面向定性属性数据聚类的学习序关系森林

Mingjie Zhao[1], Sen Feng[1], Yiqun Zhang[1,6,*], Mengke Li[2,3,6], Yang Lu[4,5,6] 和 Yiu-Ming Cheung[6]

[1]广东工业大学计算机科学与技术学院，中国广州 [2]广东省人工智能与数字经济实验室（深圳），中国深圳 [3]深圳大学计算机科学与软件工程学院，中国深圳 [4]福建省智慧城市感知与计算重点实验室，厦门大学信息学院，中国 [5]中华人民共和国教育部多媒体可信感知与高效计算重点实验室，厦门大学，中国 [6]香港浸会大学计算机科学系，中国香港

arXiv:2603.03387v1 [stat.ML] 3 Mar 2026

摘要。聚类是理解数据模式的一种基本方法，通常采用直观的欧几里得距离空间。然而，对于定性属性值所反映的隐式簇分布，情况并非如此，例如症状、婚姻状况等属性的名义值。因此，本文发现了一种树状距离结构，可灵活表示属性内定性值之间的局部序关系。具体而言，将一个值视为树的顶点，便能捕捉该顶点值与其他值之间丰富的顺序关系。为获得适用于聚类的树结构，本文提出了一种联合学习机制，通过迭代学习更合适的树结构和簇。结果表明，整个数据集的潜在距离空间可以由所学习的树构成的森林很好地表示。大量实验表明，联合学习能够使森林适应聚类任务，从而获得准确结果。在 12 个真实基准数据集上与 10 种对比方法进行比较，并通过显著性检验，验证了所提方法的优越性。所提方法的源代码可在 [39] 获取。

1 引言

由多值定性属性（也称为分类属性或名义属性）组成的数据集在聚类任务中无处不在 [13, 26, 29], , 例如客户、患者等的聚类。与所有值都分布在欧几里得距离轴上的数值属性不同，分类属性的定性值无法反映其距离结构。例如，有

属性“职业”有三个可能取值{司机, 律师, 护士}，但它们在距离轴上的最优数值嵌入尚不明确。因此，大多数相关工作致力于挖掘这种隐式距离结构 [1, 18, 10], , 并可大致分为：1）距离度量和 2）距离学习方法 [3], , 划分依据是它们是否与下游聚类任务相关联。

值得注意的是，对于定性取值的距离度量，例如汉明距离 [4], , 仅根据两个取值是否相同来执行布尔距离度量。尽管后续的度量方法 [2, 25, 22] 引入了各种数据统计信息以提升距离辨别能力，但它们仍将每个

* 通讯作者。电子邮箱：yqzhang@gdut.edu.cn

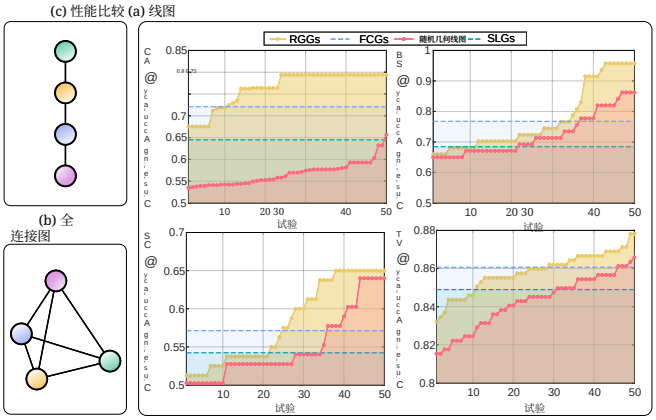

图1. 采用不同类型距离结构时聚类性能的直观比较。(a)和(b)展示了典型的线图和全连接图。(c)展示了采用以下距离结构时的 k模式 [16]聚类性能：1）随机生成图（RGG，不一定全连接，但确保所有属性值均相连），2）全连接图（全连接图），3）随机生成线图（随机几何线图），以及4）语义线图（语义线图，按照语义顺序在图中排列可能取值）。涉及随机化的RGG和随机几何线图各实现50次，并对聚类准确率进行排序，以便更好地可视化。

取值间距离孤立地处理，而未考虑所有可能取值的整体距离结构。为解决这一问题，引入信息熵以有效耦合一个属性的可能取值，并据此形成基于熵的距离度量 [27] ，从而更恰当地量化距离。近年来，更多距离度量 [37, 35, 36] 尝试将可能取值有序地嵌入距离轴，以获得类似于数值属性的距离结构。然而，它们的取值顺序依赖于取值明确的语义顺序，例如，对于有序属性“评价推荐”，其取值顺序为 {strong_接受，clear_接受，weak_接受}；而在处理名义属性时，这种顺序通常并不存在。

近年来，致力于将距离度量与聚类任务联系起来的距离学习方法受到越来越多的关注，因为它们通常能够获得更

适合聚类的距离结构。早期的一种尝试 [9]将样本-簇相似度建模为簇中可能取值的出现概率。后来，人们提出了直接对距离空间建模的方法，包括基于核函数的 [41] 和基于图的 [33, 34] 距离度量学习。然而，这些方法都基于特定假设，例如特定核函数能够很好地表示距离度量，或者距离度量遵循受特定领域知识启发的可能取值图结构。

得益于图的普适性，基于图的学习方法 [33, 34, 38, 36, 8] 已被证明能够实现更具竞争力的聚类性能。具体而言，图被用于表示某一属性的可能取值之间的关系。对于具有明确语义顺序的有序属性值，采用图 1 (a) 所示的线图，并通过学习边权重来表示距离。对于没有语义顺序的名义属性值，则采用图 1 (b) 所示的全连接图，以便进行距离学习。尽管如此，仍有一对棘手的问题亟待解决：合理的先验知识是有效进行距离学习的前提，而数据知识通常是通过在定义明确的距离度量下观察数据分布获得的。

图 1 (c) 充分揭示了先验知识带来的局限性。可以看出，与全连接图和语义线图相比，在两种随机图（即 RGG 和随机几何线图）下进行聚类，更有望获得较高的准确率。此外，RGG 的表现明显优于随机几何线图，这可能是因为 RGG 没有过度限制属性值之间的关系必须遵循某种顺序，从而为通过随机化获得潜在的最优关系奠定了基础。上述观察带来两点启示：1）距离结构的拓扑自由度越高，聚类结果越好；2）明确的语义顺序未必最适合聚类。因此，如何在不依赖值关系先验知识的情况下，针对特定聚类任务获得最优距离结构，对于突破当前聚类性能至关重要。

本文提出了一种新的定性数据学习范式，即基于序森林学习的聚类（COForest）。学习过程不再局限于调整假设值图中各取值之间的距离，而是允许图结构和距离与聚类联合学习。该方法通过迭代以下步骤进行学习：1）根据当前数据划分推断图结构；2）利用图距离结构执行聚类，从而更恰当地获得数据划分。由于推断出的图是最小生成树，因此能够简洁而灵活地表示可能取值之间的关系。事实证明，学习过程会反复改进并逼近上界，从而避开次优解，获得更高的聚类准确率。本文的主要贡献可概括为以下三点：

