大型三维体数据的聚类：一种基于抽样的方法

Thomas Lang 弗劳恩霍夫集成电路研究所，X射线技术开发中心部，Innstraße 43，德国帕绍 94032，电子邮箱：thomas.lang@iis.fraunhofer.de

摘要

在许多 X 射线计算机断层成像应用中，对重建的三维体数据进行无监督分割，是进一步研究数字化对象的图像处理流程中的重要步骤。因此，目标是在体数据上训练聚类算法，通过为每个体素分配一个簇索引来实现体素级分类。然而，聚类方法（例如 K 均值）通常具有相对于数据集大小呈多项式增长的渐近运行时间，因此很少适用于大型体数据。本文介绍了一种基于随机采样的新型聚类技术，可对任意大的体数据进行体素级分类。所提出的方法通过高效地对数据进行线性遍历，提取固定大小的代表性随机样本，并以此训练分类器。随后，再通过一次线性遍历完成分割，并为每个体素分配簇索引。定量和定性评估表明，即使样本量很小，也能取得出色的结果。因此，借助聚类对任意大的体数据进行无监督分割成为可能。

1. 引言

聚类，即对数据集进行无监督划分，是许多数据处理应用中的重要步骤[1]，计算机断层扫描（CT）也不例外。在 CT 中，聚类主要用于图像分割，尤其常见于医学成像领域[2,3]。过去几十年来，计算机断层扫描图像处理领域涌现出许多不同的技术，包括 K 均值、DBSCAN、模糊聚类等。大多数方法都会从数据集中提取信息，也就是说，它们需要经过训练，才能用于处理未见过的数据集并对其进行划分。然而，尽管近年来采样理论已应用于特征空间，例如用于降维[4]，但实际上鲜有研究考虑对数据集进行抽样，以便将聚类应用于工业计算机断层成像中出现的更大规模数据集。一项值得注意的研究也采用了分层抽样方法来提取数据的代表性子集，并通过某种（可能计算量很大的）哈希机制定义分层，据此对数据集进行聚类[5]。不过，该研究中的数据集约包含五十万个数据点，而在计算机断层扫描领域，单个体数据中就包含数十亿个体素。

本研究关注工业 CT 生成的大型体数据集，并将广为人知的随机采样方法与 CT 领域及其中的聚类相结合。除实践评估外，还将给出理论结果。

2. 大规模数据的随机采样

为了从一般的数据项总体中抽取随机样本，已有种类繁多的算法[6]。其中最著名的有伯努利抽样、均匀抽样和蓄水池抽样。后者有许多优点，包括能够从规模可能任意大的总体中抽取固定大小的均匀分布样本，而且事先甚至不必知道总体规模。因此，它尤其适合从可视为数据流的超大型体数据集中抽取随机样本。

2.1 蓄水池抽样与算法 L

本文重点关注蓄水池抽样，它通过一次线性遍历，从总体中抽取一个服从均匀分布的样本。在此过程中，标题所指的蓄水池会被填满，之后再随机选取元素并用新元素替换。遍历总体之后，该蓄水池便构成抽取出的样本[7]。后来出现了一种名为算法 L[8] 的改进算法，它通过观察发现，向蓄水池插入元素之间的“跳跃”服从几何分布，从而避免了为总体中的每个元素生成随机数这一高昂开销。因此，算法 L（如下方伪代码所示）会跳过若干元素而不考虑将其插入蓄水池；插入间隔则通过从几何分布中抽取一个样本来计算。下文中，$rand(a,b)$从区间$[a,b]$上的均匀分布中抽取一个样本，而$randi(a,b)$则从该离散范围中抽取一个整数样本。

输入：总体 $X=\{x_1,x_2,\dots\}$，样本量 M 输出：样本 S

算法 L：1。

$S=\{x_1,\dots,x_M\}$

2 $W=$ $rand\ 0,1\ /M$. exp (log (( ))

