心的程度，以及这种误差如何转化为整体聚类代价。

引理 B.7。 对于预测簇 X_i，我们有：

$$\begin{aligned}&\texttt{Cost}(X_i^*, \hat{c}_i)\\ &\leqslant \left(1 + \frac{\alpha}{1-\alpha} + \frac{4\alpha + \alpha\epsilon}{(1-2\alpha)(1-\alpha)}\right)\texttt{Cost}(X_i^*, c_i^*).\end{aligned}$$

证明。我们首先将成本$_2(X_i^*, c_i^*)$的计算分成两部分

$$\text{Cost}_2(X_i^*, c_i^*) = \text{Cost}_2(X_i^* \backslash T_i, c_i^*) + \text{Cost}_2(T_i, c_i^*). \tag{31}$$

根据命题 B.2，成本$_2(T_i,\ c_i^*)$可以写成

$$\begin{aligned}&\mathrm{Cost}_2(T_i, \mathtt{Cen}_2(T_i))\\&\quad+(1-\frac{|X_i^*\backslash T_i|}{|X_i^*|})|X_i^*|\times||c_i^*-\mathtt{Cen}_2(T_i)||_2^2.\end{aligned}\tag{32}$$

同理，我们有

$$\begin{aligned}\mathrm{Cost}_2(X_i^*/T_i, c_i^*)&=\mathrm{Cost}_2(X_i^*\backslash T_i, \mathtt{Cen}_2(X_i^*\backslash T_i))\\&\quad+\frac{|X_i^*\backslash T_i|}{|X_i^*|}|X_i^*|\times||c_i^*-\mathtt{Cen}_2(X_i^*\backslash T_i)||_2^2\end{aligned}\tag{33}$$

结合不等式 31、32 和 33，我们得到

$$\begin{aligned}\mathrm{Cost}_2(X_i^*, c_i^*)&=\mathrm{Cost}_2(X_i^*\backslash T_i, \mathtt{Cen}_2(X_i^*\backslash T_i))\\&\quad+\frac{|X_i^*\backslash T_i|}{|X_i^*|}|X_i^*|||c_i^*-\mathtt{Cen}_2(X_i^*/T_i)||_2^2\\&\quad+\mathrm{Cost}_2(T_i, \mathtt{Cen}_2(T_i))\\&\quad+(1-\frac{|X_i^*\backslash T_i|}{|X_i^*|})|X_i^*|||c_i^*-\mathtt{Cen}_2(T_i)||_2^2\\&=\frac{1-\frac{|X_i^*\backslash T_i|}{|X_i^*|}}{\frac{|X_i^*\backslash T_i|}{|X_i^*|}}|X_i^*|||c_i^*-\mathtt{Cen}_2(T_i)||_2^2\\&\quad+\mathrm{Cost}_2(X_i^*/T_i, \mathtt{Cen}_2(X_i^*/T_i))\\&\quad+\mathrm{Cost}_2(T_i, \mathtt{Cen}_2(T_i))\\&\geqslant\frac{1-\alpha}{\alpha}|X_i^*|||c_i^*-\mathtt{Cen}_2(T_i)||_2^2\\&\quad+\mathrm{Cost}_2(T_i, \mathtt{Cen}_2(T_i))\\&\geqslant\frac{1-\alpha}{\alpha}|X_i^*|||c_i^*-\mathtt{Cen}_2(T_i)||_2^2\\&\quad+\frac{1-2\alpha}{4\alpha+2\epsilon\alpha}\cdot(1-\alpha)|\tilde{X_i^*}|||\hat{c}_i-\mathtt{Cen}_2(T_i)||_2^2,\end{aligned}\tag{34}$$

最后，由柯西-施瓦茨不等式，我们得到

$$\begin{aligned}&(||c_i^*-\mathtt{Cen}_2(T_i)||_2+||\hat{c}_i-\mathtt{Cen}_2(T_i)||_2)^2\\&\leq\left(\frac{\alpha}{1-\alpha}+\frac{4\alpha+2\epsilon\alpha}{(1-2\alpha)(1-\alpha)}\right)\\&\quad\times(\frac{1-\alpha}{\alpha}|X_i^*|\times||c_i^*-\mathtt{Cen}_2(T_i)||_2^2\\&\quad+\frac{1-2\alpha}{4\alpha+2\epsilon\alpha}\cdot(1-\alpha)|X_i^*|\times||\hat{c}_i-\mathtt{Cen}_2(T_i)||_2^2)/|X_i^*|\\&\leq(\frac{\alpha}{1-\alpha}+\frac{4\alpha+2\epsilon\alpha}{(1-2\alpha)(1-\alpha)})\mathtt{Cost}_2(X_i^*, c_i^*)/|X_i^*|.\end{aligned}$$

这直接推出

$$\begin{aligned}&||\hat{c}_i-\mathtt{Cen}_2(X_i^*||_2^2\\&\leqslant\mathtt{Cost}_2(X_i^*, c_i^*)\left(\frac{\alpha}{1-\alpha}+\frac{4\alpha+2\epsilon\alpha}{(1-2\alpha)(1-\alpha)}\right)/|X_i^*|.\end{aligned}$$

根据命题 B.2，这与我们的证明目标等价。

□

现在，我们通过确立近似比和运行时间复杂度，正式证明定理 2.1。

定理 *B.1* 的证明。我们的第一步是计算该算法的近似比。在每个簇中，由引理 B.7，我们得到

$$\mathtt{Cost}_2(X_i^*, \hat{c}_i)\leqslant\left(1+\frac{\alpha}{1-\alpha}+\frac{4\alpha+\alpha\epsilon}{(1-2\alpha)(1-\alpha)}\right)\mathtt{Cost}_2(X_i^*, c_i^*).\tag{35}$$

因此，对于该实例，我们有

$$\begin{aligned}&\sum_{i\in[k]}\mathtt{Cost}_2(X_i^*, \{\widehat{c}_j\}_{j=1}^k)\\&\leqslant\left(1+\frac{\alpha}{1-\alpha}+\frac{4\alpha+\alpha\epsilon}{(1-2\alpha)(1-\alpha)}\right)\sum_{i\in[k]}\mathtt{Cost}_2(X_i^*, c_i^*).\end{aligned}\tag{36}$$

接下来，我们分析该算法的时间复杂度。首先，计算候选中心集的大小。S 中固定大小子集的总数为 $\binom{O(\varepsilon^{-1}(1-\alpha))}{O(1/\epsilon)}\leqslant 2^{O(1/\epsilon)}$，构建候选集所需的时间为$O\left(2^{O(1/\epsilon)}d\right)$。对于候选中心集中的每个候选点，计算其成本所需的时间为 n_id。因此，该算法的总体时间复杂度为

$$\sum_{i\in[k]}2^{O(1/\varepsilon)}|\tilde{X}_i|d\log\frac{k}{\delta}=2^{O(1/\varepsilon)}nd\log\frac{k}{\delta}.\tag{37}$$

接下来，我们分析该算法的成功概率。根据引理 B.5，单个簇内的成功概率为 $1-\frac{k}{\delta}$.。根据并集界，该算法的整体成功概率为 $\geqslant 1-k\times\frac{k}{\delta}=1-\delta$。

□

C Learning-Augment k-中位数的补充实验

表5-8展示了在数据集 CIFAR-10、Fashion-MNIST、PHY 和 MNIST 上的实验结果，其中 α 变化而 k 固定。表9-11展示了 k 变化而 α 固定时的结果。这两组结果均表明，我们的算法比所有竞争方法快得多，同时通常能获得更好的近似质量，尤其是在高维数据集上。

D Learning-Augment$k$-means 实验

我们在真实世界数据集上评估了算法。实验在一台配备 Intel(R) Xeon(R) Gold 6154 CPU 和 1024GB 随机存取存储器的服务器上进行。对于所有实验，我们报告了 10 次独立运行的平均聚类代价及其标准差。

