

## 基于目标特征选择和去除的改进 $K$ -means 聚类算法

杨华晖<sup>†</sup>, 孟 晨, 王 成, 姚运志

(陆军工程大学 导弹工程系, 石家庄 050003)

**摘 要:** 针对高维数据聚类中  $K$ -means 算法无法有效抑制噪声特征、实现不规则形状聚类的缺点, 提出一种基于目标点特征选择和去除的改进  $K$ -均值聚类算法. 该算法使用闵可夫斯基规度作为评价距离进行目标点的分类, 增设权重调节参数  $a$ 、重置权重系数  $\alpha$  进行特征选择和去除, 可有效减小非聚类指标特征带来的噪声影响. 算法验证实验选取 UCI 真实数据集和人工数据集进行聚类分析, 验证改进算法对抑制噪声特征的有效性, 与  $WK$ -means、 $iMWK$ -means 算法进行实验对比, 分析聚类学习时特征选择的适用性, 同时寻找最优的距离系数  $\beta$  和权重系数  $\alpha$ .

**关键词:**  $K$ -均值算法; 特征选择; 高维数据聚类; 特征赋权; 数据去噪

中图分类号: N945

文献标志码: A

## Improved $K$ -means clustering algorithm based on feature selection and removal on target point

YANG Hua-hui<sup>†</sup>, MENG Chen, WANG Cheng, YAO Yun-zhi

(Department of Missile Engineering, Army Engineering University, Shijiazhuang 050003, China)

**Abstract:** Aiming at the weakness that the  $K$ -means algorithm cannot effectively suppress the noise attributes and realize irregular shape clustering on high-dimensional data, an improved  $K$ -means clustering algorithm based on feature selection and removal on target point is proposed. In the improved  $K$ -means algorithm, the Minkowski metric is adopted as the evaluation of distance for the classification of the target point. The weighting adjustment parameter  $a$  is added and the weighting coefficient  $\alpha$  is reset for feature selection and removal, which can reduce the effect of non-clustering index noise features. The UCI real datasets and artificial datasets are used for clustering analysis in the algorithm validation experiment. And the effectiveness of suppressing the noise features is validated. Compared with the  $WK$ -means and  $iMWK$ -means algorithms in the validation experiment, the applicability of feature selection in clustering learning process is analyzed. At the same time, the optimal distance coefficient  $\beta$  and the weighting coefficient  $\alpha$  are found.

**Keywords:**  $K$ -means algorithm; feature selection; high-dimensional data clustering; feature weighting; data denoising

## 0 引 言

$K$ -means 聚类算法是一种简单高效、易于实现的统计分析方法, 在模式识别、机器学习、数据挖掘等领域被广泛应用, 且在工程实践中具有很强的实用性<sup>[1-3]</sup>. 在知识发现和挖掘过程中, 常常需要对数据库进行数据聚类分析, 由于样本数据大, 特征分量多, 运用  $K$ -means 算法仅对特征进行权重赋值<sup>[4-6]</sup>, 则无法对噪声特征进行有效抑制, 因此需要在权重赋值之前对特征进行选择.

在  $K$ -means 特征赋值方法中, 文献[4]首先进行了研究并提出了  $WK$ -means 算法, 在目标函数中给出了特征权值参数  $w$ , 通过分步更新目标点分类矩阵、簇中心点和特征权值  $w$  迭代求解最优收敛点; 其后,

文献[7-10]采用  $iMWK$ -means 算法对初始聚类中心和特征权重调节系数进行了优化, 能够对目标特征进行组合处理, 允许存在不同距离的评价标准 ( $\beta \neq 2$ ), 对不规则形状的簇能有效初始化. 另外, 还有文献[11]提出的 FWSA 算法、文献[12]提出的 FGK 算法, 文献[13]提出的 IK-P 算法等. 这些方法都是在  $WK$ -means 算法的基础上提出的, 且都能做到对所有特征进行赋权值, 但并没有在赋权值之前对特征进行选择. 针对此问题, 本文提出一种基于目标特征选择和去除的改进  $K$ -means 算法, 该算法给出目标函数, 对迭代过程进行讨论, 并给出最优参数选取值. 最后选取 UCI 机器学习真实数据集和人工数据集进行实验评估, 验证算法对噪声的抑制效果, 对比  $WK$ -means、

收稿日期: 2017-11-15; 修回日期: 2018-04-12.

