优化初始类中心的自适应 K-medoids 算法

刘金金

(河南师范大学 软件学院,河南 新乡 453007)

摘 要:针对传统的 K-medoids 聚类算法在聚类时需要随机选择初始类中心且指定聚类数目 K,及聚类结果不稳定的问题,提出了一种优化初始类中心的自适应 K-medoids 算法(adaptive K-medoids algorithm for optimizing initial class centers,CH_KD).其思想是定义了特征重要度,以此筛选出每一簇中最优的代表特征,组成特征子集,并重点研究了传统划分算法的自适应优化与改进.首先,利用特征标准差定义特征区分度,选择出区分度强的特征.其次,利用皮尔逊相关系数度量特征簇中每个特征的冗余度,选择出冗余度低的特征.最后,将特征区分度与特征冗余度之积作为特征重要度,以此筛选出每一簇中最优的代表特征,组成特征子集.实验将所提算法与其他聚类算法在 14 个 UCI 数据集上进行对比,结果验证了 CH_KD 算法的有效性与优势.

关键词:无监督;特征区分度;特征冗余度;CH 函数;特征选择

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2367(2025)01-0106-10

聚类算法是将数据集划分成不同的簇,目的是使同一个簇中的样本相异度较低,而不同簇间的样本相异度较高.对于处理大规模无标签数据,聚类算法[1]在数据挖掘领域占据了重要地位,其发展至今已有众多分支,主要分为 2 大类[2]:层次聚类算法和划分聚类算法.K-medoids 算法是其中一种划分聚类算法,由于其划分聚类结构清晰、时间效率高而得到了广泛的应用.此算法首先经过簇数量的选择,然后选取合适的初值,最后完成初始化过程后进行聚类.K-medoids 算法是基于 K-means 算法的一种改进算法.

K-medoids 聚类算法的优点是能够处理大型数据集,结果簇相当紧凑,且簇与簇之间分明清晰.但缺点是传统的 K-medoids 聚类算法随机选择初始类中心,而且需要人为指定聚类数目 K,导致聚类结果不稳定.由于聚类算法初始化对结果的影响非常大,所以现有方法大多是将其他算法与 K-medoids 结合使用,这样可以有效地提高 K-medoids 算法在聚类的准确率和效率,快速准确地找到最佳簇中心.

赵成[3]提出一种基于聚类和中心向量的快速 K 近邻分类算法.王全民等[4]将经典的果蝇优化算法与 K-medoids算法结合为一种新型的 K-medoids 算法,使得此新算法的聚类效果更好.魏霖静等[5]将 K-medoids算法与聚类簇思想结合起来,对每个聚类簇进行混合蛙跳算法优化.管雪婷等[6]提出一种优化萤火虫的 K-medoids 聚类算法且融合了云模型,可以有效地抑制 K-medoids 算法易陷入局部最优的问题.管雪婷[7]提出一种基于改进的萤火虫优化的 K-medoids 算法.杨楠[8]提出一种基于改进布谷鸟算法的 K 中心点聚类算法.刘叶等[9]将 K-means 算法与 K-medoids 算法相结合.李莲[10]提出了一种基于改进的人工蜂群的 K-medoids聚类算法.谭成兵等[11]使用布谷鸟优化的 K-medoids 算法进行聚类,通过多节点并行聚类的方式可以提高聚类效率.李欣宇等[12]将 K-medoids 算法与密度聚类算法的思想结合,减少算法执行的时间且提高聚类结果的准确度.但这些算法存在特征维度高与冗余度高的问题.

收稿日期:2023-08-22;**修回日期**:2023-11-15.

基金项目:国家自然科学基金 (62072159;U1804164;61902112).

作者简介(通信作者):刘金金(1985—),女,河南新乡人,河南师范大学讲师,研究方向为信息识别、数据处理,E-mail:hsdliujinjin@126.com.

引用本文:刘金金.优化初始类中心的自适应 K-medoids 算法[J].河南师范大学学报(自然科学版),2025,53(1):106-115.(Liu Jinjin.Adaptive K-medoids algorithm for optimizing initial class center[J].Journal of Henan Normal University(Natural Science Edition),2025,53(1):106-115.DOI:10.16366/j.cnki.1000-2367.2023.08.22.0001.)

针对 K-medoids 算法初始类中心的问题,可以利用类内误差平方和选取初始类中心的候选集[13],对 K-medoids的初始类中心进行优化.针对文献[14－15]中存在的 K 值问题,根据文献[16]中的方法,可利用基于中位数的轮廓系数或 CH 函数来确定合适的 K 值.

本文的算法思想如下.首先,设计了算法 1 和算法 2 来解决选取初始类中心候选集和 K-medoids 聚类算法最佳聚类数目 K 的问题,由此提出一种优化初始类中心的自适应 K-medoids 算法(adaptive K-medoids algorithm for optimizing the initial class center,CH_KD).然后,使用 CH_KD 算法将特征集划分簇.最后,特征选择根据定义的特征重要度来选取每个特征簇中最具有代表性的特征,组成最终的特征子集.

本文的特征选择是基于优化类中心的自适应 K-medoids 算法的无监督特征选择,其目的是降低维度,保留具有高分类信息且冗余度低的特征[17].利用优化类中心的自适应 K-medoids 算法对特征进行聚类,使相似(冗余)特征聚为一类,以便选出冗余度低且区分度强的特征子集.为了选出区分度强的特征,利用特征标准差定义特征区分度;为了选出冗余度低的特征,利用皮尔逊相关系数度量[18]特征簇中每个特征的冗余度;并以特征区分度和特征冗余度之积定义特征的重要度,以此选出每一簇中最优的代表特征,组成特征子集.