- 本文提出了一项新的见解：对于某些聚类任务，存在一种最优潜在图，可用于表示定性属性的距离结构；该图应能够灵活确定，而不受先验知识的限制。
- 提出COForest，以迭代优化距离结构与簇，从而规避次优解。与仅在给定拓扑结构下调节距离的现有方法相比，COForest进一步允许重构拓扑结构，从而带来更高的学习自由度。
- 我们开展了包括显著性检验、消融研究和定性视觉比较在内的全面实验评估，以证明.....的优越性

在无先验知识偏差的情况下学习距离结构。

2 提出方法

2.1 问题形式化

下面对采用距离学习的分类数据聚类问题进行形式化描述。给定一个分类数据集 $X = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n\}$，其中包含 n 个数据样本。每个样本 $\mathbf{x}_i$ 可表示为一个 l 维行向量 $\mathbf{x}_i = [x_{i,1}, x_{i,2}, .., x_{i,l}]^\top$，由 l 个属性 $A = \{\mathbf{a}_1, \mathbf{a}_2, ...\mathbf{a}_l\}$表示。每个属性 $\mathbf{a}_r$ 可表示为一个列向量 $\mathbf{a}_r = [x_{1,r}, x_{2,r}, ..., x_{n,r}]$，由所有 n 个样本的第 r 个取值组成。其中，这些取值可视为从有限个可能取值 $V_r = \{v_{r,1}, v_{r,2}, ..., v_{r,o_r}\}$中采样得到，$o_r$ 表示 $\mathbf{a}_r$ 的可能取值数量，而 $v_{r,\ g}$ 表示 $\mathbf{a}_r$ 的第 g 个可能取值。

划分式聚类旨在将 X 划分为 k 个互不重叠的样本子集 $C = \{C_1, C_2, ..., C_k\}$，其目标是最小化子集内不相似度 [5, 11, 30], ，通常表示为

$$L(\mathbf{Q}) = \sum_{j=1}^{k} \sum_{i=1}^{n} q_{i,j} \cdot \Gamma(\mathbf{x}_i, C_j), \tag{1}$$

其中$\mathbf{Q}$为一个 $n \times k$矩阵，其第(i, j)项$q_{i,j} \in \{0, 1\}$ 表示样本$\mathbf{x}_i$ 与簇 C_j之间的归属关系，且$\mathbf{Q}$ 的每一行（例如第 i行）满足 $\sum_{j=1}^{k} q_{i,j} = 1$。在聚类过程中，$q_{i,j}$ 的取值由

$$q_{i,j} = \begin{cases} 1, & \text{if } j = \arg\min_{y} \Gamma(\mathbf{x}_i, C_y) \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

样本-簇不相似度 $\Gamma(\mathbf{x}_i,\ C_j)$ 由不同属性的值级不相似度共同体现：

$$\Gamma(\mathbf{x}_i, C_j) = \sum_{r=1}^{l} \gamma(x_{i,r}, C_{j,r}), \tag{3}$$

其中，$\gamma(x_{i,\ r},\ C_{j,\ r})$ 表示从属性 $\mathbf{a}_r$的角度来看，样本 $\mathbf{x}_i$ 与簇 C_j 之间的差异度。

由于可能取值之间的关系尚未定义，因此计算样本与样本集合 C_j之间的差异度并非易事。因此，这项工作的关键在于如何从属性层面合理定义距离 $\gamma(x_{i,\ r},\ C_{j,\ r})$。接下来，我们首先介绍如何灵活地建模值级关系，并据此定义相应的距离度量。然后，我们提出联合学习方案，使所定义的距离结构能够与聚类联合学习。

2.2 序森林构建

如第1节所述，线图和全连接图在表示属性的距离结构方面都有明显的缺点。也就是说，线图强制所有可能取值之间呈现顺序关系，其有效性在很大程度上依赖于对可能取值的先验顺序知识。相比之下，全连接图为每一对可能取值构建多条路径，因此无法简洁、准确地反映可能取值之间的关系。

因此，我们提出构建序森林 $M = \{M_1, M_2, \ldots, M_l\}$，其中 M_r 是对应于第 r个属性$\mathbf{a}_r$的最小生成树（MST）。每棵树 M_r 记为

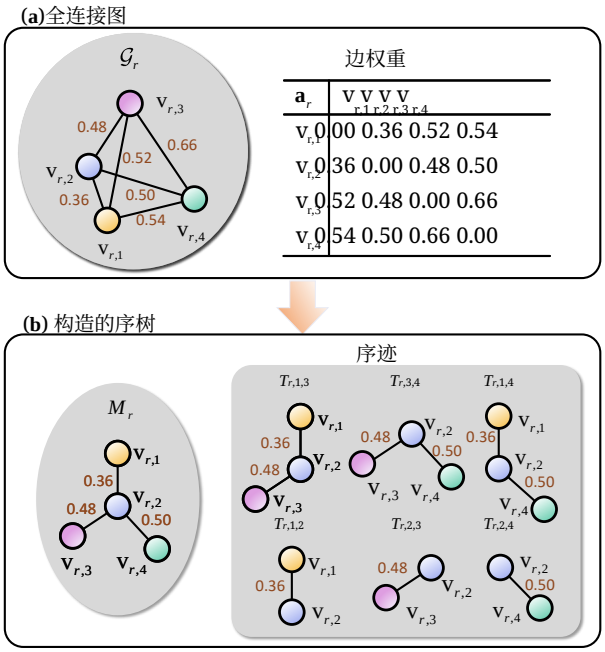

图2. 序树构造过程。(a) 构建一个全连接图 $\mathcal{G}_r$，并配有反映边权重的距离矩阵。(b) 使用 Prim 或 Kruskal算法生成一棵序树，使每对节点之间都有唯一的序迹，其定义见定义1。

一个元组 $M_r =< V_r, B_r, W_r >$，其中V_r 中的全部 o_r 个可能取值均被视为节点。 B_r 是边长度之和最小且能够连接所有 o_r 个节点的 $o_r - 1$ 条边的集合。 W_r 包含反映边长度的权重。给定具有精确边权重的全连接图，可通过Prim或Kruskal [31, 24] 算法搜索出这样的树，如图 2(a)所示。这些权重实际上是任意一对可能取值之间的距离，可使用任何现有的、针对定性值定义的距离度量来计算。从图 2(b)可以看出，所构建的最小生成树简洁而精确地反映了可能取值之间的局部序关系，因此被称为序树。

备注1 (关系图的推广)。 给定属性$\mathbf{a}_r$，构造的序树 M_r 表示不同可能取值子集之间的顺序关系，并为每对可能取值提供一条唯一轨迹，如图 *2(b)* 所示。因此， M_r 可视为以下两者的推广：*1)* 表示同一顺序中所有取值的线图，以及 *2)* 表示可能取值之间冗余关系的全连接图。

为了根据序树定义节点之间的差异度，我们首先定义两个节点之间的序迹。

定义1 (序迹)。 给定一棵序树 M_r，序迹$T_{r,\ s,\ u}$ 是一个集合，包含从节点 $v_{r,\ s}$ 到 $v_{r,\ u}$的最短路径上各节点之间的所有权重。由于序树是无向的，$T_{r,\ s,\ u} = T_{r,\ u,\ s}$ 成立。