3

. $i=K-1$. 当未完成时执行

4

i += ⌊ $rand\ 0,1\ /\ 1-W$ ⌋ + 1 log (

( )) log ( ) 如果尚未到达数据流末尾，则

$r=randi(1,K)$ $S_r=x_i$ exp (log ($rand$

$W*=$

(0,1)) $/M$) fi done 5. 返回 S

如前所述，蓄水池抽样和算法 L 的主要优点是：只需对总体进行一次线性遍历，即可获得大小固定为 M 的均匀分布随机样本。首先，将前 M 个观测值放入蓄水池。然后，以 M/i 的概率接纳索引为 i 的每个（未跳过的）总体项目，并用它替换蓄水池中的一个随机条目。最终即可获得一个均匀分布的样本[9]。

2.2 分层随机抽样

尽管算法 L 在计算开销方面效率很高，但它仍然从总体中抽取均匀分布的样本。然而，尤其是在计算机断层扫描中，灰度值分布通常偏向于零，即观测到的零值远多于表征材料体素的值。因此，视数据集而定，有时上述算法无法从某些材料中抽取任何体素值，因为这些材料出现的频率远低于零值。为弥补这一点，分层随机抽样更为可取。

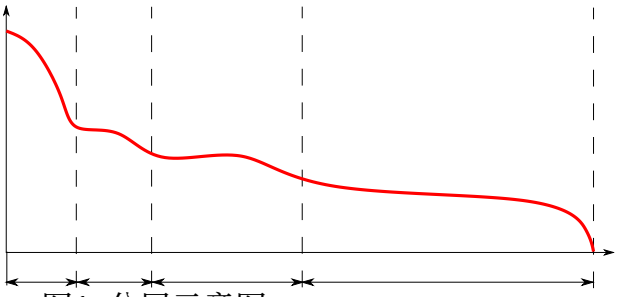

图1. 分层示意图。

其核心思想是，预先将断层扫描中可观测到的灰度值范围划分为不同的层。随后，上述随机采样从每一层中抽取具有代表性的值子集，这些样本共同构成最终样本。图1展示了这一思想：将可观测灰度值的范围（x轴）划分为四个部分，以适应观测到的灰度值分布（红线）。自然，一个尚待解答的问题是如何进行这种划分。原则上，任何如上图所示对该范围进行划分的分层方式都是有效的。此外，实践中发现，对于给定数量的K 个层，以下三种策略适用于X射线计算机断层成像。这里，这些分层策略对单位区间进行划分，之后还需缩放该区间以匹配可观测值的范围。

在记号上，形式为$(f(i))_{i=s}^{e}$的表达式表示由值$(f(s),\dots, f(e))$构成的有序元组，而$(a, b) + (c, d) = (a, b, c, d)$表示元组连接。

2.2.1 线性分层

可以说，这是最简单的策略，它会生成间距均匀的分区，即 $B = (\frac{i}{M})_{i=0}^{K}$.

2.2.2 指数分层 指数分层 s

策略遵循这样一种直觉：观察到的灰度值分布明显偏向于零，并呈现“指数”形式，参见图1。具体而言，$B = (0,) + (2^{1-K+i})_{i=0}^{K-1}$。

2.2.3 混合分层 基于前两种

针对这些可能性，可以定义一种混合策略：在值域的前一部分（直到某个依赖于具体应用的阈值）采用指数策略，随后对其余部分进行线性划分。

2.2.4 分层随机抽样 将算法 L 与其结合

这种分层形成了整体的分层随机抽样流程：首先计算分层，然后估计每层的样本量。后者会计算原始体数据中各层内数值的相对频率，并按计划抽取的元素总数进行缩放。接下来，对每一层执行算法 L，即从落入各个层的体素子集中抽取均匀随机样本。将所有样本合并后，即得到最终抽取的样本。

需要处理的一个具体实际问题是，对于某些数据分布，某一层中的样本量可能为零。当某一层在总体中包含的值相对较少时，就会出现这种情况，因为这些值的数量会按比例缩放到可能较小的总体样本量。由于样本量始终为整数，个别层的样本量可能为零。在这种情况下，不会从这些区域抽取样本，尽管下面的理论结果依赖于至少从某一层中选取一个实例。

