分类变量距离实现的通用框架

Michel van deVelden Alfonso IodiceD'Enza Angelos Markos Carlo Cavicchia

2023年1月6日

arXiv:2301.02190v1 [stat.ML] 4 Jan 2023

摘要

通过一组变量测量某些属性时，研究对象彼此之间的差异程度在许多统计方法中都发挥着重要作用。例如，分类、聚类和数据可视化方法都需要量化观测值之间的差异。我们可以将这种差异的量化称为距离。距离的恰当定义取决于数据的性质和所处理的问题。对于数值变量之间的距离，已有许多依据观测差异大小定义的方式。对于分类数据，距离的定义则更为复杂，因为观测差异的大小无法直接量化。因此，已有许多基于分类变量度量差异的方法。本文提出了一个通用框架，可高效、透明地实现分类变量观测值之间的距离。我们表明，若干已有距离可以纳入该框架。此外，该框架还能自然地引出新的距离形式，并支持实现灵活且针对具体案例和数据的距离定义。此外，在监督分类场景中，该框架可用于构建能够纳入响应变量与预测变量之间关联的距离，从而提升基于距离的分类器的性能。

1 引言

在许多统计方法中，对差异度（即对象彼此不同的程度）进行量化具有重要作用。我们可以将这种差异度量化称为距离。K近邻（KNN，Cover和Hart 1967年）等分类方法，以及K均值（MacQueen 1967年）、围绕中心点划分（PAM，Kaufman和Rousseeuw 1990年）等聚类方法、层次连接法（Gordon 1999年），还有多维尺度分析（Borg和Groenen 2005年）和双标图（Gabriel 1971年；Gower、Lubbe和Le Roux 2011年）等数据可视化方法，都需要定义个体和/或对象之间的距离。距离定义的选择方式取决于数据的性质和所要解决的问题。

数值数据的距离度量通常基于观测值之间差异的大小（不同距离度量的列表可参见例如Mardia 1978年）。然而，对于分类数据，情况更为复杂，因为我们无法直接观测、因而也无法直接量化差异的大小；我们只能直接判定是否存在差异。

在多变量情境下计算距离时，可以区分两种情况。第一种是分别为每个变量计算距离，然后将其相加。第二种是在计算距离时考虑变量之间的关联。对于数值变量，若干广为人知的距离（例如欧几里得距离或曼哈顿距离）都隐含地假设变量之间相互独立。显然，在这种"独立"情况下，测量尺度必须具有可比性。对于分类变量，也有若干度量在计算多变量距离时，会对每个变量的不相似度求和。例如，在简单匹配中，两个观测值之间的距离定义为对应变量的类别不匹配的次数。

对于数值变量，可以使用马氏距离来考虑变量间关联，即利用（样本）协方差对观测到的差异加权，以考虑变量间的相关性

。对于分类数据，存在所谓的基于关联的距离。在这类距离中，利用分类变量之间的关联来量化观测值之间的差异。如何在分类数据场景中考虑关联并非易事。近年来提出的几种分类变量距离新方法确实属于基于关联的距离（例如，Le 和 Ho 2005；Ahmad 和 Dey 2007；Jia、Cheung 和 Liu 2014；Ring 等人 2015）。

定义分类变量距离的复杂性，以及近期对这一主题的关注，可从大量综述现有距离（例如，Boriah、Chandola 和 Kumar 2008；Alves、Couceiro 和 Napoli 2019）或提出（新）距离（例如，Le 和 Ho 2005；Ahmad and Dey 2007；Jia、Cheung 和 Liu 2014；Ring 等人 2015；Šulc 和 ˇ Rezanková 2019；Bai 和 Liang 2022）的文章中得到体现。本文提出了一个用于实现分类距离的通用框架。我们的框架既可用于纳入现有的分类变量距离，也允许研究人员定义并实现新的或定制的距离。

通过在我们的框架中重新表述现有距离，可以评估它们之间的差异与相似之处。目前，由于这些方法所采用的记号和研究领域（因而目标）种类繁多，这类比较并非易事。此外，我们的框架还能构建并定义新的、可高度定制的距离。例如，在监督分类情境下，该框架可用于定义一种考虑变量与响应变量类别之间关联的距离。

由于我们的框架并非针对特定方法或应用，因此可用于计算任何需要距离计算的方法或应用所需的距离矩阵。例如，多维尺度分析、聚类分析，或在监督情境下的 K近邻。我们证明，使用该框架进行距离计算快速、高效且透明，并且相比现有实现有显著改进。具体而言，对于 Ahmad 和 Dey（2007）提出的分类变量距离，我们的实现比现有实现快得多。

关于距离度量的定义，一个重要问题是如何验证这些度量。也就是说，我们如何知道所选度量是否恰当？虽然这一新框架并未回答这个问题，但采用统一的通用表述，能够简化对不同选择进行理论和实证比较的过程。

我们通过选用若干已知的和新提出的基于关联的分类距离，将基于距离的数据分析方法应用于多个知名分类数据集，以此展示我们的方法。我们在 R软件包 catdist 中实现了基于通用框架执行所有分类距离计算的函数。该软件包可在 GitHub[1] 获取，并即将在 CRAN 上发布。

本文的组织结构如下。第2节介绍一些记号后，第3节将阐述我们的通用框架。第 4 节介绍几种常见的分类变量距离度量，并展示如何将它们纳入框架。第5节介绍基于共现的分类距离，并特别关注 Ahmad 和 Dey（2007）提出的一种距离度量。第 6 节展示如何利用我们的框架构建和实现监督距离。第7节介绍距离定义的调整，随后在第8节使用几个数据集展示我们的方法。第9节总结全文。

2 记号

假设我们有n 个观测值，涉及Q 个分类变量，并令$j \in 1 \dots Q$个变量的类别数为q_j。然后，我们可以使用指示矩阵对分类数据进行编码。也就是说，对于每个分类变量$j \in 1 \dots Q$，我们创建一个$n \times q_j$二元矩阵$\mathbf{Z}_j$，其中n行对应观测值，q^j 列对应类别。观测到的类别用 1 表示，所有其他类别均赋值为 0。此外，对于分类变量的每个观测值，恰好观测到一个类别，并且我们只纳入数据集中至少出现过一次的类别。因此，$\mathbf{Z}^j$ 的每一列至少包含一个值为 1 的元素，并且$\mathbf{Z}_j \mathbf{1}_{q_j} = \mathbf{1}_n$，其中，通常情况下，$\mathbf{1}_i$表示一个$i$ 维列向量，其元素全为 1。也就是说，各列之和为 1。

[1]https://github.com/alfonsoIodiceDE/catdist_程序包

利用这些指示矩阵，我们可以将Q个分类变量的数据编码为一个所谓的超指示矩阵，方法是将所有指示矩阵并排排列。也就是说，

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \dots & \mathbf{Z}_Q \end{pmatrix}.$$

此外，定义

$$\mathbf{P} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}, \tag{1}$$

and

$$\mathbf{P}_d = \frac{1}{n}\left(\mathbf{Z}'\mathbf{Z}\right) \odot \mathbf{I}_{Q^*}, \tag{2}$$

其中 $\odot$ 表示阿达马积，即逐元素乘法，而 $Q^* = \sum_{j=1}^{Q} q_j$。注意，$\mathbf{P}_d$ 是一个对角矩阵，其对角元素是各变量内各类别的观测相对频率。此外，令

$$\mathbf{p} = \mathbf{P}_d \mathbf{1}_{Q^*} \tag{3}$$

表示观测相对频率向量，并且

$$\mathbf{p}^- = \mathbf{P}_d^{-1} \mathbf{1}_{Q^*} \tag{4}$$

是观测相对频率倒数构成的向量。

注意，ij对应的$\mathbf{P}$ 非对角块给出了变量i 和j的各类别的共现相对频率。它们可视为变量i和j的（经验）联合概率分布。为计算第5节中的基于关联的距离，我们还定义

$$\mathbf{R} = \mathbf{P}_d^{-1}\left(\mathbf{P} - \mathbf{P}_d\right). \tag{5}$$

ij对应的$\mathbf{R}$ 非对角块的各行，给出了i变量的各类别在j变量各类别上的分布。这些可解释为（经验）条件分布。

3 基于类别差异度的分类距离计算

对于分类变量，如何量化不同类别之间的差异并不明显。例如，假设我们观察了三个人，分别来自荷兰、意大利和希腊。从地理上看，也许从文化上看，意大利和希腊比荷兰更相似。然而，如何将这种差异纳入考量并非易事。在我们的框架中，我们通过定义类别差异度来实现这一点。