基金项目: 国家自然科学基金项目 (61501493).

<sup>†</sup>通讯作者. E-mail: yanghuahui1991@163.com.

iMWK-means 算法分析特征选择的优势和适用性.

1 K-means特征赋权问题

K-means 特征赋权值算法的应用是当前面对数据“维度灾难”问题的真实反映,现行聚类算法处理的数据早已超出传统单一数值的概念,每个目标点不仅含有多维数值<sup>[14]</sup>,而且还有分类属性<sup>[15]</sup>. 在计算机中需要多个数据类型共同描述某个目标点<sup>[16]</sup>,传统的 K-means 算法已经不能满足当前数据处理的要

求. 处理多类型数据,需要对不同特征维进行分组处理和权值分配. 运用 K-means 算法对工程实践问题进行模式识别或数据挖掘时,还要注意以下 3 个问题:

- 1) 避免噪声特征的干扰;
- 2) 允许同一簇中的各特征权值不同;
- 3) 允许同一特征在不同簇中具有不同权值.

为直观描述上述 K-means 特征赋权问题,列举来自某工程测试领域的数据库聚类问题,如表 1 所示.

表 1 某测试数据库的取样目标值

| 目标点 | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 特征值 | 24.4  | 17.3  | 13.8  | 20.9  | 25.1  | 15.2  | 3.3   | 4.4   | 2.9   |
|     | 24.2  | 20.1  | 13.1  | 3.1   | 1.9   | 1.9   | 25.2  | 13.7  | 10.1  |
|     | 6.4   | 6.9   | 6.5   | 20.2  | 13.2  | 17.6  | 18.4  | 22.7  | 13.3  |

在表 1 中: $X_1 \sim X_9$  选自某测试数据库中的 9 个目标点,  $X_1 \sim X_3, X_4 \sim X_6, X_7 \sim X_9$  分别表达了被测对象的 3 种工作状态. 为直观观测目标点位置,每个目标点用 3 个数据特征值进行描述,如图 1 所示.

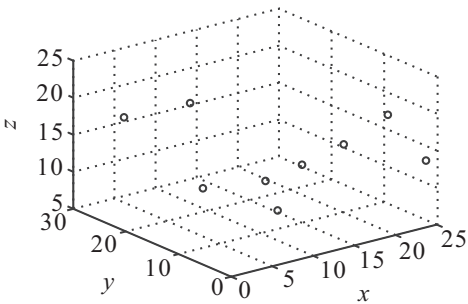
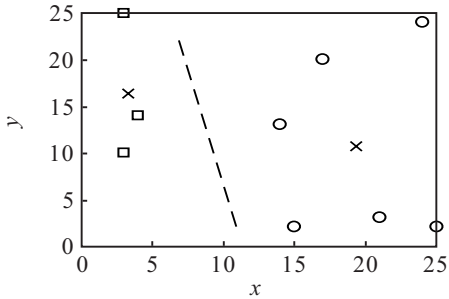


图 1 数据  $X_1 \sim X_9$  三维分布

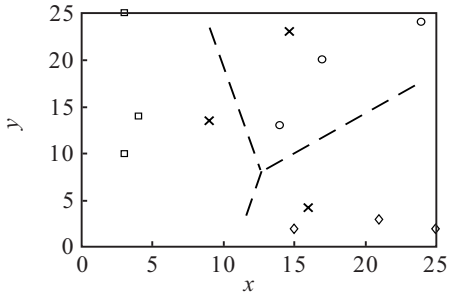
对上述目标点进行聚类,给定聚类簇数  $K = 3$ ,随机选择初始聚类中心,分别使用 K-means 算法、WK-means 算法和 iMWK-means 算法迭代求解最优收敛点,得到图 2 所示的聚类效果图,符号“ $\times$ ”表示迭代后输出的聚类中心.

