arXiv:2305.03022v1 [cs.LG] 3 May 2023

FastAMI——一种用于聚类比较指标中偶然性调整的蒙特卡洛方法

Kai Klede、Leo Schwinn、Dario Zanca、Bj¨orn Eskofier

机器学习与数据分析实验室，弗里德里希-亚历山大埃尔朗根-纽伦堡大学（FAU）¨{kai.klede、leo.schwinn、dario.zanca、bjoern.eskofier}@fau.de

摘要

聚类是机器学习的核心，其应用随着数据日益丰富而不断拓展。然而，随着数据集规模增大，对聚类进行比较并校正偶然因素在计算上会变得困难，从而妨碍无偏的真实值比较和解决方案选择。我们提出了 FastAMI，这是一种基于蒙特卡罗的方法，可高效近似调整互信息（AMI），并将其扩展至标准化互信息（SMI）。我们使用合成数据和真实数据，将该方法与精确计算以及近期提出的一种基于成对置换的 AMI 变体进行了比较。与精确计算相比，我们的方法速度足够快，能够对大型数据集进行这类校正后的信息论比较，同时比成对方法的结果准确得多。

1 引言

聚类比较指标，例如互信息或兰德指数，通常在有真实值信息时用于验证聚类（Aggarwal 和 Reddy，2014年）。在图的语境中，两个节点邻域的互信息反映了它们的相似性，并可用于预测缺失链接（Hoffman、Steinley 和 Brusco，2015；Shakibian 和 MoghadamCharkari，2017）。另一种应用是多聚类算法，它们通过衡量互信息或其某种变体，为单个数据集识别多个在定性上不同的聚类解（M¨uller 等人，2010；Wei 等人，2021）。其他用途包括分类特征选择（其中每个特征都被视为一个簇），以及共识聚类中的解选择（Lancichinetti 和 Fortunato，2012）。

这些聚类比较指标的一个众所周知的问题是，在比较两个随机聚类划分时，它们不假定基线值恒定。相反，簇的数量趋近于数据点数量时，这些指标往往会变大（ Nguyen、Epps 和 Bailey，2009），因此将其用作外部验证准则时，会偏向较小的簇。为获得恒定基线，人们根据簇标签在随机置换下的期望值对这些指标进行调整，见调整兰德指数（ ARI ）（ Hubert 和 Arabie，1985 ）和调整互信息（ AMI ）（ Nguyen 、 Epps，和 Bailey（2010）。

Romano 等人（2016）概述了上述及其他聚类比较指标。他们分析了这些指标之间的关系，并制定了指南，说明在不同情况下应使用哪种指标。直接引用摘要中的话，作者得出结论：

> 当参考聚类中存在多个大小相同的大簇时，应使用调整兰德指数（ARI）；当参考聚类不平衡且存在小簇时，应使用调整互信息（AMI）。

然而，已有研究表明，调整互信息的计算复杂度为 $\mathcal{O}(\max(R,C)N)$，其中 R,C 分别是待比较聚类中的簇数，N 是数据点数（Ro- mano 等，2014年）。对于包含许多小簇和少数较大簇的大型数据集——这正是适合使用调整互信息的场景——其计算会变得困难，往往不切实际。许多来自不同领域的数据集都具有这些特征，例如社交网络、协作网络或网络图（Leskovec 和 Krevl，2014）。在实践中，许多作者转而使用未经调整的归一化互信息（Tian 等，2019；Chunaev，2020年；Karim 等，2021年）；这种方法计算速度更快，但无法校正随机巧合。

作为一种变通方法，Lazarenko 和 Bonald（2021）提出，在偶然性调整中仅考虑成对置换，即只对两个独立样本交换其簇标签的聚类进行调整。虽然这种方法能够加快计算速度（$\mathcal{O}(RC)$），但这些成对置换会导致更高程度的共享信息。如图1所示，实际的期望互信息 （EMI）被成对期望互信息高估。因此，成对调整互信息会系统性地低估调整互信息，导致其结果无法与现有数值进行比较。如表1所示，甚至簇相似度的相对排序也会受到影响。

借助 FastAMI，我们提出一种基于蒙特卡罗的方法，使得能够对具有大量数据和较小簇的数据集进行 AMI 比较[1]。所提出的方法可对准确度进行细粒度控制，并弥补成对方法的不足。我们证明，其预期运行时间的悲观上界与成对

[1]代码位于 https://github.com/mad-lab-fau/fastami-benchmark

方法相当。我们通过合成数据和真实数据证明，FastAMI 在实践中显著更快，同时能生成更准确的结果。即使精确实现和成对实现因内存需求过高或运行时间不切实际而无法使用，FastAMI 仍然适用。在此过程中，我们还发现并修正了合成聚类比较基准测试中的一个常见缺陷。

尽管调整互信息在比较两个随机聚类时是无偏的，但已有研究表明，在将多个随机聚类与固定的真实标签进行比较时，它仍然表现出偏向选择较大簇的倾向（Romano 等，2014年）。作为解决方案，有人建议同时考虑标准化互信息（标准化互信息）中的方差；根据作者的说法，当 $N/RC < 5$ 且比较三个以上的聚类时，应采用该指标。然而，标准化互信息的应用仍落后于调整互信息，一个潜在原因可能是其高昂的计算成本 $\mathcal{O}(RCN^3)$。我们将基于蒙特卡罗的方法扩展到标准化互信息，使其成为适用于小型到中型数据集的计算可行替代方案。