矩阵 $\mathbf{\Delta}_j$ 是变量j的类别差异矩阵。该矩阵的元素 δ_{ab}，其中a 和b 表示变量j的两个类别，用于量化第j个变量中类别a 和b之间的不相似度。我们可以对不相似度矩阵施加符合典型距离定义的条件。也就是说，1）某个类别与其自身的不相似度为零（$\delta_{aa} = 0$，适用于所有类别）

2）不相似度具有对称性（$\delta_{ab} = \delta_{ba}$，适用于所有类别对）。3）不相似度满足三角不等式。也就是说，如果a、b和c表示变量j的不同类别，那么对于所有类别a、b和c，

$$\delta_{ac} \leq \delta_{ab} + \delta_{bc}.$$

如果这三个条件都满足，则可以将这些不相似度视为类别之间的度量距离。如果它们为非负值且仅满足前两个条件，则可以将其解释为类别之间的非度量距离。不过，我们将它们称为类别差异度，并将“距离”一词保留用于观测值之间的距离。

如果我们有Q个分类变量，且每个变量都对应一个类别不相似度矩阵 $\mathbf{\Delta}_j$，则可以构造一个$Q^* \times Q^*$分块对角矩阵 $\mathbf{\Delta}$，其中各对角块分别为相应的类别不相似度矩阵。

类别不相似度矩阵可用于计算观测值之间的距离矩阵，方法如下。首先，考虑与第j个分类变量对应的n行q列$_j$指示矩阵$\mathbf{Z}_j$。此外，我们有对应的类别不相似度矩阵 $\mathbf{\Delta}_j$。我们可以得出以下定理：

定理 1. 分类变量 j 的观测值之间的距离为

$$\mathbf{D}_j = \mathbf{Z}_j \mathbf{\Delta}_j \mathbf{Z}_j'.$$

证明。 将Z_j的第i行与 $\mathbf{\Delta}_j$相乘，会选出 $\mathbf{\Delta}_j$中与个体i所选类别对应的行。类似地，将该行与$\mathbf{Z}_j'$的第i'列（即第i'个观测值）相乘，会选出与个体i'所选类别对应的元素。因此，$\mathbf{D}_j$的第（i, i'）个元素是个体i与i'所选类别之间的差异度。■

定理 2。如果将 Q 个分类变量上观测值之间的距离定义为每个分类变量对应的不相似度之和，则 $n \times n$ 距离矩阵可计算为

$$\mathbf{D} = \mathbf{Z}\mathbf{\Delta}\mathbf{Z}'. \tag{6}$$

证明。

$$\begin{aligned}
\mathbf{D} &= \mathbf{Z}\mathbf{\Delta}\mathbf{Z}' \\
&= \begin{pmatrix} \mathbf{Z}_1 & \dots & \mathbf{Z}_Q \end{pmatrix} \begin{pmatrix} \mathbf{\Delta}_1 & & \\ & \ddots & \\ & & \mathbf{\Delta}_Q \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_Q' \end{pmatrix} \\
&= \sum_{j=1}^{Q} \mathbf{Z}_j \mathbf{\Delta}_j \mathbf{Z}_j' \\
&= \sum_{j=1}^{Q} \mathbf{D}_j
\end{aligned}$$

■

由 (6) 可知，分类变量观测值之间的距离取决于类别不相似度矩阵 $\mathbf{\Delta}_j$ 的选择。这为定义一组分类变量的适当距离度量提供了很大的灵活性。在下一节中，我们将简要回顾 $\mathbf{\Delta}$ 的一些选择，并说明它们与现有距离的关系。

请注意，此公式并未明确纳入分类变量之间的关联。也就是说，对于一个变量观测到的类别差异，与其他变量观测到的类别差异并无关联。不过，有一些方法可以考虑此类观测情况。例如，与其为每个分类变量构建一个指示矩阵，不如为观测值的所有可能组合（或其子集）构建一个指示矩阵。也就是说，可以针对每个类别组合（或其子集）分别创建一个指示矩阵。因此，此类矩阵的列数为 $\prod_{j=1}^{Q} q^j$，并且只需一个类别不相似度矩阵，其中每个类别都是所有 Q 个变量的类别组合。然而，当分类变量有多个时，组合总数以及最终指示矩阵中的类别数会迅速增多。此外，为这些组合找到合适的类别不相似度矩阵并非易事。

考虑分类变量之间关联的另一种方法，是在构建类别不相似度矩阵时利用这些关联。也就是说，根据某个变量在 $\mathbf{\Delta}_j$ 中的类别与其他变量之间的关联，来定义该变量各类别之间的不相似度。在第5节中，我们将给出这类类别差异度量的示例。

3.1 集合间距离

假设我们有两组分别针对相同Q 个分类变量的观测值。这两组数据可分别整理为$n_1 \times Q^*$ 和$n_2 \times Q^*$ 超级指示矩阵$\mathbf{Z}^{(1)}$ 和$\mathbf{Z}^{(2)}$。那么，对于已知的类别不相似度矩阵 $\mathbf{\Delta}$，很容易验证，两组观测值之间的距离可计算为

$$\mathbf{D}^{(12)} = \mathbf{Z}^{(1)}\mathbf{\Delta}\mathbf{Z}^{(2)\prime}. \tag{7}$$

请注意，矩阵 $\mathbf{D}^{(12)}$ 的阶数为 $n_1 \times n_2$。

在考虑基于距离的分类问题时，计算集合间距离可能会很有用。例如，在 K近邻中，需要计算“新”（未标记）观测值与带标签数据集中的观测值之间的距离。K近邻预测基于训练集中K 个最近邻的标签（通常通过多数投票得出）。类似地，在围绕中心点划分中，这是一种类似于 K均值的常用聚类算法；它不考虑围绕均值的簇内变异，而是考虑围绕实际观测值（即中心点）的变异。如果将中心点收集在$\mathbf{Z}^{(1)}$ 中，并将“新”数据点收集在$\mathbf{Z}^{(2)}$中，我们便可以通过考虑距离$\mathbf{D}^{(12)}$ 并选择最小距离，将新点分配到现有簇中。

4 独立类别差异度矩阵

我们首先考虑几种分类数据差异度的定义，这些定义不考虑变量间关联。因此，在多变量情境下，距离按变量分别计算后求和；对于每个变量，其类别差异度独立于其他变量上的观测值。不过，类别差异度可能取决于某个变量的观测频数。

我们并不打算全面涵盖不相似度的各种定义，而是选取 Šulc 和 ˇRezanková（2019）提出的定义（其中也包含了 Boriah、Chandola 和 Kumar 2008 以及 Alves、Couceiro 和 Napoli 2019 回顾的若干定义），并将其转换为差异度。我们通过定义适当的类别不相似度矩阵 $\mathbf{\Delta}$，展示如何将这些差异度纳入我们的框架。关于更详细的说明，以及在聚类分析场景下对这些不相似度定义的相对表现的研究，请参见 Šulc 和 ˇRezanková（2019）。

4.1 重叠或简单匹配

简单匹配（SM）的思路是：若两个观测值的类别不匹配，则它们之间的距离为 1；若匹配，则距离为 0。因此，所有不同类别之间的类别不相似度均为 1。对于具有 q_j 个类别的第 j 个分类变量，我们定义

$$\mathbf{\Delta}_{M_j} = \mathbf{1}_{q_j}\mathbf{1}'_{q_j} - \mathbf{I}_{q_j}.$$