实验在 MATLAB R2016b 上进行了 14 个数据集上的对比,实验结果表明,本文所提的 CH_KD 算法与 K-medoids、K-means 以及 KCOIC 进行实验对比后发现,CH_KD 算法聚类结果最优;CH_KDFS 算法与其他算法相比,特征个数越少且 AUC(Aera under the curve)值高,表明分类效果好.

1　K-medoids 算法

K-medoids 算法是一种迭代重定位的算法,基本思想是:首先,从数据集中随机选择 K 个样本作为初始类中心.其次,按距离最近原则将其余的样本划分到离其最近的类中心所在的簇中.然后,从每个簇中选择使得类内误差平方和最小的样本作为新的类中心.最后,直到类中心不再变化或者达到指定的迭代次数时,算法结束.

给定样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$,则样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的欧氏距离

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\sum_{f=1}^{q}(x_{if}-x_{jf})^2}\,, \tag{1}$$

式中,q 为样本的特征数目,x_{if} 表示样本 $\boldsymbol{x}_i$ 在第 f 个特征上的取值.样本之间的距离越近,则表明 2 样本之间越相似,反之则越相异.

类内误差平方和

$$S_{EC}=\sum_{k=1}^{K}\sum_{x\in C_k}d^2(\boldsymbol{x},\boldsymbol{O}_k), \tag{2}$$

式中 $k=1,2,\cdots,K$,K 表示类别数目,$\boldsymbol{C}_k$ 表示第 k 类样本集合,$\boldsymbol{x}$ 为 $\boldsymbol{C}_k$ 的样本,$\boldsymbol{O}_k$ 表示 $\boldsymbol{C}_k$ 的类中心,样本 $\boldsymbol{x}$ 和 $\boldsymbol{O}_k$ 的相异度 $d(\boldsymbol{x},\boldsymbol{O}_k)$ 以欧氏距离度量.S_{EC} 的值越小则表示算法的聚类效果越好.

2　K-medoids 算法自适应优化

传统的 K-medoids 聚类算法随机选择初始类中心,而且需要人为指定聚类数目 K,但选择的初始类中心和聚类数目 K 决定着聚类的结果,所以导致 K-medoids 算法的聚类结果不稳定.针对 K-medoids 算法初始类中心的问题,受文献[13]的启发,可以利用类内误差平方和选取初始类中心的候选集,对 K-medoids 的类中心进行优化.针对 K-medoids 算法 K 值的问题,受文献[16]的启发,可利用基于中位数的轮廓系数或 CH 函数来确定合适的 K 值.由此,本文设计了一种优化初始类中心的自适应 K-medoids 算法(adaptive K-medoids algorithm for optimizing initial class centers,CH_KD).

假设给定的数据集为 $X=\{\boldsymbol{x}_i\mid \boldsymbol{x}_i\in\mathbf{R}^q,1\leq i\leq n\}$,即数据集中有 n 个样本,每个样本有 q 维特征,第 i 个样本的第 f 个特征值为 $\lambda_{i,f}$;欲将数据集 X 划分为 K 个簇 $\boldsymbol{C}_k$,$1\leq k\leq K$.

定义 1　数据集 X 的平均样本距离

$$d_M(X)=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i}^{n}d(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{3}$$

式中，$d_M(X)$ 由数据集 X 中任意 2 个样本间距离和的平均值计算得到.

定义 2 簇间相似度

$$S_C=\min(d(\boldsymbol{O}_i,\boldsymbol{O}_j)), \tag{4}$$

式中，$i=1,2,\cdots,n$，$j=1,2,\cdots,n$，$\boldsymbol{O}_i$ 和 $\boldsymbol{O}_j$ 分别表示簇 $\boldsymbol{C}_i$ 和簇 $\boldsymbol{C}_j$ 的类中心，可知簇间相似度由 2 个簇的类中心欧式距离计算得到.S_C 的值越小表示 2 簇之间的距离越近，则 2 簇的相似度越高，反之则越相异.

定义 3 样本 $\boldsymbol{x}_i$ 的误差平方和

$$S_{Ei}=\sum_{j=1,j\neq i}^{n} d^2(\boldsymbol{x}_j,\boldsymbol{x}_i), \tag{5}$$

式中，样本 $\boldsymbol{x}_i$ 的误差平方和根据其与其余样本间的欧氏距离平方和计算得到.

CH(Calinski-Harabasz)函数作为内部聚类效果的衡量标准之一，其原理是通过簇间方差和簇内方差来评估聚类效果，定义如下.

$$\mathrm{CH}(K)=\frac{\mathrm{tr}(B)/(K-1)}{\mathrm{tr}(W)/(n-K)}, \tag{6}$$

$$\mathrm{tr}(B)=\sum_{l=1}^{K} d^2(O_i,z), \tag{7}$$

$$\mathrm{tr}(W)=\sum_{i=1}^{K}\sum_{x_j\in C_i} d^2(x_j,z_j), \tag{8}$$

式中，z 表示数据集的平均值，z_j 表示 $\boldsymbol{C}_j$ 簇内所有样本的平均值.可知，$\mathrm{tr}(B)$ 表示簇间的离散程度，$\mathrm{tr}(W)$ 表示簇内的紧密程度.$\mathrm{CH}(K)$ 则体现了在聚类数目为 K 时聚类质量的好坏，CH 值越大则表明聚类效果越好.

本文提出的 CH_KD 算法主要包括 2 个部分：首先利用类内误差平方和选取初始类中心候选集，其次利用中位数的轮廓系数或者 CH 函数来确定合适的 K 值.为了更直观地说明改进的 2 种算法，图 1 给出算法的大致流程图.

