计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):175−181 [doi: 10.15888/j.cnki.csa.008498]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合可信度的 k_m-means 算法①

熊君竹, 何振峰

(福州大学 计算机与大数据学院, 福州 350108)
通信作者: 熊君竹, E-mail: xiongjunzhu@126.com

摘　要: 以 K-means 为代表的聚类算法被广泛地应用在许多领域, 但是 K-means 不能直接处理不完整数据集. k_m-means 是一种处理不完整数据集的聚类算法, 通过调整局部距离计算方式, 减少不完整数据对聚类过程的影响. 然而 k_m-means 初始化阶段选取的聚类中心存在较大的不可靠性, 容易陷入局部最优解. 针对此问题, 本文引入可信度, 提出了结合可信度的 k_m-means 聚类算法, 通过可信度调整距离计算, 增大初始化过程中选取聚类中心的可靠性, 提高聚类算法的准确度. 最后, 通过 UCI 和 UCR 数据集验证算法的有效性.

关键词: 不完整数据; 聚类中心; 可信度; 局部距离; K-means

引用格式: 熊君竹,何振峰.结合可信度的 k_m-means 算法.计算机系统应用,2022,31(6):175–181. http://www.c-s-a.org.cn/1003-3254/8498.html

Clustering Algorithm of k_m-means with Credibility

XIONG Jun-Zhu, HE Zhen-Feng

(College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China)

Abstract: The clustering algorithm represented by K-means is widely used in many fields, but K-means cannot directly deal with incomplete data. k_m-means is a clustering algorithm for processing incomplete data. It reduces the impact of incomplete data on the clustering process by adjusting the calculation method of partial distance. However, the centroids selected in the initialization stage of k_m-means are unreliable, resulting in local optimal solutions. For incomplete data, a clustering algorithm that combined credibility was proposed to solve this problem. The calculation of distance was adjusted by credibility to increase the reliability of cluster centroids in the initialization stage, improving the performance of clustering algorithm. Finally, the algorithm was verified by the experimental results from the UCI and UCR dataset.

Key words: incomplete data; cluster centroids; credibility; partial distance; K-means

聚类是一种无监督学习方法, 它基于给定的相似性评价措施将数据集划分成若干簇, 使得在同一个簇中的对象彼此具有高相似性, 处于不同簇中的对象具有低相似性[1]. 聚类被广泛应用在许多领域, 如机器学习、模式识别等[2]. 对于传统的聚类算法, 属性值的缺失导致部分实例之间的距离无法有效的度量, 所以不能直接应用到不完整的数据集上. 因此, 在不完整数据集上的聚类研究是非常有意义的.

目前针对不完整数据的处理, 主要采用以下方法: (1) 在聚类之前直接删除包含缺失值的实例, 尽管它很简单, 但是当缺失比例较小时, 删除是非常有效的方法[3]. (2) 对缺失的数据集进行填充, 通过填充的方式获取完整的数据集, 然后使用传统聚类算法进行聚类分析, 算法的性能往往受限于预估填充值的准确性[4]. (3) 在聚类过程中采用非填充的方式, 在聚类之前没有对缺失部分进行填充处理, 而是在不完整状态下减少缺失值对聚类过程的影响. 如利用数据包含的背景知识, 通过在实例之间添加少量“软约束”[5] 引导包含缺失值的实

① 基金项目: 福建省自然科学基金 (2018J01794)
收稿时间: 2021-08-13; 修改时间: 2021-09-13; 采用时间: 2021-09-29; csa 在线出版时间: 2022-05-26

例进行簇的划分, 减少数据填充过程中不可靠的填充值对聚类的影响.

针对不同的问题, 研究人员提出了多种结合不完整数据的聚类算法. 文献 [6] 提出一种删除策略, 当数据集缺失比例较小时, 一般小于 10%, 直接将包含缺失值的实例删除, 数据缺失比例较小时, 对于最终的聚类分析结果不会产生较大的影响. 文献 [7] 提出 K-Pod 算法, 采用迭代填充的方式处理不完整数据, 填充的过程中使用期望最大化算法 (expectation maximization) 预估缺失属性值, 每一次迭代过程中使用 K-means 对填充后的实例进行标记, 直到算法收敛, 但是迭代过程中重复使用 K-means 方法进行更新, 消耗了大量时间. 文献 [5] 提出 SLIM 算法, 在聚类过程中添加基于距离的成对约束, 引导包含缺失值的实例进行簇的划分. 文献 [8] 提出 k_m-means 算法, 该算法将缺失值的处理结合到聚类过程中, 通过调整局部距离计算方式, 减少实例中缺失值对目标函数的影响.

k_m-means 是对 K-means 的扩展, 仍然受到 K-means 算法固有缺陷的限制[9]. 该算法采用 K-means++[10] 的初始化方式获取 k 个聚类中心, 使得中心之间距离尽可能的大. 由于缺失值的存在, 导致聚类中心的可靠性存在较大的不确定性. 主要表现在两方面: 直接不确定性和间接不确定性. 初始化过程中选择的聚类中心, 在标记阶段引导簇的划分起到十分关键的作用, 所以增大初始化阶段选取聚类中心的可靠性研究具有重要意义.

