网络首发时间：2023-02-17 10:00:05
网络首发地址：https://kns.cnki.net/kcms/detail//41.1339.T.20230216.1329.005.html

文章编号：1671-6833(2023)02-0053-08

基于新的距离度量的异构属性数据子空间聚类

邓秀勤[1]，郑丽苹[1]，张逸群[2]，刘冬冬[1]

（1.广东工业大学 数学与统计学院，广东 广州 510520；2.广东工业大学 计算机学院，广东 广州 510006）

摘 要：真实数据集中往往包含分类属性和数值属性，其中分类属性可分为有序属性和标称属性，同时具有分类属性和数值属性的数据集可称为异构属性数据。针对现有异构属性数据距离度量不区分分类属性中的有序属性导致信息缺失、聚类效果不理想这一问题，提出了一种基于新的距离度量的异构属性数据子空间聚类算法。首先，总结了现有的异构属性数据距离度量的思路和区分有序属性的解决方案；其次，利用不同属性的数据特征分别定义了有序属性、标称属性和数值属性下的属性值之间的距离公式；再次，利用簇间差异和簇内距离这 2 个因素分别给出了不同属性在聚类过程中的动态加权方案；最后，联立距离公式和加权机制得到了可适用于异构属性数据的距离度量，进而设计了一种基于新的距离度量的异构属性数据子空间聚类算法。由于该算法既统一了异构属性数据的距离度量又能在子空间中进行簇搜索，因此该算法能在异构属性数据集上取得良好的聚类效果，在 11 个真实数据集上的对比实验结果验证了此算法的有效性。

关键词：异构属性数据；有序属性；距离度量；子空间聚类算法；动态权重

中图分类号：O235；TP311.13 **文献标志码**：A **doi**：10.13705/j.issn.1671-6833.2023.02.007

聚类[1-2]的根本目的是将对象按照某种相似或距离规则进行划分，使得簇内对象距离尽可能小，簇间对象之间距离尽可能大。因此，聚类的核心问题是衡量对象之间的距离或相似度。当对象是数值数据，常用欧氏距离和马氏距离等来衡量对象之间的距离；当对象是分类数据[3]，属性值是类别属性而不是数字，常用汉明距离及其变体[4]来度量。但这些距离度量只能单独运用于数值数据或者分类数据，不能用来衡量同时具有数值属性和分类属性的异构属性数据对象之间的距离[5]。由于异构属性数据难以处理，常被编码成纯数值数据或离散化成纯分类数据进行处理，但编码和离散化过程会在一定程度上曲解数据的原始信息，进而影响聚类结果。

因此，许多学者利用异构属性数据的数据特征来定义距离，Huang[6]分别运用汉明距离和欧式距离定义分类和数值属性值之间的距离，并用一个权重来平衡 2 种属性的差异。Huang 等[7]提出的自动加权的 k-Means 聚类算法（WKM），不再用一个权重来平衡 2 种属性的差异，而是用拉格朗日子乘子法得到最优权重来衡量属性之间关系。之后，Ienco 等[8]提出了基于上下文的分类数据距离度量；Jian 等[9]对非独立同分布分类数据提出了耦合相似度量（coupled metric similarity，CMS）。这些距离度量[10-11]都可以与 k-Prototypes 算法[6]结合，运用于异构属性数据中。但是，如图 1 所示，分类属性由有序属性和标称属性构成，这些距离度量将分类属性中的有序属性[10]当作标称属性对待，没有区分有序属性，导致有序属性的等级结构信息被忽略，难以在含有有序属性的数据集上取得理想的聚类效果。

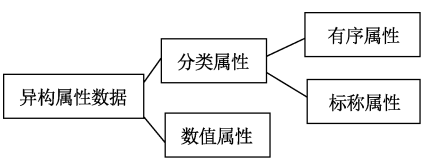

图 1 异构属性数据的属性构成

Figure 1 Attribute composition of heterogeneous-attribute data

收稿日期：2022-09-24；**修订日期**：2022-11-23

基金项目：国家自然科学基金资助项目（12101136）；广东省自然科学基金资助项目（2022A1515011592）；广东省研究生教育创新计划项目（2021SFKC030）

作者简介：邓秀勤（1966—），女，广东连州人，广东工业大学教授，主要从事机器学习、数据挖掘等研究，E-mail：dxq706@gdut.edu.cn。

引用本文：邓秀勤，郑丽苹，张逸群，等. 基于新的距离度量的异构属性数据子空间聚类[J]. 郑州大学学报（工学版），2023，44(2)：53-60.（DENG X Q，ZHENG L P，ZHANG Y Q，et al. Subspace clustering of heterogeneous-attribute data based on a new distance metric[J]. Journal of Zhengzhou University (Engineering Science)，2023，44(2)：53-60.）

将有序属性的取值按等级排列后用连续的整数 $1,2,\cdots,n$ 来编码,并利用数值属性的度量方法来度量属性值之间的距离,是一种最为常见的有序属性处理方法。然而数字编码的方式同样会丢失有序属性中的数据分布特征。近年来,Zhang 等[11-12]分别从信息熵和图论的角度定义了区分有序属性的分类数据统一距离度量(unified distance metric,UDM)和可学习加权的分类数据距离度量。虽然这些距离度量也可以与 k-Prototypes 等混合数据聚类算法结合,运用于含有数值属性的异构属性数据,但是它们并没有提出可以同时处理两种异构属性数据的办法。

与此同时,同一个属性对不同簇的重要性是不一样的,大部分的聚类算法平等地对待每一个属性,这在一定程度上具有不合理性。子空间聚类则较好地解决了这一类问题,它根据数据的子空间而不是整个数据空间将数据对象分组到集群中。例如 Cheung 等[13]利用簇内对象的信息熵来量化属性对簇的贡献;Jia 等[14]提出的属性权重 OCIL 算法(attribute-weighted OCIL algorithm,WOCIL)利用不同属性的聚类区分能力和簇内紧凑性来设定加权方案。但是,这些权重定义方案同样没有区分有序属性。