数据集。遵循 Nguyen、Chaturvedi 和 Nguyen（2023）、Ergun 等人（2022）以及 Huang 等人（2025）的研究，我们使用一系列错误率 α，在 CIFAR-10 (n =50、000、d = 3、072) (Krizhevsky 和 Hinton 2009)、PHY (n = 10、000、d = 50) (KDD 2004) 和 MNIST (n =1、797、d = 64) (Deng 2012) 数据集上评估算法。此外，我们还在 Fashion-MNIST (n = 60000、d = 784) (Xiao、Rasul 和 Vollgraf 2017) 数据集上评估了算法的表现，以衡量其在高维数据集上的效能。”

预测器生成与误差模拟为了评估我们的算法，我们首先使用以 KMeans++ 初始化的 Lloyd 算法为每个数据集计算真实划分（记作 KMe++）。随后，我们通过在每个真实簇中随机选取 α 比例的点，并将它们重新分配到随机选定的簇中，生成错误率为 α 的受损划分（记作预测器）。为确保公平比较，对于任一错误率 α，所有算法均在完全相同的一组受损标签上进行测试。

算法在实验中，我们评估所提出的采样与搜索算法，并将其性能与其他最先进的学习增强方法进行比较，包括 Ergun 等人（ 2022 ）提出的算法（记作 Erg ）、Nguyen、Chaturvedi 和 Nguyen（2023）提出的算法（记作 Ngu），以及 Huang 等人（2025）的近期工作（记作快速采样、快速估计和快速过滤）。我们将预测器生成的标签所计算出的成本作为基准。根据未受损标签计算出的成本则视为最优成本。我们的主要比较关注给定数据集上的聚类代价和算法运行时间。此外，我们还计算了标准差，以评估算法的稳定性。

结果。表 12-15 展示了在数据集 CIFAR-10、Fashion-MNIST、PHY 和 MNIST 上，固定 k 并改变 α 时的实验结果。表 16-18 展示了固定 α 并改变 k 时的结果。这两组结果均表明，我们的算法比 Fast-Filtering 方法稍慢，但通常能获得更好的近似质量，尤其是在高维数据集上。

表5：固定 $k = 10$ 并改变 α时在CIFAR-10数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	2.7492e+07	-	-	-	-	-	-	-
0.1	预测器	2.7528e+07	-	-	-	-	-	-	-
	EFSplus	2.7406e+07	1109.14	29.87	14.76	**0.6631**	0.0012	**0.5918**	0.0026
	HFHplus	2.7401e+07	3929.91	184.22	1.45	0.6562	0.0078	0.5862	0.0107
	NCN	2.7404e+07	3466.13	34.43	7.95	0.6588	0.0050	0.5862	0.0065
	Ours	**2.7397e+07**	5089.44	**14.16**	0.26	0.6538	0.0090	0.5815	0.0131
0.2	预测器	2.7626e+07	-	-	-	-	-	-	-
	EFSplus	2.7418e+07	1881.31	33.79	15.96	**0.6571**	0.0025	0.5737	0.0043
	HFHplus	2.7403e+07	6017.94	184.37	1.69	0.6548	0.0082	0.5796	0.0113
	NCN	2.7415e+07	2612.95	32.81	6.77	0.6506	0.0079	0.5671	0.0099
	Ours	**2.7400e+07**	4507.52	**14.63**	0.26	0.6560	0.0050	**0.5800**	0.0078
0.3	预测器	2.7906e+07	-	-	-	-	-	-	-
	EFSplus	2.7460e+07	2413.62	35.60	19.24	0.6519	0.0046	0.5574	0.0061
	HFHplus	2.7413e+07	7979.19	185.80	1.97	0.6498	0.0065	0.5678	0.0093
	NCN	2.7454e+07	9237.83	32.87	8.06	0.6513	0.0051	0.5633	0.0072
	Ours	**2.7409e+07**	4717.90	**14.27**	0.36	**0.6546**	0.0060	**0.5728**	0.0105
0.4	预测器	2.8316e+07	-	-	-	-	-	-	-
	EFSplus	2.7540e+07	3959.35	31.26	20.27	0.6302	0.0046	0.5344	0.0064
	HFHplus	2.7439e+07	27686.32	186.90	1.83	0.6375	0.0101	0.5444	0.0163
	NCN	2.7505e+07	18855.20	35.30	5.75	0.6286	0.0106	0.5306	0.0154
	Ours	**2.7425e+07**	10487.16	**14.22**	0.43	**0.6447**	0.0082	**0.5556**	0.0135
0.5	预测器	2.8889e+07	-	-	-	-	-	-	-
	EFSplus	2.7584e+07	4907.95	25.72	18.49	0.6179	0.0039	0.5162	0.0048
	HFHplus	2.7488e+07	43919.20	185.27	1.69	0.6300	0.0139	0.5273	0.0186
	NCN	2.7569e+07	27179.60	39.10	11.63	0.6152	0.0135	0.5130	0.0196
	Ours	**2.7481e+07**	26429.82	**13.84**	0.35	**0.6341**	0.0071	**0.5311**	0.0141

表6：固定 $k = 10$ 并改变 α时，在Fashion-MNIST数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	8.4054e+07	-	-	-	-	-	-	-
0.1	预测器	8.4259e+07	-	-	-	-	-	-	-
	EFSplus	8.4050e+07	115.17	270.47	12.97	**0.9558**	0.0003	**0.9642**	0.0003
	HFHplus	8.4049e+07	834.92	749.72	18.47	0.9528	0.0036	0.9591	0.0050
	NCN	8.4050e+07	181.80	272.22	4.37	0.9552	0.0008	0.9636	0.0007
	Ours	**8.4048e+07**	933.64	**47.37**	0.78	0.9516	0.0019	0.9581	0.0022
0.2	预测器	8.4935e+07	-	-	-	-	-	-	-
	EFSplus	8.4057e+07	287.83	283.13	24.07	**0.9490**	0.0006	**0.9573**	0.0006
	HFHplus	8.4053e+07	1598.91	751.66	25.42	0.9433	0.0065	0.9493	0.0097
	NCN	8.4057e+07	309.52	282.97	13.78	0.9486	0.0006	0.9568	0.0006
	Ours	**8.4052e+07**	961.03	**47.96**	3.33	0.9453	0.0043	0.9515	0.0066
0.3	预测器	8.6223e+07	-	-	-	-	-	-	-
	EFSplus	8.4076e+07	467.75	282.57	8.66	**0.9419**	0.0005	**0.9486**	0.0006
	HFHplus	8.4065e+07	3527.00	751.13	22.67	0.9333	0.0058	0.9387	0.0088
	NCN	8.4077e+07	695.49	299.50	22.54	0.9414	0.0008	0.9481	0.0010
	Ours	**8.4062e+07**	3848.20	**45.38**	1.33	0.9362	0.0067	0.9420	0.0105
0.4	预测器	8.8209e+07	-	-	-	-	-	-	-
	EFSplus	8.4109e+07	631.67	297.27	13.66	**0.9348**	0.0005	0.9380	0.0007
	HFHplus	8.4101e+07	11512.25	758.16	29.89	0.9331	0.0022	**0.9391**	0.0021
	NCN	8.4111e+07	1206.20	302.45	25.71	0.9330	0.0050	0.9359	0.0064
	Ours	**8.4100e+07**	12684.42	**45.29**	2.18	0.9320	0.0041	0.9374	0.0029
0.5	预测器	9.0897e+07	-	-	-	-	-	-	-
	EFSplus	8.4150e+07	1320.47	304.67	11.82	0.9185	0.0007	0.9191	0.0009
	HFHplus	8.4148e+07	12671.01	751.08	26.53	0.9176	0.0010	0.9195	0.0019
	NCN	8.4152e+07	2503.70	305.95	10.98	0.9178	0.0010	0.9183	0.0014
	Ours	**8.4145e+07**	17385.62	**47.87**	2.04	**0.9201**	0.0048	**0.9224**	0.0066