事实证明，每条序迹都会唯一地串联一定数量的紧密相连节点，同时排除距离较远的其他节点。这使序树能够灵活地表示可能具有价值的子集的局部序关系，从而在距离结构学习中获得更大的自由度。两个节点 $v_{r、s}$ 和 $v_{r、u}$ 之间的差异度可以通过它们的序迹直观地体现出来，例如，将序迹上的所有权重相加即可。迄今为止，权重的定义在构建序森林以及形成节点之间的差异度方面发挥着关键作用。由于我们关注的是聚类任务，因此在第 2.3 节中充分利用聚类信息来定义权重和差异度，并在第 2.4 节中将差异度的学习与聚类结合起来。

2.3 聚类友好的迹距离

给定聚类划分 **Q**，应首先计算全连接图的权重，然后从中提取序树，以得到值级不相似度。具体而言，两个节点$v_{r,\ u}$ 和 $v_{r,\ s}$ 之间的权重定义为从不同簇中提取的概率分布之间的距离，计算方式如下：

$$w_{r,u,s} = \left\|\mathbf{p}_{v_{r,u}} - \mathbf{p}_{v_{r,s}}\right\|_p, \tag{4}$$

其中，$\mathbf{p}_{v_{r,u}} = [p_{C_1|v_{r,s}}, p_{C_2|v_{r,s}}, \ldots, p_{C_k|v_{r,s}}]$，$\mathbf{p}_{v_{r,u}} = [p_{C_1|v_{r,u}}, p_{C_2|v_{r,u}}, \ldots, p_{C_k|v_{r,u}}]$ 和 $\mathbf{p}$，是维向量，分别表示$v_{r,u}$ 和$v_{r,s}$ 在所有 k 个簇中的概率分布。$p_{C_j|v_{r,u}} = |X_{r,u} \cap C_j|/|X_{r,u}|$ 其中 $|\cdot|$ 表示集合中的样本数量，$X_{r,u} = \{\mathbf{x}_i | x_{i,r} = v_{r,u}\}$ 是一个样本集合，收集了 X 中的所有样本，且这些样本的第 r个取值等于$v_{r,u}$。符号 $\|\cdot\|_p$ 表示 p-范数，它直观地反映了节点直接概率分布的差异；这里我们采用一种常见的设定 $p = 2$。概率分布 $\mathbf{p}_{v_{r,u}}$ 描述了某个取值在所有 k 个簇中的分布模式，因此，模式相似的两个取值被认为更为相似。因此，我们使用公式（4）中定义的权重构建序树，该序树具有产生当前划分 Q 的可能性。

对于构造的序树 M_r 以及公式（4）中定义的权重，两个可能取值 $v_{r,\ u}$ 和 $v_{r,\ s}$ 之间的差异度定义为定义 1 中所述的序迹长度，可写为

$$d_{r,u,s} = \sum_{w_{r,u,s} \in T_{r,u,s}} w_{r,u,s}. \tag{5}$$

由于这是基于公式（4）以一种便于聚类的方式定义的权重计算得出的，因此我们称其为聚类友好型轨迹距离。相应地，由顺序树结构 M_r 反映的样本-簇距离 $\gamma(x_{i,\ r},\ C_{j,\ r})$ 可基于值级迹距离定义为

$$\gamma(x_{i,r}, C_{j,r}; M_r) = \mathbf{p}_{j,r}^{\top} \mathbf{d}_{r,u}, \tag{6}$$

其中，为简化起见且不失一般性，我们假设样本值 $x_{i,\ r}$ 等于可能取值 $v_{r,u}$，$\mathbf{d}_{r,u} = [d_{r,u,1}, d_{r,u,2}, \ldots, d_{r,u,o_r}]$，$o_r$ 是一个 -维向量，包含 $v_{r,\ u}$ 与 V_r 中每个 o_r 个可能取值之间的迹距离；$\mathbf{p}_{j,\ r} = [p_{v_{r,1}|C_j}, p_{v_{r,2}|C_j}, \ldots, p_{v_{r,o_r}|C_j}]$ 是一个 o_r 维向量，描述簇 C_j 内 V_r 中可能取值的概率分布，其中 $p_{v_{r,\ u}|C_j} = |X_{r,\ u} \cap C_j|/|C_j|$。

相应地，基于整个序森林 M 定义的总体样本-簇距离 $\Gamma(\mathbf{x}_i, C_j)$可表述为

$$\Gamma(\mathbf{x}_i, C_j; M) = \sum_{r=1}^{l} \gamma(x_{i,r}, C_{j,r}; M_r). \tag{7}$$

定理 1. 在序树 M_r 语境下定义的轨迹距离度量$d_{r,u,s}$ 是一种有效的距离度量。

证明。$d_{r,\ u,\ s}$ 对于任意 $r \in \{1, 2, ..., l\}$ 和 u，s，$g \in \{1, 2, ..., o_r\}$，均满足非负性、对称性和三角不等式，如下所示。

非负性： $d_{r,\,u,\,s} \geq 0$。$d_{r,\,u,\,s}$ 是序迹的长度，根据公式（4）和（5），在范数设为 $p = 2$时，它由非负权重构成，因此始终非负；

对称性：$d_{r,\,u,\,s} = d_{r,\,s,\,u}$。由于序树是无向图，从无向图中提取的轨迹上的权重在求和时遵循交换律；

三角不等式：$d_{r,\,u,\,s} \leq d_{r,\,u,\,g} + d_{r,\,g,\,s}$。序迹是两个值之间长度为 $d_{r,\,u,\,s}$的唯一路径。从 $v_{r,\,u}$ 到$v_{r,\,s}$，绕行经过序迹 $T_{r,\,u,\,s}$ 之外的另一个节点 $v_{r,\,g}$必然会涉及来自其他序迹的额外权重。鉴于每个权重均为非负数，结论成立。 □

定理 2。 在序森林 M的上下文中定义的样本-簇距离 $\Gamma(\mathbf{x}_i, C_j; M)$ 表示一个有效的距离度量。

证明。 $\Gamma(\mathbf{x}_i, C_j; M)$ 的计算可视为 $\mathbf{p}_{j,r}$s 中由概率表示的非负权重，对 $\mathbf{d}_{r,u}$s 中一系列 $d_{r,u,s}$s 进行加权求和，这依据的是公式 (6) 和 (7)。根据定理 1，迹距离 $d_{r,u,s}$ 是一种度量，因此根据非负性、对称性和三角不等式， $\Gamma(\mathbf{x}_i, C_j; M)$ 也是一种度量。 □

2.4 联合学习算法

通过将第 2.2 节和第 2.3 节提出的序森林构建机制整合到聚类目标中，可以实现簇与序森林的联合学习。因此，根据公式（1）、（5）、（6）和（7）， $L(\mathbf{Q}\ M)$ 可进一步细化为 $L(\mathbf{Q}, M)$:

$$L(\mathbf{Q}, M) = \sum_{j=1}^{k}\sum_{i=1}^{n} q_{i,j} \cdot \sum_{r=1}^{l} \gamma(x_{i,r}, C_{j,r}; M_r). \tag{8}$$

于是，问题变为如何计算 $\mathbf{Q}$ 和 M，以使L最小化；通常通过迭代地固定其中一个并计算另一个来求解。具体而言，给定固定的距离结构 $\hat{}M$，可计算 $\mathbf{Q}$，使 $L(\mathbf{Q}, \hat{}M)$ 最小化：

$$q_{i,j} = \begin{cases} 1, & \text{if } j = \arg\min_{y} \sum_{r=1}^{l} \gamma(x_{i,r}, C_{y,r}; M_r) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