针对现有的异构属性数据距离度量没有区分分类属性中的有序属性而导致信息缺失、聚类效果不理想的问题,本文提出了一种新的异构属性数据距离度量,并基于此距离度量得到一种新的异构属性数据子空间聚类算法(subspace clustering algorithm for heterogeneous-attribute data,SCAH)。

1 相关工作

大部分数据集都包含分类属性和数值属性,其中分类属性可分为有序属性和标称属性,有序属性和标称属性的区别是有序属性的类别属性值之间具有等级关系。例如:学生成绩这一有序属性,它的属性值分别是优秀、良好、及格、不及格,很明显这些属性值之间具有等级关系。

现有的异构属性数据的距离定义和权重定义都是为标称属性和数值属性而设计,未能区分分类属性中有序属性的等级结构特点,Zhang 等[11-12]提出的分类数据距离度量,很好地解决了这个问题。其解决方案是:当给定同个属性下 2 个属性值 $a_{r,m}$ 和 $a_{r,h}$ 的距离公式为 $dist(a_{r,m},a_{r,h})$,如果这个属性为有序属性,且属性值的等级从高到低排列分别为 $a_{r,1},\cdots a_{r,m},\cdots,a_{r,h},\cdots,a_{r,v^r}$。为了突出有序属性的等级结构特征,$dist(a_{r,m},a_{r,h})$ 只需要转变为 $\sum_{s=\min(m,h)}^{g=\max(m,h)-1} dist(a_{r,s},a_{r,s+1})$ 即可。这个解决方案虽然仅在纯分类数据距离度量中运用,但是可以将其扩展到异构属性数据中。

2 本文方法

给定异构属性数据集 $\boldsymbol{X}$ 的 n 个对象和 d 个属性分别为 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 和 $A_1,A_2,\cdots,A_d$,对任意对象 $\boldsymbol{x}_m$ 可记为 $(x_m^1,x_m^2,\cdots,x_m^d)$,其中 x_m^i 表示第 m 个对象中 A_i 的属性值,$i=1,2,\cdots,d$,$m=1,2,\cdots,n$。d 个属性中前 d_c 个属性为分类属性,后 d_u 个属性为数值属性,$d=d_c+d_u$。其中分类属性前 d_o 个属性表示有序属性,后 d_n 个属性表示标称属性,$d_c=d_o+d_n$。分类属性 A_r 的属性值由 $\{a_{r,1},a_{r,2},\cdots,a_{r,v^r}\}$ 构成。当 A_r 为有序属性,其属性值之间等级关系从高到低可表示成:$a_{r,1}\succ a_{r,2}\succ\cdots\succ a_{r,v^r}$,符号“$\succ$”表示左边属性值等级高于右边属性值。若 n 个对象被划分成 k 个簇,分别用 $C_1,C_2,\cdots,C_k$ 表示,$\boldsymbol{X}$ 的每个对象都属于 k 个簇中的一个。对任意一个簇的簇中心记为 $(c_j^1,c_j^2,\cdots,c_j^d)$,其中,$c_j^m$ 表示第 j 个簇第 m 个属性的中心。当 $0<m\leqslant d_c$,c_j^m 是第 j 个簇中第 m 个属性中频率最高的属性值;当 $d_c<m\leqslant d_u$,c_j^m 是第 j 个簇中第 m 个属性的均值。

如表 1 所示,对象是 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_6$,其中 $\boldsymbol{x}_1=(\uparrow,\square,1.0)$。属性分别是 A_1,A_2,A_3,它们分别代表有序属性、标称属性和数值属性,且 A_1 的属性值由 $\{\uparrow,\updownarrow,\downarrow\}$ 构成,令属性值的等级关系为 $\uparrow\succ\updownarrow\succ\downarrow$,$A_2$ 的属性值由 $\{\square,\triangle\}$ 构成。

表 1 异构属性数据集实例

Table 1 Example of heterogeneous-attribute data

对象	A_1(有序)	A_2(标称)	A_3(数值)
$\boldsymbol{x}_1$	$\uparrow$	$\square$	1.0
$\boldsymbol{x}_2$	$\uparrow$	$\square$	0.2
$\boldsymbol{x}_3$	$\uparrow$	$\triangle$	0.4
$\boldsymbol{x}_4$	$\updownarrow$	$\square$	0.6
$\boldsymbol{x}_5$	$\updownarrow$	$\square$	0.8
$\boldsymbol{x}_6$	$\downarrow$	$\triangle$	0.7

聚类的目的是通过式(1)的最小化目标函数得到最优的划分 $\boldsymbol{Q}^*$:

$$\boldsymbol{Q}^*=\arg\min_{Q}\Big[\sum_{j=1}^{k}\sum_{i=1}^{n}q_{ij}dist(\boldsymbol{x}_i,C_j)\Big]。\quad(1)$$

其中 $\boldsymbol{Q}=q_{ij}$ 是一个 $n\times k$ 的矩阵,满足条件:$\sum_{j=1}^{k}q_{ij}=$

$1,0 < \sum_{i=1}^{n} q_{ij} < n$, 且 $q_{ij} \in \{0,1\}$, $i = 1,2,\cdots,n$, $j = 1,2,\cdots,k$。$dist(\boldsymbol{x}_i, C_j)$ 为 $\boldsymbol{x}_i$ 与 C_j 的距离, $\boldsymbol{x}_i$ 与 C_j 的距离相当于 $\boldsymbol{x}_i$ 的每一个属性值与 C_j 簇中心的距离和。由于分类数据的属性值代表样本的重要特征,数值数据更关注数据的整体效果[14],所以利用权重来衡量不同属性对簇的影响,则有

$$dist(\boldsymbol{x}_i, C_j) = \frac{d_c}{d_c + 1}\Big[\sum_{r=1}^{d_c} w_j^r dist^c(x_i^r, c_j^r)\Big] + \frac{1}{d_c + 1}\Big[\sum_{s=1}^{d_u} w_j^{d_c+s} dist^u(x_i^{d_c+s}, c_j^{d_c+s})\Big]。\tag{2}$$