表7：固定 $k = 10$ 并改变 α 时在 PHY数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	1.9797e+08	-	-	-	-	-	-	-
0.1	预测器	1.9849e+08	-	-	-	-	-	-	-
	EFSplus	1.9800e+08	**478.11**	392.68	99.57	0.9882	0.0001	0.9900	0.0001
	HFHplus	1.9799e+08	6778.35	39.54	6.32	**0.9896**	0.0017	**0.9913**	0.0017
	NCN	1.9800e+08	4363.70	376.60	102.88	0.9883	0.0005	0.9900	0.0005
	Ours	**1.9798e+08**	7427.30	**34.05**	2.86	0.9867	0.0049	0.9884	0.0053
0.2	预测器	2.0034e+08	-	-	-	-	-	-	-
	EFSplus	1.9813e+08	610.96	360.84	114.79	0.9789	0.0002	0.9799	0.0002
	HFHplus	1.9797e+08	1449.20	35.91	7.18	0.9934	0.0010	0.9950	0.0009
	NCN	1.9807e+08	26763.49	262.61	94.40	0.9816	0.0039	0.9824	0.0049
	Ours	**1.9797e+08**	**522.96**	**32.72**	3.64	**0.9940**	0.0005	**0.9954**	0.0004
0.3	预测器	2.0446e+08	-	-	-	-	-	-	-
	EFSplus	1.9844e+08	**2913.39**	357.00	93.67	0.9710	0.0002	0.9690	0.0002
	HFHplus	1.9798e+08	7618.53	37.66	5.30	0.9889	0.0032	0.9903	0.0036
	NCN	1.9832e+08	72445.21	234.39	82.27	0.9739	0.0027	0.9724	0.0031
	Ours	**1.9798e+08**	3555.61	**27.97**	1.55	**0.9919**	0.0022	**0.9931**	0.0025
0.4	预测器	2.1242e+08	-	-	-	-	-	-	-
	EFSplus	1.9896e+08	**4654.51**	304.39	97.99	0.9614	0.0002	0.9558	0.0002
	HFHplus	1.9799e+08	7119.24	35.93	5.46	0.9885	0.0029	0.9903	0.0031
	NCN	1.9874e+08	146423.75	225.87	80.83	0.9633	0.0027	0.9588	0.0036
	Ours	**1.9798e+08**	5888.86	**28.25**	1.95	**0.9890**	0.0031	**0.9908**	0.0034
0.5	预测器	2.2831e+08	-	-	-	-	-	-	-
	EFSplus	2.0004e+08	8404.95	306.93	83.25	0.9500	0.0002	0.9380	0.0002
	HFHplus	1.9799e+08	15837.90	35.58	2.82	0.9879	0.0029	0.9896	0.0026
	NCN	1.9976e+08	399881.55	262.76	91.23	0.9525	0.0042	0.9422	0.0062
	Ours	**1.9798e+08**	**3841.72**	**26.50**	0.81	**0.9893**	0.0014	**0.9908**	0.0012

表8：固定 $k = 10$ 并改变 α时在MNIST数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	44696.31	-	-	-	-	-	-	-
0.1	预测器	44795.92	-	-	-	-	-	-	-
	EFSplus	44695.47	0.78	25.23	6.42	0.9721	0.0030	0.9722	0.0038
	HFHplus	44695.66	0.40	3.58	1.09	0.9682	0.0069	0.9688	0.0071
	NCN	44695.55	0.47	24.97	6.58	**0.9739**	0.0053	**0.9746**	0.0056
	Ours	**44695.31**	0.71	**3.36**	0.91	0.9688	0.0072	0.9692	0.0072
0.2	预测器	45104.16	-	-	-	-	-	-	-
	EFSplus	44698.38	0.57	26.28	7.15	0.9695	0.0056	0.9718	0.0053
	HFHplus	44697.47	0.45	3.56	1.29	0.9705	0.0031	0.9737	0.0036
	NCN	44698.24	0.18	27.18	7.91	0.9690	0.0014	0.9718	0.0017
	Ours	**44697.18**	0.30	**3.40**	0.87	**0.9727**	0.0021	**0.9761**	0.0024
0.3	预测器	45654.02	-	-	-	-	-	-	-
	EFSplus	44706.18	1.20	25.54	7.29	**0.9639**	0.0059	**0.9662**	0.0058
	HFHplus	44700.07	1.48	3.47	1.09	0.9618	0.0059	0.9633	0.0057
	NCN	44706.96	1.15	25.83	8.28	0.9618	0.0036	0.9639	0.0036
	Ours	**44699.50**	1.24	**3.05**	0.55	0.9614	0.0033	0.9633	0.0031
0.4	预测器	46620.32	-	-	-	-	-	-	-
	EFSplus	**44704.53**	1.16	27.35	6.00	**0.9681**	0.0036	**0.9687**	0.0048
	HFHplus	44704.70	1.73	3.82	1.16	0.9560	0.0086	0.9554	0.0106
	NCN	44706.67	1.27	27.65	7.45	0.9656	0.0049	0.9673	0.0058
	Ours	44701.89	1.63	**3.39**	1.03	0.9606	0.0066	0.9608	0.0078
0.5	预测器	48022.97	-	-	-	-	-	-	-
	EFSplus	44717.07	3.38	28.34	7.28	0.9538	0.0037	**0.9537**	0.0042
	HFHplus	44718.66	2.15	3.63	1.16	0.9451	0.0065	0.9446	0.0080
	NCN	44720.35	1.46	26.36	7.25	**0.9501**	0.0031	0.9502	0.0039
	Ours	**44716.27**	5.17	**3.40**	0.96	0.9446	0.0081	0.9439	0.0102

表9：固定 $\alpha = 0.2$ 并改变 k时，在Fashion-MNIST数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
k	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	8.4058e+07	-	-	-	-	-	-	-
	预测器	8.4938e+07	-	-	-	-	-	-	-
	EFSplus	8.4059e+07	305.75	281.82	11.34	0.9460	0.0005	0.9547	0.0006
	HFHplus	8.4053e+07	1129.16	733.73	24.48	0.9419	0.0035	0.9489	0.0057
	NCN	8.4059e+07	811.26	285.90	2.63	0.9455	0.0015	0.9541	0.0016
	Ours	**8.4052e+07**	1275.81	**60.78**	3.40	**0.9473**	0.0025	**0.9557**	0.0030
20	K-Med++	7.6802e+07	-	-	-	-	-	-	-
	预测器	7.7606e+07	-	-	-	-	-	-	-
	EFSplus	7.6802e+07	294.01	244.01	2.30	0.9466	0.0006	0.9348	0.0008
	HFHplus	7.6792e+07	1249.97	945.64	3.66	0.9507	0.0028	0.9402	0.0048
	NCN	7.6802e+07	234.77	268.17	21.22	0.9458	0.0005	0.9335	0.0008
	Ours	**7.6791e+07**	951.50	**78.03**	0.57	**0.9510**	0.0022	**0.9400**	0.0035
30	K-Med++	7.3922e+07	-	-	-	-	-	-	-
	预测器	7.4732e+07	-	-	-	-	-	-	-
	EFSplus	7.3935e+07	339.73	238.19	2.97	0.9476	0.0006	0.9308	0.0010
	HFHplus	7.3915e+07	1969.80	1185.44	16.58	0.9516	0.0034	0.9359	0.0066
	NCN	7.3935e+07	1675.89	230.32	44.44	0.9446	0.0043	0.9258	0.0070
	Ours	**7.3914e+07**	388.43	**100.95**	1.95	**0.9526**	0.0011	**0.9371**	0.0018
40	K-Med++	7.1839e+07	-	-	-	-	-	-	-
	预测器	7.2598e+07	-	-	-	-	-	-	-
	EFSplus	7.1853e+07	621.20	237.52	3.37	0.9430	0.0005	0.9212	0.0006
	HFHplus	7.1829e+07	902.75	1451.52	18.02	0.9507	0.0020	0.9309	0.0039
	NCN	7.1853e+07	299.48	245.99	3.36	0.9425	0.0003	0.9203	0.0004
	Ours	**7.1828e+07**	905.20	**119.70**	0.57	**0.9522**	0.0016	**0.9322**	0.0035
50	K-Med++	7.0470e+07	-	-	-	-	-	-	-
	预测器	7.1242e+07	-	-	-	-	-	-	-
	EFSplus	7.0494e+07	200.76	232.55	2.20	0.9465	0.0012	0.9207	0.0025
	HFHplus	7.0466e+07	654.43	1759.73	20.54	0.9544	0.0009	0.9337	0.0013
	NCN	7.0495e+07	530.36	257.05	26.50	0.9457	0.0002	0.9196	0.0004
	Ours	**7.0465e+07**	820.99	**140.33**	1.37	**0.9574**	0.0012	**0.9377**	0.0017