可信度来源于 He 在 EKM 算法[11] 中, 描述成对约束在簇划分过程中的满足度, 即某个簇中, 实例的划分结果满足成对约束的个数占成对约束总个数的比例. EKM 算法中认为满足度高, 则该簇可信度高; 满足度低, 则该簇的可信度低. 因此, 受 He 工作的启发, 为了描述不完整数据集中实例之间距离的可信度, 结合实例的完整性, 本文将实例中属性值的完整性称为实例可信度, 实例中缺失值的比例越小, 则计算出的局部距离可信度的越高; 反之, 实例中缺失值比例越高, 则计算出的局部距离可信度的越低.

为了解决 k_m-means 初始化阶段选取聚类中心的可靠性问题, 本文在初始化过程中引入可信度, 通过可信度调整距离的计算, 减小选取聚类中心的直接不确定性, 增大初始化完成后选取聚类中心的可靠性, 使得聚类中心能够更好地引导标记阶段的簇划分. 本文第 1 节介绍不完整数据的缺失机制和符号定义, 第 2 节介绍 k_m-means 算法, 第 3 节介绍改进的 k_m-means 算法, 第 4 节先对初始化阶段选取聚类中心的可靠性问题进行实验与分析说明, 然后对改进后的 k_m-means 进行实验并分析结果, 第 5 节对所做的工作进行总结.

1 不完整数据集

1.1 不完整数据的缺失机制

不完整数据的缺失机制[12] 有 3 种不同的类型. 完全随机缺失, 是指缺失部分独立于本身, 与数据集的其他属性无关; 随机缺失, 是指缺失部分独立于本身, 与数据集中其他属性有关; 非随机不可忽略缺失, 是指缺失部分与本身有关, 与数据集中的其他属性也有关. 完全随机缺失和随机缺失被称作可忽略缺失, 目前处理不完整数据集主要针对可忽略缺失.

1.2 符号定义

现有文献中有许多关于缺失度 (missing rate) [12,13] 的定义. 本文采用属性值缺失度和实例缺失度, 从不同维度描述缺失程度. 属性值缺失度衡量数据集中缺失的属性值比例, 实例缺失度衡量含有缺失属性值的实例比例.

设数据集$D=\{X_1,X_2,\cdots,X_n\}\in R^{n\times p}$, n 是数据集 D 的规模, p 是属性个数. X_i的部分属性值可能会出现缺失, 用?表示缺失值. 假设$Y=\{Y_1,Y_2,\cdots,Y_n\}$是一个 p 维的缺省信息矩阵, Y_i的第 j 个属性$Y_{ij}=I(X_{ij}\text{ is recorded})$, X_{ij}是X_i的第 j 个属性, $I(\cdot)$是一个指示函数, 当自变量为真时为 1, 否则为 0.

定义 1. 属性值缺失度 (value missing rate, VMR). 设 D 中属性值出现缺失的个数为$n_v=\sum_{i=1}^{n}\sum_{j=1}^{p}(1-Y_{ij})$, 则数据集 D 的属性值缺失度为$VMR=n_v/np$.

定义 2. 实例缺失度 (instance missing rate, IMR). 设 D 中含有缺失属性值的实例个数为n_i, 则数据集 D 的实例缺失度为$IMR=n_i/n$.

在不完整数据集中, 属性值缺失度和实例缺失度描述数据集的整体缺失程度.

2 k_m-means 算法简介

k_m-means 算法又被称为结合不完整数据集处理的 K-means 类型算法, 在 Hartigan 等人[14] 提出的 K-means 算法框架基础上改进, 使其能够高效的结合缺失值的处理. 该算法的主要思想是: 通过修正实例与聚类中心之间的相似性度量方式, 将缺失值的处理结合到算法

中, 减少实例中的缺失值对目标函数的影响, 使得算法准确度有不错的提升.