式中: $dist^c(x_i^r, c_j^r)$ 表示分类属性中第 r 个属性值与 c_j^r 的距离公式; $dist^u(x_i^{d_c+s}, c_j^{d_c+s})$ 表示数值属性中第 s 个属性值与 $c_j^{d_c+s}$ 的距离。权重 $\boldsymbol{W} = w_j^r$ 为一个 $k \times d$ 的矩阵, $j = 1,2,\cdots,k$, $r = 1,2,\cdots,d$。式(2)中 w_j^r 表示 C_j 中第 r 个属性的影响, $w_j^{d_c+s}$ 表示 C_j 第 s 个数值属性的影响,且满足条件:

$$\sum_{r=1}^{d_c} w_j^r + \sum_{s=1}^{d_u} w_j^{d_c+s} = 1。\tag{3}$$

因此问题的关键有两个:①定义分类属性 2 个值之间的距离 $dist^c(x_i^r, c_j^r)$ 和数值属性 2 个值之间的距离 $dist^u(x_i^{d_c+s}, c_j^{d_c+s})$; ②定义权重 $\boldsymbol{W}$。对于这两个问题的解决方案分别在 2.1 和 2.2 节中进行介绍。

2.1　属性值之间的距离定义

首先定义分类属性下的距离 $dist^c(x_i^r, c_j^r)$, 若 $x_i^r = a_{r,m}$, $c_j^r = a_{r,h}$, 则有

$$dist^c(x_i^r, c_j^r) = dist^c(a_{r,m}, a_{r,h})。\tag{4}$$

因为 $dist^c(a_{r,m}, a_{r,h})$ 需要定义同一属性下 2 个属性的属性值之间的距离,而同一属性下的 2 个属性值之间的距离不仅与这个属性有关,还与其他分类属性有关系[8,11],也就是说 $a_{r,m}$ 和 $a_{r,h}$ 的距离不仅与 A_r 有关系,还与其他属性有关系。用 $\psi_{r,t}(a_{r,m}, a_{r,h})$ 表示 A_r 属性值 $a_{r,m}$ 和 $a_{r,h}$ 在 A_t 属性中的距离, $t = 1,2,\cdots,d_c$。当 $a_{r,m}$ 和 $a_{r,h}$ 相同时, $\psi_{r,t}(a_{r,m}, a_{r,h}) = 0$; 当 $a_{r,m}$ 和 $a_{r,h}$ 不相同时,用 $[p_1^{a_{r,m}}, p_2^{a_{r,m}}, \cdots, p_{v^t}^{a_{r,m}}]$ 和 $[p_1^{a_{r,h}}, p_2^{a_{r,h}}, \cdots, p_{v^t}^{a_{r,h}}]$ 分别表示 A_r 属性值为 $a_{r,m}$ 和 $a_{r,h}$ 的条件下,在 A_t 属性值为 $a_{t,1}, a_{t,2}, \cdots, a_{t,v^t}$ 的概率分布列,其中 $p_{v^t}^{a_{r,m}}$ 表示 A_r 的属性值为 $a_{r,m}$ 的条件下 A_t 属性值为 v^t 的概率。

当 A_t 为标称属性时,其距离可用其差异和 $\sum_{i=1}^{v^t} |p_i^{a_{r,m}} - p_i^{a_{r,h}}|$ 来表示,同时为了使距离的取值范围在 $[0,1]$,应除以 2,因此当 A_t 为标称属性时,这 2 个概率分布列的距离如定义 1 所示。

定义 1　若 A_t 为标称属性, $[p_1^{a_{r,m}}, p_2^{a_{r,m}}, \cdots, p_{v^t}^{a_{r,m}}]$ 与 $[p_1^{a_{r,h}}, p_2^{a_{r,h}}, \cdots, p_{v^t}^{a_{r,h}}]$ 这 2 个概率分布列的距离可定义为 $\sum_{i=1}^{v^t} |p_i^{a_{r,m}} - p_i^{a_{r,h}}| /2$。

当 A_t 为有序属性时,若用定义 1 的公式来定义距离,则 $[1,0,0,0]$ 和 $[0,1,0,0]$ 之间的距离与 $[1,0,0,0]$ 和 $[0,0,1,0]$ 之间的距离都等于 1,也就是说当这个概率分布列分别代表着有序属性值的优秀、良好、及格和不及格时,它意味着优秀与良好的差距和优秀与及格的差距是一样的,很明显定义 1 中的公式并不能用来衡量有序属性值之间的距离,所以文献[12]定义并解释了有序属性的 2 个概率分布列的距离,有序属性的概率分布列的距离如定义 2 所示。

定义 2　若 A_t 为有序属性, $[p_1^{a_{r,m}}, p_2^{a_{r,m}}, \cdots, p_{v^t}^{a_{r,m}}]$ 与 $[p_1^{a_{r,h}}, p_2^{a_{r,h}}, \cdots, p_{v^t}^{a_{r,h}}]$ 这 2 个概率分布列的距离可定义为 $\dfrac{\sum_{i=1}^{v^t-1} \left| \sum_{w=1}^{i} (p_w^{a_{r,m}} - p_w^{a_{r,h}}) \right|}{v^t - 1}$。