2.3 理论结果

关于蓄水池抽样，文献中给出了必要样本量以及样本相对于总体分布的近似质量的上界。假设总体大小为 N 个元素，样本量为 M，并将从总体 X中抽取的样本记为 S 。那么，相对频率（其构成了对实际数据分布的近似）的近似误差有界为[10]

$$\sup_{E\in\mathcal{E}}\left|\frac{|S\cap E|}{M}-\frac{|X\cap E|}{N}\right|\leq\mathcal{O}\left(\sqrt{\frac{d+\log(1/\delta)}{M}}\right)$$

概率至少为 $1-\delta$，其中 E 是 X 的子集族，d 是 E 的 Littlestone 维数。简单来说，样本量越大，近似误差越小，也就是说，样本越能准确代表总体。

上述结论的证明过程相当复杂，且依赖一些较弱的假设[10]。建议读者参阅相关内容。本文将给出应用分层时的类似界限。

引理 1. 考虑上述参数，并假设满足相同的前提条件[10]。记 c_k 为层 $X_k\subseteq X, k=1,\ldots,K$ 中各值的相对频率。那么，分层蓄水池抽样的抽样近似误差有如下界：

$$\sup_{E\in\mathcal{E}}\left|\frac{|S\cap E|}{M}-\frac{|X\cap E|}{N}\right|\leq\mathcal{O}\left(\sqrt{\frac{d+\log(1/\delta)}{M\min\limits_{1\leq k\leq K}c_k}}\right)$$

概率至少为 $1-\delta$。

证明。根据定义，$c_k = N_k/N$ 其中 N_k 是层 X_k、 $k = 1$、……、 K中包含的项目数。据此，计算每层的样本量为 $n_k = Mc_k$。定义 $\mathrm{E}_k \subseteq \mathrm{E}$，其中 $\mathrm{E}_k \cap 2^{X_k} \neq \emptyset$ ，E 如上所述。

首先，注意到由于各层两两不相交

$$|X \cap E| = \left|\bigcup_{k=1}^{K} X_k \cap E\right| = \left|\bigcup_{k=1}^{K} (X_k \cap E)\right| = \sum_{k=1}^{K} |X_k \cap E|$$

并且对每个 $E \in$E，层 S_k 中的样本也同理。因此，

$$\sup_{E\in\mathcal{E}} \left|\frac{|S \cap E|}{M} - \frac{|X \cap E|}{N}\right| \leq \sum_{k=1}^{K} \sup_{E\in\mathcal{E}} \left|\frac{|S_k \cap E|}{M} - \frac{|X_k \cap E|}{N}\right|$$

由于各层（以及每层中的样本）与该层以外的总体子集的交集为空集，右侧可简化为将 E 改为 E$_k$。此外，有

$$\left|\frac{|S_k \cap E|}{M} - \frac{|X_k \cap E|}{N}\right| = \left|\frac{|S_k \cap E|}{n_k} c_k - \frac{|X_k \cap E|}{N_k} c_k\right| = |c_k| \left|\frac{|S_k \cap E|}{n_k} - \frac{|X_k \cap E|}{N_k}\right|$$

因此，

$$\sup_{E\in\mathcal{E}} \left|\frac{|S \cap E|}{M} - \frac{|X \cap E|}{N}\right| \leq \sum_{k=1}^{K} \underbrace{|c_k|}_{\leq 1} \sup_{E\in\mathcal{E}_k} \left|\frac{|S_k \cap E|}{n_k} - \frac{|X_k \cap E|}{N_k}\right| \leq \mathcal{O}\left(\sqrt{\frac{d(\mathcal{E}_k) + \log(1/\delta)}{\min_{1\leq k\leq K} n_k}}\right)$$

以至少 $1 - \delta$ 的概率成立，其中 d(E$_k$) 是集合 E$_k$ 的 Littlestone 维数。最后，由于对于这些子集，d(E$_k$) $< d$ (10, 式 15)，我们得到

$$\sup_{E\in\mathcal{E}} \left|\frac{|S \cap E|}{M} - \frac{|X \cap E|}{N}\right| \leq \mathcal{O}\left(\sqrt{\frac{d + \log(1/\delta)}{M \min_{1\leq k\leq K} c_k}}\right)$$