也就是说，每个类别之间的距离恰好为 1。如果我们有 Q 个分类变量，并希望计算所有变量的简单匹配，只需将所有 $\mathbf{\Delta}_{M_j}$ 收集到分块对角矩阵 $\mathbf{\Delta}_M$中。

$$\mathbf{\Delta}_M = \mathbf{K}_b - \mathbf{I}_{Q^*},$$

其中，$\mathbf{K}_b$是一个$Q^* \times Q^*$分块对角矩阵；对于$j = 1 \ldots Q$，其对角块为$q_j \times q_j$ </style}的全 1 矩阵（Kb_j）。

4.2 Eskin

对于 Eskin距离（Eskin等人（2002年）），类别差异度取决于类别数。类别数较多的变量，其不相似度小于类别数较少的变量。具体而言，对于具有q_j 个类别的变量，不同类别之间的差异度为

$2/q_j^2$。因此，对于具有q_j个类别的第j个分类变量，其类别差异矩阵定义为

$$\mathbf{\Delta}_{E_j} = \left(2/q_j^2\right)\left(\mathbf{1}_{q_j}\mathbf{1}'_{q_j} - \mathbf{I}_{q_j}\right).$$

将所有Q 差异度矩阵收集到一个分块对角矩阵中，可得到 Eskin 类别差异矩阵 $\mathbf{\Delta}_E$。如果所有变量具有相同的类别数，Eskin 仅对简单匹配差异度进行重新缩放。

4.3 Lin

Lin（1998年）提出了一种信息论度量：对频繁取值上的匹配赋予更高权重，并对不频繁取值上的不匹配赋予更低权重。我们按如下方式实现 Lin 的方案：定义$\widetilde{\ }\mathbf{P}_r = \mathbf{p}${v7>以及 $\mathbf{P}_c = \mathbf{1}_{Q^*}\mathbf{p}'$，其中$\mathbf{p}$ 的定义见第 3 节。此外，令 $\mathbf{P} = \mathbf{P}_r + \mathbf{P}_c - \mathbf{P}_d$。于是，类别差异矩阵可定义为

$$\mathbf{\Delta}_{Lin} = \left[\log(\mathbf{P}_r) + \log(\mathbf{P}_c) - 2\log(\widetilde{\mathbf{P}})\right] \oslash 2\log(\mathbf{P}_r + \mathbf{P}_c),$$

其中，$\oslash$ 表示 Hadamard 除法（即逐元素除法），而 log(·) 对括号内对象的各元素取对数，并将结果收集为一个大小相同的对象。

请注意，Lin 对某个类别与其自身之间的差异度定义为零。此外，在我们的实现中，对于每个变量，两个不同类别（例如类别a 和b）之间的不相似度为

$$[\log(p_a) + \log(p_b) - 2\log(p_a + p_b)]/2\log(p_a + p_b),$$

其中，p_a 和p_b 分别是类别a 和b的相对频率。

4.4 逆出现频率

对于逆出现频率（IOF，Boriah、Chandola 和 Kumar 2008），类别被观察得越频繁，赋予的差异度越高。具体而言，类别差异矩阵定义为

$$\mathbf{\Delta}_{IOF} = [\log(n\mathbf{p})]\,[\log(n\mathbf{p})]' - [\log(n\mathbf{p})]\,[\log(n\mathbf{p})]' \odot \mathbf{I}_{Q^*}.$$

值得注意的是，对于每个变量，IOF 对某个类别与其自身之间的差异度为零，而两个不同类别（例如a 和b）之间的差异度对应于

$$\log(np_a)\log(np_b).$$

IOF 度量与信息检索中的逆文档频率（TF-IDF）概念相关，用于衡量给定术语对文档的相关性（Spärck Jones 1972年）。换言之，由于罕见术语比更常见的术语贡献更多信息，IOF 度量反映了术语的罕见程度，而较低的 IOF 不相似度对应于更罕见的术语。使用对数频率可以降低出现频率极高的术语所产生的影响。

4.5 出现频率

对于出现频率（OF）差异度，类别出现得越少，不相似度就越高。类别差异矩阵定义如下：

$$\mathbf{\Delta}_{OF} = \left[\log\left(\mathbf{p}^-\right)\right]\left[\log\left(\mathbf{p}^-\right)\right]' - \left[\log\left(\mathbf{p}^-\right)\right]\left[\log\left(\mathbf{p}^-\right)\right]' \odot \mathbf{I}_{Q^*}.$$

因此，某个类别与自身之间的 OF 不相似度为零，而两个不同类别a 和b 之间的差异度为

$$\log(p_a)\log(p_b).$$

4.6 Goodall不相似度

在 Boriah、Chandola和Kumar（2008）中，探讨了 Goodall相似度的四种变体。这些变体基于 Goodall 的原始提议（Goodall 1966）。将相似度转换为不相似度后（不相似度为 1–相似度），这四种度量有一个共同点：不同类别之间的不相似度，如同简单匹配的情况一样，始终等于一。然而，某个类别与自身之间的差异度取决于各类别的观测比例。下面给出 Goodall 3 和 Goodall 4 的类别不相似度矩阵。对于 Goodall 1 和 Goodall 2，也可以构造此类矩阵。不过，这些定义需要用到比例的条件和。具体而言，对于 Goodall 1，类别a 与自身之间的差异度定义为观测比例小于或等于类别a观测比例的各类别观测比例的平方和。对于 Goodall 2，则定义为观测比例大于或等于类别a观测比例的各类别观测比例的平方和。

Goodall 3 和 4 度量无需计算（条件）总和，并且在 $\mathbf{\Delta}$ 的对角块上分别使用类别比例的平方和 1 减去类别比例的平方。也就是说，

$$\mathbf{\Delta}_{G_3} = \mathbf{K}_b - \mathbf{I}_{Q^*} + \mathbf{P}_d^2,$$

and

$$\mathbf{\Delta}_{G_4} = \mathbf{K}_b - \mathbf{P}_d^2.$$

按照这些定义，一个类别与其自身的类别不相似度并非为零。因此，所得“距离”不满足距离的典型要求。请注意，对于 Goodall 1 和 3 度量，当匹配的类别较常见时，赋予较高的不相似度；而对于 Goodall 2 和 4 度量，当匹配的类别较少见时，赋予较高的不相似度。

4.7 变量熵和变量易变性不相似度

Šulc 和 ˇRezanková（2019年）提出了两种分别与 Goodall 1 和 2 相关的基于变异性的不相似度度量。当类别不匹配时，这些不相似度等于 1；而当类别匹配时，变量熵（VE）度量使用熵，变量易变性（VM）度量使用基尼系数来量化“不相似度”。具体而言，对于第j个分类变量（含q_j个类别），类别不相似度矩阵定义如下：

$$\mathbf{\Delta}_{VE_j} = \mathbf{K}_{b_j} + \left(\frac{1}{\log q_j} \sum_{l=1}^{q_j} p_l \log p_l \right) \mathbf{I}_{q_j}$$

and

$$\mathbf{\Delta}_{VM_j} = \mathbf{K}_{b_j} - \left[\frac{q_j}{q_j - 1} \left(1 - \sum_{l=1}^{q_j} p_l^2 \right) \right] \mathbf{I}_{q_j},$$

分别如此。将所有Q 差异度矩阵汇集到一个分块对角矩阵中，即可得到 VE 和 VM 类别不相似度矩阵 $\mathbf{\Delta}_{VE}$和 $\mathbf{\Delta}_{VM}$。

4.8 有序类别

如果类别有序，则可以在不相似度中体现其顺序。一种简单的做法是将不相似度定义为类别编号之差。也就是说，类别a与b之间的差异度就是$b - a$。如果数据是等级顺序数据或评分（例如，李克特）量表数据，这一定义就意味着将数据视为区间数据。不过，实现其他自定义的有序类别间距离定义也很简单。例如，如果类别对应数值尺度上的分箱（例如年龄组或收入组），则可以使用分箱中点之间的差值