给定数据集$D=\{X_1,X_2,\cdots,X_n\}\in R^{n\times p}$. 假设 K 是已知的, 算法的目标寻找划分集$C=\{C_1,C_2,\cdots,C_K\}$和聚类中心$\mu=\{\mu_1,\mu_2,\cdots,\mu_K\}$, 优化目标是使得每个实例到它所在聚类中心的距离总和最小, 优化目标函数如式 (1):

$$W_\kappa=\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{p}I(X_i\in C_k)Y_{ij}\left(X_{ij}-\hat{\mu}_{kj}\right)^2 \quad (1)$$

对于每一个划分集合 C, 聚类中心的更新如式 (2):

$$\hat{\mu}_{kj}=\frac{\sum_{i=1}^{n}I(X_i\in C_k)Y_{ij}X_{ij}}{\sum_{i=1}^{n}I(X_i\in C_k)Y_{ij}} \quad (2)$$

实例X_i和聚类中心$\hat{\mu}_k$之间的距离可表示为式 (3):

$$\delta_{i,C_k}^2=\|X_i-\hat{\mu}_{\mathrm{k}}\|^2=\sum_{j=1}^{p}Y_{ij}\left(X_{ij}-\hat{\mu}_{kj}\right)^2 \quad (3)$$

k_m-means 在不完整数据集上对实例标记时依赖于指标$\Delta_{k,i}^-$和$\Delta_{l,i}^+$, 分别表示将X_i从簇C_k中移除 WSS (within-cluster sum of squares) 的减少值和将X_i加入簇C_l中 WSS 的增加值, 其中$\Delta_{k,i}^-$和$\Delta_{l,i}^+$的计算如式 (4) 和式 (5) 所示:

$$\Delta_{k,i}^-=\sum_{j=1}^{p}\frac{n_{kj}}{n_{kj}-Y_{ij}}\delta_{i,C_k}^2 \quad (4)$$

$$\Delta_{l,i}^+=\sum_{j=1}^{p}\frac{n_{lj}}{n_{lj}+Y_{ij}}\delta_{i,C_l}^2 \quad (5)$$

其中, $n_{kj}=\sum_{i=1}^{n}I(X_i\in C_k)Y_{ij}$, n_{kj}表示簇C_k的所有实例中第 j 个属性值未缺失的个数. 同时 k_m-means 算法引入辅助向量$\xi_i^{(t)}$ (t 表示迭代次数), 表示实例X_i到所有聚类中心的最短距离时, 对应的簇索引, 具体计算如式 (6):

$$\xi_{\mathrm{i}}^{(t)}=\arg\min_{1\leqslant k\leqslant K}\delta_{i,C_k}^2 \quad (6)$$

k_m-means 类似于 K-means 的思想, 首先通过算法 2 初始化得到 K 个聚类中心, 使用式 (6) 对实例的所属簇进行初始化标记, 然后比较将X_i移除簇C_k和加入簇C_l后对W_κ的贡献度大小, 分别用$\Delta_{k,i}^-$和$\Delta_{l,i}^+$表示, 确定X_i应该归属于那个簇使得目标函数 (式 (1)) 最小, 每次迭代完 n 个实例后, 再通过式 (2) 更新聚类中心, 如此循环直到最后聚类中心不再改变, 算法终止. 其中, 活动集L表示上一次迭代过程发生变化的簇, 结合不完整数据集的聚类中心初始化过程本文将在算法 2 中讨论, k_m-means 算法框架如算法 1 所示.

算法 1. k_m-means 算法

输入: 数据集 D, 聚类数 K
输出: 聚类簇的划分$C=\{C_1,C_2,\cdots,C_K\}$, 聚类中心$\mu=\{\mu_1,\mu_2,\cdots,\mu_K\}$

1) 采用算法 2 初始化得到 K 个簇中心, $\{\hat{\mu}_k^{(0)};k=1,2,\cdots,K\}$
2) 采用式 (6) 初始化每一个实例的所属簇, $\xi^{(0)}=\{\xi_1^{(0)},\xi_2^{(0)},\cdots,\xi_n^{(0)}\}$
3) 初始化活动集$\mathcal{L}=\{1,2,\cdots,K\}$和迭代次数$t=0$
4) while $\mathcal{L}\neq\varnothing$ do
5) for each $X_i\in D$
6) 获取距离X_i最近的簇$k=\xi_i^{(t)}$, 采用式 (4) 计算$\Delta_{k,i}^-$. // t 表示迭代次数
7) if $k\in\mathcal{L}$
8) 计算$l=\arg\min_{b\neq k}\Delta_{b,i}^+$
9) else
10) 计算$l=\arg\min_{b\in L}\Delta_{b,i}^+$
11) if $\Delta_{l,i}^+<\Delta_{k,i}^-$
12) 将X_i从C_k移入C_l, 更新$\xi_i^{(t+1)}=l$
13) 采用式 (2) 更新聚类中心$\hat{\mu}_k^{(t+1)}$和$\hat{\mu}_l^{(t+1)}$, 并将k和l放入$\mathcal{L}$中
14) else
15) X_i仍属于C_k, 设置$\xi_i^{(t+1)}=\xi_i^{(t)}$
16) end for
17) 从$\mathcal{L}$中移除本轮迭代过程中没有更新的簇索引, 更新迭代次数$t=t+1$
18) end while