其中 $i = \{1, 2, ..., n\}$ 和 $j = \{1, 2, ..., k\}$。公式（9）是通过采用序森林 $\hat{}M$ 作为其距离结构，从公式（2）严格推导而来。计算出 $\mathbf{Q}$ 后，将其固定为 $\mathbf{Q}$，然后根据图2和公式（4）-（5）重构 M。

总之，通过迭代求解以下两个最小化问题来优化 L：1）固定 M，运行 k-modes [16]，迭代计算$\mathbf{Q}$ 直至收敛；2）固定$\mathbf{Q}$，重构 M 以更新距离度量。由于 M的状态空间有限，在迭代搜索过程中，所有状态最终都会逐渐遍历完毕，因此可以保证算法收敛。整个算法总结如算法 1 所示。

定理 3. *COForest* 的时间复杂度为 $O(nlk\mathcal{I}\mathcal{E})$

证明。 为更直观地给出证明，我们首先定义概率$\mathbf{P}$ 和迹距离矩阵 $\mathbf{D}$，并假设在最坏情况分析中， $\varsigma = \max(o_1, o_2, ..., o_l)$。$\mathbf{P}$是一个 $k \times l$概率矩阵，其(j, r) $_{j\,r} = [p_{v_{r,1}|C_j}, p_{v_{r,2}|C_j}, ..., p_{v_{r,o_r}|C_j}]$ $l \times \varsigma$)项为 $\mathbf{p}_{\cdot}$。 $\mathbf{D}$是一个迹距离矩阵，其(r, u)项为$\mathbf{d}_{r,u} = [d_{r,u,1}, d_{r,u,2}, ..., d_{r,u,o_r}]$。

假设求解一次$L(\mathbf{Q},$ 以更新$\mathbf{D}$，以求解 $L(\hat{}\mathbf{Q}\ M)$。

算法1： COForest：基于序森林学习的聚类

要求： 数据集 X，所求簇数 k
确保： 划分$\mathbf{Q}$，序森林 M
初始化：根据 $\mathcal{E} \leftarrow 0$ 和设置外层与内层循环计数器 $\mathcal{I} \leftarrow 0$；运行 k-modes [16] 以获得相对稳定的初始 $\mathbf{Q}^{\{\mathcal{E}\}}$；根据$\mathbf{Q}^{\{\mathcal{E}\}}$构造初始 $M^{\{\mathcal{E}\}}$；设置外层循环的收敛标记为 $Conv\_\mathcal{E} \leftarrow False$。
当 $Conv\_\mathcal{E} = False$ **执行**
  $Conv\_\mathcal{I} \leftarrow False$;
  当 $Conv\_\mathcal{I} = False$ **执行**
    $\mathcal{I} \leftarrow \mathcal{I} + 1$；通过公式（9）计算 $\mathbf{Q}^{\{\mathcal{I}\}}$；
    如果 $\mathbf{Q}^{\{\mathcal{I}\}} = \mathbf{Q}^{\{\mathcal{I}-1\}}$，**则**
      $Conv\_\mathcal{I} \leftarrow True$;
    end
  end
  如果 $\mathbf{Q}^{\{\mathcal{E}\}} = \mathbf{Q}^{\{\mathcal{I}\}}$，**则**
    $Conv\_\mathcal{E} \leftarrow True$;
  else
    $\mathcal{E} \leftarrow \mathcal{E} + 1$；$\mathbf{Q}^{\{\mathcal{E}\}} \leftarrow \mathbf{Q}^{\{\mathcal{I}\}}$；重构 $M^{\{\mathcal{E}\}}$;
  end
结束

对于 $\mathcal{I}$的每次迭代，都需要遍历所有 n 个数据样本一次来准备$\mathbf{P}$，其复杂度为 $O(nkl\varsigma)$；还需要在n 个样本上遍历 l 个属性的所有 ς 个值一次来准备$\mathbf{D}$，其复杂度为 $O(nl\varsigma)$。然后，根据公式（9），通过考虑 l 个属性的 ς 个值，将 n 个样本聚类为 k 个簇，时间复杂度为 $O(nkl\varsigma)$。因此，总共进行 $\mathcal{I}$ 次迭代时，求解 $L(\mathbf{Q}, \hat{}M)$的时间复杂度为 $O(\mathcal{I}nkl\varsigma)$。

对于 $\mathcal{E}$的每一次迭代，由于$\mathbf{P}$ 和$\mathbf{D}$ 已经准备好，可通过构建 M_r 来搜索 l 个属性中每一个属性的 ς 个可能取值的序树，时间复杂度为 $O(nl\varsigma^2)$。对于整个COForest算法的 $\mathcal{E}$ 次迭代，考虑到 $\mathcal{I}$ 次内部迭代，COForest的总体时间复杂度为 $O(\mathcal{E}(\mathcal{I}nkl\varsigma + nl\varsigma^2))$。

由于 ς 在大多数情况下是一个介于2到8之间的小整数，可将其视为常数，总体时间复杂度可简化为 $O(nlk\mathcal{I}\mathcal{E})$，其与 n 和 l呈线性关系。 2

3 实验

设计了五项实验，在12个真实公开数据集上使用三种有效性指标，将所提出的COForest与10种对比方法进行比较，以评估COForest。实验总结如下:

- 带有显著性检验的聚类性能比较表明，COForest显著优于传统对比方法和最先进的对比方法（第3.2节）。
- 通过比较 COForest 的五个消融版本，消融研究证实了 CO-Forest 各核心组件的有效性（第 3.3 节）。
- 通过分别绘制不同数据集规模下学习过程中的目标函数值和执行时间，验证了 COForest 的收敛性和效率（第 3.4 节）。
- 通过使用 t-SNE 对不同方法的簇判别能力进行定性比较，充分验证了所学习距离结构的合理性（第 3.5 节）。
- 通过将 COForest 的聚类性能与专为混合数据提出的方法进行比较，验证了将 COForest 扩展到包含数值型和分类型属性的混合数据的潜力（请参阅 [39] 提供的“补充材料”）。

表1. 10种对比方法的信息。“类型”表示一种方法是分别学习距离定义与聚类，还是联合学习两者。

No.	对比方法	Year	Type
1	KMD [16]	1998	分开
2	LSM [27]	1998	分开
3	JDM [20]	2016	分离
4	CBDM [17]	2012	分离
5	OCIL [9]	2013	联合
6	UDMC [35]	2022	分离
7	DLC [33]	2020	联合
8	H2H [38]	2022	联合
9	HDC [34]	2022	联合
10	ADC [36]	2023	独立

表2. 12 个数据集的统计信息。l 和 n 分别表示属性数和样本数。k^* 表示真实簇数。

No.	数据集	缩写	l	n	k^*
1	Hayes-Roth	HR	4	132	3
2	汽车评估	CE	6	1728	4
3	澳大利亚信用	AC	8	690	2
4	国会投票	VT	16	435	2
5	剖宫产	CS	4	80	2
6	大豆（小型）	SB	35	47	4
7	托儿所	NS	8	12960	4
8	Zoo	ZO	16	101	7
9	胸外科	TS	13	470	2
10	心力衰竭	HF	5	299	2
11	炎症诊断	DS	5	120	2
12	镜片	LS	4	24	3