来定义能更准确反映底层数值的不相似度。更一般地，令 $\mathbf{\Delta}_o^i$ 表示 $\mathbf{\Delta}_o^*$的第i个对角块， δ_{ab}^i 表示其中的ab元素。然后，可以通过令 $\delta_{iab} \leq \delta_{iab^*}$，来设定有序类别之间的不相似度，其中$a$，$b \in 1 \ldots i_{q_i}$、$b^* \neq b$且$b^* > a$。请注意，此定义无法保证满足三角不等式。也就是说，如果没有额外约束，两个类别之间的直接距离可能大于它们之间的间接距离。

5 基于关联的类别差异矩阵

一些作者（例如 Ahmad and Dey 2007；Jia、Cheung 和 Liu 2014；Le 和 Ho 2005；Ring 等人 2015）提出了考虑分类变量间关联的分类变量距离度量。其基本思想是，与数值变量的马氏距离类似，与变量间关联一致的差异所含信息较少（即应对应较小的差异度值），而与总体关联不一致的差异所含信息较多。如何具体实现这一思想，取决于如何计算分类变量间的关联，以及如何将这种关联纳入类别差异度。

这里，我们给出一种计算并汇集基于关联的差异度的通用形式，可直接在第3节的通用框架中实现。随后，我们介绍若干具体变体，并将其与近期提出的方法联系起来。

5.1 基于关联的差异度的一般形式

基于关联的距离的计算核心在于共现比例矩阵**P** 以及相应的剖面矩阵**R**，其定义见公式（1）和（5）。特别要注意，**P** 和**R** 的非对角块分别可解释为（经验）联合概率分布和条件概率分布。通过考虑不同方式来量化条件分布之间的不相似度（即**R**的非对角块的各行之间的不相似度），我们可以构造不同的类别不相似度矩阵 $\mathbf{\Delta}$，再应用公式（6）即可获得观测间距离。

令$\mathbf{R}^{ij}$ 表示**R**的第ij个非对角块，并令$\mathbf{r}_a^{ij}$表示其第a行。注意，$\mathbf{R}^{ij}$ 每一行的元素之和均为 1。因此，这些元素可视为（经验）条件概率。对于变量i的所有类别对（其中$i = 1 \ldots Q$），我们根据其与变量j （其中$j \neq i$）的关联，定义类别之间的不相似度如下：

$$\delta^{ij}(a,b) = \Phi^{ij}\left(\mathbf{r}_a^{ij}, \mathbf{r}_b^{ij}\right), \tag{8}$$

其中，一般而言，a 和b 表示变量i 的类别，而 Φ^{ij} 是不相似度函数，用于量化基于变量i 和j之间关联的剖面差异。变量i 所有类别对之间的总体类别差异度可定义为

$$\delta^i(a,b) = \sum_{j \neq i}^{Q} w_{ij}\delta^{ij}(a,b) = \sum_{j \neq i}^{Q} w_{ij}\Phi^{ij}\left(\mathbf{r}_a^{ij}, \mathbf{r}_b^{ij}\right). \tag{9}$$

公式（9）中的权重w_{ij} 使我们能够灵活调整不同变量在计算基于关联的类别差异度时的重要性，如 Φ^{ij}所定义。对于每个变量i，收集类别差异矩阵 $\mathbf{\Delta}_\Phi^i$中的元素 $\delta_i(a,b)$，并将它们排列在分块对角矩阵的对角线上，即可得到不相似度矩阵 $\mathbf{\Delta}_\Phi$，进而可使用公式（6）计算观测值之间的距离。

公式 9 提供了一种非常通用的方法，可使用成对特定权重和不相似度函数来定义类别差异度。

对于基于关联的不相似度函数 Φ^{ij}，任何能够量化两个分布之间差异的函数都可以使用。Drost (2018) 简要介绍了 46 种不同函数及其在 R 软件包 philentropy 中的实现；Cha (2007) 则对这些函数进行了更全面的综述，并将它们划分为不同的类型和类别。

关于权重 w^{ij} 的选择，我们区分两种方案。第一种是所有权重都相等，通常取 1 或 $1/(Q-1)$，从而得到总和或平均值。或者，也可以为不同的变量对设置不同的权重。这些权重既可以依据专家知识来选择（例如，根据研究人员的经验和偏好），也可以采用数据驱动方法。例如，可以根据预先确定的、依赖于数据的标准（如 Cramér's V 这样的关联度量），将某些权重设为零，而将其他权重设为某个常数。权重非零的变量对随后可以称为"上下文"变量。Ienco、Pensa 和 Meo (2009)、Jia、Cheung 和 Liu (2014) 以及 Ring 等人 (2015) 介绍了采用此类基于上下文的不相似度的方法。

如果有衡量解整体拟合度的客观指标，可以将w^{ij} 和 Φ^{ij} 视为调节参数，并搜索这些参数的不同组合以作出选择。在第7节中，我们将进一步探讨w^{ij} 和 Φ^{ij}的调节。

在接下来的小节中，我们将介绍一些特定的不相似度函数选择。具体而言，我们将基于变量 i 和j之间的关联，给出变量i的类别a 和b 之间的类别差异度定义。也就是说，我们将介绍公式（8）中 Φ^{ij} 的具体选择。将这些定义代入公式（9），即可得到不相似度矩阵 $\mathbf{\Delta}_\Phi$，并可用于计算观测间距离。为简化记号，我们现在省略上标ij。

5.2 剖面间总变差距离

两个离散概率分布之间的总变差距离（TVD）可定义为分布间L_1 范数的 $1/2$ 倍。我们可以通过将类别不相似度函数 Φ 定义为以下形式，在我们的框架中实现这一点：

$$\Phi(\mathbf{r}_a, \mathbf{r}_b) = \frac{1}{2}\sum_{l=1}^{q_j} |r_{al} - r_{bl}|, \tag{10}$$

其中r_{al} 和r_{bl} 分别表示$\mathbf{r}_a$ 和$\mathbf{r}_b$的第l个元素。

使用此定义计算 Φ 的类别差异度，等价于 Ahmad 和 Dey（2007）提出的方法（针对分类变量）。然而，由于这种关系并不显而易见，且似乎鲜为人知，因此我们在此对此作一些详细说明。

5.2.1 Ahmad 和 Dey 的分类变量距离

Ahmad 和 Dey（2007）认为，类别之间的差异度应根据它们在整体数据集中的分布以及与其他类别的共现情况计算，而不应孤立地计算。其思路是在构建距离时考虑类别的共现情况。具体而言，他们考察一个变量的所有类别组合，并选择一种划分（即类别组合），使两个互补划分中这两个类别所占比例之和最大。按照 Ahmad 和 Dey（2007）的方法，我们可以将变量i的类别a 和b 之间相对于变量j各类别分布的差异度定义为

$$\delta(a,b) = \max_{\omega_j} \left(P(\omega_j|a) + P(\bar{\omega}_j|b) - 1\right), \tag{11}$$

其中， ω_j 及其补集 $\bar{\omega}_j$ 针对变量j的类别构成二元划分，且$P(\omega_j|a)$表示变量<style id='i'>i中类别 a 对应的观测比例；该比例对应于变量j的类别集合，如 ω_j所定义。注意，引入术语 -1 仅为将不相似度的上限固定为 1。变量j的二元划分数量（不包括包含全部或不含任何类别的划分）等于 $2^{q_j}-2$，其中q_j表示变量j的类别数，因此，随着变量类别数增加，该数量呈指数增长。

Ahmad 和 Dey（2007）提出了一种计算其距离的算法。该算法的复杂度为$O(Q^{*2}n + Q^{*2}\bar{q}^3)$，其中$Q^*$ 表示类别总数，n 表示观测值数量，$\bar{q}$ 表示每个变量的平均类别数。然而，正如下文及附录 A 中更详细的说明，对于分类变量之间的距离，Ahmad 和 Dey（2007）的距离等价于全变差距离，而使用公式 10 进行计算的效率要高得多。