对 3 种算法得到的聚类结果分析可知,由于 K-means 随机选取了初始点且无法对目标特征进行权重判断, K-means 算法易陷入局部收敛,从而无法得到正确的聚类结果,即  $K = 3$  簇内目标点个数为零. WK-means( $\beta = 1.8$ )和 iMWK-means( $\beta = 2$ )可以实现正确聚类,但最后得到的聚类中心效果有所差别,即对目标特征值的权重判断有偏差,且就最后迭代求解的次数而言: WK-means 在设置实验中的迭代更新次数  $N_{WK-means} = 23$  次,计算时间  $T_{WK-means} = 0.14\text{ ms}$ ; iMWK-means 迭代更新次数  $N_{iMWK-means} = 7$  次,计算时间  $T_{iMWK-means} = 0.05\text{ ms}$ . 在计算效率上 iMWK-means 更高.

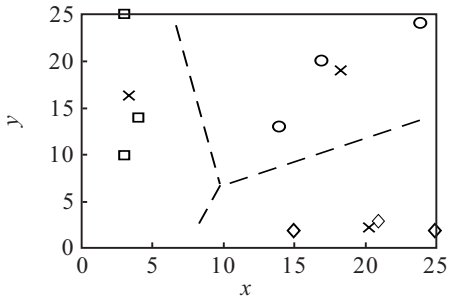
通过表 1 分析的 3 个簇的数值描述可以判断:簇 1



(a) K-means 算法



(b) WK-means 算法



(c) iMWK-means 算法

图 2 3 种算法聚类效果(二维)

的第 3 维特征、簇 2 的第 2 维特征、簇 3 的第 1 维特征相比其他特征的簇内散度较小、簇间散度大,因此可以很快实现数据的归类判断,得到正确的聚类结果.

针对上述问题,本文提出一种基于目标特征选择和去除的改进 K-means 算法 (Feature selection K-means, FSK-means),主要针对高维数据聚类中的特

征选择问题进行研究,给出需要优化的目标函数以及迭代过程中各分量的求解公式,最后使用UCI机器学习数据集和随机数据集进行验证,对比分析  $K$ -means 算法、WK-means 算法、iMWK-means 算法,验证特征选择对噪声去除的有效性和算法自身的适用性。

## 2 特征赋权的 $K$ -means 算法

### 2.1 WK-means 算法

WK-means 算法是在  $K$ -means 算法基础上对目标点的不同特征类型、不同特征权重进行描述的改进算法,  $K$ -means 算法给出的目标函数为

$$F(s, c) = \sum_{k=1}^K \sum_{i \in S_k} \sum_{v=1}^M (y_{iv} - c_{kv})^2. \quad (1)$$

其中:  $y_{iv}$  表示各目标点,  $c_{kv}$  表示各聚类中心,  $K$  表示簇数,  $S_k$  表示各聚类簇,  $M$  表示数据维数。

改进的 WK-means 算法对各维特征进行权重赋值后,增加了  $\omega_{kv}$  参数,对不同簇的不同特征进行了权重分配,给出的目标函数是

$$F(s, c, w) = \sum_{k=1}^K \sum_{i \in S_k} \sum_{v=1}^M w_{kv}^\beta (y_{iv} - c_{kv})^2. \quad (2)$$

其中:  $\beta$  表示权重调节系数,经讨论后得到  $\beta \geq 1$  的取值范围。已有文献验证了  $\beta = 1.8$  时聚类计算的良好效果,但在实际运用算法进行聚类学习时仍需要对  $\beta$  的值进行多次试验。另外,文献[4]给出了在目标特征为分类属性和数值类型时的  $\omega_{kv}$  取值情况。