3.1 实验设置

下面简要介绍实验设置。

10种对比方法的超参数（如有）设为相应论文推荐的值。每种方法均运行 10 次，并报告平均性能。

12 个数据集来自不同领域，并用于实验。所有数据集均为从 UCI 机器学习库 [23], 收集的真实公开数据集，统计信息见表2。实验前，我们通过删除存在缺失值的样本对数据集进行预处理。由于我们关注的是类别型数据聚类，因此省略 AC、TS、HF 和 DS 数据集中的数值属性。对于所有对比方法，我们将 $k = k^*$ 设为待求簇数。

采用了三个评估指标：聚类准确率（CA） [15], 调整兰德指数（ARI） [32, 14],和标准化互信息（NMI） [12], ，从不同角度评估聚类性能。其中，CA 是一种传统指标，它基于所得簇与真实类别之间的最佳排列映射计算匹配率。相比之下，ARI 和 NMI 的区分能力更强，其取值范围分别为 $[-1, 1]$和 $[0, 1]$, 。对于所有指标，数值越高表示聚类性能越好。NMI 结果见“补充材料” [39]。

3.2 聚类性能

本节考察不同算法的聚类性能，并对 COForest 的优越性进行统计分析。

不同方法的聚类性能分别在表3和表4中就CA和调整兰德指数进行了比较。每个数据集上的最佳和次佳结果分别以**粗体**和下划线标出。观察结果包括以下三个方面：1）总体而言，COForest在几乎所有数据集上表现最佳，表明其在聚类方面的优越性。2）COForest在TS和HF数据集上的性能并未明显优于次佳方法。然而，这些数据集上的次佳方法各不相同，表明了COForest的稳健性。3）尽管COForest在VT数据集上的CA和调整兰德指数性能并非最佳，但它保持次佳，且并未被优胜方法大幅超越。此外，CE、NS和LS数据集上的CBDM结果未予报告，因为这些数据集的属性彼此独立，导致CBDM无法根据相关属性度量距离。

显著性检验首先对表 3 和表 4 最后一行报告的平均性能排名分别进行弗里德曼检验。对应的 p 值分别为 0.00020 和 0.00002，均通过了 99% 置信区间下的检验（即 p 值 = 0.01）。在此基础上，进行了 Bonferroni-Dunn（BD）事后检验。对于在 12 个数据集上比较 11 种方法，双侧 BD 检验在 95%（α= 0.05）和 90%（α = 0.1）置信区间下的临界差异（CD）区间分别为 3.8048 和 3.5204。从表 3 和表 4 的“调整兰德指数”行可以看出，除在 $\alpha = 0.05$ 下 DLC 方法的 ARI 性能外，所有对比方法均落在 CD 区间右边界之外。不过值得一提的是，DLC 非常接近 $\alpha = 0.05$ 的边界，且仍处于 $\alpha = 0.1$ 的边界之外。总的来说，检验结果表明，所提出的 COForest 显著优于其他对比方法。

3.3 消融实验

为明确展示 COForest 核心组件的有效性，我们在图3中比较了它的若干消融版本。为评估所提出的序森林学习机制，我们将 COForest 与 COF^{I}进行比较；后者不进行迭代学习，仅构建一次序森林。为评估所提出的序森林结构，我们将 COF^I 中的序森林替换为线图和全连接图，分别得到 COF^{II} 和 COF^{III}。此外，为验证公式（4）中用于计算权重的基于概率分布的度量，我们还令 COF^{III} 采用传统的汉明距离，得到版本 COF^{IV}。

从图3可以看出，COForest 的表现优于其四个变体，整体上体现了其有效性。更具体的观察可归纳为四点：1）在所有数据集上，COForest 的表现均不逊于 COF^I，验证了联合学习序森林与聚类的必要性。2）在 12 个数据集中，有 10 个数据集上 COF^I 的表现不逊于 COF^{II} 和 COF^{III}。这表明，即使不经过学习，我们构建的序森林在反映距离结构方面也更加合理。原因可能是，正如备注1所分析的，序树是一种广义距离结构，能够更灵活地表示多种局部序关系。3）COF^{II} 和 COF^{III}在 12 个数据集上的表现互有胜负，这揭示了线图和全连接图各自存在局限性。4）采用基于概率分布的度量的 COF^{III} 在 10 个数据集上的表现优于采用汉明距离的 COF^{IV}，表明公式（4）中使用概率分布是合理的。

3.4 收敛性与效率评估

为评估 COForest 的收敛性，我们在图4中绘制了其在全部12个数据集上学习期间的目标函数值 L。横轴和纵轴分别表示学习

表 3. 通过 CA 评估的聚类性能。“调整兰德指数”行报告平均性能排名。

Data	KMD	LSM	JDM	CBDM	OCIL	UDMC	DLC	H2H	HDC	ADC	COForest（我们的方法）
HR	0.3795±0.02	0.3826±0.03	0.3841±0.03	0.4083±0.06	0.3621±0.05	0.3886±0.01	0.3659±0.03	0.3333±0.00	0.3758±0.02	0.3970±0.05	**0.4530±0.07**
CE	0.3730±0.04	0.3587±0.04	0.3597±0.04	-	0.3659±0.05	0.3505±0.03	0.3746±0.04	0.3354±0.06	0.3730±0.04	0.3730±0.04	**0.4261±0.06**
AC	0.7494±0.05	0.7823±0.04	0.6858±0.12	0.7417±0.08	0.7781±0.10	0.7674±0.08	0.7499±0.14	0.7942±0.00	0.7484±0.09	0.7709±0.09	**0.8307±0.05**
VT	0.8621±0.01	0.8662±0.00	0.8662±0.00	0.8749±0.00	**0.8763±0.00**	0.8639±0.00	0.8540±0.08	0.8736±0.00	0.8736±0.00	0.8713±0.00	0.8761±0.00
CS	0.5475±0.02	0.5425±0.05	0.5475±0.05	0.5787±0.03	0.5037±0.18	0.5788±0.03	0.6013±0.04	0.6050±0.02	0.5862±0.03	0.5875±0.02	**0.6450±0.02**
SB	0.8191±0.18	0.8553±0.19	0.7830±0.16	0.8213±0.15	0.7936±0.33	0.8426±0.17	0.8723±0.17	0.9511±0.10	0.8128±0.13	0.8191±0.16	**0.9723±0.09**
NS	0.3454±0.04	0.3171±0.04	0.3064±0.03	-	0.3454±0.08	0.3235±0.04	0.3301±0.06	0.3441±0.05	0.3454±0.04	0.3454±0.04	**0.3626±0.09**
ZO	0.6564±0.10	0.6594±0.10	0.7149±0.09	0.6921±0.08	0.5663±0.31	0.6564±0.09	0.7020±0.10	0.6980±0.04	0.6713±0.12	0.6812±0.11	**0.7832±0.12**
TS	0.7083±0.08	0.6717±0.08	0.7023±0.10	0.6957±0.09	0.6689±0.08	0.7087±0.09	0.6868±0.08	0.5723±0.03	0.7104±0.08	0.6947±0.10	**0.7232±0.09**
HF	0.5344±0.03	0.5344±0.03	0.5421±0.03	0.5498±0.02	0.4880±0.17	0.5378±0.02	0.5381±0.02	0.5441±0.05	0.5388±0.02	0.5378±0.02	**0.5532±0.03**
DS	0.6833±0.11	0.6833±0.11	0.6975±0.11	0.7142±0.12	0.7242±0.16	0.6725±0.11	0.7242±0.16	0.6267±0.04	0.6725±0.11	0.6975±0.11	**0.7617±0.08**
LS	0.5250±0.07	0.5417±0.10	0.5417±0.10	-	0.5417±0.08	0.5792±0.14	0.5500±0.09	0.5167±0.09	0.5250±0.07	0.5250±0.07	**0.6833±0.14**
$\overline{AR}$	7.2500	7.2083	6.7917	6.4583	7.0833	6.5833	5.3750	6.3750	6.2083	5.5833	1.0833