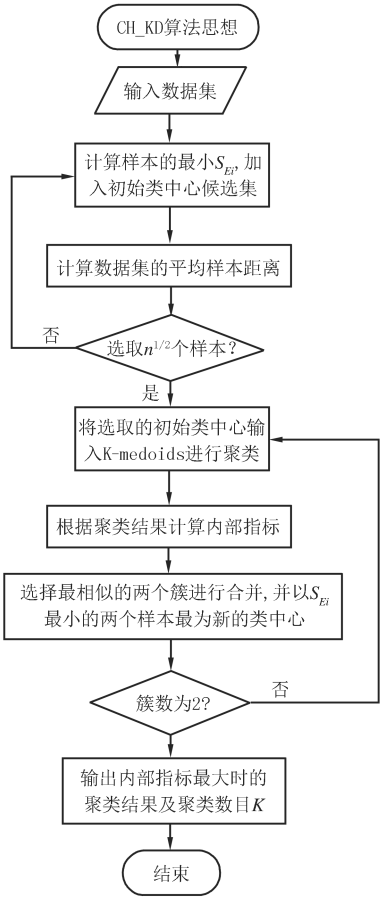

图1 CH_KD算法的流程图

Fig.1 CH_ The flowchart of KD algorithm

依据图 1 可知，为解决 K-medoids 聚类算法初始类中心的问题，设计算法 1.

算法 1 选取初始类中心候选集算法伪代码如下.

输入 数据集 X；

输出 $\mathrm{int}(\sqrt{n})$个初始类中心候选集.

1) 初始化类中心候选集矩阵 cen_id；

2) 根据式(1)计算样本间的欧式距离矩阵 dist_matrix；

3) for $i=1\cdots\mathrm{int}(\sqrt{n})$ do %选取 $\mathrm{int}(\sqrt{n})$个类中心；

4) 根据式(5)计算每个样本的 S_{Ei}；

5) 记 S_{Ei} 最小的样本索引为 pot；

6) cen_id(i)=pot；

7) 根据式(3)计算样本集的平均距离 d_M；

8) for $j=1\cdots n$ do；

9) if dist_matrix($\boldsymbol{x}_{\mathrm{pot}},\boldsymbol{x}_j$) $<d_M$；

10) 删除 $\boldsymbol{x}_i$ 样本 d_M 内的样本；

11) end if；

12) end for；

13) end for；

14) 返回 $\mathrm{int}(\sqrt{n})$个初始类中心的集合 cen_id.

为解决 K-medoids 聚类算法聚类数目 K 的问题，设计算法 2.

算法 2　确定最佳聚类数目算法的伪代码如下.

输入　int($\sqrt{n}$)个类中心候选集;

输出　最佳 K 值及其对应的聚类结果.

1)　初始化内部指标 CH_int =−1;

2)　for i=1…(int($\sqrt{n}$)−1) do %合并到 2 个簇为止;

3)　将选取的类中心输入 K-medoids 算法,获得聚类结果 cx 和类中心 cen_id;

4)　根据式(6)计算此状态下的内部评价指标的值,记为 CH;

5)　if CH > CH_int %记录内部评价指标最大时的聚类结果;

6)　CH_int = CH;

7)　record_cx = cx %记录聚类结果;

8)　ecord_cen_id = cen_id %记录类中心,即类数目 K;

9)　end if;

10)　根据式(4)计算簇间的相似度,查找最相似的 2 个簇;

11)　合并最相似的 2 个簇,并以 SEC 最小的样本作为新的类中心;

12)　end for;

13)　返回聚类结果和聚类数目 K.

算法 1 实现初始类中心的选取,其中步骤 4)～14)的时间复杂度为 $O(n^{1/2}(n+n+n))$,则算法 1 的总时间复杂度为 $O(n^{3/2})$.算法 2 实现聚类数目的确定,其中步骤 2)～12)的时间复杂度为 $O(n^{1/2}(tK(n-K)^2+n^{1/2}+n))$,其中 t 为迭代次数,K 为聚类数目,则算法 2 的时间复杂度可估计为 $O(n^{5/2})$.总体上来看,本问所提算法 CH_KD 的时间复杂度可估计为 $O(n^{5/2})$.

3　特征重要度确定

特征选择的目的是降低维度,保留具有高分类信息且冗余度低的特征.利用优化类中心的自适应 K-medoids 算法对特征进行聚类,使相似(冗余)特征聚为一类,以便选出冗余度低且区分度强的特征子集.为了选出区分度强的特征,利用特征标准差定义特征区分度,为了选出冗余度低的特征,利用皮尔逊相关系数度量特征簇中每个特征的冗余度[19],并以特征区分度和特征冗余度之积定义特征的重要度,以此选出每一簇中最优的代表特征,组成特征子集.

现给定数据集为 $X=\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathbf{R}^q, 1 \leq i \leq n\}$,$n$、$q$ 分别表示数据集的样本数量和特征数量.用 $\boldsymbol{f}_i=(f_{1i}, f_{2i}, \cdots, f_{mi})$ 表示第 i 个特征向量,则有 $X=\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_q\}$.

定义 4　特征 $\boldsymbol{f}_i$ 的区分度

$$d_{\mathrm{dis},i}=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(f_{ji}-\frac{1}{n}\sum_{j=1}^{n}f_{ji}\right)^2}, \tag{9}$$

式中,$i=1,2,\cdots,q$,f_{ji} 表示特征 $\boldsymbol{f}_i$ 在第 j 个样本上的取值.由于区分能力强的特征其方差较大,所以以特征的标准差来度量特征的区分度 $d_{\mathrm{dis},i}$ 是合理的.