### 2.2 iMWK-means 算法

iMWK-means 算法给出的目标函数为

$$F(s, c, w) = \sum_{k=1}^K \sum_{i \in S_k} \sum_{v=1}^M w_{kv}^\beta |y_{iv} - c_{kv}|^\beta. \quad (3)$$

就给定的目标函数而言, iMWK-means 算法使用闵可夫斯基规度作为评价距离,拓展了欧氏距离的限制。另外, iMWK-means 对不规则的簇形状有较好的聚类效果,是因为在进行聚类迭代之前对数据进行预处理,智能选取初始聚类中心<sup>[17]</sup>,给定初始中心点  $c_{kv}$  的位置。但从  $K$ -means 算法实现的步骤上分析,使用搜索算法预选聚类中心与迭代分配权值是两个不相关的步骤,初始中心选择可作为  $K$ -means 算法的另外研究内容<sup>[18-19]</sup>,因此下述算法仅从特征维度的权重分配进行探讨。

## 3 FSK-means 算法

### 3.1 基本思路

基于目标特征选择的  $K$ -means 算法,基本实现思路是将特定簇内的某特征的权重系数置为零,即

$$w_{kv} = 0, k \in K, v \in M. \quad (4)$$

因此,应当给出某判别标准,用以判定是否实现算法中权重系数  $\omega_{kv}$  的置零, FSK-means 算法给出的目标函数为

$$F(s, c, w) = \sum_{k=1}^K \sum_{i=1}^N \sum_{v=1}^M R_{ik} (w_{kv} \pm a)^\alpha d_{ik},$$

$$d_{ik} = |y_{iv} - c_{kv}|^\beta, \quad (5)$$

且满足

$$\sum_{v=1}^M w_{kv} = 1, 0 \leq w_{kv} \leq 1, k \in K, \quad (6)$$

$$R_{ik} \in \{0, 1\}, k \in K, i \in N. \quad (7)$$

其中:  $y_{iv}$  为各目标点;  $c_{kv}$  为聚类中心点;  $R_{ik}$  为目标点分配矩阵,  $R_{ik} = 1$  表示第  $i$  个目标点聚类在  $k$  簇内;  $N$  为目标点总数;  $a$  为特征的选择判断参数;  $w$  为权重系数;  $\alpha$  和  $\beta$  分别为权重和距离的调节系数。与 iWMK-means 给出的式(3)相比,上述目标函数添加了特征选择的判断阈值  $a$ ,并使用“ $\pm$ ”运算对不同簇内不同特征权重进行调节。对于小于特征选择阈值  $a$  的权重  $w$ ,通过比较后在算法实现中将  $w$  置零,从而实现去除聚类高维数据中的噪声特征分量。另外,式(5)将权重与距离调节参数分离,以满足不同数据形状、不同数据结构的聚类要求。上述目标函数的最优求解采用分步求最优值的方法,主要分为3个优化步骤:

- 1) 固定权重参数  $\omega_{kv}$  和聚类中心  $c_{kv}$ ,更新目标点分配  $R_{ik}$ ;
- 2) 固定分配矩阵  $R_{ik}$  和权重参数  $w_{kv}$ ,更新聚类中心  $c_{kv}$ ;
- 3) 固定聚类中心  $c_{kv}$  和分配矩阵  $R_{ik}$ ,更新权重参数  $\omega_{kv}$ 。

**定理1** 目标点分配矩阵  $R_{ik}$  由下式更新:

$$R_{ik} = \begin{cases} 1, & \sum_{v=1}^M (w_{kv} \pm a)^\alpha |y_{iv} - c_{kv}|^\beta \leq \sum_{v=1}^M (w_{k'v} \pm a)^\alpha |y_{iv} - c_{k'v}|^\beta, \\ & k' = 1, 2, \dots, K, k \neq k'; \\ 0, & \text{else.} \end{cases} \quad (8)$$