2 背景与相关工作

令 S 为一组 N数据点 $\{s_1, s_2, \dots s_N\}$，并令两个聚类分别为满射 $u\colon\ S \to \{1, 2, \dots, R\}$ 和 $v\colon\ S \to \{1, 2, \dots, C\}$，它们将数据集 S 划分为 R 个和 C 个簇。我们用 $a_i := |u^{-1}(i)|$，$b_j := |v^{-1}(j)|$ 表示簇大小，用 $n_{ij} := |u^{-1}(i) \cap v^{-1}(j)|$ 表示两个簇之间共享的数据点数量。

两个聚类的互信息（Edwards 2008）定义为

$$I(u, v) := \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log\left(\frac{N \cdot n_{ij}}{a_i \cdot b_j}\right), \tag{1}$$

其中，对于$n_{ij} = 0$，$n_{ij}/N \log\left(N n_{ij}/(a_i b_j)\right)$定义为零。例如，如果两个聚类将数据集中的每个点都分配到同一个簇，则它们的互信息为零。相反，如果两个聚类中的每个点都各自构成一个簇，则互信息达到最大值 $\log N$。这些简单的例子已经说明了将互信息用作聚类评估指标时存在的问题：互信息预计会随着簇的数量增加而增大（图1），从而导致其偏向于更细粒度的聚类。

因此，在随机置换聚类分配的情况下，互信息会减去其期望值，同时保持簇的数量和大小 $A = \{a_1, \dots a_R\}$，$B = \{b_1, \dots b_C\}$ 不变（随机置换模型）（Nguyen、Epps 和 Bailey 2009）

$$\mathrm{E}\{I \mid A, B\} = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{\substack{n_{ij}=\max(0,\\ a_i+b_j-N)}}^{\min(a_i,b_j)} \frac{n_{ij}}{N} \log\left(\frac{N n_{ij}}{a_i b_j}\right) P\{n_{ij} | a_i, b_j, N\}. \tag{2}$$

这里，$P\{n_{ij} | a_i, b_j, N\}$ 表示 n_{ij} 个a_i 数据点中也位于簇 $u^{-1}(i)$ 中的概率

随机置换后大小为$v^{-1}(j)$ 的 b_j，其分布由超几何分布给出。

随后可将调整互信息定义为

$$\mathrm{AMI}(u, v) := \frac{I(u, v) - \mathrm{E}\{I | A, B\}}{\mathrm{avg}(H(u), H(v)) - \mathrm{E}\{I | A, B\}}, \tag{3}$$

其中，avg($H(u)$ $H(v)$) 是给定 A，B 互信息的上界，充当归一化因子，其熵为

$$H(u) := \sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N}. \tag{4}$$

算术平均数和几何平均数、最小值及最大值均可用于 avg（Nguyen、Epps 和 Bailey，2010）。在本研究中，我们遵循 scikit-learn（Pedregosa 等，2011）的默认选择，采用算术平均数。

为降低期望互信息的计算复杂度，Lazarenko 和 Bonald（2021）将其计算限制为仅使用成对置换

$$\mathrm{EMI}_{\text{pair}}(u, v) := \mathrm{E}\{I(u, \sigma \circ v) \mid \sigma \in S_N : \sigma \text{ pairwise}\}. \tag{5}$$

当存在 i，$j \in \{1, \dots, N\}$ 使得 $\sigma(i) = j$、$\sigma(j) = i$ 以及 $\sigma(t) = t\ \forall t \neq i$，$j$ 时，置换 $\sigma \in S_N$ 是成对的。尽管原作者定义的成对调整互信息未进行归一化，但我们按照公式 3 对其进行归一化，以便与精确结果进行直接比较。

直接比较两个聚类时，调整互信息具有恒定的基线。然而，通过恒定的真实值参考来比较聚类时，调整互信息会偏向簇数较多的聚类（Romano 等人，2014年）。标准化互信息的定义为标准化互信息，其

$$\mathrm{SMI}(u, v) := \frac{I(u, v) - \mathrm{E}\{I \mid A, B\}}{\sqrt{\mathrm{Var}\{I \mid A, B\}}} \tag{6}$$

考虑了这种偏差。标准化互信息表示在随机且相互独立聚类的假设下，两个聚类之间的差异相当于多少个标准差。

3 方法

调整互信息的运行时间较长，主要是因为期望互信息的计算成本很高。标准化互信息所需的 $\mathrm{Var}\{I|A, B\}$ 计算甚至更为复杂。因此，我们通过蒙特卡洛抽样近似计算期望互信息和互信息的方差。

3.1 期望互信息的蒙特卡罗估计

我们观察到，两个簇的共享信息，即公式 2 中的各项求和项，完全由簇大小 a，b、它们的重叠 n 以及样本总数 N 决定。因此，期望互信息可以根据边缘分布 A 中簇大小为 a 的概率 $P\{a|A\}$ 重新表述。

$$E\{I|A, B\} = RC \sum_{a} \sum_{b} P\{a|A\} P\{b|B\} \sum_{n} \frac{n}{N} \log\left(\frac{Nn}{ab}\right) P\{n|a, b, N\}. \tag{7}$$