表 10：固定 $\alpha = 0.2$ 、改变 k 时在 PHY数据集上的性能比较。

条件		Cost		时间（秒）		NMI		ARI	
k	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	2.0224e+08	-	-	-	-	-	-	-
	预测器	2.0428e+08	-	-	-	-	-	-	-
	EFSplus	2.0205e+08	4444.80	362.31	52.30	**0.9369**	0.0002	**0.9192**	0.0003
	HFHplus	2.0147e+08	105661.96	42.33	1.91	0.9289	0.0139	0.9036	0.0240
	NCN	2.0163e+08	109131.95	160.15	14.33	0.9173	0.0097	0.8866	0.0150
	Ours	**2.0135e+08**	82812.18	**20.72**	0.59	0.8934	0.0222	0.8446	0.0360
20	K-Med++	1.0910e+08	-	-	-	-	-	-	-
	预测器	1.1004e+08	-	-	-	-	-	-	-
	EFSplus	1.0927e+08	1096.15	335.96	46.37	0.9590	0.0002	0.9416	0.0004
	HFHplus	1.0910e+08	1141.88	40.96	2.95	**0.9892**	0.0095	**0.9873**	0.0148
	NCN	1.0923e+08	23983.20	225.00	47.90	0.9630	0.0011	0.9473	0.0016
	Ours	**1.0910e+08**	4120.98	**25.49**	0.90	0.9807	0.0075	0.9737	0.0117
30	K-Med++	8.4404e+07	-	-	-	-	-	-	-
	预测器	8.5019e+07	-	-	-	-	-	-	-
	EFSplus	8.4491e+07	721.59	294.21	54.13	0.9514	0.0003	0.9202	0.0007
	HFHplus	8.4404e+07	4372.98	42.26	2.05	0.9856	0.0048	**0.9810**	0.0107
	NCN	8.4480e+07	14266.80	221.81	49.30	0.9541	0.0037	0.9240	0.0079
	Ours	**8.4404e+07**	3043.38	**27.08**	3.26	**0.9857**	0.0024	0.9809	0.0051
40	K-Med++	6.9910e+07	-	-	-	-	-	-	-
	预测器	7.0361e+07	-	-	-	-	-	-	-
	EFSplus	7.0002e+07	527.34	287.23	23.10	0.9535	0.0003	0.9153	0.0007
	HFHplus	6.9910e+07	4461.84	44.65	2.59	0.9894	0.0027	0.9852	0.0062
	NCN	6.9989e+07	9213.02	209.77	35.93	0.9571	0.0028	0.9224	0.0066
	Ours	**6.9909e+07**	**1497.41**	**32.51**	0.63	**0.9901**	0.0006	**0.9867**	0.0010
50	K-Med++	6.2759e+07	-	-	-	-	-	-	-
	预测器	6.3112e+07	-	-	-	-	-	-	-
	EFSplus	6.2797e+07	**503.13**	285.51	22.48	0.9510	0.0004	0.9025	0.0007
	HFHplus	6.2756e+07	1072.71	44.69	1.34	0.9834	0.0020	0.9758	0.0038
	NCN	6.2792e+07	5662.71	208.89	29.26	0.9528	0.0013	0.9062	0.0030
	Ours	**6.2755e+07**	677.03	**36.87**	0.73	**0.9842**	0.0009	**0.9776**	0.0016

表11：固定 $\alpha = 0.2$ 并改变 k 时在 MNIST数据集上的性能比较。

	条件	Cost		时间（秒）		NMI		ARI	
k	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	44695.78	-	-	-	1.0000	-	1.0000	-
	预测器	45101.25	-	-	-	-	-	-	-
	NCN	44700.26	0.66	4.68	0.36	0.9738	0.0031	0.9768	0.0026
	快速估计	44698.33	0.78	2.18	0.15	0.9655	0.0037	0.9670	0.0044
	Ngu	44701.73	0.97	4.53	0.44	**0.9745**	0.0016	**0.9774**	0.0013
	Ours	**44697.49**	0.58	**0.28**	0.03	0.9659	0.0028	0.9677	0.0024
20	K-Med++	40348.81	-	-	-	1.0000	-	1.0000	-
	预测器	40937.18	-	-	-	-	-	-	-
	NCN	40347.92	1.70	5.10	0.34	**0.9643**	0.0039	**0.9619**	0.0043
	快速估计	40344.74	0.81	3.65	0.10	0.9546	0.0177	0.9466	0.0246
	Ngu	40345.16	0.76	4.83	0.37	0.9633	0.0009	0.9605	0.0012
	Ours	**40344.54**	1.02	**0.35**	0.04	0.9640	0.0021	0.9599	0.0026
30	K-Med++	37920.28	-	-	-	1.0000	-	1.0000	-
	预测器	38506.77	-	-	-	-	-	-	-
	NCN	37900.65	2.17	5.60	0.46	0.9506	0.0019	0.9376	0.0029
	快速估计	**37885.34**	7.12	5.02	0.34	0.9439	0.0131	0.9236	0.0223
	Ngu	37898.98	1.35	5.00	0.51	**0.9523**	0.0053	**0.9415**	0.0075
	Ours	37893.81	3.94	**0.39**	0.06	0.9481	0.0016	0.9309	0.0041
40	K-Med++	36419.10	-	-	-	1.0000	-	1.0000	-
	预测器	36981.30	-	-	-	-	-	-	-
	NCN	36398.94	4.92	6.28	0.46	**0.9498**	0.0049	**0.9185**	0.0086
	快速估计	**36381.18**	7.13	6.23	1.08	0.9232	0.0083	0.8667	0.0135
	Ngu	36405.05	8.05	4.74	0.47	0.9458	0.0098	0.9120	0.0190
	Ours	36394.45	3.43	**0.43**	0.05	0.9360	0.0193	0.8887	0.0350
50	K-Med++	35078.76	-	-	-	1.0000	-	1.0000	-
	预测器	35676.22	-	-	-	-	-	-	-
	NCN	35058.31	5.67	6.68	0.44	0.9561	0.0016	0.9127	0.0051
	快速估计	35047.58	10.19	7.64	1.05	0.9472	0.0135	0.8874	0.0317
	Ngu	35059.27	7.08	5.88	0.89	**0.9603**	0.0079	**0.9188**	0.0190
	Ours	**35044.96**	13.82	**0.51**	0.04	0.9448	0.0136	0.8852	0.0277