表 4. 根据 ARI 评估的聚类性能。“调整兰德指数”行报告平均性能排名。

Data	KMD	LSM	JDM	CBDM	OCIL	UDMC	DLC	H2H	HDC	ADC	COForest（本文）
HR	-0.0064±0.01	-0.0051±0.01	-0.0048±0.01	0.0127±0.03	-0.0073±0.02	-0.0037±0.00	-0.0068±0.01	-0.0149±0.00	-0.0056±0.01	0.0043±0.03	**0.0429±0.04**
CE	0.0229±0.03	0.0314±0.02	0.0321±0.02	-	0.0501±0.06	0.0289±0.02	0.0676±0.03	0.0140±0.03	0.0229±0.03	0.0229±0.03	**0.1016±0.07**
AC	0.2575±0.10	0.3228±0.07	0.1892±0.17	0.2569±0.11	0.3421±0.15	0.3107±0.11	0.3178±0.22	0.3453±0.00	0.2714±0.12	0.3225±0.12	**0.4462±0.12**
VT	0.5233±0.02	0.5354±0.01	0.5354±0.01	0.5613±0.01	**0.5655±0.01**	0.5287±0.01	0.5208±0.18	0.5572±0.00	0.5572±0.00	0.5503±0.00	0.5647±0.00
CS	-0.0033±0.01	0.0017±0.03	0.0038±0.03	0.0137±0.01	0.0070±0.03	0.0140±0.02	0.0342±0.03	0.0319±0.02	0.0191±0.02	0.0190±0.01	**0.0732±0.02**
SB	0.7657±0.20	0.8164±0.24	0.6826±0.22	0.7652±0.21	0.7902±0.33	0.8232±0.19	0.8500±0.20	0.9271±0.16	0.7595±0.18	0.7863±0.19	**0.9562±0.14**
NS	0.0630±0.02	0.0556±0.02	0.0457±0.02	-	0.1146±0.10	0.0617±0.03	0.0886±0.08	0.0847±0.09	0.0630±0.02	0.0630±0.02	**0.1352±0.13**
ZO	0.5707±0.13	0.5872±0.15	0.6496±0.14	0.6187±0.12	0.5093±0.29	0.5937±0.15	0.6315±0.12	0.6255±0.06	0.6010±0.15	0.6128±0.15	**0.7511±0.18**
TS	0.0054±0.05	0.0123±0.05	0.0188±0.05	0.0171±0.04	-0.0048±0.05	0.0198±0.05	-0.0034±0.04	-0.0249±0.00	0.0084±0.05	0.0031±0.04	**0.0220±0.04**
HF	-0.0067±0.00	-0.0067±0.00	-0.0013±0.01	-0.0009±0.01	-0.0002±0.00	-0.0023±0.00	-0.0005±0.00	-0.0043±0.00	-0.0019±0.00	-0.0023±0.00	**0.0045±0.01**
DS	0.1697±0.18	0.1697±0.18	0.1944±0.19	0.2280±0.20	0.2839±0.32	0.1543±0.18	0.2839±0.32	0.0615±0.04	0.1543±0.18	0.1944±0.19	**0.2901±0.16**
LS	0.0756±0.10	0.1180±0.15	0.1180±0.15	-	0.1287±0.12	0.1919±0.22	0.1379±0.15	0.0786±0.11	0.0756±0.10	0.0756±0.10	**0.3359±0.22**
$\overline{AR}$	8.6667	7.0000	6.5417	6.7500	5.2083	6.0833	4.7083	6.5417	7.0833	6.3333	1.0833

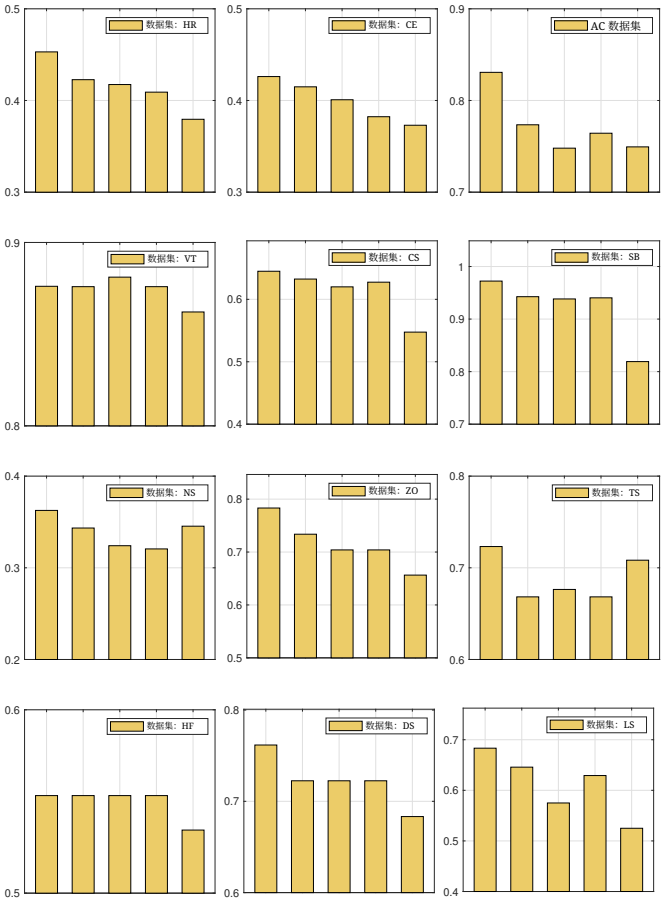

图3. 不同消融版 COForest 的 CA 性能　　版本。

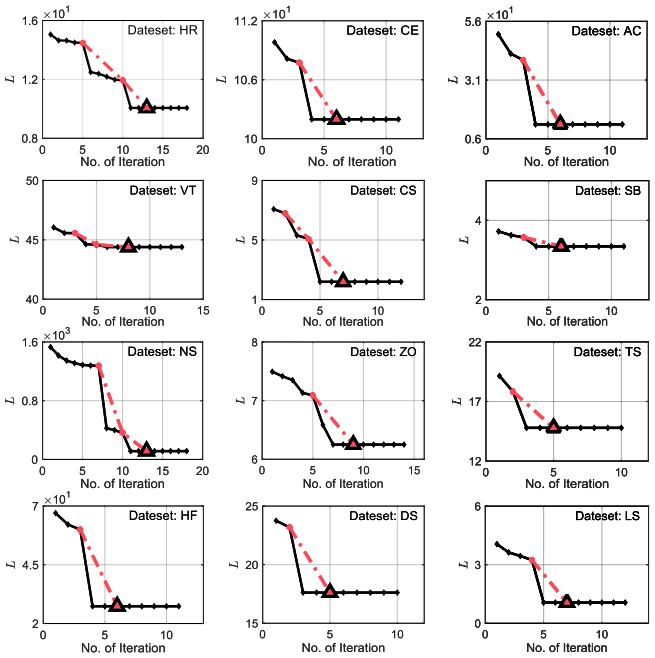

图4. COForest 在不同数据集上的收敛曲线。 L 表示目标函数的值。

迭代次数和 L 的值。曲线上的三角形标记表示 COForest 收敛时的迭代次数，红点标记序森林重构的迭代次数。可以观察到，每次更新序森林后， L 都会下降，这表明森林重构与 L 的最小化相一致。此外，在大多数情况下，COForest 在15次迭代内收敛；对于一个通过迭代重构距离结构并学习数据划分的过程而言，这种效率相当高。