定义 5　特征 $\boldsymbol{f}_i$ 的冗余度

$$d_{\mathrm{red},i}=\sum_{j\in C_i}(1-|r_{ij}|), \tag{10}$$

$$r_{ij}=\frac{\sum_{t=1}^{n}\left(f_{ti}-\frac{1}{n}\sum_{t=1}^{n}f_{ti}\right)\left(f_{tj}-\frac{1}{n}\sum_{t=1}^{n}f_{tj}\right)}{\sqrt{\sum_{t=1}^{n}\left(f_{ti}-\frac{1}{n}\sum_{t=1}^{n}f_{ti}\right)^2\sum_{t=1}^{n}\left(f_{tj}-\frac{1}{n}\sum_{t=1}^{n}f_{tj}\right)^2}}, \tag{11}$$

式中,$i=1,2,\cdots,q$,$\boldsymbol{C}_i$ 表示第 i 个特征簇,r_{ij} 表示特征 $\boldsymbol{f}_i$ 和特征 $\boldsymbol{f}_j$ 间的皮尔逊相关系数,r_{ij} 绝对值越小,则表示特征 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 之间越不相关.$d_{\mathrm{red},i}$ 表示特征 $\boldsymbol{f}_i$ 在特征簇 $\boldsymbol{C}_i$ 中的冗余度,$d_{\mathrm{red},i}$ 的值越大则表明 $\boldsymbol{f}_i$ 特征冗余度越低.

定义 6　特征 $\boldsymbol{f}_i$ 的重要度

$$d_{\mathrm{imp},i}=d_{\mathrm{dis},i}d_{\mathrm{red},i}. \tag{12}$$

由式(12)可知,特征 $\boldsymbol{f}_i$ 的重要度 $d_{\mathrm{imp},i}$ 定义为特征 $\boldsymbol{f}_i$ 的区分度与特征 $\boldsymbol{f}_i$ 的冗余度之积,特征 $\boldsymbol{f}_i$ 的重要度 $d_{\mathrm{imp},i}$ 的值越大,则表明特征 $\boldsymbol{f}_i$ 越重要.并且由此设计了基于优化类中心的自适应 K-medoids 算法的无监督特征选择(unsupervised feature selection of adaptive K-medoids algorithm based on optimal class center CH_KDFS),称算法 3.

算法 3 CH_KDFS 算法的伪代码如下.

输入 数据集 X;

输出 特征子集 F_c.

1) 初始化特征集 $F_s=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_q\}$,特征子集 $F_c=\mathrm{NULL}$;

2) 采用优化类中心的自适应 K-medoids 算法对 F_s 进行聚类,得到 K_f 个特征簇;

3) 根据式(9)计算每个特征的区分度;

4) for $i=1\cdots K_f$ do;

5) 根据式(10)计算第 i 个特征簇中每个特征的冗余度;

6) 根据式(12)计算第 i 个特征簇中每个特征的重要度;

7) 选出最优代表特征 $\boldsymbol{f}_{\mathrm{best}}$,加入特征子集 $F_c(i)=\boldsymbol{f}_{\mathrm{best}}$;

8) end for;

9) 输出特征子集.

算法 3 实现特征选择,步骤 2 的时间复杂度为 $O(q^{5/2})$,步骤 3 的时间复杂度为 $O(nq)$,步骤 4) ~ 7) 的时间复杂度为 $O(nq^2)$.则算法 3 总的时间复杂度可估计为 $O(q^{5/2})$.

4 实验结果与分析

4.1 实验准备

为验证所提聚类算法的聚类效果及其特征选择效果的好坏,在 15 个数据集上进行实验对比.表 1 给出了实验所使用的 15 个数据集的详细描述.其中前 10 个数据集为 UCI 数据集,UCI 数据库是用于机器学习的数据库.后 5 个数据集为 ASU 数据集(Academic Search Ultimate,综合学科参考类全文数据库).

使用准确率(Accuracy,ACC)、兰德系数(Rand Index,RI)、F-measure(F_1)、调整互信息(Adjusted Mutual Information,AMI)、归一化互信息(Normal Mututal Information,NMI)和调整兰德系数(Adjusted Rand Index,ARI)这 6 种指标对聚类结果进行评价,这 6 种评价指标均为值越大则表示算法性能越好.在 SVM、NB、和 KNN 分类器下,使用分类精度 precision 和 AUC 这 2 种指标对分类效果进行评价,分类精度和 AUC 的值越大,则表明分类效果越好.为避免特征量纲带来的影响,提高算法的准确率,使用文献[20]的方式对数据进行标准化.

实验结果中加粗的数值表示实验结果的最优值.实验的主要环境是 Windows10 64 位操作系统,处理器为 Inter(R)Core(TM)i5-8500 和 3.00 GHz 8.0 GB 内存;实验在 MATLAB R2016b 上进行.

表 1 15 个数据集的描述

Tab. 1 Description of 15 datasets

序号	数据集	样本数	属性数	类别数	数据来源	序号	数据集	样本数	属性数	类别数	数据来源
1	Seeds	210	7	3	UCI	9	DLBCL	77	5 469	2	UCI
2	Waveform21	5 000	21	3	UCI	10	Iris	150	4	3	UCI
3	Statlog	2 000	36	6	UCI	11	colon	62	2 000	2	ASU
4	Segmentation-test	2 310	19	7	UCI	12	lung	203	3 312	5	ASU
5	Heart	270	13	2	UCI	13	ORL	400	1 024	40	ASU
6	Zoo	101	16	7	UCI	14	COIL20	1 440	1 024	20	ASU
7	Soybean-small	47	35	4	UCI	15	SMK-CAN-187	187	19 993	2	ASU
8	Ionosphere	351	33	2	UCI						

4.2　CH_KD 算法聚类结果分析

将 CH_KD 算法与 K-medoids 聚类算法、K-means 聚类算法及 KCOIC 聚类算法进行实验对比,验证所提 CH_KD 聚类算法的有效性.本节实验在常用的 9 个 UCI 数据集 Seeds、Waveform21、Statlog、Segmentation-test、Heart、Zoo、Soybean-small、Ionosphere 和 Iris 上进行,为充分体现算法的有效性,采用 6 种评价指标 ACC、RI、F_1、AMI、NMI 和 ARI 对算法的聚类效果进行评价.表 2 给出 K 值自适应聚类算法 CH_KD 算法和 KCOIC 算法的聚类数目,K-means 算法和 K-medoids 算法的聚类数目与真实类别数相等.