以至少 $1 - \delta$ 的概率成立，证明至此完成。 □

直观而言，这个新的界限表达了同一个观点：样本量越大，每个层中的样本量也越大，因此对整体数据分布的近似就越好。在这一数学表述中，未考虑层中样本为空的情况，这种情况也不合规范，最好通过程序来处理。

3. 基于随机采样的聚类

先前的算法很容易应用于三维体数据集，因为这类数据集常由计算机断层扫描生成。只需将每个体素及其对应的灰度值视为整体总体（即该体积）中的一个独立成员即可。
基于提取的样本，可以使用任何特定于应用的方法来获取信息，或训练 AI 模型执行逐体素分割等任务。尤其是后一种情况，本文重点关注聚类算法，因为其训练过程相对于训练数据规模通常具有多项式运行时间复杂度。需要先说明的是，以下技术不仅适用于普通灰度值，也适用于从数据中获得的特征向量，只需将模型替换为相应的多维模型即可。不过，本文关注的是仅基于体素强度值的简单聚类。

3.1 K 均值

K 均值聚类或许是最常用的聚类方法，其基本思想是通过最小化总体 X 中每个值到最近聚类中心的距离，将总体划分为 X_1、……、 $X_K \subseteq X$ ，即

$$\underset{X_1,\ldots,X_K}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{x \in X_j} \|x - \mu_j\|_2^2, \quad \text{s.t.} \quad X = \bigcup_{j=1}^{K} X_j$$

其中，所有子集 X_j 两两不相交，μ_j 表示每个子集的均值。
一般而言，这个问题已知是 NP-困难，因此存在许多启发式算法来较好地求解它。其中一个例子是 Lloyd算法[11]，其运行时间复杂度为多项式[12]。

训练完成后，通过将每个体素归入中心与其数值最接近的子集，并将该子集的索引赋给该体素，对其进行分类，这与上述表述类似。如果出现并列，则选择最小的索引。

3.2 小批量 K 均值

小批量 K 均值是经典 K 均值聚类算法的多种改进方法之一。具体而言，在内部训练循环的每次迭代中，都会从总体中随机抽取一批总体项目，批次大小可配置。基于该子集执行常规的 K 均值训练，并持续运行整个算法，直到满足某个收敛准则[13]。
此处不再展开详述，但需要指出的是，与我们的方法不同，内部采样既不考虑实际的数据分布，也无法得出任何理论结果。此外，该方法假定整个总体均已存在，并且要求支持随机访问，而我们的方法则不需要。在所进行的实验中，我们研究了两种方法的结合，即使用本工作介绍的技术提取随机样本，并以此训练小批量 K 均值算法。

3.3 使用高斯混合模型进行聚类

经典 K 均值聚类的一种常见替代方法是使用高斯混合模型进行逐体素分配。高斯混合模型是一种概率模型，试图通过多个高斯分布的加权叠加来近似整体灰度值分布；也就是说，该方法假设模型的形式为

$$\sum_{j=1}^{K} \pi_j \, \mathcal{N}\left(\cdot \,|\mu_j, \Sigma_j\right), \quad \text{where} \quad 0 \le \pi_j \le 1, \; \sum_{j=1}^{K} \pi_j = 1$$

更具体地说，此处的目标是求出分布参数 μ_j 和 Σ_j，以及各个分布对应的权重因子 π_j 。通常采用期望最大化算法来确定模型参数。该算法保证收敛，但可能需要经过许多次迭代。每次迭代都需要进行若干次矩阵乘法，其计算量随矩阵大小呈立方级增长，因此计算成本非常高。

通过为总体中的每个元素分配使其似然函数达到最大值的高斯分布索引，来完成划分。或者，也可以根据此类模型为单变量情形[14]推导出一种阈值方法。

3.4 将聚类与随机采样相结合

将前文介绍的随机采样技术与上述聚类方法相结合，过程十分直接：

1. 从该体积中抽取固定大小的随机样本。 2. 使用该样本训练逐体素分类器。
3. 根据模型，为该体积中的每个体素分配一个簇索引。

由于随机采样只在线性遍历体积时处理局部信息，分类步骤也只涉及局部信息，因此只要样本能够放入内存以训练模型，该方法就适用于任意大小的体积。更重要的是，整个过程的运行时间随体积中的体素数量线性增长，而训练模型所需的多项式运行时间仅用于样本；样本大小是固定的，且通常与体积大小无关。