则 $\psi_{r,t}(a_{r,m}, a_{r,h})$ 可记为

$$\psi_{r,t}(a_{r,m}, a_{r,h}) = \begin{cases} \dfrac{\sum_{i=1}^{v^t-1} \left| \sum_{w=1}^{i} (p_w^{a_{r,m}} - p_w^{a_{r,h}}) \right|}{v^t - 1}, & a_{r,m} \neq a_{r,h}, t \leqslant d_o; \\ 0, & a_{r,m} = a_{r,h}; \\ \dfrac{\sum_{i=1}^{v^t} |p_i^{a_{r,m}} - p_i^{a_{r,h}}|}{2}, & a_{r,m} \neq a_{r,h}, t > d_o。 \end{cases}\tag{5}$$

$\psi_{r,t}(a_{r,m}, a_{r,h})$ 代表 A_r 的属性值 $a_{r,m}$ 和 $a_{r,h}$ 在 A_t 属性中的影响。对 $\psi_{r,t}(a_{r,m}, a_{r,h})$ 求和并求均值就是 A_r 的属性值 $a_{r,m}$ 和 $a_{r,h}$ 之间的距离。同时为了突出 A_r 为有序属性时的等级结构,利用相关工作中的解决方案,则 $dist^c(a_{r,m}, a_{r,h})$ 可表示为

$$dist^c(a_{r,m}, a_{r,h}) = \begin{cases} \dfrac{1}{d_c} \sum_{l=\min(t,h)}^{\max(t,h)-1} \sum_{t=1}^{d_c} \psi_{r,t}(a_{r,l}, a_{r,l+1}), & r \leqslant d_o; \\ \dfrac{1}{d_c} \sum_{t=1}^{d_c} \psi_{r,t}(a_{r,m}, a_{r,h}), & r > d_o。 \end{cases}\tag{6}$$

对于 $dist^u(x_i^{d_c+s}, c_j^{d_c+s})$ 可直接运用经典的欧氏距离来计算两个数值 $x_i^{d_c+s}$ 和 $c_j^{d_c+s}$ 的距离。则

$$dist^u(x_i^{d_c+s},c_j^{d_c+s})=\sqrt{(x_i^{d_c+s}-c_j^{d_c+s})^2}。\quad(7)$$

如表1数据集,A_1 属性值为“↑”的条件下 A_2 属性值分别为“□”和“△”的概率分布列为[2/3,1/3];A_1 属性值为“↕”的条件下 A_2 属性值分别为“□”和“△”的概率分布列为[1,0]。因 A_2 为标称属性,则由式(5)可得 $\psi_{1,2}(\uparrow,\updownarrow)=1/3,\psi_{1,1}(\uparrow,\updownarrow)=1/2$,同理有 $\psi_{2,1}(\square,\triangle)=1/4,\psi_{2,2}(\square,\triangle)=1$。则由式(6)可得 $dist^c(\uparrow,\updownarrow)=5/12,dist^c(\square,\triangle)=5/8$。

2.2 动态权重的定义

在同一个数据集中,一个属性对不同簇的重要程度可能不一样,不同的属性对同一个簇的重要程度也可能不一样。所以权重矩阵 $\boldsymbol{W}$ 需要研究第 r 个属性对簇 C_j 贡献,记为 H_j^r。在量化第 r 个属性对簇 C_j 贡献时,考虑簇间差异和簇内距离这2个因素[14]。令第 r 个属性对簇 C_j 的簇间差异为 F_j^r,它衡量了属性 A_r 区分簇 C_j 与其他簇的能力;令第 r 个属性对簇 C_j 的簇内距离为 M_j^r,它评估了属性 A_r 的簇 C_j 是否具有紧凑的结构。

首先测量簇间差异 F_j^r,令 $P_1(r,j)$ 和 $P_2(r,j)$ 分别为 A_r 在簇 C_j 内外的统计信息。

当 A_r 为分类属性,常用簇内外的概率分布列的距离来表示簇间差异,定义1和定义2分别区分了标称属性和有序属性的概率分布列的距离,所以定义1和定义2同样可以量化分类属性簇内外概率分布的差异。令 A_r 在簇 C_j 内外的概率分布列分别为 $P_1(r,j)=[p_1^1,p_2^1,\cdots,p_{v^r}^1],P_2(r,j)=[p_1^2,p_2^2,\cdots,p_{v^r}^2]$。用 $H^o(P_1(r,j),P_1(r,j))$ 表示 A_r 为有序属性时簇 C_j 的内外差异,则

$$H^o(P_1(r,j),P_1(r,j))=\frac{\sum_{i=1}^{v^r-1}\left|\sum_{w=1}^{i}(p_w^1-p_w^2)\right|}{v^r-1}。\quad(8)$$

用 $H^n(P_1(r,j),P_1(r,j))$ 表示 A_r 是标称属性时簇 C_j 的内外差异,则

$$H^n(P_1(r,j),P_1(r,j))=\frac{\sum_{i=1}^{v^r}|p_i^1-p_i^2|}{2}。\quad(9)$$

当 A_r 为数值属性,海林格距离[14-15]可以量化数值属性的簇间差异,令簇 C_j 内外的统计信息分别是 $P_1(r,j)\sim N(\mu_1,\sigma_1^2),P_2(r,j)\sim N(\mu_2,\sigma_2^2)$,其中 $\mu_1=\frac{1}{N_j}\sum_{\boldsymbol{x}_i\in C_j}\boldsymbol{x}_i^r,\mu_2=\frac{1}{N-N_j}\sum_{\boldsymbol{x}_i\notin C_j}\boldsymbol{x}_i^r$ 分别表示簇 C_j 内外 A_r 属性的均值。$\sigma_1^2=\frac{1}{N_j-1}\sum_{\boldsymbol{x}_i\in C_j}(x_i^r-\mu_1)^2$ 和 $\sigma_2^2=\frac{1}{N-N_j-1}\sum_{\boldsymbol{x}_i\notin C_j}(x_i^r-\mu_2)^2$ 分别表示簇 C_j 内外 A_r 属性的方差,用 $H^u(P_1(r,j),P_1(r,j))$ 表示当 A_r 为数值属性时簇 C_j 的内外差异,则

$$H^u(P_1(r,j),P_1(r,j))=\sqrt{1-\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\exp\left(-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right)}。\quad(10)$$