5.2.2 EqAhmad 和 Dey 距离与剖面间全变差距离的等价性 s

从 $\delta(a,\ b)$变为 $\delta(b,\ a)$时，最优划分 ω_j（即使总和最大化的类别组合）只需翻转（也就是说，取其补集），因此 Ahmad 和 Dey 距离具有对称性：

$$\delta(a,b) = \max_{\omega}\left(P(\omega|a) + P(\bar{\omega}|b) - 1\right) = \max_{\omega}\left(P(\omega|b) + P(\bar{\omega}|a) - 1\right) = \delta(b,a),$$

为方便起见，我们省略了各个 ω 的下标 j。

由于 $P(\bar{\omega}|b) = 1 - P(\omega|b)$，且 $P(\bar{\omega}|a) = 1 - P(\omega|a)$，因此我们有

$$\begin{aligned}\delta(a,b) &= \max_{\omega}\left(P(\omega|a) - P(\omega|b)\right)\\ &= \max_{\omega}\left(P(\omega|b) - P(\omega|a)\right)\\ &= \max_{\omega}\left|P(\omega|a) - P(\omega|b)\right|.\end{aligned} \tag{12}$$

公式（12）表明，Ahmad 和 Dey 距离等于在所有观测比例组合中寻找最大差值。这意味着，我们可以将该距离表示为概率差向量的上确界范数。公式（10）中定义的全变差距离，也可以定义为两个概率分布中可分配给同一事件的概率之间的最大差值。因此，Ahmad 和 Dey（2007）提出的分类变量距离等价于全变差距离。为完整起见，我们在附录 A 中给出了这一等价性的完整证明。

5.3 剖面之间的 Kullback-Leibler 散度

Kullback-Leibler 散度（KL，Kullback 和 Leibler，1951年；Kullback，1959）是一种基于熵的度量，用于衡量概率分布之间的不相似度。Le 和 Ho（2005年）通过计算所有其他变量上的（经验）条件概率分布之间的 KL 散度之和，为变量i 的各类别定义类别差异度。采用与前文类似的记号，我们可以将此散度实现为：将所有权重 w^{ij} 设为 1，并将 Φ 定义为

$$\Phi(\mathbf{r}_a, \mathbf{r}_b) = \sum_{l=1}^{q_j}\left[r_{al}\log\left(\frac{r_{al}}{r_{bl}}\right) + r_{bl}\log\left(\frac{r_{bl}}{r_{al}}\right)\right],$$

其中，log() 表示二进制对数；与前文一样，r_{al} 和 r_{bl} 分别表示$\mathbf{r}_a$ 和$\mathbf{r}_b$的第l个元素。需要注意的是，KL 并不对称。因此，使用 $\mathbf{\Delta}_{KL}$ 计算距离可能会得到非对称距离。

5.4 χ^2-剖面之间的距离

一种与数据可视化技术——对应分析——密切相关的分类数据距离是卡方距离。卡方距离有多种形式和实现方式，其区别在于所采用的标准化方法。也就是说，卡方距离考虑比例之差的平方，并将其除以期望比例。对于一个 $n \times p$ 列联矩阵 $\mathbf{F}j = \mathbf{Z}_i'\mathbf{Z}_j$，行 a和 b之间的平方 χ^2距离可定义为

$$s\sum_{l=1}^{p}\frac{1}{f_{\bullet l}}\left(\frac{f_{al}}{f_{a\bullet}} - \frac{f_{bl}}{f_{b\bullet}}\right)^2, \tag{13}$$

其中，s 是$\mathbf{F}$ 中所有元素的总和，$\bullet$ 表示对矩阵相应维度（行或列）求和（例如，参见 Gifi 1990年，第266页）。用我们的记号，可以通过定义以下内容，将卡方距离作为类别差异度来实现：

$$\Phi(\mathbf{r}_a, \mathbf{r}_b) = \sum_{l=1}^{q_j}\frac{1}{p_l}\left(r_{al} - r_{bl}\right)^2,$$

其中，我们省略了常数s，p_l 对应于$\mathbf{P}_d$ 的第j个块中的第l个元素；与之前一样，r_{al} 和r_{bl} 分别表示$\mathbf{r}_a$ 和$\mathbf{r}_b$的第l个元素。

6 基于关联的监督距离

在有监督设置中，我们希望根据分类变量 x_j（其中 $j = 1, \ldots, Q$）的观测值，将观测分配到某个变量（例如 y）的类别（即类别）中。为此，我们可以定义基于关联的分类变量距离的有监督变体。也就是说，我们可以定义能够考虑变量 y 与 x之间关联的类别差异度。接下来，我们可以使用 K 最近邻或基于距离的聚类方法进行预测，并将簇的数量设定为类别数量，再对簇和类别进行事后比较。也就是说，我们将簇与真实类别匹配，并据此分配标签。

为了定义基于关联的监督距离，我们创建一个 $n \times c$ 指示矩阵 $\mathbf{Z}_y$，其中 c对应于 y的类别数量。如果将该指示矩阵添加到 $\mathbf{Z}$ 的右侧，并将补充后的 $\mathbf{Z}$ 代入公式 (1) 至 (5)，就可以利用公式 (8) 和 (9) 计算类别差异度。请注意，在这一新设置中，我们有 $Q+1$ 个变量，因此有 $Q+1$ 个基于关联的类别差异矩阵 $\mathbf{\Delta}^i$。不过，第 $(Q+1)$ 个对角块给出了变量 y 各类别之间的类别差异度。在有监督设置中，需要根据其他 Q 个变量的数据预测 y 的某个类别。因此，不应使用 y 的类别差异度。只需在整体分块对角类别不相似度矩阵中忽略这些差异度$\mathbf{\Delta}$，即可轻松实现这一点。也就是说，我们通过仅收集对角线上前Q 个类别不相似度矩阵来构造 $\mathbf{\Delta}$。

和之前一样，我们的框架允许以极大的灵活性纳入变量 y 的信息。具体而言，不相似度函数 Φ^{ij} 和权重 w^{ij} 均针对变量对分别设定。正如第 5 节所建议的，也可以将所有权重 w^{ij} 设为 1，使类别差异度考虑所有关联。我们将这种选择称为“完全监督”差异度。

或者，在有监督设置中，可以选择让类别差异度仅取决于与变量 y 的关联。这对应于将 $w_{ij} = 1$ 设为 $j = Q+1$，并将所有其他变量对的权重设为 0。此时，类别差异度或许能更好地区分 y 的类别。我们将这种选择称为 “ 监督 ” 差异度。请注意，这两种监督变体都要求为所有变量对选择基于关联的不相似度函数 Φ^{ij}。

7 聚合与差异度调优

我们在第3节中提出的通用框架，为实现分类变量之间的距离提供了很大的灵活性。具体而言，在前文各节中，我们介绍了若干类别差异度量。不过，还有许多其他选择。例如，Drost (2018) 所述的 R 软件包 philentropy 中提供的全部 46 个函数，均可用于基于关联的函数。此外，正如定义（9）所示，所有独立的类别差异度量都可以根据研究者的偏好进行组合和聚合。如何准确确定最合适的类别差异度定义和聚合策略并非易事，这种选择可能取决于数据的属性和研究目标。

在若干基于距离的方法中，例如聚类分析和多维尺度分析，并没有明确的拟合度量，因为这类方法往往主要用于探索。也就是说，其目标是发现并解释数据中的模式。由于解决方案的可解释性难以量化，验证通常并非易事。然而，如果能够计算拟合度量，我们就可以用它来选择聚合方式和类别差异度定义。也就是说，我们可以采用多种聚合方式和类别差异度定义，并通过考察所选度量来比较每种方案的拟合效果。