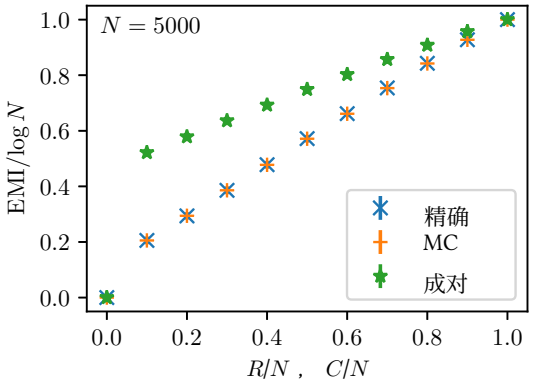

图1：随着待比较聚类中的簇数 $R = C$ 接近数据点数 N，互信息的期望值也会增大。成对置换下的期望值高估了期望互信息，而我们的蒙特卡罗（MC）方法则能复现精确结果。图中显示了 200 对聚类边缘分布 A， B的平均值，其中每个边缘分布均从将 $N = 5000$划分为 $R = C$ 个部分的所有划分中均匀随机选取。

现在，我们可以为 a， b 和 n 抽取样本，并通过样本均值近似求和。经过初始设置（耗时 $\mathcal{O}(R\log R + C\log C)$）以确定 A和 B的簇大小频数后，我们可以通过 Walker 别名法（Walker 1974, 1977）在恒定时间内轻松生成 a 和 b 的随机变量。

直接从超几何分布中抽取两个簇 n 的重叠，在簇大小较小的情况下会导致无重叠 $n = 0$ 频繁出现。然而，这些项对互信息没有贡献。为了利用这种稀疏性，我们将互信息中的线性项吸收到概率密度函数中

$$nP\{n|a,b,N\} = n\frac{\binom{b}{n}\binom{N-b}{a-n}}{\binom{N}{a}}. \tag{8}$$

二项式系数可改写为

$$n\binom{b}{n} = \frac{b!}{(n-1)!\,(b-n)!} = b\binom{b-1}{n-1}, \tag{9}$$

使我们能够根据超几何分布抽取变元 $n-1$，其中 $a-1$, $b-1$ 和 $N-1$

$$\begin{aligned} nP\{n|a,b,N\} =& \frac{a\cdot b}{N}\frac{\binom{b-1}{n-1}\binom{N-b}{a-n}}{\binom{N-1}{a-1}} \\ =& \frac{a\cdot b}{N}P\{n-1|a-1,b-1,N-1\}. \end{aligned} \tag{10}$$

已有几种方法可以在平均恒定时间内生成超几何随机变量（Kachitvichyanukul and Schmeiser 1985；H¨ormann 1994），本文选择了 Stadlober（1990）提出的均匀比方法。The

算法1 FastAMI 的核心是通过蒙特卡罗方法对期望互信息进行近似计算。我们采用数值稳定的 Welford 在线算法变体来计算均值和样本方差，并在达到所需精度时定义停止准则。

要求： 簇大小 A， B，期望互信息精度 p，最小样本数 $i_{\min} > 1$ **确保：** 期望互信息 $\sim \mathcal{N}\left(E\{I|A,B\},\ s^2_{\mathrm{EMI}}\right)$ 其中 min $\{s_{\mathrm{EMI}}/$ $s_{\mathrm{EMI}}\} \le p$ 期望互信息，期望互信息，为 A 设置 Walker 随机抽样， B**当** $i < i_{\min}$ 或 $M_2 > (p\max\{1,$期望互信息$\})^2 \cdot i \cdot (i-1)$ **时执行** $i \leftarrow i+1$ 通过 Walker 别名法抽取 a, b 通过 Stadlober 方法抽取 $m \sim P\{m|a-1,b-1,N-1\}\, x \leftarrow ab\log$ $[\ m+1\ \ N/\ ab\ ]$ ( ) ( ) 期望互信息 期望互信息 $\Delta_0 \leftarrow x -$ $+\Delta_0/i$ $\leftarrow$期望互信息 $\Delta_1 \leftarrow x -$ 期望互信息 $M_2 \leftarrow M_2 + \Delta_0\Delta_1$**结束循环** $s_{\mathrm{EMI}} = \sqrt{M_2/(i\cdot(i-1))}$

当 EMI 的相对误差估计达到所需精度s_{EMI}/EMI $\le p$ 时，方程 7 的近似计算终止。为确保 EMI 接近零时算法仍能终止，当 EMI < 1时，算法会切换到绝对误差判据 $s_{\mathrm{EMI}} \le p$。

总结上述所有步骤，我们制定了算法 1，即期望互信息的蒙特卡洛估计。完成 $\mathcal{O}(R\log R + C\log C)$ 的初始设置成本后，即可在常数时间内抽取样本，因此在实际应用中，蒙特卡罗样本的数量决定了运行时间。

3.2 期望互信息运行时间的上界

鉴于平均而言可以在常数时间内抽取样本，我们推导出收敛到期望互信息所需样本数期望值的渐近上界。

定理 1。 算法 *1* 在精度为 p、具有 R 和 C 个簇以及 $N \ge 3$ 个数据点的两个聚类上终止所需的样本数期望渐近上界为 $\mathcal{O}\left(RC/p^2\right)$。

证明。 对于给定样本数 N_{samples}， 蒙特卡洛近似s_{EMI} 的绝对误差为

$$s^2_{\mathrm{EMI}} = \frac{\mathrm{Var}\left\{\frac{RC}{N^2}ab\log\left(\frac{Nn}{ab}\right)|A,B\right\}}{N_{\text{samples}}}, \tag{11}$$