表12：固定 $k = 10$ 并改变 α 时在 CIFAR-10 数据集上的性能比较。

	条件	Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	KMe++	3.9481e+11	-	-	-	-	-	-	-
0.1	预测器	3.9680e+11	-	-	-	-	-	-	-
	EFSplus	4.2240e+11	6.5569e+07	7.86	0.22	0.8930	0.0004	0.8853	0.0005
	NCN	3.9869e+11	0	127.44	6.20	0.8980	0.0000	0.8943	0.0000
	快速估计	3.9747e+11	5.5532e+08	7.62	0.29	0.8777	0.0093	0.8681	0.0122
	快速过滤	3.9680e+11	2.5731e+08	**0.73**	0.07	0.9201	0.0064	0.9245	0.0073
	快速采样	4.0572e+11	5.2472e+07	61.94	4.30	0.9132	0.0003	0.9149	0.0003
	Ours	**3.9522e+11**	2.8906e+07	1.86	0.08	**0.9315**	0.0018	**0.9366**	0.0024
0.2	预测器	4.0312e+11	-	-	-	-	-	-	-
	EFSplus	4.0320e+11	-e+07	7.70	0.49	0.8539	0.0004	0.8348	0.0005
	NCN	4.0163e+11	0	243.57	10.32	0.8463	0.0000	0.8260	0.0000
	快速估计	4.0309e+11	3.2826e+09	7.23	0.23	0.8068	0.0259	0.7661	0.0394
	快速过滤	4.0181e+11	4.4826e+08	**0.72**	0.05	0.8634	0.0156	0.8548	0.0198
	快速采样	4.1658e+11	6.9052e+07	61.60	1.58	0.8220	0.0009	0.7897	0.0012
	Ours	**3.9625e+11**	1.6161e+08	1.56	0.10	**0.8901**	0.0051	**0.8888**	0.0064
0.3	预测器	4.1417e+11	-	-	-	-	-	-	-
	EFSplus	4.1822e+11	7.2262e+07	8.03	0.22	0.7860	0.0004	0.7272	0.0006
	NCN	4.0506e+11	0	349.53	10.41	0.7959	0.0000	0.7530	0.0000
	快速估计	4.1693e+11	5.0344e+09	7.38	0.18	0.7432	0.0294	0.6692	0.0466
	快速过滤	4.0731e+11	5.4559e+08	**0.74**	0.04	0.8202	0.0178	0.7967	0.0257
	快速采样	4.3372e+11	1.9210e+08	65.39	2.72	0.8064	0.0007	0.7695	0.0010
	Ours	**3.9831e+11**	4.9532e+08	1.70	0.14	**0.8500**	0.0071	**0.8350**	0.0095
0.4	预测器	4.2775e+11	-	-	-	-	-	-	-
	EFSplus	2.8665e+12	1.1421e+08	7.71	0.37	0.7448	0.0003	0.6573	0.0006
	NCN	4.1171e+11	0	466.82	11.00	0.7494	0.0000	0.6767	0.0000
	快速估计	4.2390e+11	3.2667e+09	7.36	0.27	0.6466	0.0356	0.5131	0.0592
	快速过滤	4.1778e+11	4.8258e+09	**0.73**	0.05	0.7700	0.0187	0.7180	0.0312
	快速采样	4.6080e+11	3.8768e+08	61.84	3.07	0.7606	0.0006	0.6852	0.0008
	Ours	**4.0041e+11**	5.0524e+08	1.65	0.21	**0.8065**	0.0071	**0.7701**	0.0110
0.5	预测器	4.4860e+11	-	-	-	-	-	-	-
	EFSplus	2.8674e+12	8.4072e+07	7.76	0.30	0.7032	0.0003	0.5797	0.0006
	NCN	4.2824e+11	0	583.26	10.26	0.7093	0.0000	0.5972	0.0000
	快速估计	4.4518e+11	4.4860e+09	7.54	0.08	0.6389	0.0450	0.5000	0.0636
	快速过滤	4.2962e+11	7.4925e+09	**0.74**	0.05	0.7040	0.0276	0.6160	0.0475
	快速采样	4.4860e+11	0	63.12	1.12	0.7136	0.0006	0.5971	0.0009
	Ours	**4.1955e+11**	7.6168e+09	1.70	0.14	**0.7346**	0.0113	**0.6551**	0.0213

表13：固定 $k = 10$ 并改变 α 时，在 Fashion-MNIST数据集上的性能比较。

	条件	Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	1.2450e+11	-	-	-	-	-	-	-
0.1	预测器	1.2613e+11	-	-	-	-	-	-	-
	EFSplus	1.2669e+11	7.7786e+06	2.917	0.556	0.8991	0.0002	0.8898	0.0004
	NCN	1.2567e+11	0.00	3.367	0.814	0.9510	0.0000	0.9570	0.0000
	快速估计	1.2575e+11	8.3179e+07	2.637	0.791	0.9325	0.0115	0.9339	0.0142
	快速过滤	1.2481e+11	3.0078e+07	**0.731**	0.319	0.9588	0.0046	0.9650	0.0056
	快速采样	1.2541e+11	2.4393e+06	16.926	2.813	0.9514	0.0002	0.9580	0.0002
	Ours	**1.2470e+11**	5.6813e+06	0.909	0.316	**0.9658**	0.0010	**0.9724**	0.0010
0.2	预测器	1.3058e+11	-	-	-	-	-	-	-
	EFSplus	1.3062e+11	1.0415e+07	2.900	0.528	0.8823	0.0002	0.8713	0.0003
	NCN	1.2814e+11	0.00	7.826	2.057	0.9028	0.0000	0.8978	0.0000
	快速估计	1.2863e+11	4.7944e+08	2.599	0.933	0.9040	0.0149	0.8999	0.0220
	快速过滤	1.2551e+11	1.4703e+08	**0.695**	0.319	0.9312	0.0076	0.9363	0.0097
	快速采样	1.2883e+11	6.1020e+06	17.190	3.196	0.9213	0.0003	0.9227	0.0003
	Ours	**1.2524e+11**	4.3570e+07	0.802	0.247	**0.9427**	0.0014	**0.9497**	0.0017
0.3	预测器	1.3752e+11	-	-	-	-	-	-	-
	EFSplus	1.3721e+11	1.7567e+07	3.031	0.608	0.8544	0.0004	0.8328	0.0005
	NCN	1.3255e+11	0.00	7.497	1.342	0.8886	0.0000	0.8817	0.0000
	快速估计	1.3293e+11	5.8739e+08	2.744	0.714	0.8656	0.0123	0.8473	0.0181
	快速过滤	1.2639e+11	3.5963e+08	**0.713**	0.384	0.9110	0.0081	0.9139	0.0117
	快速采样	1.3347e+11	6.5379e+07	15.654	2.612	0.8726	0.0012	0.8529	0.0023
	Ours	**1.2608e+11**	5.0719e+07	0.746	0.349	**0.9240**	0.0055	**0.9300**	0.0072
0.4	预测器	1.4681e+11	-	-	-	-	-	-	-
	EFSplus	1.4641e+11	3.1038e+07	3.008	0.505	0.8405	0.0004	0.8136	0.0007
	NCN	1.3952e+11	0.00	11.137	1.466	0.8567	0.0000	0.8329	0.0000
	快速估计	1.3915e+11	1.3205e+09	2.469	0.498	0.8398	0.0153	0.8101	0.0234
	快速过滤	1.2779e+11	9.4167e+08	0.708	0.376	0.8825	0.0169	0.8772	0.0241
	快速采样	1.3831e+11	1.0812e+08	15.036	3.219	0.8629	0.0009	0.8422	0.0014
	Ours	**1.2741e+11**	2.8672e+08	**0.706**	0.264	**0.9000**	0.0075	**0.9019**	0.0109
0.5	预测器	1.5849e+11	-	-	-	-	-	-	-
	EFSplus	1.5808e+11	2.9992e+07	2.845	0.640	0.8122	0.0003	0.7728	0.0005
	NCN	1.4889e+11	0.00	11.449	1.345	0.8329	0.0000	0.7991	0.0000
	快速估计	1.4781e+11	1.4452e+09	2.508	0.530	0.8090	0.0270	0.7711	0.0386
	快速过滤	1.3227e+11	1.7529e+09	**0.663**	0.283	0.8445	0.0185	0.8349	0.0252
	快速采样	1.4682e+11	1.2533e+08	15.540	2.515	0.8441	0.0012	0.8168	0.0021
	Ours	**1.3183e+11**	6.7374e+08	0.721	0.188	**0.8555**	0.0043	**0.8510**	0.0060