在监督分类场景中，如果我们有一个已知真实类别的数据集，就可以通过比较真实类别与“预测”类别来评估拟合效果。随后，可以根据两者之间的差异来选择聚合方式和类别差异度定义。因此，聚合状态（即权重 w^{ij}）和类别差异函数（即 Φ^{ij}）可以

视为调优参数。不过请注意，公式 9 允许许多种组合，因此需要做出一些选择，以限制总搜索空间。

8 应用

为说明我们的通用框架如何在实践中应用，我们考虑用于监督学习和无监督学习的基于距离的方法。在有监督设置中，一种基于距离的方法是K近邻平均法，可用于回归和分类问题：每个新观测值都根据一组K 个相近的 训练点（邻居）进行标记。在无监督设置中，基于距离的聚类分析旨在将观测值分配到各个组（簇）中，使组内距离较小，而簇间距离较大。

作为一般设置，我们考虑九个不同的带标签数据集（见表 1），这些数据集均可通过 UCI 机器学习仓库[2]获取。每个数据集被划分为五折，用于交叉验证。在训练折上，针对每种已评述的不相似度度量（见表 2），计算一个由成对类别差异度构成的分块对角矩阵 $\mathbf{\Delta}$。测试折用于评估所考虑方法的性能。性能指标取决于所考虑的方法：

- 准确率 最近邻分类器的准确率：正确分类的测试观测值所占的比例（Metz 1978）；

- 调整兰德指数 （调整兰德指数，Hubert 和 Arabie 1985），用于比较测试观测值的簇分配与 真实 簇分配（标签）。

重复该过程，直到每一折都用作测试集。整个交叉验证过程针对不同的随机划分重复 10 次。

表 1：数据集信息

数据集	n	p	簇数
澳大利亚信用数据集	690	8	2
平衡	625	4	3
cars	1728	6	4
淋巴瘤	148	18	4
大豆（大型）	307	35	19
tae	151	5	3
井字棋	958	9	2
vote	435	16	2
wbcd	699	9	2

表 2：实验中考察的类别差异度量

独立	基于关联
简单匹配（第 4.1 节）	总变差距离（第 5.2 节）
Eskin（第 4.2 节）	KL（第 5.3 节）
Lin（第 4.3 节）	KL（第 5.3 节）
IOF（第 4.4 节）	有监督 TVD、有监督 TVD-full（第 6 节）
OF（第 4.5 节）	
Goodall 3 和 4（第 4.6 节）	
VE、VM（第 4.7 节）	

[2]https://archive.ics.uci.edu/ml/index.php

8.1 K近邻分类数据

测试观测值的 K近邻分类基于计算每个测试观测值与训练观测值之间的距离，具体如第 3.1 节所述。令 $\mathbf{\Delta}_{train}$ 表示类别差异矩阵，其中下标$train$ 表示：如果类别差异度取决于数据，则仅使用训练集中的观测值。此外，$\mathbf{Z}_{test}$ 和$\mathbf{Z}_{train}$ 分别是测试观测值和训练观测值的指示矩阵。我们关注的距离位于以下各列中

$$\mathbf{Z}_{train}\mathbf{\Delta}_{train}\mathbf{Z}'_{test}$$

第j个测试观测的最近邻，是j列中最小的K 个值所对应的观测。

考虑的邻居数量越少，分类器的灵活性越高，因此K 是一个超参数。我们使用第8节所述的重复交叉验证程序来调整该超参数。具体而言，我们考虑$K \in \{1,3,5,9,15,21\}$ 的取值；对于每种数据集／类别差异度组合，选择使分类器测试准确率的交叉验证估计最小的K 值。

图1展示了针对每个所考虑的数据集，以及每种所考虑的类别差异度定义，调整后的KNN分类器的准确率评估结果。在每个面板中，每个点的位置对应于使用所示类别差异度定义得到的准确率；点的大小与调整后的超参数K值成正比。以每个点为中心的线段表示准确率在10次重复试验中的两个标准差范围。对于每个数据集，距离按降序排列，以突出表现最佳的距离。对于某些数据集，例如澳大利亚信用数据集、投票和$wbcd$，无论选择哪种距离，准确率都很高，并且在10次交叉验证重复试验中没有变异。对于较小的数据集，例如淋巴瘤 和助教评估数据集，10次交叉验证重复试验中的变异更大，这符合预期。

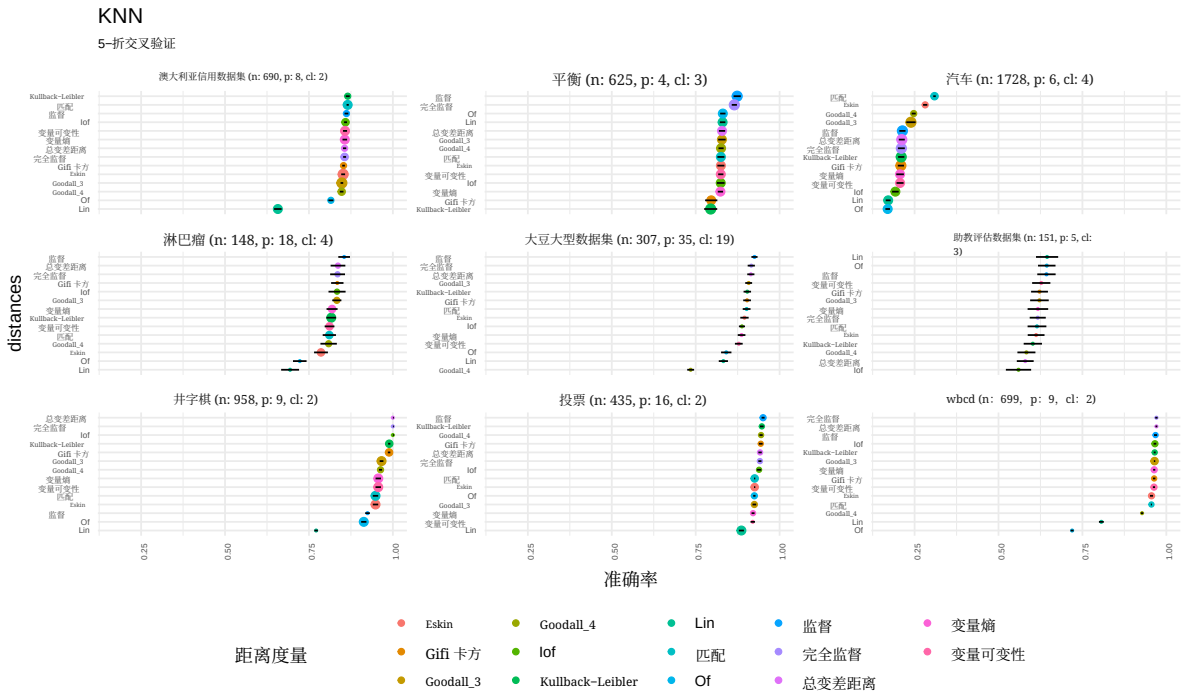

图 1：每种所考察的距离度量和数据集对应的 K近邻分类准确率 t

8.2 围绕中心点划分

围绕中心点划分（PAM）是一种迭代聚类过程，以一组观测值之间的成对距离矩阵作为输入。在一个簇中，中心点对应于中位数

观测值，就像 K 均值中的质心对应于均值一样。K 个中心点的初始集合是随机选取的，每个观测值都会被分配给距离最近的中心点；根据所得簇的分配结果，相应地更新中心点。当中心点集合不再发生变化时，迭代过程便停止。尽管我们是在无监督情境下进行分析，但可以通过基于有监督关联不相似度计算距离，并利用交叉验证评估 PAM 在每种数据集/类别不相似度组合上的表现；这也使其与基于 KNN 的应用保持一致。