根据中心极限定理，在样本数量趋于无穷时。反过来，达到特定误差所需的期望样本数为

$$N_{\text{samples}} = \frac{\mathrm{Var}\left\{\frac{RC}{N^2}ab\log\left(\frac{Nn}{ab}\right)|A,B\right\}}{s^2_{\mathrm{EMI}}}. \tag{12}$$

Bhatia-Davis 不等式（Bhatia 和 Davis，2000）在变量有界时，为方差提供了一个上界 $m \leq \frac{RC}{N^2} ab \log \binom{N \cdot n}{ab} \leq M$

$$\operatorname{Var}\left\{\frac{RC}{N^2} ab \log\left(\frac{Nn}{ab}\right) | A, B\right\} \leq (M - \mathrm{EMI})\,(\mathrm{EMI} - m)\,. \tag{13}$$

当概率质量函数集中在区间的两个极端值时，等号成立。两个簇重叠部分所包含的信息量大于或等于零（$m = 0$），当其中一个聚类仅包含一个簇时，等号成立（ $a = N$ 或 $b = N$）。对于上界，注意到两个簇的重叠部分最大为较小簇的大小 $n \leq \min\{a, b\}$。不失一般性，假设 $a \leq b$

$$ab \log \frac{nN}{ab} \leq Nb \log \frac{N}{b} \leq \frac{N^2}{e} = M \quad \text{for } N \geq e, \tag{14}$$

由方程 13 可得所需蒙特卡罗样本数期望值的如下上界

$$N_{\text{samples}} \leq \frac{(RC - \mathrm{EMI})\,\mathrm{EMI}}{s^2_{\mathrm{EMI}}} \tag{15}$$

算法1中的收敛准则要求 $s_{\mathrm{EMI}} \leq p$对于期望互信息 ≤ 1 和 $s_{\mathrm{EMI}} \leq p$期望互信息对于期望互信息 > 1，使得

$$N_{\text{samples}} \leq \mathcal{O}\left(\frac{RC}{p^2}\right). \tag{16}$$

□

这相当于在精度为 p 时，与成对调整期望互信息具有相同的渐近运行时间。不过，Bhatia-Davis 不等式仅给出了粗略的理论上限，实验将进一步揭示该算法的运行时间。

3.3 蒙特卡罗方差估计

我们比较了两种近似计算互信息方差的方法:

- 分离蒙特卡罗——我们将方差拆分为两项 Var $\{I \mid A, B\} = \mathrm{E}\{I \mid A, B\}^2 - \mathrm{E}\{I^2 \mid A, B\}$，其中第一项按照算法1进行近似。我们将公式10及相关恒等式应用于（Romano 等人 2014年）定理1中给出的 $\mathrm{E}\{I^2 \mid A, B\}$ 显式公式，并通过类似的算法对其进行近似。

- 直接蒙特卡罗 - 不再为 n_{ij} 生成单个值，而是使用 Patefield（1981）提出的方法，一次性随机抽样所有列联矩阵的空间 $(n_{ij})_{\substack{i \in \{1,\dots,R\} \\ j \in \{1,\dots,C\}}}$，并观察互信息。EMI 和方差的估计量随后就是样本均值和方差。

直接的蒙特卡罗方法也可以应用于 EMI。然而，整个列联表都需要保存在内存中。对于 AMI 的近似而言，这并不实用，因为内存占用正是我们优化的限制因素之一（见表 2 ）。

4 实验

在上一节中，我们介绍了一种通过蒙特卡洛抽样近似 EMI 和方差的方法。现在，我们将继续评估这些方法的收敛速度。我们使用合成数据探索 R、C 和 N 的不同参数范围，并通过来自众多领域的真实数据集展示该方法的实际应用。

合成实验和聚类算法基准测试套件上的实验使用了一台配备英特尔 i7-10750H 和 32GB 内存的笔记本电脑。针对斯坦福大型网络数据集合集中较大数据集的实验则在一台配备 512GB 内存的英特尔至强 E5-2680 v4 系统上进行。

4.1 合成数据

以往研究通常假设簇的数量固定，然后将每个数据点均匀地分配到一个随机簇中，从而生成随机聚类（Nguyen、Epps 和 Bailey 2009；Romano 等人 2014年）。这种方法过度强调平衡聚类，即所有簇的大小都相近。使用随机概率分布而非均匀分配，可以使簇大小的方差更大（ Lazarenko 和 Bonald 2021）。然而，这两种方法都无法保证生成预期数量的簇。第一种情况下，恰好有一个簇被分配到零个数据点的概率为$C\,((C-1)/C)^N$，利用容斥原理，至少有一个空簇的概率为

$$P\{\exists i : i \notin \operatorname{Im} v\} = \sum_{i=1}^{C} (-1)^{i-1} \binom{C}{i} \left(\frac{C-i}{C}\right)^N. \tag{17}$$

$\operatorname{Im} v$ 表示聚类 v的聚类像，即所有标签的集合。对于固定数量的数据点 N，空簇的概率会随着簇数量的增加而增大，例如对于 $N = 100$ 和 $C = 30$，该概率已经达到 ≈ 66.5 %。当随机选择概率时，这种可能性会进一步增加，因为簇标签的概率可能任意低。