步骤 (8) 和步骤 (10) 中的$\Delta_{b,i}^+$采用式 (5) 计算得出. 下面我们将讨论结合不完整数据集的聚类中心初始化过程, 本文将 k_m-means 算法中初始化聚类中心的部分称作 KMIWMD (K-means++ initialization with missing data) 具体流程如算法 2 所示.

算法 2. KMIWMD 算法

输入: 数据集 D, 聚类数量 K
输出: K 个聚类中心

1) $\mu\leftarrow$随机选取$\hat{\mu}_1=X_i$作为第 1 个聚类中心
2) while $|\mu|<K$ do // $|\cdot|$表示集合的个数
3) 计算概率$pro_i=d_i^2/\sum_{y=1}^{n}d_y^2$, $i\in\{1,2,\cdots,n\}$
4) 通过概率pro_i选取第 k 个 (k=2, 3, $\cdots$, K) 聚类中心$\hat{\mu}_k$
5) $\mu\leftarrow\mu\cup\hat{\mu}_k$
6) end while

步骤 (3) 中$d_i^2=\min_{l=1,2,\cdots,k-1}\tilde{d}_{i,C_l}^2$, $\tilde{d}_{i,C_l}^2$通过式 (7) 计算. KMIWMD 初始化聚类中心的过程中使用 K-means++的框架, 在实例之间的距离计算中结合了缺失值的处理. 在结合缺失值的处理的过程中, 重新定义了初始化聚类中心过程中距离的计算方式, 如式 (7) 所示:

$$\tilde{d}_{i,C_k}^2=\tilde{\delta}_{i,C_k}^2=\delta_{i,C_k}^2\Big/\sum_{j=1}^{p}Y_{ij}Y_{kj} \quad (7)$$

在 KMIWMD 算法初始化完成后, 选取的聚类中

心存在不可靠性, 这种不可靠性具体表现在两方面: (1) 直接不确定性, 被选为簇中心的实例存在缺失值时, 它所处的空间位置并不确定, 在缺省值被填充为某些值时, 它不适合作为簇中心. (2) 间接不确定性, 在度量实例之间的距离时, 由于缺失值的存在, 可能出现原本距离较近的实例, 得出的距离较大, 导致相距较远的实例不会被选作下一个聚类中心, 如图 1 所示. 我们将其他包含缺失值的实例, 导致聚类中心的不确定性称作间接不确定性. 上述两种不确定性会导致聚类中心在标记阶段对簇的划分产生错误的引导, 容易陷入局部最优解.