具体而言，对于每个数据集和每种类别差异度定义，将 PAM 应用于训练数据，以获得一组中心点。也就是说，

$$\mathbf{D} = \mathbf{Z}_{train}\mathbf{\Delta}_{train}\mathbf{Z}'_{train}$$

其中，对于 $\mathbf{\Delta}_{train}$，我们考虑第 4 节和第 5 节中介绍并列于表 2 的所有类别不相似度矩阵。接下来，我们按如下方式计算测试观测值到中心点的距离矩阵，

$$\mathbf{Z}_{medoid}\mathbf{\Delta}_{train}\mathbf{Z}'_{test},$$

并将每个测试观测分配到与最近中心点对应的簇。

结果见图 2。我们观察到，除 汽车 数据集外（所有方法在该数据集上的表现都不佳），有监督的基于关联的距离总体表现良好。

总体而言，K近邻和 PAM 的结果得出以下结论。

- 在有监督设置下分类准确率较高的数据集，在无监督设置下也具有较高的调整兰德指数。

- 当方法的整体表现很差（例如，对于 汽车）或很好（例如，对于 *wbcd*）时，类别不相似度的选择似乎不会影响分类准确率。

- 在无监督设置中，基于关联的度量似乎更具优势：在 *wbcd* 中，ARI 值高于 0.75 的六种度量中，有五种是基于关联的。

扩展结果及用于复现这些结果的 R 代码可在线获取 [3]。

9 结论

本文提出了一个通用框架，用于以灵活、高效且透明的方式实现分类变量之间的距离。具体而言，我们展示了独立距离和基于关联的距离均可纳入该框架。因此，该框架既可用于实现若干现有度量，也能轻松引入高度可定制的新度量。

从理论视角来看，我们的方案颇具价值，因为它通过简化文献中使用的各种记号和定义，使人们能够评估不同不相似度度量之间的差异，从而使这些度量的实现更加透明。从应用角度来看，由于所提出的框架不局限于特定方法或应用，因此能够定义针对具体问题的不相似度度量。对此需要指出的是，在有监督情境下，我们的方案可用于定义考虑与响应变量之间关联的度量。

对于独立类别差异度，第 4 节中介绍的定义（有序类别差异度除外）已在 nomclust R 包（Šulc and ˇRezanková 2015）中实现。不过，该软件包未实现基于关联的度量。在 catdist 包 [4] 中，本文介绍的独立度量和基于关联的度量均已实现。

[3] https://alfonsoiodicede.github.io/blogposts_archive/distances_experiment_superv_unsuperv.html [4] 可在 GitHub 上获取，网址为 https://github.com/alfonsoIodiceDE/catdist_ 软件包，并且很快将在 CRAN 上重新发布。

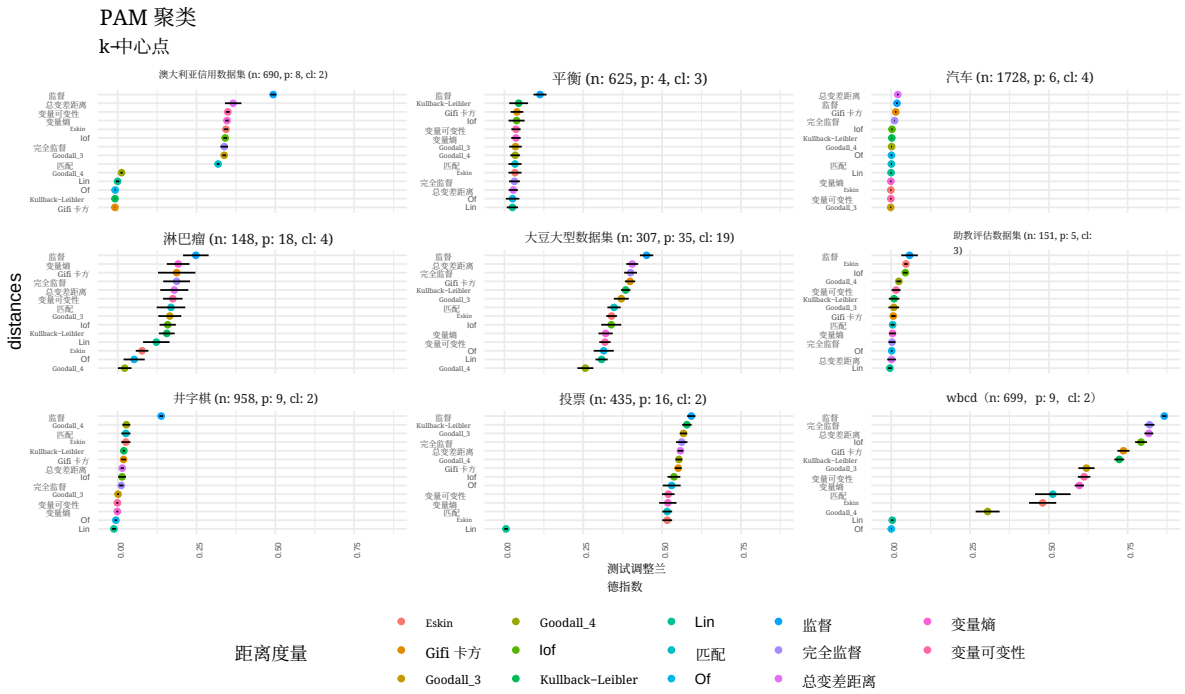

图 2：K-中心点的调整兰德指数结果

为说明针对具体问题选择“最佳”或“最合适”距离的重要性，我们在监督（通过 K近邻）和无监督（通过 PAM）场景中都使用了我们的框架。对真实世界数据集的应用表明，如果变量不具备判别能力、数据中不存在明显的簇结构，或者 K近邻/PAM 并非解决当前问题的合适方法，那么选择特定度量分别不会影响分类准确率或聚类性能。同样，也存在所有度量表现都同样出色的情况。撇开极端情形不谈，选择“最合适”的度量可以提升分类/聚类效果；在许多情况下，基于关联的度量优于独立类别度量。要评估这一论断，还需要开展一项更有系统性的研究，使用合成数据和更多实证数据集。这项研究超出了本文的范围。

借助本文提出的通用框架，可以轻松实现新的或定制的距离。事实上，第 6 节中介绍的监督全变差距离就是“新”的度量。不过，与其介绍和评估新的度量，不如系统地比较文献中已有的、适用于分类变量的不同不相似度度量各自的优缺点，这或许更有意义。

附录 A

这里我们证明，采用总方差定义类别差异度（如公式（10）所示），等价于 Ahmad 和 Dey（2007）提出的类别差异度定义。回顾一下，正如第 5.2.2 节所述，Ahmad 和 Dey（2007）将变量i的类别a 和b 之间的差异度定义为相对于变量j的类别分布的差异度，具体如下：

$$\delta(a,b) = \max_{\omega_j} \left(P(\omega_j|a) + P(\bar{\omega}_j|b) - 1 \right), \tag{14}$$

其中， ω_j 及其补集 $\bar{\omega}_j$ 相对于变量j的类别构成二元划分，而$P(\omega_j|a)$表示变量i中类别a 对应的观测比例，该类别对应于集合

类别的　　变量j的s，如 ω_j所定义。在第 5.2.2 节中，我们证明了公式（14）等价于　　t to

$$\delta(a,b)=\max_{\omega_j}\left|P(\omega_j|a)-P(\omega_j|b)\right|. \tag{15}$$

设 $\mathbf{K}^{q_j}$ 为一个设计矩阵，它定义了变量 j的q^j个类别的所有二元划分，但不包括空划分和完整划分。因此，$\mathbf{K}^{q_j}$是一个由 0 和 1 构成的$q_j\times q_j^\star$矩阵，其中 $q_j^\star=\sum^{q_j-1}{}_{l=1}\ \binom{q_j}{l}=2^{q_j}-2$。我们可以将公式（15）改写为

$$\delta(a,b)=\left\|\mathbf{K}'_{q_j}(\mathbf{r}_a-\mathbf{r}_b)\right\|_\infty=\left\|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\right\|_\infty,$$