因此 F_j^r 可记为

$$F_j^r=\begin{cases}H^o(P_1(r,j),P_1(r,j)),0<r\leqslant d_o;\\H^n(P_1(r,j),P_1(r,j)),d_o<r\leqslant d_c;\\H^u(P_1(r,j),P_1(r,j)),d_c<r\leqslant d。\end{cases}\quad(11)$$

然后,属性 A_r 在簇 C_j 的簇内距离 M_j^r,可以通过属性 A_r 的簇 C_j 的平均对象簇距离来估计,平均对象簇距离越小,代表了这个簇划分效果的越好,则

$$M_j^r=\frac{1}{N_j}\sum_{\boldsymbol{x}_i\in C_j}dist(x_i^r,c_i^r)。\quad(12)$$

其中:

$$dist(x_i^r,c_i^r)=\begin{cases}dist^c(x_i^r,c_i^r),0<r\leqslant d_c;\\dist^u(x_i^r,c_i^r),d_c<r\leqslant d。\end{cases}\quad(13)$$

最后,对于 F_j^r 和 M_j^r,当 F_j^r 越大,M_j^r 越小时代表了第 r 个属性对簇 C_j 贡献越大;当 F_j^r 越小,M_j^r 越大时代表了第 r 个属性对簇 C_j 贡献越小,则

$$H_j^r=\frac{F_j^r}{M_j^r+1}。\quad(14)$$

式中:F_j^r 和 M_j^r 的取值范围都是[0,1],M_j^r+1 是为了防止分母等于0,所以 H_j^r 的取值范围为[0,1]。则属性权重 w_j^r 可定义为

$$w_j^r=\frac{H_j^r}{\sum_{s=1}^{d}H_j^s}。\quad(15)$$

如表1所示,若所有对象分为 C_1、C_2 两簇,分别为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$ 和 $\{\boldsymbol{x}_4,\boldsymbol{x}_5,\boldsymbol{x}_6\}$,则这两簇的簇中心分别为(↑,□,8/15)和(↕,□,7/10),A_1 在簇 C_1 内外的统计信息分别是 $P_1(1,1)=[1,0,0],P_2(1,1)=[0,2/3,1/3]$,则利用式(8)、(11)求得簇间差异 $F_1^1=2/3$,利用公式(12)求得簇内距离 $M_1^1=187/360$,并利用式(14)得属性 A_1 对簇 C_1 贡献 $H_1^1=240/547$。

2.3 SCAH算法

在2.1节中定义了 $dist^c(x_i^r,c_j^r)$ 和 $dist^u(x_i^{d_c+s},$

$c_j^{d_c+s}$) 距离公式,2.2 节中的权重定义利用了簇间差异 F_j^r 和簇内距离 M_j^r 这两个因素量化了属性对簇的贡献 H_j^r。将属性值之间的距离和权重联立后得到一种新的异构属性数据距离度量:

$$dist(\boldsymbol{x}_i, C_j) = \frac{d_c}{d_c+1}\Big[\sum_{r=1}^{d_c} w_j^r dist^c(x_i^r, c_j^r)\Big] + \frac{1}{d_c+1}\Big[\sum_{s=1}^{d_u} w_j^{d_c+s} dist^u(x_i^{d_c+s}, c_j^{d_c+s})\Big]。\quad (16)$$

其中,$dist^c(x_i^r, c_j^r)$ 和 $dist^u(x_i^{d_c+s}, c_j^{d_c+s})$ 如式(6)和(7)所示,权重 w_j^r 和 $w_j^{d_c+s}$ 如式(15)所示。

现讨论式(6)和式(16)的性质。对于同一属性的属性值 $a_{r,m}, a_{r,h}, a_{r,s}$,式(6)满足以下性质:

(1) $dist^c(a_{r,m}, a_{r,h}) \geqslant 0$;

(2)当 $a_{r,m} = a_{r,h}$ 时,$dist^c(a_{r,m}, a_{r,h}) = 0$;

(3) $dist^c(a_{r,m}, a_{r,h}) = dist^c(a_{r,h}, a_{r,m})$;

(4) $dist^c(a_{r,m}, a_{r,h}) \leqslant dist^c(a_{r,m}, a_{r,s}) + dist^c(a_{r,s}, a_{r,h})$。

对于任意数据对象 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$、$\boldsymbol{x}_t$,式(16)也满足以下性质:

(1) $dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \geqslant 0$;

(2)当 $\boldsymbol{x}_i = \boldsymbol{x}_j$ 时,$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$;

(3) $dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = dist(\boldsymbol{x}_j, \boldsymbol{x}_i)$;

(4) $dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant dist(\boldsymbol{x}_i, \boldsymbol{x}_t) + dist(\boldsymbol{x}_t, \boldsymbol{x}_j)$。

由于式(6)满足距离度量的条件,式(16)也满足距离度量的条件,因此式(16)中的 $dist(\boldsymbol{x}_i, C_j)$ 是一个距离度量。

最后利用式(16)提出了新的子空间聚类算法 SCAH。SCAH 的算法的迭代步骤如下。

步骤 1　输入数据集 $\boldsymbol{X}$ 和 k、d_o、d_c、d;

步骤 2　从 $\boldsymbol{X}$ 随机选取 k 个对象 $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_k$ 作为簇中心,初始化权重 $w_j^r = 1/d, r = 1,2,\cdots,d, j = 1,2,\cdots,k$;

步骤 3　令 $C_j = \varnothing, j = 1,2,\cdots,k$;