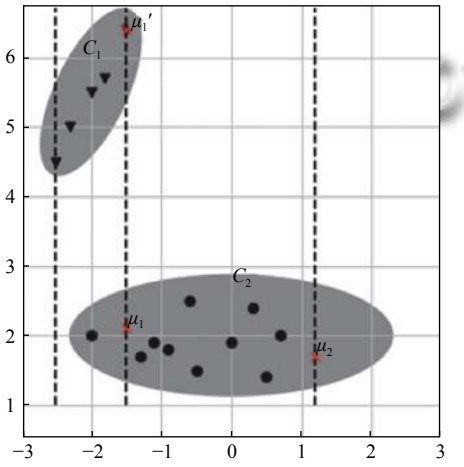

图 1 不可靠的聚类中心影响示意图

如图 1 所示, 存在两个大小不同的簇C_1和C_2, 假设使用 KMIWMD 初始化选择的第一个聚类中心μ_1. 当$\mu_1=(-1.5,?)$存在缺失值时, 由于缺失值的存在, 导致原本二维平面上的点与点之间的距离变成了点与直线之间的距离, 即其他实例到μ_1的距离变成了到直线 x=–1.5 的距离. 此时μ_2与直线 x=–1.5 距离最远, 由算法 2 中的步骤 (3) 计算得, 概率pro_2最大, 所以选择μ_2为下一个聚类中心. 由图 1 可知此时选择的两个聚类中心均在簇C_2中, 这会导致在接下来的实例标记过程中, 初始中心无法有效的引导簇的划分. 当$\mu_1=(-1.5,2.1)$不存在缺失值时, 距离μ_1最远的实例为μ_1', 通过计算可知pro_1'最大, μ_1'作为下一个聚类中心. 因为μ_1'和μ_1分布在不同的簇中, 所以和μ_2相比μ_1'是一个更好的聚类中心. 上述过程中, 我们将μ_1是否包含缺失值称作直接不确定性; 将μ_1存在缺失值时, 导致选取下一个聚类中心的发生变化称作间接不确定性. 针对上述存在的问题, 我们将在第 3 节讨论解决方案.

3 结合可信度的 k_m-means 聚类算法

针对 k_m-means 算法初始化 (KMIWMD 算法) 过程中, 选取聚类中心的不可靠性问题, 受 He 等人的工作启发[11,15], 本文在式 (7) 的计算过程中引入可信度, 如定义 3 和定义 4, 通过可信度优化不完整数据集的初始化过程, 减少初始化完成后, 多个聚类中心位于同一个簇中, 使得聚类中心能够更好地引导簇的划分. 结合式 (7) 和可信度, 推出新的结合缺失值处理的初始化方式, 式 (8) 为结合实例可信度的距离计算, 当实例X_i中存在缺失值, 即$IC_i<1$, 在计算实例与聚类中心之间的距离时, 实例可信度通过减少该实例被选做下一个聚类中心的概率pro_i, 尽量保证选取较完整的实例作为下一个聚类中心, 从而增大选取聚类中心可靠性. 式 (9) 为结合公共属性可信度的距离计算, 式 (9) 与公共属性可信度的结合方式和式 (8) 类似.

定义 3. 实例可信度 (instance credibility, IC). 设 D 中某个实例X_i的属性值未缺失的个数为$n_{in}=\sum_{j=1}^{p}Y_{ij}$, 则$X_i$的实例可信度为$IC=n_{in}/p$.

定义 4. 公共属性可信度 (public attribute credibility, PAC). 设实例X_i与实例X_k, 两实例中同一个属性均未缺失的属性值个数为$n_c=\sum_{j=1}^{p}Y_{ij}Y_{kj}$, 则$X_i$与$X_k$的公共属性可信度为$PAC=n_c/p$.

在不完整数据集中, 实例可信度描述单个实例的缺失程度, 实例可信度越小, 表示该实例包含缺失属性值越多, 实例越不可信. 公共属性可信度描述任意两个实例之间, 同一个属性均未缺失的属性值比例, 公共属性可信度越小, 表示两个实例之间公共缺失的属性值越多, 实例越不可信.

$$\tilde{d}^2_{i,C_k}=IC_i\frac{\delta^2_{i,C_k}}{\sum_{j=1}^{p}Y_{ij}Y_{kj}} \tag{8}$$

$$\tilde{d}^2_{i,C_k}=PAC_{i,k}\frac{\delta^2_{i,C_k}}{\sum_{j=1}^{p}Y_{ij}Y_{kj}} \tag{9}$$

式 (8) 中的IC_i表示X_i的实例可信度, 式 (9) 中的$PAC_{i,k}$表示X_i与C_k的公共属性可信度. 结合上文对算法 2 的分析与改进, 本文提出优化后的 KMIWMD 算法, 即结合可信度的不完整数据集聚类中心初始化算法, 记作 KMIWMD++, 具体的流程如算法 3 所示.

算法 3. KMIWMD++算法

输入: 数据集 D, 聚类数 K, 可信度阈值θ'
输出: K 个聚类中心

1) 采用定义 3 计算各个实例$IC=\{IC_1,IC_2,\cdots,IC_n\}$
2) $\mu\leftarrow$随机选取$\hat{\mu}_1=X_i$作为第一个聚类中心, 其中, $IC_i>\theta'$
3) while $|\mu|<K$ do // $|\cdot|$表示集合的个数
4) 使用式 (8) 计算$d_i^2=\min_{l=1,2,\cdots,k-1}d_{i,C_l}^2$
5) 计算概率$pro_i=d_i^2\Big/\sum_{y=1}^{n}d_y^2$, $i\in\{1,2,\cdots,n\}$
6) 通过概率pro_i选取第 k 个 (k=2, 3, …, K) 聚类中心$\hat{\mu}_k$
7) $\mu\leftarrow\mu\cup\hat{\mu}_k$
8) end while