其中 $\mathbf{d}_{ab}=(\mathbf{r}_a-\mathbf{r}_b)$，且 $\|\mathbf{x}\|_\infty$表示向量 $\mathbf{x}$的上确界范数，即$\mathbf{x}$ 中绝对值最大的元素。

$\mathbf{K}^{q_j}$ 的列数，也就是需要计算范数的向量的大小，会随着类别数呈指数增长。例如，当$q_j=4$时，$q_j^\star=14$；但当$q_j=8$ 时，我们有$q_j^\star=254$。考虑二元划分时，只需考虑这些组合中的一半。尽管如此，当分类变量的类别数不太少时，考虑所有组合在计算上会很昂贵。

使用以下关系，可以更高效地计算类别a 与b 之间的距离

$$\left\|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\right\|_\infty=\frac{1}{2}\|\mathbf{d}_{ab}\|_1, \tag{16}$$

其中，$\|\mathbf{x}\|_1$ 表示向量$\mathbf{x}$的L_1 范数，即

$$\|\mathbf{x}\|_1=\sum_i|x_i|.$$

要证明公式（16）成立，请注意，$\mathbf{d}_{ab}$ 各元素组合的绝对值最大值，可通过选取符号相同的元素组成组合得到。此外，由于$\mathbf{d}_{ab}$ 各元素之和为零，即:

$$\mathbf{d}_{ab}\mathbf{1}_{q_j}=(\mathbf{r}_a-\mathbf{r}_b)\mathbf{1}_{q_j}=0,$$

因此立　即得即正数元素之和等于负数值之和。因此，

$$\left\|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\right\|_\infty=\sum_{l:d_l>0}|d_l|=\sum_{l:d_l<0}|d_l|,$$

其中，d_l 表示$\mathbf{d}_{ab}$的第l个元素。最后，由于

$$\|\mathbf{d}_{ab}\|_1=\sum_l|d_l|=\sum_{l:d_l>0}|d_l|+\sum_{l:d_l<0}|d_l|$$

公式（16）中的等价关系可直接得出，因此我们可以将相对于变量j的类别，类别a 和b 之间的 Ahmad 和 Dey 距离表示为:

$$\delta(a,b)=\frac{1}{2}\|\mathbf{d}_{ab}\|_1=\frac{1}{2}\|\mathbf{d}_{ab}\|_1=\frac{1}{2}\sum_{l=1}^{q_j}|r_{al}-r_{bl}|. \tag{17}$$

比较公式 10 和 17 可知，Ahmad 和 Dey 距离等价于第 5.2 节中引入的全变差距离。

经费资助

作者　　研究、撰写和/或发表本文未获得任何经济资助　　.

利益冲突

作者声明不存在利益冲突。

数据可用性

扩展结果及其复现代码可在线获取：https://alfonsoiodicede. github.io/blogposts_ archive/distances_experiment_superv_unsuperv.html。本研究使用的数据下载自 UCI 仓库（Dua and Graff 2017），也可在 GitHub 上的 catdist 包中获取：https://github.com/alfonsoIodiceDE/ catdist_package。

参考文献

Ahmad, A. 和 L. Dey（2007年）。“一种用于混合数值型和类别型数据的 k 均值聚类算法”。载于：《数据与知识工程》63.2，第 503–527 页。Alves, G.、M. Couceiro 和 A. Napoli（2019年12月）。“类别数据聚类的相似性度量选择”。工作论文或预印本。网址：https://hal.archives-ouvertes.fr/hal- 02399640。Bai, L. 和 J. Liang（2022年）。“一种基于图表示的类别数据聚类框架”。载于：《模式识别》，第 108694 页。Borg, I. 和 P.J.F. Groenen（2005年）。《现代多维尺度分析：理论与应用》。施普林格科学与商业媒体。Boriah, S.、V. Chandola 和 V. Kumar（2008年）。“类别数据的相似性度量：比较评估”。载于：*2008*年 *SIAM* 国际数据挖掘会议论文集。SIAM，第 243–254 页。Cha, S.H.（2007年）。“概率密度函数之间距离/相似性度量的综合综述”。载于：《应用科学中的数学模型与方法国际期刊》1.4，第 300–307 页。Cover, T. 和 P. Hart （ 1967 年）。“ 最近邻模式分类 ”。载于：《 *IEEE* 信息论汇刊》13.1，第 21–27 页。Drost, H.G.（2018年）。“Philentropy：使用 R 进行信息论分析与距离量化”。载于：《开放源代码软件期刊》3.26，第 765 页。Dua, D. 和 C. Graff（2017年）。*UCI* 机器学习仓库。网址：http://archive.ics.uci. edu/ml。Eskin, E. 等（2002年）。“一种用于无监督异常检测的几何框架”。载于：数据挖掘在计算机安全中的应用。施普林格，第 77–101 页。Gabriel, K. R.（1971年）。“矩阵的双标图形显示及其在主成分分析中的应用”。载于：《生物计量学》58.3，第 453–467 页。Gifi, A.（1990年）。《非线性多变量分析》。约翰·威立父子有限公司。Goodall, D.W.（1966年）。“一种基于概率的新相似性指数”。载于：《生物统计学》，第 882–907 页。Gordon, A.D.（1999年）。《分类》。第 2 版。查普曼与霍尔/CRC。Gower, J. C.、S. G. Lubbe 和 N. J Le Roux（2011年）。《理解双标图》。约翰·威立父子。Hubert, L. 和 P. Arabie（1985年）。“比较划分”。载于：《分类学期刊》2.1，第 193–218 页。Ienco, D.、R.G. Pensa 和 R. Meo（2009年）。“用于类别数据聚类的基于上下文的距离学习”。载于：国际智能数据分析研讨会。施普林格，第 83–94 页。Jia, H.、Y. Cheung 和 J. Liu（2014年）。“一种用于类别数据无监督学习的新距离度量”。载于：《*IEEE* 神经网络与学习系统汇刊》27，第 1065–1079 页。Kaufman, L. 和 P.J. Rousseeuw（1990年）。《发现数据中的群组：聚类分析导论》。约翰·威立父子，纽约。Kullback, S.（1959年）。《信息论与统计学》。纽约：威立。

Kullback, S. 和 R.A. Leibler（1951年）。“论信息与充分性”。载于：《数学统计年刊》22.1，第 79–86 页。Le, S.Q. 和 T.B. Ho（2005年）。“一种用于类别数据的基于关联的不相似度量”。载于：《模式识别快报》26.16，第 2549–2557 页。ISSN：0167-8655。Lin, D.（1998年）。“相似性的理论信息定义”。载于：第十五届国际机器学习会议论文集。美国加利福尼亚州旧金山：Morgan Kaufmann 出版社，第 296–304 页。ISBN：1558605568。MacQueen, J.（1967年）。“多变量观测分类与分析的若干方法”。载于：第五届伯克利数学统计与概率研讨会论文集，第一卷：统计学。加利福尼亚州伯克利：加利福尼亚大学出版社，第 281–297 页。Mardia, K.V.（1978年）。“经典多维尺度分析的一些性质”。载于：《统计通讯：理论与方法》7.13，第 1233–1241 页。Metz, C. E.（1978年）。“ROC 分析的基本原理”。载于：《核医学研讨会》8.4，第 283–298 页。Ring, M. 等（2015年）。“ConDist：一种上下文驱动的分类距离度量”。载于：《机器学习与数据库知识发现》。A. Appice 等编。Cham：Springer国际出版公司，第 251–266 页。Spärck Jones, K.（1972年）。“词项特异性的统计解释及其在检索中的应用”。载于：《文献学杂志》28，第 11–21 页。Šulc, Z. 和 H. ˇRezanková（2015年）。“nomclust：一个用于对以名义变量为特征的对象进行层次聚类的 R 软件包”。载于：第九届国际统计与经济学日会议论文集。布拉格：Slaný：Melandrium，第 1581–1590 页。—（2019年）。“层次聚类中类别数据相似性度量的比较”。载于：《分类学期刊》36.1，第 58–72 页。