步骤 4　利用式(16)计算对象 $\boldsymbol{x}_i, i = 1,2,\cdots,n$ 与 $\boldsymbol{u}_j$ $(j = 1,2,\cdots,k)$ 的距离 $dist(\boldsymbol{x}_i, \boldsymbol{u}_j)$;

步骤 5　记 $d_{ih} = \min\limits_{j=1,2,\cdots,k} dist(\boldsymbol{x}_i, \boldsymbol{u}_j)$,$C_j = C_j \cup \{\boldsymbol{x}_i\}$;

步骤 6　重复步骤 4~5,直到所有样本划分完成;

步骤 7　更新 $C_1, C_2, \cdots, C_k$ 的簇中心;

步骤 8　利用式(8)~(15)更新权重 $\boldsymbol{W}$;

步骤 9　重复步骤 4~8,直到 $C_1, C_2, \cdots, C_k$ 不再改变,输出簇 $C_1, C_2, \cdots, C_k$。

3　实验部分

为了验证 SCAH 的有效性,使用了 11 个真实数据集,在 3 个聚类效果评估指标上与 6 个异构数据聚类算法进行比较,设计了 3 个实验来验证 SCAH 在异构属性数据集的有效性。

3.1　实验设置

11 个数据集都来自 UCI 机器学习数据库,其中包含 5 个纯分类数据、3 个纯数值数据和 3 个异构属性数据,所有数据集缺失的对象都被删除。d_o、d_n、d_u 分别表示有序属性、标称属性和数值属性的属性个数,n 表示数据集的对象个数。数据集的具体信息如表 2 所示。

表 2　数据集属性说明

Table 2　Statistics of the 11 data sets

数据集	d_o	d_n	d_u	n
Qualitative	6	0	0	250
PE	4	0	0	66
Blance	4	0	0	625
Tic	0	9	0	958
CEE	3	8	0	666
Wine	0	0	13	178
Glass	0	0	9	214
HCV	0	0	11	581
Heart	1	6	6	270
Ganpanes	0	9	6	647
Adult	0	8	6	30 162

分别利用 3 个聚类指标来衡量聚类的效果,它们是聚类精度 (CA),调整的兰德指数 (ARI) 和归一化信息 (NMI)。公式分别为

$$CA = \sum_{i=1}^{n} \delta(c_i, map(l_i))/n; \quad (17)$$

$$ARI = (RI - E(RI))/(\max(RI) - E(RI)); \quad (18)$$

$$NMI = \frac{\sum_{r=1}^{k}\sum_{t=1}^{k} c_{r,t}\lg(n \cdot c_{r,t})/(c_r \cdot c_t)}{(\sum_{r=1}^{k} c_r \lg c_r/N)(\sum_{t=1}^{k} c_t \lg c_t/N)}。\quad (19)$$

式中:c_i 表示第 i 个数据的真实标签;$map(l_i)$ 为映射到真实标签后的预测标签值,当 $c_i = map(l_i)$,$\delta(c_i, map(l_i)) = 1$,当 $c_i \neq map(l_i)$,$\delta(c_i, map(l_i)) = 0$;$E(RI)$、$\max(RI)$ 为 RI 的均值和最大值;k、c_r、c_t 和 $c_{r,t}$ 分别表示类标签的数量,预测标签值为 r 的数量,真实标签值为 t 的数量和同时满足预测标签值为 r,真实标签值为 t 的数量。NMI 从信息理论角度来衡量预测标签和真实标签的一致性。CA 和 NMI 的取值范围都为 $[0,1]$;ARI 的取值范围为 $[-1,1]$。当指标值越大,代表聚类的效果越好。

为了验证 SCAH 的聚类效果,将 SCAH 分别与

传统的聚类算法、经典的聚类算法和前沿的聚类算法进行比较。其中 WKM[7] 是异构属性数据聚类算法里比较常用的算法,可作为传统的聚类算法来比较;WOCIL[14] 是经典的异构属性数据子空间聚类算法;CMS[9] 和 UDM[11] 是较为前沿的分类数据距离度量,将它们分别与 k-Prototypes 算法和 WOCIL 算法结合得到的 KP+CMS 算法、KP+UDM 算法、WOCIL+CMS 算法和 WOCIL+UDM 算法,且 KP+UDM 和 WOCIL+UDM 是区分了有序属性的聚类算法。

实验一利用纯分类属性数据来验证 SCAH 在分类数据集中的有效性;实验二利用纯数值属性数据验证 SCAH 在数值数据集中的有效性;实验三利用异构属性数据集来验证 SCAH 在异构属性数据集中的有效性。为了保证实验的公平,这 6 个聚类算法都随机选取 k 个聚类中心,其中 k 是从数据集的标签中获取的,并重复 10 次后用聚类指标的平均值和标准差作为实验结果,用"平均值±标准差"来表示,其中效果最好的数值用黑体来表示。当 *ARI* 为负数时,用"-"来表示。由于 KP+CMS 和 KP+UDM 在纯数值数据中的结果是一样的,同时 WOCIL+ CMS 和 WOCIL+UDM 在纯数值数据中的结果与 WOCIL 是一样的,所以 KP+UDM、WOCIL+ CMS 和 WOCIL+UDM 的实验结果不在表 4 列出。具体实验结果如表 3~5 所示。

3.2 实验分析

实验一的结果如表 3 所示,在只有有序属性的分类数据集 Qualitative、PE 和 Blance 中,整体上来说 SCAH 的聚类效果是最好的,特别是与 WKM、WOCIL、KP+CMS 和 WOCIL+ CMS 的实验结果相比,SCAH 的效果非常显著,这表明了在分类数据中区分有序属性是有效的。在只有标称属性的分类数据集 Tic 中,SCAH 的聚类效果也较好,但是 CA 的结果没有 KP+CMS 好,这是因为 SCAH 的最大优势是衡量有序属性的距离。在 CEE 中,SCAH 的聚类效果仍是最好的。综上所述,SCAH 区分有序属性和标称属性的距离定义和权重定义在纯分类数据中是有效的。