相反，我们提出一种最大熵方法：给定固定数量的数据点 N 和固定数量的簇 R, C，我们从所有可能的簇大小分布 A 和 B 的空间中进行随机采样，使其满足这些约束，即对 R 和 C 的整数划分进行均匀采样。我们采用了 Arratia 和 Desalvo（2016）提出、最初由 Fristedt（ 1993）开发的拒绝采样法。对于期望互信息（EMI）而言，对簇大小分布进行采样已经足够，因为 EMI 仅取决于 A, B; 而标准化互信息（SMI）则取决于两个具体的聚类。在这种情况下，我们根据边缘分布创建一个有序簇，并对其进行随机打乱，从而保证固定的 R, C。

期望互信息 计算调整互信息中计算成本高昂的部分，是获取互信息的期望值（期望互信息）。因此，我们比较了计算 $R = C$ 和固定 $N = 5000$ 的期望互信息的方法。我们重点关注 $R = C$，因为渐近

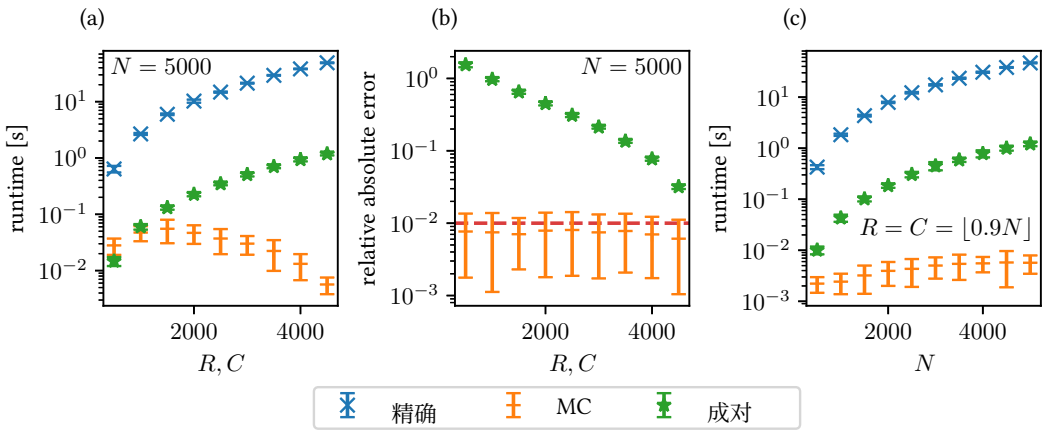

图 2：我们计算了 $R = C$ 个簇和 N 个数据点下 200 对随机簇大小分布的期望互信息，并比较了三种不同方法。(a) 随着簇数 $R = C$ 接近数据点数 N，scikit-learn（Pedregosa 等，2011）中实现的精确期望互信息和成对期望互信息的运行时间越来越长。我们的蒙特卡罗方法不仅优于后两者，而且在簇数较多时表现最佳。(b) 蒙特卡罗实现平均能够达到规定的精度 0.01（红色虚线），而成对调整会引入依赖于模型的误差。(c) 我们重复实验，但这次改变 N，同时固定 $R = C = \lfloor 0.9N \rfloor$。随着数据集规模增大，性能差距也随之扩大。

成对算法和我们的蒙特卡罗算法的运行时间在这种配置下增长最为显著。随着簇的数量从 500 增加到 4500，精确计算和成对期望互信息的运行时间均增加了近 100 倍（图 2a ）。这种性能下降令人担忧，因为正如 Romano 等人（2016）所指出的，在这种情况下，调整互信息优于兰德指数。另一方面，我们的蒙特卡罗方法在簇数量较多的情况下表现最佳。原因在于，随着 $R = C$ 增大，较小的簇逐渐占据主导地位，簇之间共享数据点 n 数量的方差随之减小，因此只需更少的蒙特卡罗采样即可实现收敛。当只有少数较大的簇时，变量 a 和 b 能够取值的簇大小数量较少，因此运行时间峰值出现在这两种极端情况之间。该实证结果证实，定理 1 中的上界高估了运行时间。

在单例簇较多的情况下，即 R，C 较高时，许多置换不会影响列联矩阵（n_{ij}），因此也不会影响互信息。这些置换会导致精确方法和成对方法的运行速度变慢，但这也意味着，省略这些置换对成对期望互信息的影响很小，因此在这种情况下相对误差较低（图 2b）。另一方面，成对方法具有较高的模型依赖误差，而在这种情况下，该方法的运行速度较快。我们的方法能够在整个参数范围内灵活调节相对误差（图 2b 中的红色虚线）。

在第二个实验中，我们展示了随着数据点数量 N 增加，精确方法和成对方法的运行时间会进一步变长（图 2c）。

SMI我们将 Romano 等人（2014）实现的精确 SMI，与用于估计方差的独立和直接蒙特卡洛方法进行比较。当估计的绝对误差或相对误差估计值达到.....时，我们终止计算

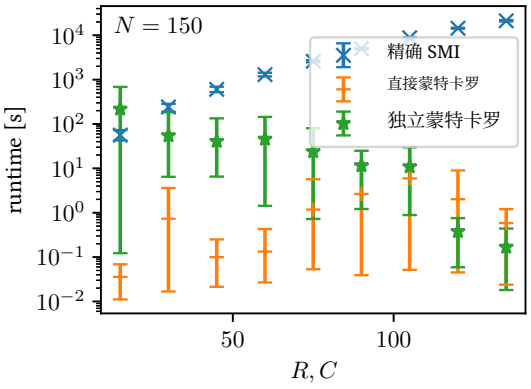

图 3：我们计算 10 对具有给定簇数量 $R = C$ 的聚类的 SMI，精度为 0.1。两种蒙特卡洛方法的表现都优于精确计算，但在簇数量多且每个簇较小的情况下，直接方法的速度快了几个数量级。