由表 2 可知,在 Seeds、Waveform21、Segmentation-test、Heart、Ionosphere 和 Iris 这 6 个数据集上,CH_KD算法确定的类数目与真实类数目相等,与其他算法相比准确率更高.在 Zoo 和 Soybean-small 数据集上,CH_KD 算法确定的类数目最接近真实类数目.在 Statlog 数据集上,CH_KD 算法确定的类数目明显多于真实类数目.从本节实验所选的 9 个常用 UCI 数据集的实验结果来看,CH_KD 算法确定最佳聚类数目的性能优于 KCOIC 算法.

表 2　2 种自适应聚类算法的聚类数目

Tab. 2　Number of clusters for two adaptive clustering algorithms

数据集	真实类别数	KCOIC	CH_KD	数据集	真实类别数	KCOIC	CH_KD
Seeds	3	2	3	Zoo	7	4	5
Waveform21	3	3	3	Soybean-small	4	6	5
Statlog	6	2	14	Ionosphere	2	8	2
Segmentation-test	7	3	7	Iris	3	7	3
Heart	2	13	2				

注:下划线标出的数值与真实类别数相同.

由表 3 可知,在 Ionosphere 数据集上,CH_KD 算法获得了准确的聚类数目,但其聚类效果低于 K-medoids 算法,分析其原因可能是 Ionosphere 数据集簇间离散程度相差较大,导致未选到更为合理的类中心,使得聚类结果略差.在 Waveform21 数据集上,CH_KD 算法的聚类效果与 K-means 算法和 K-medoids 算法相差不多.在 Seeds、Statlog、Segmentation-test、Heart、Zoo、Soybean-small 和 Iris 这 7 个数据集上的聚类效果明显优于 K-means 算法和 K-medoids 算法.在 Waveform21 数据集上,4 种算法的聚类效果相差不大,且 CH_KD 算法的聚类效果优于 KCOIC 算法.特别是在 4 种聚类算法表现均良好的 Heart 数据集、Zoo 数据集和 Iris 数据集上,CH_KD 算法在 6 个指标上均取得了最优值.在 Heart 数据集上,CH_KD 算法在 ACC、RI、F_1、AMI、NMI 和 AMI 指标上分别高出其他对比算法 0.74%～20.55%、13.83%～15.6%、20.54%～40.96%、18.38%～23.89%、10.96%～24%、26.99%～31.23%.在 Zoo 数据集上,CH_KD 算法在 ACC、RI、F_1、AMI、NMI 和 AMI 指标上分别高出其他对比算法 4.95%～8.81%、1.7%～12.7%、3.76%～17.33%、5.6%～18.5%、1.56%～19.11%、4.14%～38.23%.在 Iris 数据集上,CH_KD 算法在 ACC、RI、F_1、AMI、NMI 和 AMI 指标上分别高出其他对比算法 0.96%～23.45%、4.98%～10.42%、14.12%～17.7%、3.58%～23.58%、5.02%～14.92%、10.29%～15.6%.

总体上来看,CH_KD 算法在 Seeds 数据集、Statlog 数据集、Segmentation-test 数据集、Heart 数据集、Zoo 数据集、Soybean-small 数据集和 Iris 数据集上的聚类结果最优.则有效地验证了所提算法的可行性及有效性.

4.3　CH_KDFS(CH_KD feature selection)算法的实验结果与分析

CH_KDFS 算法与 CMIM 算法、mRMR 算法、JMIM 算法、MRI 算法、DCSF 算法以及 MRMSR 算法在 SVM、NB 和 KNN($K=1$)这 3 个分类器下的分类精度如表 4 所示.分类器的各参数根据文献[21－22]进行设置.

由表 4 可知,在 lung 数据集上,CH_KDFS 算法在 SVM、NB 和 KNN 这 3 个分类器上的分类精度均最优.在 COIL20 数据集上,CH_KDFS 算法在 SVM 和 KNN 这 2 个分类器上的分类精度最优.在 SVM 分类器上高出其他 6 种算法 8.94%～13.33%,在 KNN 分类器上高出其他 6 种算法 3.33%～7.76%,但在 NB 分类器上的分类精度略低于部分算法.在 SMK-CAN-187 数据集上,CH_KDFS 算法的分类精度均略低于部分算

法.上述分析表明,CH_KDFS 算法易受分类器的影响,但整体来看 CH_KDFS 算法具有更好的分类性能.