To e评估 COForest 在大型合成数据上的效率　　集合

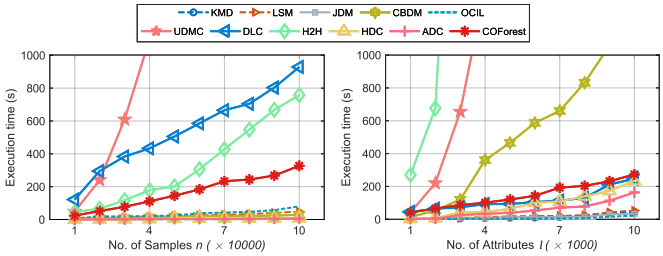

图 5. 合成数据集上的执行时间。

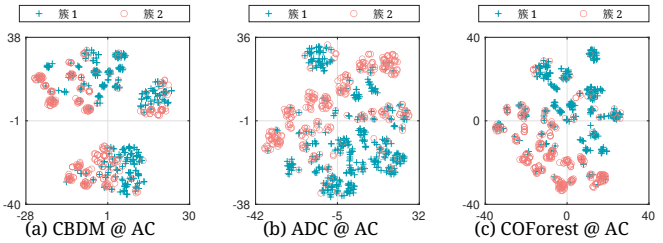

图 6. AC 数据集的 t-SNE 可视化。

是在不同属性和样本规模下随机生成的。具体而言，我们通过以下方式生成：1）将属性数量固定为 l =20，并将样本数量 n 从 10k 增至 100k，步长为 10k；2）将样本数量固定为 $n = 2k$，并将属性数量 l 从 1k 增至 10k，步长为 1k，其中“k”表示“千”。请注意，每个属性有五种可能取值，簇的数量 k 始终设为五。图 5 展示了全部 11 种方法的执行时间。可以看出，COForest 的执行时间低于或接近当前最先进的 UDMC、DLC 和 H2H。此外，COForest 的执行时间随 n 和 l 增长的趋势几乎呈线性，这与定理 3 的时间复杂度分析一致。总之，与当前最先进的方法相比，COForest 效率较高；与最简单的方法相比，它不会带来过多额外的计算成本。

3.5 定性评估

为展示 COForest 的簇区分能力及其所获得的距离结构的直观性，我们使用 COForest、CBDM 和 ADC 学习得到的属性值间距离，对 AC 数据集的属性进行编码。随后通过 t-SNE [28] 将编码后的数据降至二维空间，并在图 6 中将数据点标记为数据集提供的“真实”标签进行可视化。如果可视化中具有相同标签的数据点聚集得越多，则表明相应的距离度量越能有效地区分不同的簇。可以看出，在对比方法中，COForest 的簇区分能力显著更强，这表明其基于树状距离结构得到的簇具有直观性。

4 相关工作

本节概述现有的基于距离度量和基于距离学习的类别数据聚类方法。

距离度量适用于分类数据，包括通过编码策略得到的度量以及直接定义的距离度量。传统的数据编码技术（如独热编码）使用汉明距离，将每个可能取值编码为一个新的属性。然而，由于其布尔性质，它无法充分捕捉可能取值之间的差异程度 [10, 6]。为克服这些局限，人们提出了基于统计的度量方法，这些方法考虑属性内取值的频率，从而体现相似取值具有较低的信息熵，并得出更合理的距离度量 [27, 9]。此后，该领域进一步发展出考虑属性间依赖关系的度量方法，从而更全面地呈现数据关系 [2, 25, 17]。此外，对取值顺序的考量区分了名义属性与有序属性；一些特定方法通过融入语义顺序来定义距离，从而显著提高了有序数据的距离准确性 [37, 35]。

距离学习方法将距离的定义纳入聚类学习过程。这包括能够动态编码类别数据的先进表示学习技术。例如，一些研究采用各种核函数，更有效地解开属性耦合 [41]，；另一些研究则为数值属性和类别属性开发混合编码策略 [21, 40]。这些方法通常需要对超参数进行细致调优。该领域的一项重大进展是引入了无参数方法，能够学习最优簇数和距离 [19, 7]。另一种创新策略是通过几何投影将名义属性转换为有序属性，以优化潜在距离；通过统一处理不同类型的类别属性，显著提升聚类性能 [38]。此外， [33] 将带有顺序信息的值视为线图，并学习图权重。后续研究 [34, 36] 进一步统一名义属性与有序属性的距离，并使其能够与聚类一同学习。

这些进展显著提升了分类数据的聚类性能。然而，仍有相互关联的棘手问题亟待解决：合理的先验知识 是有效距离学习的前提，而数据知识 通常是通过在定义明确的距离度量下观察数据分布而获得的。

5 总结

本文展示并分析了制约当前定性数据聚类性能的关键问题，即距离学习受到距离结构先验知识的限制。为此，本文提出了一种名为 COForest 的新型学习范式，将距离结构的构建纳入学习过程，并与聚类实现联合优化。给定期望簇数 k 后，COForest 的学习过程无需参数，且易于应用于各种数据集。此外，学习得到的树状距离结构简洁且具有很强的可解释性，非常适合表示定性数据的隐含分布。大量实验展示了 COForest 的优越性，以及其关键技术组件的有效性。

所提出的方法在给定“真实”簇数的情况下，针对静态定性数据展现出卓越的聚类性能。未来，可以考虑将其扩展到更复杂的实际情境，例如从由定量属性和定性属性混合构成的流式数据中学习 ，并处理簇数未知 且不平衡 的情况。

致谢

本研究部分获得国家自然科学基金委员会（NSFC）项目资助：62476063、62102097、62376233 和 62306181，以及国家自然科学基金委员会/研究资助局

（RGC）联合研究计划项目资助，项目编号为 N_HKBU214/21；广东省自然科学基金项目资助：2024A1515010163 和 2023A1515012855；研究资助局优配研究金项目资助：12201321、12202622 和 12201323；研究资助局高级研究学者计划项目资助，项目编号为 SRFS2324- 2S02；深圳市科技计划项目资助，项目编号为 RCBS20231211090659101；以及小米青年人才计划资助。