步骤 (4) 中我们可以使用式 (8) 或式 (9) 结合不同的可信度, 调整初始化过程中的距离计算. KMIWND++算法确定下一个聚类中心的时, 在局部距离计算的过程中引入可信度 (实例可信度或公共属性可信度), 步骤 (2) 中通过实例可信度选择包含较完整信息的实例作为第一个聚类中心, 减小簇中心包含缺失值对第 k (k=2, 3, …, K) 个中心选取的影响. 步骤 (4)、步骤 (5) 通过实例可信度 (或者公共属性可信度) 调整距离计算, 减少包含缺失值的实例被选作下一个中心的概率, 尽可能地减少直接不确定性对中心选取的影响, 使得初始化完成后多个中心分布在不同的簇中. 最后, 本文将使用算法 3 初始化的算法 1 称作 KMMC (k_m-means with credibility), 接下来在第 4 节通过实验分析对比 k_m-means 和 KMMC 在结合不完整数据集的处理效果.

4 实验与分析

如表 1 所示, 实验阶段使用 7 个 UCI 数据集和 3 个 UCR 数据集. Seeds, Ceramic, Wine, Wdbc, CCBR, Iris 和 Column 来自于 UCI 数据集. Plane, CBF 和 Trace 来自于 UCR 数据集. Wdbc 表示的是 Breast Cancer Wisconsin (Diagnostic) 数据集, CCBR 表示 Cervical Cancer Behavior Risk 数据集, Ceramic 表示 Chemical Composition of Ceramic Samples 数据集, Column 表示的是 Vertebral Column 数据集. 每组数据都采用了 Z-Score 标准化.

本文对实验需要的缺失数据集进行如下处理, 构建随机缺失机制下的不完整数据集. 在完整数据集基础上, 分别构建不同实例缺失度 (*IMR*) 的不完整数据集, 构建过程中分别取 *IMR* 为 0、10% 和 20%. 首先通过随机数发生器在 n 个实例中随机选取 $miss=IMR\times n$ 个实例作为缺失部分, 然后通过随机数发生器依次在 $miss$ 个实例中随机选取 m 个属性, 将该属性对应的值设置为空值, 每个实例的随机种子为该实例在数据集中的序号, 最后将 $miss$ 个包含缺失值的部分和 $n-miss$ 个完整部分组合成实验所需的不完整数据集. 公共属性可信度存在为 0 的情况, 如$X_1=(?,?,6),X_2=(1,5,?)$, 此时$PAC_{1,2}=0$, 为了避免这种情况, 构建过程中保证每个实例中属性值缺失的个数 $m<p/2$. 针对不同的实例缺失度进行实验分析, 聚类准确性使用调整兰德里系数 ARI (adjusted rand index) 衡量, 实验结果为 5 折交叉验证的平均值. 通过实验发现算法 3 中可信度阈值θ'取 0.8 时, 初始化选择的聚类中心效果较好, 故本文实验中θ'统一取 0.8.

表 1 数据集信息

数据集	实例个数	属性个数	聚类数
Seeds	210	7	3
Ceramic	88	17	2
Wine	178	13	3
Wdbc	569	30	2
CCBR	72	19	2
Iris	150	4	3
Column	310	6	3
Trace	200	275	4
CBF	930	128	3
Plane	210	144	7

第一组实验, 如表 2 和表 3 所示, 通过实验分析 KMIWMD (算法 2) 和 KMIWMD++ (算法 3) 初始化完成后聚类中心包含缺失值的情况和初始完成后聚类中心出现在同一个簇中的情况, 其中算法 2 为 k_m-means 算法中的初始化聚类中心部分.

表 2 算法 2 与算法 3 初始化完成后聚类中心包含缺失值的情况

算法	实例缺失度 (%)	m_i	出现的次数
KMIWMD	10	m_1	239
		m_2	29
		m_3	2
	20	m_1	362
		m_2	59
		m_3	2
	30	m_1	390
		m_2	107
		m_3	4
KMIWMD++	10	m_1	**213**
		m_2	**13**
		m_3	1
	20	m_1	**317**
		m_2	**41**
		m_3	3
	30	m_1	**338**
		m_2	**54**
		m_3	2

表 3 算法 2 与算法 3 初始化聚类中心完成后聚类中心在同一个簇中的情况

算法	实例缺失度 (%)	s_i	出现的次数
KMIWMD	10	s_2	427
		s_3	3
	20	s_2	479
		s_3	9
	30	s_2	515
		s_3	13
KMIWMD++	10	s_2	**400**
		s_3	3
	20	s_2	**411**
		s_3	8
	30	s_2	**421**
		s_3	9

第 2 组实验, 如表 4 所示, 通过实验对比 k_m-means 算法结合不同的聚类中心初始化算法 (算法 2 和算法 3) 对聚类准确度的影响.