SMI 的精度低于 $p = 0.1$。两种近似算法的表现都优于精确 SMI，但分离式方法会随着簇数量增加而变慢。相比之下，直接方法在整个参数范围内都带来了数个数量级的加速（图 3）。一种可能的解释是，直接方法天然地纳入了方差大于或等于零这一约束。互信息的样本方差始终为正，但对 $\mathrm{E}\{I \mid A, B\}^2$ 和 $\mathrm{E}\{I^2 \mid A, B\}$ 分别进行估计后求差，也可能得到负值。接下来，我们为 FastSMI 选择直接方法。

变体	结果返回	Mean 运行时间 [s]	斯皮尔曼（不完整）
sklearn AMI	100%	$3.6 \cdot 10^{-3}$	-
FastAMI	100%	$1.3 \cdot 10^{-2}$	1.000
成对调整互信息	100%	$1.6 \cdot 10^{-3}$	0.577
精确 SMI	38.3 %	4.8	-
FastSMI	99.2 %	$2.4 \cdot 10^{-2}$	(0.997)

表1：FastAMI（精度为 $p = 0.01$ ）和 FastSMI（精度为 p =0.1）在聚类算法基准测试套件 （Gagolewski 等，2020）上的基准测试结果。成对调整互信息和 FastAMI 均缩短了精确版本的运行时间；FastAMI 的相对排名则几乎与精确结果完全一致。然而，最大 R/N 为 0.03 时，基准测试套件并未处于不平衡领域，而在该领域中，精确 AMI 的计算并不切实际（图 2a）。精确 SMI 在基准测试中的 66004 次比较里，有 61.7 % 超时，耗时超过 20 秒（“返回结果”列）。FastSMI 在每次比较耗时更短的情况下，在超时前返回了更多结果；对于能够获得精确结果的比较，FastSMI 与精确结果之间也实现了较高的斯皮尔曼相关系数。

4.2 真实数据

上一节展示了蒙特卡罗近似如何加快随机聚类的期望互信息和标准化互信息计算。本节将说明，这种加速如何应用于实际中更相关的调整互信息，并用于由各领域真实数据集产生的聚类。

Lazarenko 和 Bonald（2021）使用来自聚类算法基准测试套件 （Gagolewski 等，2020）的79个数据集，通过调整互信息和成对调整互信息比较多种聚类算法的解。他们采用斯皮尔曼等级相关系数来衡量结果的质量。作为第一个基准测试，我们在表1中复现了这些结果，并将其扩展以评估 FastAMI。与成对调整互信息相比，我们的方法与精确解具有显著更高的斯皮尔曼相关系数，但其运行速度较慢。另一方面，对于标准化互信息，61.7 % 的比较在20 s后超时。FastSMI（$p = 0.1$）完成了基准测试中99.2 % 的任务，且与精确结果具有较高的斯皮尔曼相关系数，从而使中等规模数据集的标准化互信息比较成为可能。然而，每次比较的平均运行时间仅为0.013 s，因此，该基准测试套件并不属于因运行时间限制而需要对调整互信息进行近似的应用领域。

相反，我们提出了一种基准，该基准基于一组大型真实世界数据集，这些数据集取自斯坦福大型网络数据集合集（ Leskovec 和 Krevl，2014年）。我们从该合集中选取被指定用于社区检测的数据集（见表 2），并应用以下六种方法来寻找聚类（Staudt、Sazonovs 和 Meyerhenke，2016）：1）连通分量；2）按度排序的标签传播；3）标签传播；4）Leiden 算法；5）Louvain 算法；以及 6）Louvain 图方程。

然后，基准测试包括对这六种聚类进行成对比较，使用调整互信息、成对调整互信息和 FastAMI。

我们测量每个数据集中的 15 组比较的总运行时间和峰值内存消耗，并将每组比较的最长运行时间限制为 2000 秒，内存限制为 503.6 吉比字节。如果超出任一限制，我们会在表 2 中报告相应的下界，并在后续计算中省略该结果。遵循 Lazarenko 和 Bonald（2021）的方法，我们报告精确结果与相应近似结果之间的斯皮尔曼等级相关系数，并排除精确计算中止的情况。此外，对于存在精确值的情况，我们报告平均绝对误差；否则，则以相应的蒙特卡洛估计代替。

当聚类不平衡时，调整互信息优于调整兰德指数（ Romano 等人，2016）。在合成实验中，我们只固定了簇的数量，而没有明确规定不平衡程度。对于真实数据集，我们使用平均归一化熵来衡量平衡性

$$\operatorname{Balance}(U) = \sum_{u_k \in U} \frac{H(u_k)}{|U| \log |\operatorname{Im} u_k|}, \tag{18}$$

其中，U 表示针对同一数据集，通过六种不同方法获得的聚类解集合。值为 1 表示所有聚类都完全平衡，即聚类中的每个簇 i 都具有相同的大小。另一方面，值为 0 表示所有聚类都由一个大型簇组成，因此是不平衡程度最大的聚类。