参考文献

[1] A. Agresti. 有序分类数据分析. 约翰·威立父子公司, 2010.[2] A. Ahmad和L. Dey. 一种在分类数据集无监督学习中计算同一属性的两个类别值之间距离的方法. 《模式识别快报》, 28(1):110–118, 2007. [3] M. Alamuri、B. R. Surampudi和A. Negi. 分类数据的距离/相似性度量综述. 见*2014*年国际神经网络联合会议论文集, 第1907–1914页, 2014. [4] P. Arabie、N. D. Baier、C. F. Critchley和M. Keynes. 分类、数据分析与知识组织研究. 分类、数据分析与知识组织研究，沙姆，瑞士：施普林格, 2006. [5] L. Bai和J. Liang. 基于熵范数的稀疏子空间聚类. 见第*37*届国际机器学习会议论文集, 第119卷, 第561–568页, 2020. [6] S. Boriah、V. Chandola和V. Kumar. 分类数据的相似性度量：比较评估. 见*2008*年*SIAM*国际数据挖掘会议论文集, 第243–254页, 2008. [7] S. Cai、Y. Zhang、X. Luo、Y.-M. Cheung、H. Jia和P. Liu. 多粒度竞争学习引导的鲁棒分类数据聚类. 见第*44*届*IEEE*分布式计算系统国际会议论文集, 2024. [8] J. Chen、Y. Ji、R. Zou、Y. Zhang和Y.-M. Cheung. Qgrl：面向异构特征数据聚类的四元数图表示学习. 见第*30*届*ACM SIGKDD*知识发现与数据挖掘会议论文集, 2024. [9] Y.-M. Cheung 和 H. Jia. 基于统一相似性度量且无需已知簇数的类别与数值属性数据聚类. 《模式识别》, 46(8):2228–2238, 2013. [10] T. R. dos Santos和L. E. Zárate. 分类数据聚类：推荐何种相似性度量？ 《专家系统与应用》, 42(3):1247–1260, 2015. [11] N. El Malki、R. Cugny、O. Teste和F. Ravat. Decwa：使用Wasserstein距离的基于密度的聚类. 见第*29*届*ACM*国际信息与知识管理会议论文集, 第20052008页, 2020. [12] P. A. Estévez、M. Tesmer、C. A. Perez和J. M. Zurada. 标准化互信息特征选择. 《*IEEE*神经网络汇刊》, 20(2):189–201, 2009. [13] A. E. Ezugwu、A. M. Ikotun、O. O. Oyelade、L. Abualigah、J. O. Agushaka、C. I. Eke和A. A. Akinyelu. 聚类算法综合综述：前沿机器学习应用、分类体系、挑战与未来研究展望. 《人工智能工程应用》, 110:104743, 2022. [14] A. J. Gates和Y.-Y. Ahn. 随机模型对聚类相似性的影响. 《机器学习研究杂志》, 18(1):3049–3076, 2017. [15] X. He 、 D. Cai 和 P. Niyogi. 用于特征选择的拉普拉斯分数. 见第*18*届神经信息处理系统国际会议论文集, 第507–514页, 2005. [16] Z. Huang. 面向含类别值的大型数据集聚类的k均值算法扩展. 《数据挖掘与知识发现》, 2 (3):283–304, 1998. [17] D. Ienco、R. G. Pensa和R. Meo. 从上下文到距离：面向分类数据聚类的差异度学习. 《*ACM*数据知识发现汇刊》, 6(1):1–25, 2012. [18] A. K. Jain、M. N. Murty和P. J. Flynn. 数据聚类：综述. 《*ACM*计算调查》, 31(3):264–323, 1999. [19] H. Jia 和 Y.-M. Cheung. 未知簇数下类别与数值数据的子空间聚类. *IEEE*神经网络与学习系统汇刊, 29(8):3308–3325, 2018. [20] H. Jia、Y.-M. Cheung和J. Liu. 一种用于分类数据无监督学习的新距离度量. *IEEE*神经网络与学习系统汇刊, 27(5):1065–1079, 2016. [21] S. Jian、L. Hu、L. Cao和K. Lu. 用于学习混合数据表示的基于度量的自动指导器. 见第32届AAAI人工智能会议论文集, 第3318–3325页, 2018. [22] S. Jian 、 G. Pang 、 L. Cao 、 K. Lu 和 H. Gao. Cure：通过层次耦合学习实现灵活的类别数据表示. IEEE知识与数据工程汇刊, 31(5):853–866, 2018. [23] M. Kelly、R. Longjohn和K. Nottingham. UCI机器学习知识库. [24] J. B. Kruskal. 论图的最短生成子树与旅行商问题. 见美国数学学会会刊, 7(1):48–50, 1956. [25] S. Q. Le和T. B. Ho. 一种面向分类数据的基于关联的相异度度量. 《模式识别快报》, 26(16):2549–2557, 2005. [26] T. Leng、L. Zhao、X. Xiong、P. Cheng和J. Zhou. 面向深度嵌入的基于自表达网络的子空间聚类. 见第26届欧洲人工智能会议论文集, 第1357–1364页, 2023. [27] D. Lin. 相似性的信息论定义. 见第15届国际机器学习会议论文集, 第296–304页, 1998. [28] L. v. d. Maaten和G. Hinton. 使用t-SNE可视化数据. 机器学习研究期刊, 9(11):2579–2605, 2008. [29] J. Makkar、S. Jain、S. Gupta等. Mfc：一种面向联邦数据聚类的多轮方法. 见第26届欧洲人工智能会议论文集, 第1672–1679页. 2023. [30] F. Nie、J. Xue、W. Yu和X. Li. 基于锚点引导的快速聚类. IEEE模式分析与机器智能汇刊, 第1898–1912页, 2024. [31] R. C. Prim. 最短连接网络及其若干推广. 贝尔系统技术期刊, 36(6):1389–1401, 1957. [32] W. M. Rand. 聚类方法评价的客观准则. 美国统计协会期刊, 66 850, 1971 336 846( ): – . [33] Y. Zhang和Y.-M. Cheung. 一种具有自动距离学习的有序数据聚类算法. 见第 34 届 AAAI 人工智能会议论文集 , 第 6869–6876 页 , 2020. [34] Y. Zhang和Y.-M. Cheung. 面向含名义属性与有序属性的分类数据聚类的属性内距离可学习加权. IEEE模式分析与机器智能汇刊, 44(7):3560–3576, 2022. [35] Y. Zhang和Y.-M. Cheung. 一种利用异构属性间关系的新距离度量，用于有序与名义属性数据聚类. IEEE控制论汇刊, 52 771, 2022 2 758( ): – . [36] Y. Zhang和Y.-M. Cheung. 面向任意类型属性数据聚类分析的基于图的相异度度量. IEEE神经网络与学习系统汇刊, 34(9):6530–6544, 2023. [37] Y. Zhang 、 Y.-M. Cheung 和 K. Tan. 一种用于有序与名义属性数据聚类的统一基于熵的距离度量. IEEE神经网络与学习系统汇刊, 31(1):39–52, 2020. [38] Y. Zhang、Y.-M. Cheung和A. Zeng. Het2Hom：将异构属性表示到同质概念空间中，用于类别与数值属性数据聚类. 见国际人工智能联合会议, 第3758–3765页, 2022. [39] M. Zhao、S. Feng、Y. Zhang、M. Li、Y. Lu和Y.-M. Cheung. “面向定性属性数据聚类的学习序关系森林”的代码与补充材料. 见第27届欧洲人工智能会议论文集, 2024. URL https://github.com/ZMJ-lucky/ ECAI-24-COForest.git. [40] C. Zhu、Q. Zhang、L. Cao和A. Abrahamyan. Mix2vec：无监督混合数据表示. 见第7届数据科学与高级分析国际会议论文集, 第118 127页, 2020 – . [41] C. Zhu、L. Cao和J. Yin. 面向类别表示的无监督异构耦合学习. IEEE模式分析与机器智能汇刊, 44(1):553–549, 2022.