从技术层面来说，随机采样的实现封装在一个 C库中。此外，还提供了 Python 绑定，因此在提取出的样本将被输入 Python模块的场景下，可以使用这个快速库轻松地从数值数据流（即数据的灰度值）中进行采样。

4. 结果

本研究开展的实验分为两部分：定量测量和定性结果。

对于定量结果，分别使用 K 均值、高斯混合模型和小批量 K 均值对每个数据集进行聚类，并将其作为基线。接下来，按照所提出的方案从数据集中提取（分层）随机样本，并使用这些方法进行另一次聚类

。样本量有所不同，既包括较小的固定样本量，也包括占数据体素数量一定比例的样本量。对于每种聚类结果，记录抽样和训练聚类方法所需的时间，并与基线的训练时间进行比较。此外，使用两种聚类相似性指标（福克斯-马洛斯指数[15]和归一化互信息[16]，后者按平均熵归一化）对基线结果与基于随机抽样支持的聚类结果进行比较。这两种指标计算聚类相似度，取值范围为零（随机信息，没有关联）到一（信息完全相同，但聚类索引可能不同）。

为进行定性评估，我们将基于分层随机样本（采用指数分层方案）训练的高斯混合模型聚类应用于选定的大型计算机断层扫描数据集。需要注意的是，由于数据集规模庞大，本例无法进行额外的定量评估。
所有结果均采用 Python 库 scikit-learn 中上述方法当前最先进的聚类实现，并结合以 C 语言实现、通过配套 Python 绑定调用的随机采样获得。

4.1 定量结果

定量分析针对 15 个选定的计算机断层扫描图像进行，其中包括工业物件（“Kolben”、“KGH”、“Leiterplatte”、“Piston”、“Tipex”、“TPA”）、文化遗产物件（“Tafelklavier”、“TRex”、“Yorick”）、在同步辐射光束线上获取的高分辨率扫描图像（“Bamboo”、“Battery”、“Brain”），以及其他呈现有趣结构属性的扫描图像（“Ei”、“Leberkas”、“WurstMitGabel”）。

图 2、图 3 和图 4 分别以热图形式展示了高斯混合模型（GMM）、K 均值和小批量 K 均值聚类的全部结果。其中，“简单”分层对应于仅应用算法 L，即不进行任何分层。同样，图 5、图 6 和图 7 展示了相较于经典算法所实现的运行时间改进。