与合成实验的结果一致，我们的方法在运行时间方面优于其他方法。与 scikit-learn 的实现和成对调整互信息相比，FastAMI 的内存占用也低得多，因为它不会将（稀疏）列联表保留在内存中。虽然随着数据集中的节点数量增加 N，FastAMI 的运行速度会变慢，但与 scikit-learn 和成对调整互信息不同，它不会在处理不平衡数据集时遇到困难。在斯皮尔曼相关系数方面，FastAMI 的表现与成对调整相当。就质量而言，真正的优势在于，FastAMI 的结果可以直接与传统的调整互信息进行比较，平均绝对误差也印证了这一点；而成对调整互信息的结果则系统性地偏低，也更难解释。

5 结论

本文提出了一种有效且实用的方法，用于近似调整互信息。FastAMI 利用了聚类列联表的稀疏性和低方差，使得此前在计算上难以实现的数据集上的调整互信息比较成为可能。我们基于合成数据分析了近似方案的行为，并在一组真实数据集上展示了先进性能。在未来的工作中，可以轻松地并行化算法1，以进一步提升其性能。我们将该算法扩展至标准化互信息，使中等规模数据集上的方差调整聚类比较在计算上成为可行。

数据集	节点	平衡	方法	结果返回	总计时间 [秒]	Peak 内存	MAE（估计值）	斯皮尔曼（不完整）
电子邮件	$1.0 \cdot 10^3$	$3.5 \cdot 10^{-2}$	scikit-learn	15/15	0.6	208 KiB	-	-
			成对	15/15	**0.4**	184 KiB	0.357	0.943
			FastAMI	15/15	0.6	**164 KiB**	**0.001**	**0.989**
DBLP	$3.2 \cdot 10^5$	$4.3 \cdot 10^{-3}$	scikit-learn	13/15	> 7539.9	-	-	-
			成对	15/15	361.0	89.6 吉比字节	(0.310)	(0.972)
			FastAMI	15/15	**4.5**	**106.4 兆字节**	(**0.003**)	(**0.986**)
亚马逊	$3.3 \cdot 10^5$	$3.8 \cdot 10^{-3}$	scikit-learn	13/15	> 8988.4	-	-	-
			成对	15/15	329.9	57.6 吉比字节	(0.421)	(0.966)
			FastAMI	15/15	**4.5**	**26.8 兆字节**	(**0.005**)	(**0.972**)
YouTube	$1.1 \cdot 10^6$	$2.4 \cdot 10^{-4}$	scikit-learn	9/15	$> 15\,967.7$	-	-	-
			成对	13/15	-	> 503.6吉比字节	(0.260)	(0.913)
			FastAMI	**15/15**	**7.4**	**119.4 兆字节**	(**0.005**)	(**0.957**)
维基百科	$1.8 \cdot 10^6$	$2.6 \cdot 10^{-2}$	scikit-learn	15/15	714.3	12.0 吉比字节	-	-
			成对	15/15	32.7	8.4 吉比字节	0.210	**0.975**
			FastAMI	15/15	**7.8**	**140.7 兆字节**	**0.001**	0.973
Orkut	$3.1 \cdot 10^6$	$2.7 \cdot 10^{-2}$	scikit-learn	15/15	2072.0	27.3 吉比字节	-	-
			成对	15/15	69.9	19.2 吉比字节	0.276	0.967
			FastAMI	15/15	**15.4**	**341.0 兆字节**	**0.001**	**0.991**
LiveJournal	$4.0 \cdot 10^6$	$2.7 \cdot 10^{-4}$	scikit-learn	10/15	$> 14\,496.1$	-	-	-
			成对	13/15	-	> 503.6吉比字节	(0.149)	(0.937)
			FastAMI	**15/15**	**23.9**	**340.8 兆字节**	(**0.003**)	(0.937)
Friendster	$6.6 \cdot 10^7$	$1.3 \cdot 10^{-5}$	scikit-learn	0/15	-	> 503.6吉比字节	-	-
			成对	6/15	-	> 503.6吉比字节	(0.031)	-
			FastAMI	**15/15**	**439.3**	**4.6 吉比字节**	(**0.003**)	-

表 2：调整互信息（scikit-learn）、成对调整互信息和 FastAMI 在运行时间与空间复杂度方面的基准比较。我们使用来自斯坦福大型网络数据集合集（Leskovec 和 Krevl，2014年）的大型图数据集，对六种不同社区检测算法的结果进行了评估。在返回结果列中，我们报告算法成功终止的次数，以及超出 2000 秒运行时间或 503。6 吉比字节内存限制的次数。无论在运行时间还是内存占用方面，我们的方法都优于 scikit-learn 和 Lazarenko 的成对算法，使得此前无法处理的数据集规模也能进行调整互信息比较。尽管成对调整互信息在斯皮尔曼相关系数方面表现相当，但我们的方法还具有绝对误差可调这一优势，因此能够直接与精确度量进行比较。平均绝对误差以蒙特卡洛估计的形式给出；当精确解不可用时，则在一个子集上计算斯皮尔曼相关系数。

致谢

本研究由巴伐利亚联合资助计划（*BayVFP*）——数字化资助项目——信息与通信技术资助领域（巴伐利亚州经济事务、区域发展与能源部）资助，并得到*Bayern Innovativ*——巴伐利亚创新与知识转移有限公司的支持。

参考文献

Aggarwal, C. C.；和 Reddy, C. K.，编者。2014年。《数据簇分析：算法与应用》。CRC出版社。ISBN 978-1-46-655821-2。