表14：固定 $k = 10$ 并改变 α时在MNIST数据集上的性能比较。

	条件	Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	KMe++	1.1653e+06	-	-	-	-	-	-	-
0.1	预测器	1.1790e+06	-	-	-	-	-	-	-
	EFSplus	1.1867e+06	1057.86	0.080	0.050	0.9687	0.0034	0.9714	0.0033
	NCN	1.1747e+06	0.00	0.057	0.035	0.9778	0.0000	0.9801	0.0000
	快速估计	1.1757e+06	315.60	0.087	0.037	0.9770	0.0017	0.9793	0.0021
	快速过滤	1.1683e+06	251.37	**0.013**	0.013	0.9775	0.0043	0.9803	0.0043
	快速采样	1.1737e+06	216.39	2.199	0.545	0.9782	0.0027	0.9805	0.0024
	Ours	**1.1677e+06**	0.00	0.016	0.014	**0.9825**	0.0000	**0.9856**	0.0000
0.2	预测器	1.2104e+06	-	-	-	-	-	-	-
	EFSplus	1.2036e+06	1605.93	0.070	0.032	0.9492	0.0035	0.9523	0.0041
	NCN	1.1937e+06	0.00	0.124	0.042	0.9545	0.0000	0.9579	0.0000
	快速估计	1.1946e+06	1497.59	0.100	0.032	0.9574	0.0049	0.9609	0.0057
	快速过滤	1.1719e+06	1187.77	**0.013**	0.013	0.9677	0.0069	0.9702	0.0066
	快速采样	1.1974e+06	637.34	2.269	0.463	0.9583	0.0023	0.9610	0.0024
	Ours	**1.1703e+06**	0.00	0.023	0.018	**0.9687**	0.0000	**0.9713**	0.0000
0.3	预测器	1.2643e+06	-	-	-	-	-	-	-
	EFSplus	1.2493e+06	3918.83	0.087	0.040	0.9294	0.0051	0.9295	0.0048
	NCN	1.2223e+06	0.00	0.129	0.037	0.9383	0.0000	0.9391	0.0000
	快速估计	1.2309e+06	4058.38	0.106	0.024	0.9336	0.0077	0.9345	0.0084
	快速过滤	1.1786e+06	1892.68	**0.020**	0.012	0.9411	0.0066	0.9438	0.0074
	快速采样	1.2438e+06	547.48	2.234	0.428	0.9383	0.0018	0.9390	0.0023
	Ours	**1.1770e+06**	0.00	0.029	0.011	**0.9432**	0.0000	**0.9455**	0.0000
0.4	预测器	1.3324e+06	-	-	-	-	-	-	-
	EFSplus	1.2890e+06	4008.71	0.103	0.042	0.9208	0.0057	0.9186	0.0079
	NCN	1.2663e+06	0.00	0.141	0.030	0.9209	0.0000	0.9198	0.0000
	快速估计	1.2804e+06	5819.88	0.096	0.037	0.9231	0.0085	0.9227	0.0098
	快速过滤	1.1837e+06	1949.92	**0.011**	0.011	0.9340	0.0097	0.9333	0.0121
	快速采样	1.2920e+06	7189.59	1.681	0.544	0.9160	0.0054	0.9151	0.0056
	Ours	**1.1815e+06**	0.00	0.012	0.014	**0.9490**	0.0000	**0.9503**	0.0000
0.5	预测器	1.4340e+06	-	-	-	-	-	-	-
	EFSplus	1.3886e+06	5775.49	0.122	0.026	0.8939	0.0057	0.8887	0.0077
	NCN	1.3516e+06	0.00	0.153	0.030	0.9056	0.0000	0.9035	0.0000
	快速估计	1.3712e+06	6201.06	0.094	0.021	0.8984	0.0142	0.8924	0.0188
	快速过滤	1.2064e+06	12137.27	**0.016**	0.017	0.8928	0.0156	0.8839	0.0202
	快速采样	1.3790e+06	7845.73	1.658	0.614	0.8867	0.0082	0.8791	0.0094
	Ours	**1.2051e+06**	0.00	0.022	0.019	**0.9150**	0.0000	**0.9078**	0.0000

表15：固定 $k = 10$ 并改变 α时，PHY数据集上的性能比较。

	条件	Cost		时间（秒）		NMI		ARI	
α	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
0	K-Med++	1.0147e+12	-	-	-	-	-	-	-
0.1	预测器	1.1078e+12	-	-	-	-	-	-	-
	EFSplus	1.0230e+12	6.5840e+08	0.357	0.138	0.9769	0.0052	0.9758	0.0066
	NCN	1.0167e+12	0.00	1.083	0.274	0.9937	0.0000	0.9953	0.0000
	快速估计	1.0239e+12	3.2252e+09	0.284	0.090	0.9776	0.0063	0.9778	0.0079
	快速过滤	1.0171e+12	5.2853e+08	**0.091**	0.050	0.9913	0.0050	0.9924	0.0060
	快速采样	1.0258e+12	1.4804e+09	1.931	0.497	0.9749	0.0055	0.9727	0.0069
	Ours	**1.0161e+12**	5.6604e+08	0.206	0.072	**0.9859**	0.0041	**0.9857**	0.0051
0.2	预测器	1.3930e+12	-	-	-	-	-	-	-
	EFSplus	1.0204e+12	2.4351e+08	0.340	0.083	0.9896	0.0025	0.9913	0.0025
	NCN	1.0219e+12	0.00	1.560	0.385	0.9859	0.0000	0.9884	0.0000
	快速估计	1.0532e+12	1.6020e+10	0.321	0.119	0.9527	0.0144	0.9420	0.0228
	快速过滤	1.0176e+12	3.4455e+08	**0.097**	0.039	0.9913	0.0019	0.9923	0.0023
	快速采样	1.0219e+12	7.6533e+08	1.847	0.523	0.9825	0.0034	0.9832	0.0042
	Ours	**1.0173e+12**	2.2068e+08	0.189	0.063	**0.9924**	0.0016	**0.9930**	0.0020
0.3	预测器	1.8631e+12	-	-	-	-	-	-	-
	EFSplus	1.0528e+12	8.8879e+09	0.438	0.143	0.9646	0.0059	0.9594	0.0085
	NCN	1.0300e+12	0.00	2.333	0.418	**0.9874**	0.0000	**0.9859**	0.0000
	快速估计	1.1067e+12	2.2806e+10	0.275	0.080	0.9309	0.0148	0.9074	0.0290
	快速过滤	1.0227e+12	2.4148e+09	**0.095**	0.035	0.9812	0.0066	0.9794	0.0093
	快速采样	1.0418e+12	6.5694e+09	1.734	0.574	0.9693	0.0068	0.9656	0.0107
	Ours	**1.0193e+12**	1.2542e+09	0.184	0.055	0.9832	0.0043	0.9816	0.0061
0.4	预测器	2.6240e+12	-	-	-	-	-	-	-
	EFSplus	1.0446e+12	2.3424e+09	0.438	0.083	0.9815	0.0037	0.9793	0.0049
	NCN	1.0367e+12	0.00	2.554	0.609	**0.9787**	0.0000	**0.9770**	0.0000
	快速估计	1.1720e+12	6.0011e+10	0.338	0.105	0.9207	0.0199	0.8932	0.0328
	快速过滤	1.0313e+12	6.8650e+09	**0.090**	0.031	0.9780	0.0098	0.9753	0.0151
	快速采样	1.0510e+12	5.7277e+09	1.740	0.408	0.9609	0.0052	0.9568	0.0081
	Ours	**1.0257e+12**	2.5217e+09	0.215	0.036	0.9759	0.0106	0.9724	0.0140
0.5	预测器	3.5556e+12	-	-	-	-	-	-	-
	EFSplus	1.0512e+12	4.1080e+09	0.407	0.070	0.9733	0.0053	0.9726	0.0068
	NCN	1.0449e+12	0.00	2.910	0.861	**0.9794**	0.0000	**0.9797**	0.0000
	快速估计	1.2459e+12	1.0526e+11	0.277	0.109	0.9056	0.0410	0.8617	0.0765
	快速过滤	1.0389e+12	4.5319e+09	**0.081**	0.024	0.9714	0.0089	0.9679	0.0125
	快速采样	1.0835e+12	2.6368e+10	1.711	0.410	0.9453	0.0147	0.9278	0.0254
	Ours	**1.0278e+12**	1.8072e+09	0.215	0.045	0.9758	0.0067	0.9728	0.0094