实验二是验证 SCAH 在纯数值数据中的有效性,如表 4 所示。其中 WKM、KP+CMS 和 SCAH 的区别是权重的定义。SCAH 的聚类效果总体上来说是比较好的,这是因为 WKM 的最优权重容易陷入局部最优,并不能真正表示属性的重要性。SCAH 的效果比 WOCIL 好的原因是 SCAH 同时利用数值属性数据的信息和对象与簇内外之间的统计信息这 2 个角度来定义的距离,比 WOCIL 更能衡量数值属性值的真实距离。所以 SCAH 适用于纯数值属性的数据集。

实验三的结果如表 5 所示,在既有分类属性和数值属性的数据集中,整体上来说 SCAH 的聚类效果是最好的,这表明了 SCAH 区分有序属性、标称属性、数值属性的距离定义和动态权重的结合能更好地量化异构属性数据的距离,更能为异构数据集找到最佳的分区。

综上所述,实验结果表明了 SCAH 不仅适用于异构属性数据集,也适用于纯分类属性和纯数值属性的数据集。同时实验也表明:SCAH 中区分有序属性、标称属性、数值属性的距离定义和动态权重对于异构属性数据聚类的有效性。

表 3 不同算法在纯分类属性数据集上的聚类效果

Table 3 Experimental results of different heterogeneous-attribute data clustering algorithms on categorical data sets

评价指标	数据集	WKM	WOCIL	KP+CMS	KP+UDM	WOCIL+CMS	WOCIL+UDM	SCAH
CA	Qualitative	0.971±0.01	0.897±0.18	0.720±0.00	0.921±0.01	0.666±0.11	0.892±0.14	**0.996±0.00**
	PE	0.488±0.08	0.544±0.09	0.538±0.06	0.612±0.04	0.470±0.08	0.589±0.10	**0.630±0.07**
	Blance	0.447±0.05	0.468±0.06	0.472±0.03	0.486±0.09	0.475±0.04	**0.509±0.02**	0.501±0.04
	Tic	0.540±0.03	0.555±0.03	**0.648±0.01**	0.525±0.02	0.625±0.03	0.567±0.03	0.571±0.05
	CEE	0.307±0.01	0.303±0.02	**0.313±0.01**	0.310±0.02	0.312±0.01	0.310±0.02	0.312±0.02
ARI	Qualitative	0.886±0.02	0.743±0.40	0.184±0.00	0.707±0.02	0.140±0.21	0.682±0.35	**0.984±0.00**
	PE	0.069±0.08	0.121±0.09	0.076±0.05	0.225±0.07	0.032±0.05	0.207±0.13	**0.255±0.11**
	Blance	0.044±0.04	0.051±0.05	0.004±0.01	0.095±0.10	0.013±0.02	0.122±0.03	**0.129±0.03**
	Tic	0.008±0.01	0.012±0.01	-	0.004±0.01	-	0.020±0.02	**0.027±0.03**
	CEE	0.001±0.00	0.002±0.00	0.001±0.00	0.008±0.01	0.001±0.00	0.007±0.01	**0.009±0.01**
NMI	Qualitative	0.838±0.02	0.728±0.38	0.268±0.00	0.673±0.02	0.168±0.19	0.646±0.31	**0.966±0.00**
	PE	0.113±0.10	0.175±0.11	0.138±0.06	0.276±0.07	0.067±0.08	0.264±0.12	**0.290±0.09**
	Blance	0.037±0.03	0.050±0.03	0.006±0.02	0.086±0.09	0.020±0.03	0.108±0.03	**0.125±0.03**
	Tic	0.007±0.01	0.008±0.00	0.001±0.00	0.004±0.01	0.001±0.00	0.017±0.02	**0.020±0.02**
	CEE	0.009±0.00	0.013±0.01	0.007±0.01	0.016±0.01	0.008±0.01	0.016±0.01	**0.027±0.03**

表 4　不同算法在纯数值属性数据集上的聚类效果

Table 4　Experimental results of different heterogeneous-attribute data clustering algorithms on numerical data sets

评价指标	数据集	WKM	WOCIL	KP+CMS	SCAH
CA	Wine	0.683±0.05	0.618±0.07	0.666±0.07	**0.713±0.04**
	Glass	0.451±0.03	0.433±0.03	0.453±0.04	**0.474±0.05**
	HCV	0.439±0.06	0.483±0.08	0.388±0.04	**0.835±0.15**
ARI	Wine	**0.355±0.07**	0.225±0.03	0.268±0.05	0.318±0.07
	Glass	**0.210±0.05**	0.163±0.06	0.193±0.04	0.178±0.07
	HCV	0.110±0.03	0.122±0.04	0.098±0.01	**0.545±0.25**
NMI	Wine	0.340±0.06	0.271±0.02	0.314±0.06	**0.358±0.05**
	Glass	0.329±0.03	0.287±0.06	0.341±0.04	**0.347±0.06**
	HCV	0.230±0.03	0.231±0.05	0.243±0.01	**0.428±0.13**

表 5　不同算法在异构属性数据集上的聚类效果

Table 5　Experimental results of different heterogeneous-attribute data clustering algorithms on heterogeneous-attribute data sets