表 3 4 种聚类算法在 9 个 UCI 数据集上的实验结果

Tab. 3 Experimental results of four clustering algorithms on 9 UCI datasets

数据集	算法	ACC	RI	F_1	AMI	NMI	ARI
Seeds	K-medoids	0.881 0	0.864 0	0.880 3	0.693 3	**0.700 3**	0.696 3
	K-means	0.891 1	0.871 5	0.890 6	**0.693 5**	0.698 4	0.710 7
	KCOIC	0.657 1	0.736 6	0.741 9	0.446 1	0.581 2	0.485 8
	CH_KD	**0.895 2**	**0.873 8**	**0.895 7**	0.685 4	0.698 2	**0.715 1**
Waveform21	K-medoids	**0.580 1**	**0.672 4**	**0.573 6**	0.350 0	0.350 0	**0.273 8**
	K-means	0.535 7	0.669 1	0.539 5	**0.364 8**	**0.364 8**	0.258 1
	KCOIC	0.525 2	0.666 4	0.530 7	0.361 6	0.361 6	0.251 1
	CH_KD	0.559 4	0.668 7	0.544 2	0.350 1	0.350 1	0.258 3
Statlog	K-medoids	0.648 3	0.816 2	0.641 1	0.513 0	0.522 1	0.405 2
	K-means	0.663 5	0.829 1	0.656 8	0.542 5	0.548 7	0.435 4
	KCOIC	0.335 5	0.364 2	0.417 7	0.162 2	0.372 4	0.089 7
	CH_KD	**0.806 5**	**0.874 3**	**0.735 7**	**0.557 8**	**0.618 9**	**0.544 3**
Segmentation-test	K-medoids	**0.569 1**	**0.831 2**	0.565 9	0.477 2	0.488 1	0.348 8
	K-means	0.553 9	0.823 9	0.568 4	0.491 1	0.516 1	0.354 7
	KCOIC	0.290 9	0.407 6	0.379 3	0.415 3	0.415 3	0.102 5
	CH_KD	0.563 6	0.817 1	**0.646 1**	**0.550 4**	**0.664 1**	**0.472 1**
Heart	K-medoids	0.587 1	0.514 0	0.583 5	0.017 8	0.020 8	0.027 5
	K-means	0.590 3	0.514 5	0.586 0	0.016 0	0.019 0	0.028 6
	KCOIC	0.785 2	0.531 7	0.381 8	0.071 1	0.149 4	0.069 9
	CH_KD	**0.792 6**	**0.670 0**	**0.791 4**	**0.254 9**	**0.259 0**	**0.339 8**
Zoo	K-medoids	0.793 1	0.839 7	0.699 5	0.607 7	0.669 2	0.527 5
	K-means	0.804 1	0.853 1	0.723 0	0.647 7	0.707 6	0.576 2
	KCOIC	0.831 7	0.949 7	0.835 2	0.736 7	0.844 7	0.868 4
	CH_KD	**0.881 2**	**0.966 7**	**0.872 8**	**0.792 7**	**0.860 3**	**0.909 8**
Soybean-small	K-medoids	0.753 2	0.805 3	0.739 2	0.626 4	0.676 9	0.507 0
	K-means	0.787 2	0.834 9	0.766 3	0.698 6	0.749 2	0.584 4
	KCOIC	0.829 8	0.815 0	0.719 7	0.500 2	0.635 8	0.433 9
	CH_KD	**0.957 4**	**0.913 0**	**0.898 9**	**0.764 6**	**0.855 2**	**0.746 8**
Ionosphere	K-medoids	0.696 2	0.575 5	0.704 9	0.105 0	0.114 7	0.140 5
	K-means	0.711 3	0.588 1	**0.716 8**	0.128 5	0.134 0	0.176 0
	KCOIC	**0.774 9**	**0.591 7**	0.674 6	**0.138 0**	**0.200 2**	**0.197 3**
	CH_KD	0.643 9	0.540 1	0.690 2	0.001 6	0.025 9	0.004 5
Iris	K-medoids	0.789 5	0.679 4	0.563 2	0.532 3	0.602 0	0.601 2
	K-means	0.670 5	0.658 2	0.533 2	0.650 6	0.670 9	0.612 0
	KCOIC	0.895 4	0.625 0	0.569 0	0.733 2	0.701 0	0.654 3
	CH_KD	**0.905 0**	**0.729 2**	**0.710 2**	**0.769 0**	**0.751 2**	**0.757 2**

CH_KDFS 算法、FSFC 算法和 WFSFC 算法在 5 个数据集 DLBCL、colon、lung、ORL 和 COIL20 数据集上所选特征个数,以及在 SVM、NB 这 2 个分类器下的 AUC 值如表 5 和表 6 所示.特征个数越少且 AUC 值越高,表明分类效果越好.实验中各分类器使用 5 折交叉验证.

表 4　7 种算法的分类精度

Tab. 4　Classification accuracy of 7 algorithms　　%

分类器	数据集	CMIM	mRMR	JMIM	MRI	DCSF	MRMSR	CH_KD
SVM	COIL20	69.07	72.04	71.02	69.78	68.37	72.76	**81.70**
	lung	82.99	86.35	85.64	83.97	84.07	87.67	**90.30**
	SMK-CAN-187	55.26	54.05	56.96	59.87	59.43	**60.33**	59.40
NB	COIL20	75.89	76.67	82.29	79.21	77.11	**82.64**	81.70
	lung	70.02	40.17	83.55	82.61	80.26	85.03	**87.00**
	SMK-CAN-187	59.85	52.06	54.82	52.42	51.17	**61.40**	59.00
KNN	COIL20	92.94	90.54	94.97	92.48	90.97	94.94	**98.30**
	lung	78.81	81.31	82.25	81.28	77.63	82.25	**85.90**
	SMK-CAN-187	59.67	58.84	61.89	57.80	55.77	**62.61**	58.80

表 5　3 种算法在 SVM 分类器下的特征选择个数和 AUC

Tab. 5　Number of feature selections and AUC of three algorithms under SVM classifier

数据集	FSFC		WFSFC		CH_KDFS	
	特征个数	AUC/%	特征个数	AUC/%	特征个数	AUC/%
DLBCL	64	**50.00**	**1**	**50.00**	15	**50.00**
colon	31	**50.00**	159	**50.00**	**6**	**50.00**
lung	180	83.60	328	81.8	**9**	**85.30**
ORL	**1**	51.80	98	64.50	13	**75.00**
COIL20	295	**90.30**	**2**	59.20	32	90.20