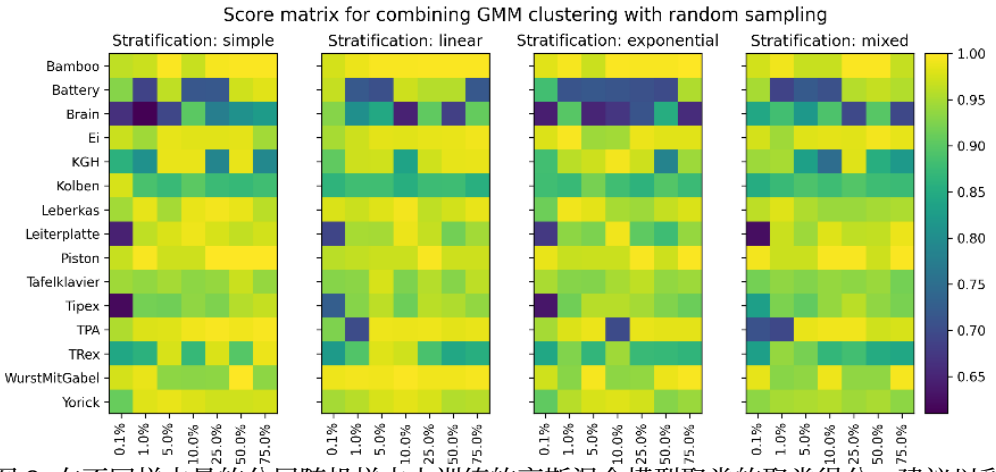

图 2. 在不同样本量的分层随机样本上训练的高斯混合模型聚类的聚类得分。建议以彩色方式欣赏此图。

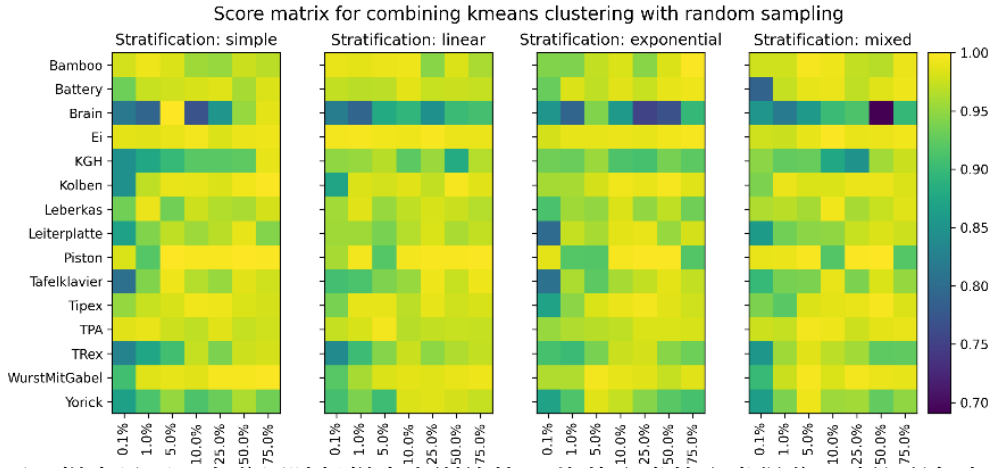

图3. 不同样本量下，在分层随机样本上训练的 K 均值聚类的聚类得分。建议彩色查看。



图4. 不同样本量下，在分层随机样本上训练的小批量 K 均值聚类的聚类得分。建议彩色查看。

总体而言，即使样本量非常小，仅为体素数量的 0.1%，也能获得良好的聚类结果；而且随着样本量（即训练集大小）的增加，得分大多会提高。自然地，在某些扫描中会出现一定波动，但需要注意的是，高斯混合模型、K 均值和小批量 K 均值的最低得分分别为 69%、61% 和 56%。此外，小批量 K 均值通常与我们的采样方法结合使用时表现不佳。这可能是因为该方法本身已经进行了一些内部采样，因此额外的采样会进一步丢失信息。不过，该变体需要对整个数据集进行随机访问，而所提出的采样方案不需要，因此在这里应优先采用所提出的方法。