评价指标	数据集	WKM	WOCIL	KP+CMS	KP+UDM	WOCIL+CMS	WOCIL+UDM	SCAH
CA	Heart	0.624±0.03	0.762±0.07	0.635±0.05	0.739±0.10	0.669±0.06	0.714±0.09	**0.787±0.07**
	Ganpanes	0.618±0.05	0.596±0.09	0.630±0.05	0.673±0.13	0.600±0.04	0.593±0.11	**0.710±0.11**
	Adult	0.642±0.03	0.683±0.06	0.680±0.02	0.576±0.00	0.653±0.08	0.599±0.11	**0.714±0.01**
ARI	Heart	0.062±0.04	0.288±0.12	0.075±0.06	0.263±0.16	0.122±0.07	0.211±0.16	**0.345±0.11**
	Ganpanes	0.060±0.04	0.064±0.10	0.072±0.06	0.182±0.18	0.043±0.04	0.071±0.16	**0.221±0.15**
	Adult	0.013±0.06	0.146±0.06	0.126±0.03	0.017±0.00	0.050±0.08	0.042±0.10	**0.175±0.00**
NMI	Heart	0.053±0.03	0.220±0.09	0.103±0.04	0.206±0.13	0.148±0.08	0.166±0.12	**0.267±0.09**
	Ganpanes	0.086±0.05	0.056±0.07	0.155±0.04	0.143±0.14	0.072±0.06	0.071±0.14	**0.176±0.11**
	Adult	0.009±0.03	**0.156±0.04**	0.103±0.01	0.063±0.01	0.054±0.03	0.073±0.05	0.153±0.04

4　结论

本文基于新的距离度量提出了一种新的异构属性数据子空间聚类算法 SCAH。与现有的异构属性聚类算法相比,①SCAH 的属性值之间的距离定义和动态权重的定义都区分了有序属性的结构特征,更能充分地利用异构属性数据的统计信息;②将距离公式和动态权重联合后,使 SCAH 在聚类过程能更好地利用距离公式和动态权重协作搜索到更优的样本划分;③SCAH 不仅在异构属性数据中有优异的聚类效果,在纯分类数据和数值数据中也有优异的聚类效果,通过在 11 个种类各异的真实数据集上的对比实验,验证了 SCAH 的有效性。

参考文献:

[1] 姜鸣, 赵红宇, 刘学良. 一种基于聚类分析的自适应步态检测方法[J]. 郑州大学学报(工学版), 2017, 38(3): 63-67.
JIANG M, ZHAO H Y, LIU X L. An adaptive gait detection method based on clustering analysis[J]. Journal of Zhengzhou University (Engineering Science), 2017, 38(3): 63-67.

[2] 王军芬,刘培跃,董建彬,等. 用于分割无损检测图像的快速模糊 C 均值算法[J]. 郑州大学学报(工学版), 2022,43(6):42-48.
WANG J F, LIU P Y, DONG J B, et al. Fast fuzzy C means algorithm for segmentation of non-destructive testing image[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(6):42-48.

[3] AGRESTI A. An introduction to categorical data analysis[M]. New York:John Wiley & Sons, 2018.

[4] HAMMING R W. Error detecting and error correcting codes[J]. The Bell System Technical Journal, 1950, 29(2): 147-160.

[5] AHMAD A, KHAN S S. Survey of state-of-the-art mixed data clustering algorithms[J]. IEEE Access, 2019, 7: 31883-31902.

[6] HUANG Z X. Clustering large data sets with mixed numeric and categorical values[C]//Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York:Springer, 1997: 21-34.

[7] HUANG J Z, NG M K, RONG H Q, et al. Automated variable weighting in *k*-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[8] IENCO D, PENSA R, MEO R. From context to dis-

tance: learning dissimilarity for categorical data clustering [J]. ACM Transactions on Knowledge Discovery From Data, 2012, 6,(1): 1-25.

[9] JIAN S L, CAO L B, LU K, et al. Unsupervised coupled metric similarity for non-IID categorical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1810-1823.

[10] AGRESTI A. Analysis of ordinal categorical data[M]. Hoboken: Wiley, 2010.

[11] ZHANG Y Q, CHEUNG Y M. A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering [J]. IEEE Transactions on Cybernetics, 2022, 52(2): 758-771.

[12] ZHANG Y Q, CHEUNG Y M. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3560-3576.

[13] CHEUNG Y M, JIA H. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number[J]. Pattern Recognition, 2013, 46(8): 2228-2238.

[14] JIA H, CHEUNG Y M. Subspace clustering of categorical and numerical data with an unknown number of clusters [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3308-3325.

[15] OOSTERHOFF J, VAN ZWET W R. A note on contiguity and hellinger distance[EB/OL]. (2011-01-01)[2022-03-12]. https://doi.org/10.1007/978-1-4614-1314-1_6.

Subspace Clustering of Heterogeneous-attribute Data Based on a New Distance Metric

DENG Xiuqin[1], ZHENG Liping[1], ZHANG Yiqun[2], LIU Dongdong[1]

(1. School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou 510520, China; 2. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Real datasets often contain categorical and numerical attributes, and categorical attributes can be divided into ordinal and nominal attributes. Datasets with both categorical and numerical attributes can be called heterogeneous-attribute data. To solve the problem that the existing distance metrics of heterogeneous-attribute data can not distinguish ordinal attributes in the categorical attributes resulting in missing information and poor clustering effect, a new subspace clustering algorithm based on distance metric was proposed. Firstly, this study summarized the existing progress of distance metric of heterogeneous-attribute data and the solutions to distinguish ordinal attribute. Then the distance formulas were defined for the attribute values of ordinal, nominal, and numerical attributes from the perspective of their natural characteristics. Subsequently, a dynamic weighting scheme was proposed to weight different attributes according to their contributed inter-and intra-cluster distances during clustering. Finally, the distance formula and dynamic weighting scheme were combined to form the distance metric applicable to heterogeneous-attribute data, and a subspace clustering algorithm for heterogeneous-attribute data was thus proposed. Because the algorithm unified the distance metric of heterogeneous-attribute data and could search clusters in subspace, it could achieve good clustering effect on heterogeneous-attribute data. Experimental results on 11 real data sets showed the effectiveness of the algorithm.

Keywords: heterogeneous-attribute data; ordinal attribute; distance metric; subspace clustering algorithm; dynamic weighting