表 6　3 种算法在 NB 分类器下的特征选择个数和 AUC

Tab. 6　The number of feature selection and AUC of three algorithms under NB classifier

数据集	FSFC		WFSFC		CH_KDFS	
	特征个数	AUC/%	特征个数	AUC/%	特征个数	AUC/%
DLBCL	64	62.30	**1**	54.60	15	**84.80**
colon	31	50.30	159	77.30	**6**	**84.00**
lung	180	86.30	328	86.60	**9**	**94.30**
ORL	**1**	63.70	98	**97.20**	13	92.80
COIL20	295	**97.50**	**2**	84.50	32	**97.50**

由表 5 和表 6 可知,在 SVM 分类器上,CH_KDFS 算法在 DLBCL、colon、lung 和 ORL 数据集上均以较少的特征取得了最优的分类效果.特别是在 ORL 数据集上,CH_KDFS 算法分别比 FSFC、WFSFC 算法高出 23.2%、10.5%.在 COIL20 数据集上,CH_KDFS 算法能以较少的特征个数达到略低于对比算法的分类效果.在 NB 分类器上,CH_KDFS 算法在 colon 和 lung 数据集上均取得了最少的特征个数且分类效果最优.在 colon 数据集上分别比 FSFC、WFSFC 算法高出 33.7%、6.7%,在 lung 数据集上分别比 FSFC、WFSFC 算法高出 8.0%、7.7%.在 ORL 数据集上,尽管 CH_KDFS 算法的分类效果低于 WFSFC 算法,但所选的特征个数远少于 WFSFC 算法.综上所述,CH_KDFS 算法在特征选择的个数以及分类效果上优于对比算法.

5　结　论

针对传统 K-medoids 算法的缺点,本文提出了 CH_KD 算法,其目的是优化 K-medoids 算法的随机选择初始类中心且需要人为指定聚类数目 K 而导致聚类结果不稳定的问题.此算法为自适应优化初始类中心的 K-medoids 算法,利用特征标准差定义特征区分度,利用皮尔逊相关系数度量特征簇中每个特征的冗余度,

将特征区分度和特征冗余度的乘积定义为特征的重要度,以此选出每一簇中最优代表特征,组成特征子集.在 MATLAB R2016b 实验表明,在 8 个常用 UCI 数据集上,CH_KD 算法确定最佳聚类数目的性能优于 KCOIC 算法.在 Seeds 数据集、Statlog 数据集、Segmentation-test 数据集、Heart 数据集、Zoo 数据集和 Soybean-small 数据集上 CH_KD 算法的聚类结果最优.CH_KDFS 算法与 FSFC 算法和 WFSFC 算法相对比,在 DLBCL、colon、lung、ORL 和 COIL20 这 5 个数据集上所选特征个数,以及在 SVM、NB 这 2 个分类器下的 AUC 值相比,特征个数越少且 AUC 值越高,表明 CH_KDFS 算法分类效果越好,即 CH_KDFS 算法在特征选择的个数以及分类效果上优于对比算法.综上所述,CH_KD 算法具有可行性和有效性,实验聚类效果较好,且算法分类效果较好.

参 考 文 献

[1] LI H L.Multivariate time series clustering based on common principal component analysis[J].Neurocomputing,2019,349:239-247.

[2] 黄晓辉,王成,熊李艳,等.一种集成族内和族间距离的加权 k-means 聚类方法[J].计算机学报,2019,42(12):2836-2848.
HUANG X H,WANG C,XIONG L Y,et al.A Weighting k-Means Clustering Approach by Integrating Intra-Cluster and Inter-Cluster Distances[J].Chinese Journal of Computers,2019,42(12):2836-2848.

[3] 赵成.基于萤火虫算法和改进 K 近邻的文本分类研究[D].重庆:重庆邮电大学,2020.
ZHAO C.Research on text classification based on firefly algorithm and improved K nearest neighbor[D].Chongqing:Chongqing University of Posts and Telecommunications,2020.

[4] 王全民,杨晶,张帅帅.一种基于改进果蝇优化的 K-mediods 聚类算法[J].计算机技术与发展,2018,28(12):17-22.
WANG Q M,YANG J,ZHANG S S.A new K-mediods clustering algorithm based on improved fruit fly optimization algorithm[J].Computer Technology and Development,2018,28(12):17-22.

[5] 魏霖静,宁璐璐,郭斌,等.基于混合蛙跳算法的 K-mediods 聚类挖掘与并行优化[J].计算机科学,2020,47(10):126-129.
WEI L J,NING L L,GUO B,et al.K-mediods cluster mining and parallel optimization based on shuffled frog leaping algorithm[J].Computer Science,2020,47(10):126-129.

[6] 管雪婷,石鸿雁.融合云模型优化萤火虫的 K-mediods 聚类算法[J].统计与决策,2021,37(5):34-39.
GUAN X T,SHI H Y.K-mediods clustering algorithm of glowworm swarm optimization combined with cloud model[J].Statistics & Decision,2021,37(5):34-39.

[7] 管雪婷.基于改进的萤火虫优化的 K 中心点算法[D].沈阳:沈阳工业大学,2021.
GUAN X T.K-center algorithm based on improved firefly optimization[D].Shenyang:Shenyang University of Technology,2021.

[8] 杨楠.基于改进布谷鸟算法的 K 中心点聚类分析及并行实现[D].兰州:西北师范大学,2018.
YANG N.K-center clustering analysis and parallel implementation based on improved cuckoo algorithm[D].Lanzhou:Northwest Normal University,2018.