表16：固定 $\alpha = 0.2$ 并改变 k时在 CIFAR-10 数据集上的性能比较。

k	条件 / 方法	Cost		时间（秒）		NMI		ARI	
		Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	7.8584e+10	-	-	-	-	-	-	-
	预测器	8.0164e+10	-	-	-	-	-	-	-
	EFSplus	7.9181e+10	2.64e+06	1.347	0.025	0.8625	0.0008	0.8403	0.0008
	NCN	7.9408e+10	0.00	3.678	0.030	0.8282	0.0000	0.7948	0.0000
	快速估计	7.9677e+10	2.54e+08	1.345	0.022	0.7905	0.0183	0.7350	0.0259
	快速过滤	7.8971e+10	3.99e+07	0.163	0.005	0.8786	0.0039	0.8673	0.0052
	快速采样	7.9523e+10	3.50e+05	28.082	0.664	0.8311	0.0003	0.7883	0.0007
	Ours	**7.8896e+10**	3.46e+07	**0.101**	0.001	**0.8889**	0.0073	**0.8816**	0.0106
20	K-Med++	7.3059e+10	-	-	-	-	-	-	-
	预测器	7.4801e+10	-	-	-	-	-	-	-
	EFSplus	7.3875e+10	2.71e+06	1.784	0.002	0.8740	0.0003	0.8398	0.0004
	NCN	7.3838e+10	0.00	3.846	0.007	0.8384	0.0000	0.7678	0.0000
	快速估计	7.4010e+10	3.28e+08	1.650	0.007	0.7990	0.0218	0.7156	0.0321
	快速过滤	7.3470e+10	4.43e+07	0.136	0.002	0.8836	0.0084	0.8578	0.0121
	快速采样	7.4160e+10	1.73e+06	52.049	0.157	0.8401	0.0007	0.7770	0.0010
	Ours	**7.3405e+10**	5.11e+07	**0.087**	0.001	**0.8990**	0.0093	**0.8803**	0.0121
30	K-Med++	7.0352e+10	-	-	-	-	-	-	-
	预测器	7.2285e+10	-	-	-	-	-	-	-
	EFSplus	7.1369e+10	3.01e+06	2.263	0.005	0.8714	0.0015	0.8257	0.0023
	NCN	7.1273e+10	0.00	4.083	0.011	0.8401	0.0000	0.7589	0.0000
	快速估计	7.1387e+10	1.31e+08	1.997	0.010	0.7977	0.0101	0.7013	0.0158
	快速过滤	7.0864e+10	6.16e+07	0.130	0.001	0.8793	0.0057	0.8381	0.0083
	快速采样	7.1625e+10	7.08e+05	72.828	0.130	0.8316	0.0008	0.7495	0.0011
	Ours	**7.0763e+10**	2.71e+07	**0.086**	0.000	**0.8963**	0.0059	**0.8636**	0.0086
40	K-Med++	6.8488e+10	-	-	-	-	-	-	-
	预测器	7.0455e+10	-	-	-	-	-	-	-
	EFSplus	6.9651e+10	4.65e+06	2.741	0.007	0.8784	0.0008	0.8265	0.0011
	NCN	6.9464e+10	0.00	4.332	0.005	0.8581	0.0000	0.7825	0.0000
	快速估计	6.9656e+10	8.30e+07	2.345	0.003	0.8035	0.0097	0.6949	0.0230
	快速过滤	6.9034e+10	4.77e+07	0.124	0.002	0.8881	0.0029	0.8463	0.0037
	快速采样	6.9830e+10	1.44e+06	92.965	0.099	0.8377	0.0004	0.7471	0.0006
	Ours	**6.8911e+10**	3.58e+07	**0.087**	0.000	**0.9042**	0.0023	**0.8716**	0.0028
50	K-Med++	6.7185e+10	-	-	-	-	-	-	-
	预测器	6.9137e+10	-	-	-	-	-	-	-
	EFSplus	6.8442e+10	2.30e+06	3.214	0.009	0.8793	0.0007	0.8159	0.0009
	NCN	6.8172e+10	0.00	4.586	0.026	0.8549	0.0000	0.7636	0.0000
	快速估计	6.8313e+10	1.05e+08	2.712	0.018	0.8011	0.0164	0.6804	0.0276
	快速过滤	6.7753e+10	1.15e+08	0.121	0.002	0.8843	0.0161	0.8327	0.0256
	快速采样	6.8527e+10	1.20e+06	113.940	0.304	0.8326	0.0004	0.7240	0.0007
	Ours	**6.7642e+10**	2.57e+07	**0.087**	0.000	**0.9023**	0.0030	**0.8602**	0.0041