Arratia, R.；和 Desalvo, S. 2016。《概率分治法：一种新的精确模拟方法，采用整数划分示例。组合概率与计算，25(3)：324–351。Bhatia, R.；以及 Davis, C. 2000年。关于方差的更优界。美国数学月刊，107(4)：353–357。Chunaev, P. 2020年。节点属性社交网络中的社区检测：综述。计算机科学评论，37：100286。Edwards, S. 2008年。Thomas M. Cover 和 Joy A. Thomas，《信息论基础》（第2版），约翰·威立父子公司（2006年）。信息处理与管理，44(1)：400–401。Fristedt, B. 1993年。大整数随机划分的结构。美国数学会会刊，337(2)：703–735。Gagolewski, M.；等。2020年。聚类算法基准套件 – 第1版。https://github.com/gagolews/clustering benchmarks v1。访问日期：2021年10月15日。Hoffman, M.；Steinley, D. L.；以及 Brusco, M. J. 2015年。A

关于在网络链接预测中使用调整兰德指数的说明。*Social Networks*, 42: 72–79. H¨ormann, W. 1994. 一种用于离散对数凹分布的通用生成器。*Computing*, 52(1): 89–96. Hubert, L.; and Arabie, P. 1985. 比较划分。*Journal of Classification*, 2(1): 193–218. Kachitvichyanukul, V.; and Schmeiser, B. 1985. 超几何随机变量的计算机生成。*Journal ofStatistical Computationand Simulation*, 22(2): 127–145. Karim, M. R.; Beyan, O.; Zappa, A.; Costa, I. G.; Rebholz- Schuhmann, D.; Cochez, M.; and Decker, S. 2021. 用于生物信息学的基于深度学习的聚类方法。*Briefings in Bioinformatics*, 22(1): 393–415. Lancichinetti, A.; and Fortunato, S. 2012. 复杂网络中的共识聚类。*Scientific Reports*, 2(1): 336. Lazarenko, D.; and Bonald, T. 2021. 成对调整互信息。arXiv:2103.12641. Leskovec, J.; and Krevl, A. 2014. SNAP 数据集：斯坦福大型网络数据集合集。https://snap.stanford. edu/data/index.html。访问日期：2022-03-22。M¨uller, E.; G¨unnemann, S.; F¨arber, I.; and Seidl, T. 2010. 发现多个聚类解：从数据的不同视角对对象进行分组。载于 Webb, G. I.; Liu, B.; Zhang, C.; Gunopulos, D.; and Wu, X. 编，*ICDM 2010*，第十届*IEEE* 国际数据挖掘会议，澳大利亚悉尼，*2010*年*12*月*14*日至*17*日，1220。IEEE 计算机学会。Nguyen, X. V.; Epps, J.; and Bailey, J. 2009. 用于聚类比较的信息论度量：是否有必要进行偶然性校正？载于 Danyluk, A. P.; Bottou, L.; and Littman, M. L. 编，第*26*届机器学习国际年会论文集，*ICML 2009*，加拿大魁北克省蒙特利尔，*2009*年*6*月*14*日至*18*日，第382卷 ACM 国际会议论文集系列，1073–1080。ACM。Nguyen, X. V.; Epps, J.; a nd Bailey, J. 2010. 用于聚类比较的信息论度量：变体、性质、归一化与偶然性校正。*Journal of Machine Learning Research*, 11: 2837–2854. Patefield, W. M. 1981. 算法 AS 159：一种根据给定行列总计生成随机 R × C 表的高效方法。*Journal ofthe Royal StatisticalSociety. Series C(Applied Statistics)*, 30(1): 91–97. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-Learn：Python 机器学习。*Journalof MachineLearningResearch*, 12: 2825–2830. Romano, S.; Bailey, J.; Nguyen, X. V.; and Verspoor, K. 2014. 用于聚类比较的标准化互信息：在偶然性调整方面更进一步。载于第*31*届机器学习国际会议论文集，*ICML 2014*，中国北京，*2014*年*6*月*21*日至*26*日，第32卷*JMLR Workshop and Conference Proceedings*，1143–1151。Romano, S.; Nguyen, X. V.; Bailey, J.; and Verspoor, K. 2016. 针对偶然因素校正聚类比较度量。*Journal of Machine Learning Research*, 17: 134:1–134:32. Shakibian, H.; and Moghadam Charkari, N. 2017. 异构复杂网络中用于链接预测的互信息模型。*Scientific Reports*, 7(1): 44981. Stadlober, E. 1990. 用于生成离散随机变量的均匀比方法。计算与应用数学杂志, 31(1): 181–189. Staudt, C. L.; Sazonovs, A.; and Meyerhenke, H. 2016. NetworKit：用于大规模复杂网络分析的工具套件。网络科学, 4(4): 508–530. Tian, T.; Wan, J.; Song, Q.; and Wei, Z. 2019. 使用基于模型的深度学习方法对单细胞RNA测序数据进行聚类。自然机器智能, 1(4): 191–198. Walker, A. J. 1974. 一种用于生成具有任意频率分布的离散随机数的新型快速方法。电子快报, 10(8): 127–128. Walker, A. J. 1977. 一种用于生成一般分布离散随机变量的高效方法。*ACM*数学软件汇刊, 3(3): 253–256. Wei, S.; Han, G.; Wang, R.; Yang, Y.; Zhang, H.; and Li, S. 2021. 归纳式多视图多重聚类。载于*2021*年第七届大数据与信息分析国际会议（*BigDIA*）, 308–315.