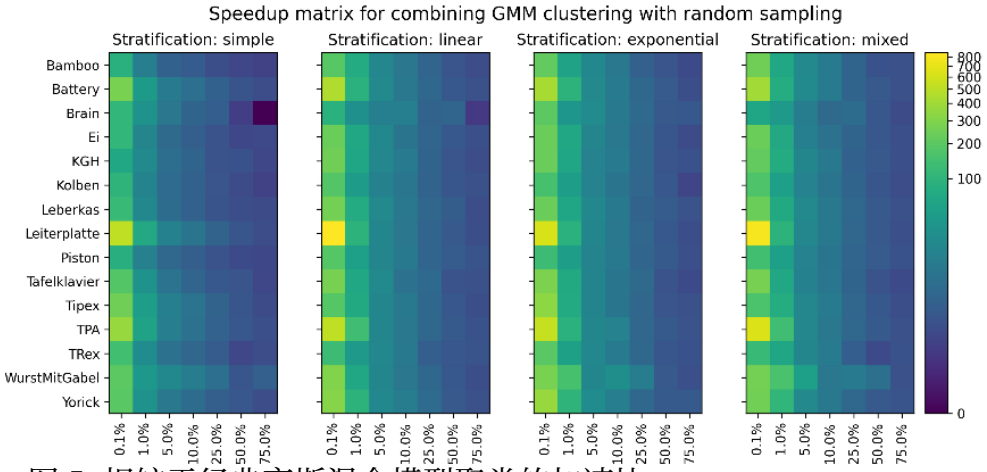

图 5. 相较于经典高斯混合模型聚类的加速比。



图 6. 相较于经典 K 均值聚类的加速比。

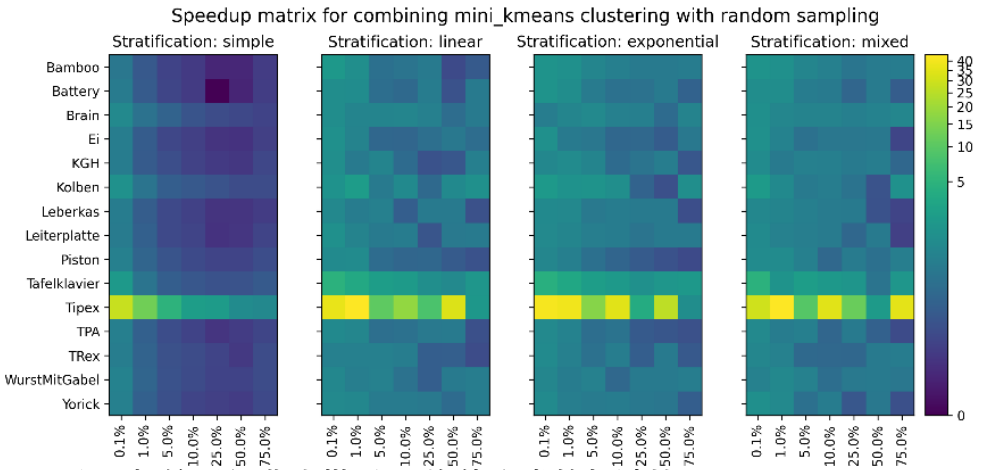

图 7. 相较于经典小批量 K 均值聚类的加速比。

在加速效果方面，无论是使用高斯混合模型还是 K 均值进行聚类，都实现了显著加速（此处最高分别为：高斯混合模型 851 倍，K 均值 541 倍），且加速比会随样本量增加而降低。例外情况与之前相同，即小批量 K 均值聚类：在大多数情况下，加速效果较小，甚至会出现“减速”。

4.2 定性结果

图8展示了一些定性聚类结果，呈现了该聚类方法在“Brain”（30.80 GB）、“Kidney”（7.26 GB）和“Leberkassemmel”（4.50 GB）扫描上的应用。后者由弗劳恩霍夫X射线技术开发中心提供，而“Brain”和“Kidney”取自人体器官图谱项目[17]。对于每项结果，均采用指数分层策略抽取仅含 4096 个元素的样本。抽样分别耗时 77 秒、14 秒和 9 秒。图8左侧展示原始数据集，右侧展示聚类结果，聚类分配效果良好。即使对于含噪声且低对比度的扫描图像，事实证明，样本量不受体积大小影响的小样本也包含了足够的信息，可用于训练鲁棒模型。该方法还应用于规模大得多的数据集（目前最大达 250 GB），但这些数据集缺少所展示扫描图像所具有的有趣结构属性。