Ta表 17：固定 $\alpha = 0.2$ 并改变 k 时，在 Fashion-MNIST数据集上的性能比较。

k	条件	Cost		时间（秒）		NMI		ARI	
	方法	Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	1.2454e+11	-	-	-	-	-	-	-
	预测器	1.3036e+11	-	-	-	-	-	-	-
	EFSplus	1.3029e+11	1.07e+07	1.227	0.004	0.8876	0.0005	0.8857	0.0006
	NCN	1.2941e+11	0.00	1.446	0.002	0.9215	0.0000	0.9266	0.0000
	快速估计	1.2933e+11	6.37e+08	1.203	0.003	0.9091	0.0214	0.9105	0.0267
	快速过滤	1.2576e+11	1.35e+08	0.272	0.006	0.9279	0.0148	0.9338	0.0211
	快速采样	1.2864e+11	4.18e+06	6.673	0.013	0.9258	0.0001	0.9338	0.0001
	Ours	**1.2541e+11**	5.86e+07	**0.205**	0.000	**0.9353**	0.0039	**0.9430**	0.0049
20	K-Med++	1.0408e+11	-	-	-	-	-	-	-
	预测器	1.1107e+11	-	-	-	-	-	-	-
	EFSplus	1.1060e+11	8.42e+06	1.364	0.003	0.8941	0.0004	0.8549	0.0005
	NCN	1.0993e+11	0.00	1.490	0.004	0.9208	0.0000	0.9002	0.0000
	快速估计	1.0858e+11	3.34e+08	1.275	0.004	0.8897	0.0009	0.8484	0.0051
	快速过滤	1.0462e+11	9.59e+07	0.248	0.001	0.9464	0.0036	0.9400	0.0045
	快速采样	1.0900e+11	2.67e+06	12.059	0.032	0.9281	0.0002	0.9122	0.0003
	Ours	**1.0446e+11**	1.13e+07	**0.201**	0.000	**0.9572**	0.0011	**0.9552**	0.0017
30	K-Med++	9.6166e+10	-	-	-	-	-	-	-
	预测器	1.0296e+11	-	-	-	-	-	-	-
	EFSplus	1.0253e+11	2.13e+07	1.506	0.006	0.8814	0.0003	0.8134	0.0007
	NCN	1.0172e+11	0.00	5.078	0.010	0.8925	0.0000	0.8337	0.0000
	快速估计	1.0050e+11	3.14e+08	1.363	0.003	0.8874	0.0091	0.8322	0.0196
	快速过滤	9.6482e+10	3.18e+07	0.224	0.001	0.9546	0.0030	0.9435	0.0060
	快速采样	1.0091e+11	2.98e+06	16.786	0.136	0.9120	0.0001	0.8736	0.0002
	Ours	**9.6401e+10**	1.04e+07	**0.199**	0.001	**0.9645**	0.0013	**0.9576**	0.0022
40	K-Med++	9.1325e+10	-	-	-	-	-	-	-
	预测器	9.8290e+10	-	-	-	-	-	-	-
	EFSplus	9.7755e+10	9.39e+06	1.593	0.007	0.8784	0.0003	0.7985	0.0007
	NCN	9.6438e+10	0.00	5.042	0.005	0.8950	0.0000	0.8224	0.0000
	快速估计	9.4824e+10	1.85e+08	1.408	0.002	0.8880	0.0037	0.8184	0.0116
	快速过滤	9.1581e+10	4.81e+07	0.224	0.001	0.9534	0.0039	0.9397	0.0054
	快速采样	9.6159e+10	1.67e+06	21.250	0.065	0.9117	0.0002	0.8683	0.0002
	Ours	**9.1515e+10**	4.14e+06	**0.201**	0.001	**0.9621**	0.0019	**0.9515**	0.0031
50	K-Med++	8.7954e+10	-	-	-	-	-	-	-
	预测器	9.5100e+10	-	-	-	-	-	-	-
	EFSplus	9.4367e+10	2.08e+07	1.714	0.006	0.8737	0.0005	0.7765	0.0011
	NCN	9.2909e+10	0.00	5.102	0.016	0.8878	0.0000	0.7986	0.0000
	快速估计	9.1920e+10	1.39e+08	1.492	0.001	0.8759	0.0026	0.7837	0.0067
	快速过滤	8.8252e+10	4.96e+07	0.211	0.001	0.9506	0.0010	0.9300	0.0014
	快速采样	9.2943e+10	3.48e+06	25.838	0.054	0.9072	0.0001	0.8537	0.0002
	Ours	**8.8179e+10**	6.09e+06	**0.204**	0.002	**0.9609**	0.0011	**0.9486**	0.0019

表18：固定 $\alpha = 0.2$ 并改变 k时在 PHY数据集上的性能比较。

k	条件 / 方法	Cost		时间（秒）		NMI		ARI	
		Avg.	标准差	Avg.	标准差	Avg.	标准差	Avg.	标准差
10	K-Med++	1.0147e+12	-	-	-	-	-	-	-
	预测器	1.3965e+12	-	-	-	-	-	-	-
	EFSplus	1.0562e+12	4.32e+09	0.109	0.000	0.9591	0.0045	0.9523	0.0068
	NCN	1.0209e+12	0.00	0.411	0.001	0.9829	0.0000	0.9834	0.0000
	快速估计	1.0657e+12	1.48e+10	0.096	0.000	0.9530	0.0149	0.9439	0.0218
	快速过滤	1.0246e+12	1.43e+08	**0.020**	0.000	**0.9882**	0.0007	**0.9902**	0.0008
	快速采样	1.0620e+12	2.00e+10	0.471	0.004	0.9590	0.0065	0.9516	0.0120
	Ours	**1.0206e+12**	1.47e+09	0.043	0.000	0.9737	0.0074	0.9705	0.0104
20	K-Med++	2.9135e+11	-	-	-	-	-	-	-
	预测器	6.1176e+11	-	-	-	-	-	-	-
	EFSplus	3.0869e+11	6.94e+08	0.126	0.000	0.9654	0.0024	0.9488	0.0039
	NCN	2.9964e+11	0.00	0.414	0.001	**0.9858**	0.0000	**0.9827**	0.0000
	快速估计	3.1909e+11	1.05e+10	0.102	0.001	0.9311	0.0123	0.8793	0.0287
	快速过滤	2.9358e+11	4.09e+08	**0.022**	0.000	0.9720	0.0015	0.9592	0.0026
	快速采样	3.0647e+11	1.16e+09	0.821	0.006	0.9617	0.0066	0.9415	0.0116
	Ours	**2.9357e+11**	4.20e+08	0.046	0.000	0.9768	0.0041	0.9675	0.0074
30	K-Med++	1.6400e+11	-	-	-	-	-	-	-
	预测器	4.3770e+11	-	-	-	-	-	-	-
	EFSplus	1.7918e+11	3.16e+08	0.141	0.000	0.9695	0.0052	0.9509	0.0095
	NCN	1.7292e+11	0.00	0.416	0.002	**0.9797**	0.0000	**0.9695**	0.0000
	快速估计	1.9011e+11	7.78e+09	0.111	0.001	0.9341	0.0056	0.8711	0.0146
	快速过滤	1.6490e+11	1.16e+08	**0.023**	0.000	0.9775	0.0017	0.9655	0.0035
	快速采样	1.7574e+11	4.23e+08	1.189	0.003	0.9709	0.0037	0.9552	0.0076
	Ours	**1.6468e+11**	9.90e+07	0.051	0.000	0.9778	0.0021	0.9666	0.0034
40	K-Med++	1.1922e+11	-	-	-	-	-	-	-
	预测器	3.6763e+11	-	-	-	-	-	-	-
	EFSplus	1.3394e+11	1.26e+08	0.161	0.000	0.9725	0.0015	0.9527	0.0031
	NCN	1.2603e+11	0.00	0.421	0.001	0.9736	0.0000	0.9549	0.0000
	快速估计	1.4017e+11	3.93e+09	0.119	0.000	0.9291	0.0051	0.8526	0.0115
	快速过滤	1.2006e+11	5.77e+07	**0.024**	0.000	0.9838	0.0024	0.9748	0.0049
	快速采样	1.3034e+11	4.65e+08	1.573	0.007	0.9705	0.0053	0.9482	0.0110
	Ours	**1.1993e+11**	3.56e+07	0.057	0.000	**0.9862**	0.0033	**0.9787**	0.0061
50	K-Med++	9.9887e+10	-	-	-	-	-	-	-
	预测器	3.1986e+11	-	-	-	-	-	-	-
	EFSplus	1.1631e+11	1.98e+08	0.175	0.000	0.9736	0.0020	0.9562	0.0045
	NCN	1.0962e+11	0.00	0.428	0.002	0.9668	0.0000	0.9365	0.0000
	快速估计	1.2197e+11	5.73e+09	0.126	0.001	0.9228	0.0091	0.8259	0.0259
	快速过滤	1.0261e+11	2.44e+07	**0.025**	0.000	0.9803	0.0009	0.9704	0.0018
	快速采样	1.1200e+11	1.96e+08	1.896	0.011	0.9689	0.0033	0.9463	0.0074
	Ours	**1.0254e+11**	1.70e+07	0.062	0.000	**0.9826**	0.0016	**0.9735**	0.0029