表 2 和表 3 分别表示对 Iris 数据集, 采用 KMIWMD 和 KMIWMD++算法进行 1 000 次实验, 初始化完成后选取的聚类中心包含缺失值的次数和聚类中心在同一个簇中的次数 (表 2 和表 3 中的数据统计在同一组实验中完成). 表 2 中的$m_i\ (i=1,2,\cdots,k)$表示选择 k 个聚类中心包含缺失值的次数, 如$m_1=239$表示 1 000 次实验中, 选取的 k 聚类中心里有 1 个聚类中心包含缺失值出现 239 次, 表 3 中$s_i\ (i=2,3,\cdots,k)$表示选择的 k 个聚类中心在同一个簇中的次数, 如$s_3=8$表示 1 000 次试验中, 选取的 k 聚类中心有 3 个聚类中心在同一个簇中出现了 8 次.

结合表 2 和表 3 中统计数据分析可知, 随着数据集中的实例缺失度增大, 选取的聚类中心包含缺失值的概率在升高, 同时, 每一次实验选取的聚类中心在同一个簇中的概率也在升高. 由算法 2 可知, 第一个聚类中心是随机从 n 个实例中选取的, 随着实例缺失度的增大, 选取的第一个聚类中心包含缺失值的可能自然在升高. 从式 (7) 计算实例与聚类中心之间的距离可知, 当聚类中心包含缺失值时, 式 (7) 得出的局部距离可能比真实的距离大. 如某一次选取的聚类中心点 $\mu_1=(6,?,8)$, 真实的聚类中心为$\mu_1'=(6,2,8)$, 使用公式 (7) 计算实例 $X_1=(2,4,3)$与μ_1之间局部距离$\tilde{d}^2_{1,C_1}=[(6-2)^2+0+(8-3)^2]/2=20.5$, 而实际中真实的距离为 $d^2_{1,C_1}=[(6-2)^2+(2-4)^2+(8-3)^2]/3=15$. 由于聚类中心中缺失值的存在, 增大了该实例被选做下一个聚类中心的可能, 而实际该实例距离上一个聚类中心较近, 属于同一簇内的可能大. 随着数据集的实例缺失度增大, 多个聚类中心在同一个簇的概率增大.

表 4 KMMC 和 k_m-means 算法 ARI 系数对比

数据集	k_m-means			KMMC		
	0	10%	20%	0%	10%	20%
Seeds	0.730	0.662	0.559	0.730/0.730	0.675/0.686	**0.583/0.561**
Ceramic	0.926	0.841	0.813	0.926/0.926	**0.896/0.852**	0.825/0.811
Wine	0.877	0.814	0.648	0.877/0.877	**0.834/0.829**	0.651/0.649
Wdbc	0.723	0.683	0.651	0.723/0.723	**0.717/0.690**	0.664/0.667
CCBR	0.461	0.378	0.269	0.461/0.461	**0.382**/0.377	0.306/**0.305**
Iris	0.769	0.754	0.671	0.769/0.769	0.763/0.773	**0.743/0.717**
Column	0.313	0.259	0.190	0.313/0.313	**0.294/0.291**	0.219/0.221
Trace	0.432	0.405	0.393	0.432/0.432	**0.422/0.420**	0.396/0.394
CBF	0.361	0.299	0.290	0.361/0.361	0.302/**0.336**	**0.291**/0.266
Plane	0.736	0.667	0.628	0.736/0.736	**0.684/0.680**	0.621/0.636