[9] 刘叶,吴晟,周海河,等.基于 K-means 聚类算法优化方法的研究[J].信息技术,2019,43(1):66-70.
LIU Y,WU S,ZHOU H H,et al.Research on optimization method based on K-means clustering algorithm[J].Information Technology,2019,43(1):66-70.

[10] 李莲.基于蜂群和粗糙集的聚类算法研究[D].长沙:长沙理工大学,2014.
LI L.Research on clustering algorithm based on bee colony and rough set[D].Changsha:Changsha University of Science & Technology,2014.

[11] 谭成兵,刘源,徐健.基于布谷鸟算法的 K-medoids 聚类挖掘与并行优化[J].台州学院学报,2021,43(03):7-12.
TAN C B,LIU Y,XU J,et al.K-medoids clustering mining and parallel optimization based on the cuckoo algorithm[J].Journal of Taizhou University,2021-43(03):7-12.

[12] 李欣宇,傅彦.改进型的 K-mediods 算法[J].成都信息工程学院学报,2006,21(4):532-534.
LI X Y,FU Y.Improved K-mediods algorithm[J].Journal of Chengdu University of Information Technology,2006,21(4):532-534.

[13] 钟志峰,李明辉,张艳.机器学习中自适应 k 值的 k 均值算法改进[J].计算机工程与设计,2021,42(1):136-141.
ZHONG Z F,LI M H,ZHANG Y.Improved k-means clustering algorithm for adaptive k value in machine learning[J].Computer Engineering and Design,2021,42(1):136-141.

[14] 陈江勇.基于平衡性的无监督特征选择算法研究[D].合肥:安徽大学,2021.
CHEN J Y.Research on unsupervised feature selection algorithm based on balance[D].Hefei:Anhui University,2021.

[15] 裴华欣.自适应密度峰划分聚类算法研究及应用[D].杭州:浙江工业大学,2018.
PEI H X.Research and application of adaptive density peak division clustering algorithm[D].Hangzhou:Zhejiang University of Technology,2018.

[16] 吴礼福,姬广慎,胡秋岑.强混响环境下基于 K-medoids 特征聚类的话者计数[J].南京大学学报(自然科学),2021,57(5):875-880.

[17] 胡军,王海峰.基于加权信息粒化的多标记数据特征选择算法[J/OL].智能系统学报:1-10[2023-04-12].http://kns.cnki.net/kcms/detail/23.1538.tp.20230317.1408.004.html.
HU J,WANG H F.Feature selection algorithm for multi tag data based on weighted information granulation[J/OL].Journal of Intelligent Systems:1-10[2023-04-12].http://kns.cnki.net/kcms/detail/ 23.1538.tp.20230317.1408.004.html.

[18] 赵源上,林伟芳.基于皮尔逊相关系数融合密度峰值和熵权法的典型新能源出力场景研究[J/OL].中国电力:1-10[2023-04-13].http://kns.cnki.net/kcms/detail/11.3265.TM.20230227.0856.006.html.
ZHAO Y S,LIN W F.Research on typical new energy output scenarios based on Pearson correlation coefficient fusion densitypeak value and entropy weight method[J/OL].China Power:1-10[2023-04-13].http://kns.cnki.net/kcms/detail/11.3265.TM.20230227.0856.006.html.

[19] 徐久成,黄方舟,穆辉宇,等.基于 PCA 和信息增益的肿瘤特征基因选择方法[J].河南师范大学学报(自然科学版),2018,46(2):104-110.
XU J C,HUANG F Z,MU H Y,et al.Tumor feature gene selection method based on PCA and information gain[J].Journal of Henan Normal University(Natural Science Edition),2018,46(2):104-110.

[20] 战庆亮,葛耀君,白春锦.流场特征识别的无量纲时程深度学习方法[J].工程力学,2023,40(02):17-24.
ZHAN Q L,GE Y J,BAI C J,et al.A dimensionless time history in-depth learning meth od for flow field feature recognition[J].Engineering Mechanics,2023,40(02):17-24.

[21] 雍菊亚,周忠眉.基于互信息的多级特征选择算法[J].计算机应用,2020,40(12):3478-3484.
YONG J Y,ZHOU Z M.Multi-level feature selection algorithm based on mutual information[J].Journal of Computer Applications,2020,40(12):3478-3484.

[22] 刘艳,程璐,孙林.基于 K-S 检验和邻域粗糙集的特征选择方法[J].河南师范大学学报(自然科学版),2019,47(2):21-28.
LIU Y,CHENG L,SUN L.Feature selection method based on K-S test and neighborhood rough sets[J].Journal of Henan Normal University(Natural Science Edition),2019,47(2):21-28.

Adaptive K-medoids algorithm for optimizing initial class center

Liu Jinjin

(College of Software, Henan Normal University, Xinxiang 453007, China)

Abstract: To solve the problem that the traditional K-medoids clustering algorithm needs to randomly select the initial cluster center and specify the number of clusters K, and the clustering results are unstable, this paper proposes an adaptive K-medoids algorithm to optimize the initial cluster center(CH_KD). The purpose is to define the feature importance, so as to screen out the best representative features in each cluster and form a feature subset, and focus on the adaptive optimization and improvement of the traditional partition algorithm. First, the feature discrimination is defined by the feature standard deviation, and the features with strong discrimination are selected. Secondly, Pearson correlation coefficient is used to measure the redundancy of each feature in the feature cluster, and the features with low redundancy are selected. Finally, the product of feature discrimination and feature redundancy is taken as the feature importance to screen out the best representative features in each cluster and form a feature subset. The experiment compares the proposed algorithm with other clustering algorithms on 14 UCI datasets, and the results verify that CH_KD the effectiveness and advantages of algorithm.

Keywords: unsupervised; feature differentiation; feature redundancy; CH function; feature selection

[责任编校　杨浦　陈留院]