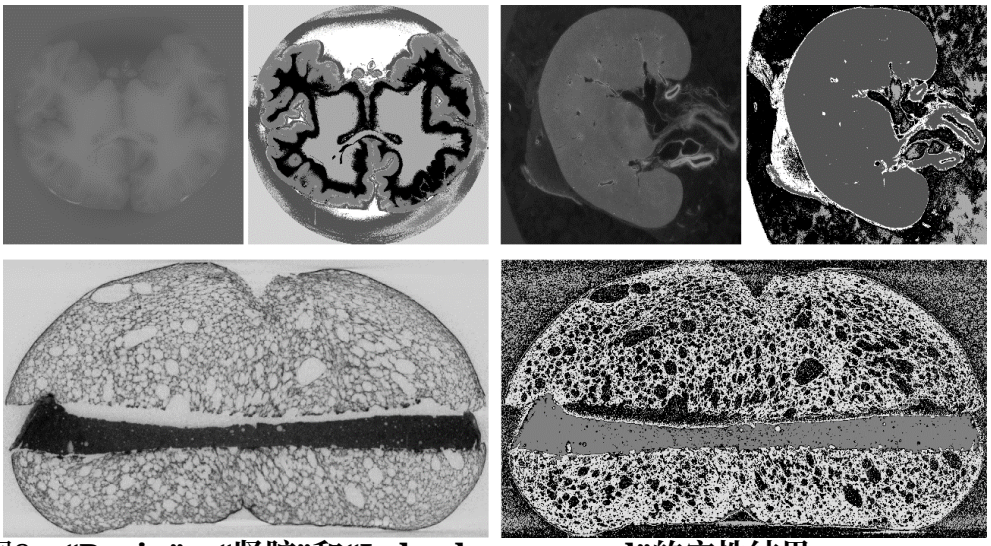

图8。**“Brain”、“肾脏”和“Leberkassemmel”的定性结果**。

5. 结论

本研究提出将聚类算法与随机采样相结合，从而实现对任意大型数据集进行体素级分类。分层可将有关灰度值分布的先验信息纳入分析，以增强鲁棒性。定性和定量分析均表明，即使提取的样本非常小，也能取得很好的结果。

参考文献和脚注

1. A Jain、M Murty 和 P Flynn，《数据聚类：综述》，ACM Comput. Surv.，第31卷，第3期，第264-323页，1999年

2. M Sutton, J Bezdek, T Cahoon，《模糊聚类图像分割：方法与问题》，载于：Bankman（编），《医学成像手册：处理与分析》，第87-106页，2000年

3. A Meyer-Baese 等，《无监督聚类方法在医学成像中的应用》，5th 自组织映射研讨会，2005年 4. N Tremblay 和 A Loukas，《通过抽样近似谱聚类：综述》，载于：F Ros、S Guillaume（编）《监督与无监督任务的抽样技术》，《无监督与半监督学习》，2020年 5. X Zhao、J Liang 和 C Dang，《一种基于分层抽样的大规模数据聚类算法》，Knowledge-Based Systems，第163卷，第416-428页，2019年 6. B Ghojogh 等，《抽样算法：从调查抽样到蒙特卡罗方法——教程与文献综述》，arXiv，2020年 7. J Vitter，《使用蓄水池进行随机采样》，ACM数学软件汇刊，第11卷，第1期，第37-57页，1985年。8. K Li，《时间复杂度为 O(n(1+log(N/n))) 的蓄水池抽样算法》，ACM数学软件汇刊，第20卷，第4期，第481-493页，1994年。9. T Roughgarden 和 G Valiant，《CS168：现代算法工具箱，第13讲：抽样与估计》，在线发表于 http://timroughgarden.org/s17/l/l13.pdf，2017年，访问日期：7月7th，2022年10. N Alon 等，《在线分类中的对抗性大数定律与最优遗憾》，载于第53rd 届 ACM SIGACT 计算理论年会（STOC）论文集，第447-455页，2021年 11. S Lloyd，《PCM 中的最小二乘量化》，IEEE信息论汇刊，第28卷，第2期，第129-137页，1982年。12. D Arthur、B Manthey 和 H Röglin，《K 均值具有多项式平滑复杂度》，arXiv：0904.1113，2009年。13. D Sculley，《网络规模K均值聚类》，载于第19th 届万维网国际会议论文集，第1177-1178页，2010年。14. T Lang，《人工智能辅助交互式分割三维体数据》，博士论文，帕绍大学，2021年。15. E Fowlkes 和 C Mallows，《比较两种层次聚类的方法》，美国统计学会杂志，第78卷，第383期，第553-569页，1983年。16. T Kvalseth，《熵与相关性：一些评论》，IEEE系统、人类与控制论汇刊，第17卷，第3期，第517-519页，1987年。17. C Walsh、P Tafforeau、W Wagner 等，《利用分层相衬断层成像对完整人体器官进行成像，细胞结构的局部分辨率可达细胞级》，Nature Methods，2021年