表 4 是 KMMC 与 k_m-means 对比实验结果, KMMC 算法中每一行两个值分别表示采用实例可信度和公共属性可信度优化的实验结果. 从实验数据分析可知, 当数据集不存在缺失值时, 通过定义 3 和定义 4 可知, 实例可信度$IC_i\equiv 1,\forall i\in(1,n)$, 公共属性可信度$PAC_{i,k}\equiv 1$, $\forall i,k\in(1,n),i\neq k$, 文中引入的可信度对于 KMMC 算法与 k_m-means 算法初始化聚类中心不存在影响, 所以表 4 中 KMMC 算法与 k_m-means 算法在完整数据集的情况下 ARI 值均相等. 在相同的缺失机制和实例缺失度的情况下, KMMC 的 ARI 值普遍要比 k_m-means 的值要高, 说明了减少初始化中心的直接不确定性, 即减少包含缺失属性值的实例作为初始样本中心, 可以提高结合缺失值处理的聚类算法性能. 但是在高维数据集 (CBF), 传统的距离度量方式难以准确找出实例之间的差异性, 导致聚类效果不佳; 同样, 在 Trace 数据集上, 当实例缺失度为 20% 时, 优化初始中对聚类结果几乎没有提升. 在高维数据集中随着实例缺失度增大, 通过优化初始化聚类中心, 无法有效的改善聚类准确度.

5 总结与展望

针对不完整数据集初始化聚类中心问题, 提出了结合可信度的不完整数据集聚类算法 KMMC, 将实例可信度和公共属性可信度运用到聚类中心初始化过程中, 减少实例中属性值的缺失对实例之间距离度量的影响, 增大初始化阶段选取聚类中心的可靠性. 通过可信度调整距离计算, 有效减少了簇划分过程中, 不可靠的聚类中心对实例标记阶段产生的错误引导. 最后, 通过 UCI 和 UCR 数据集对比 KMMC 与 k_m-means 算法的聚类准确度, 实验结果表明, 改进初始化聚类中心的 KMMC 算法的准确度优于 k_m-means 算法, 验证了 KMMC 算法的有效性. 未来工作将致力于如何在不完整数据集上引入成对约束, 引导聚类过程中的簇划分, 减少不完整数据对实例标记阶段的影响.

参考文献

1 Saxena A, Prasad M, Gupta A, *et al*. A review of clustering techniques and developments. Neurocomputing, 2017, 267: 664–681. [doi: 10.1016/j.neucom.2017.06.053]

2 Pérez-Suárez A, Martínez-Trinidad JF, Carrasco-Ochoa JA. A review of conceptual clustering algorithms. Artificial Intelligence Review, 2019, 52(2): 1267–1296. [doi: 10.1007/s10462-018-9627-1]

3 Poddar S, Jacob M. Clustering of data with missing entries using non-convex fusion penalties. IEEE Transactions on Signal Processing, 2019, 67(22): 5865–5880. [doi: 10.1109/TSP.2019.2944758]

4 Wei YH, Tang Y, McNicholas PD. Flexible high-dimensional unsupervised learning with missing data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(3): 610–621. [doi: 10.1109/TPAMI.2018.2885760]

5 Yang Y, Zhan DC, Wu YF, *et al*. Semi-supervised multi-modal clustering and classification with incomplete modalities. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(2): 682–695.

6 Strike K, Emam KE, Madhavji N. Software cost estimation with incomplete data. IEEE Transactions on Software Engineering, 2001, 27(10): 890–908. [doi: 10.1109/32.962560]

7 Chi JT, Chi EC, Baraniuk RG. *k*-POD: A method for *k*-means clustering of missing data. American Statistician, 2016, 70(1): 91–99. [doi: 10.1080/00031305.2015.1086685]

8 Lithio A, Maitra R. An efficient *k*-means-type algorithm for clustering datasets with incomplete records. Statistical Analysis and Data Mining, 2018, 11(6): 296–311. [doi: 10.1002/sam.11392]

9 Jain AK. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 2010, 31(8): 651–666. [doi: 10.1016/j.patrec.2009.09.011]

10 Brunsch T, Röglin H. A bad instance for *k*-means++. Theoretical Computer Science, 2013, 505: 19–26. [doi: 10.1016/j.tcs.2012.02.028]

11 He ZF. Evolutionary *k*-means with pair-wise constraints. Soft Computing, 2016, 20(1): 287–301. [doi: 10.1007/s00500-014-1503-6]

12 Santos MS, Pereira RC, Costa AF, *et al*. Generating synthetic missing data: A review by missing mechanism. IEEE Access, 2019, 7: 11651–11667. [doi: 10.1109/ACCESS.2019.2891360]

13 龚奇源, 杨明, 罗军舟. 面向缺失数据的数据匿名方法. 软件学报, 2013, 24(12): 2883–2896.

14 Hartigan JA, Wong MA. A *k*-means clustering algorithm. Journal of the Royal Statistical Society, 1979, 28(1): 100–108.

15 Zhou J, Wang QN, Hung CC, *et al*. Credibilistic clustering: The model and algorithms. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2015, 23(4): 545–564. [doi: 10.1142/S0218488515500245]