**证明** 目标点分配到定义距离最小的聚类簇内,当  $y_i$  与  $c_k$  的距离最小时,  $y_i$  被分配到  $k$  簇内<sup>[4]</sup>。□

**定理2** 聚类中心  $c_{kv}$  更新公式为

$$c_{kv} = \frac{y_{iv} \cdot R_{ik}}{\sum_{i=1}^N R_{ik}}, 1 \leq k \leq K, 1 \leq v \leq M. \quad (9)$$

**证明** 聚类中心更新由各簇内所有目标点求取

各维度的均值得到<sup>[7]</sup>. □

**定理3** 权重参数 $\omega_{kv}$ 由下式给出:

$$w_{kv} = \begin{cases} \frac{Ma+1}{\left(\sum_{u=1}^V \frac{D_{kv\beta}}{D_{ku\beta}}\right)^{1/\alpha-1}} - a, & D_{kv\beta} \neq 0; \\ 0, & D_{kv\beta} = 0, \sum_{u=1}^V D_{u\beta} \neq 0. \end{cases} \quad (10)$$

其中

$$D_{kv\beta} = \sum_{i \in S_k} |y_{iv} - c_{kv}|^\beta. \quad (11)$$

**证明** 由式(5)和条件(6)构建新目标函数,即

$$L = \sum_{k=1}^K \sum_{i=1}^N \sum_{v=1}^M R_{ik} (w_{kv} \pm a)^\alpha |y_{iv} - c_{kv}|^\beta + \lambda \left(1 - \sum_{v=1}^M w_{kv}\right). \quad (12)$$

令

$$D_{kv\beta} = \sum_{k=1}^K \sum_{i \in S_k} |y_{iv} - c_{kv}|^\beta R_{ik}. \quad (13)$$

采用式(12)对参量 $\omega$ 求偏导数,可得

$$\frac{\partial L}{\partial w} = \alpha (w_{kv} \pm a)^{\alpha-1} D_{kv\beta} - \lambda. \quad (14)$$

将式(14)置零可得

$$\alpha (w_{kv} \pm a)^{\alpha-1} D_{kv\beta} - \lambda = 0. \quad (15)$$

设此时迭代次数为 $x$ ,则有如下情况.

**情况1** 若 $W_{kv}^{(x-1)} < a$ ,则表明 $k$ 簇内 $v$ 特征的权重分量 $\omega_{kv}$ 小于设定阈值 $a$ ,此时 $W_{kv}^{(x)}$ 对 $a$ 取“-”号,有 $W_{kv}^{(x-1)} - a < 0$ ,又由于权重分量不能取负值,此时置 $\omega_{kv} = 0$ ;

**情况2** 若 $W_{kv}^{(x-1)} \geq a$ ,则表明 $k$ 簇内 $v$ 特征的权重分量 $\omega_{kv}$ 大于设定阈值 $a$ ,此时 $W_{kv}^{(x)}$ 对 $a$ 取“+”号,此时式(15)有

$$\alpha (w_{kv} + a)^{\alpha-1} D_{kv\beta} - \lambda = 0, \quad (16)$$

整理可得

$$\omega_{kv} + a = \left[ \frac{\lambda}{D_{kv\beta}} \right]^{1/\alpha-1}. \quad (17)$$

式(17)两边对簇内所有特征 $v$ 取和可得

$$1 + M_a = \left( \frac{\lambda}{\alpha} \right)^{1/\alpha-1} \sum_{v=1}^m \left( \frac{1}{D_{kv\beta}} \right)^{1/\alpha-1}. \quad (18)$$

对式(17)和(18)整理可得

$$w_{kv}^{(x)} = \frac{Ma+1}{\left(\sum_{u=1}^V \frac{d_{kv\beta}}{D_{ku\beta}}\right)^{1/\alpha-1}} - a, \quad (19)$$

且此时恒存在 $\omega_{kv} \geq 0$ . □

### 3.2 算法迭代过程

输入: 数据集 $Y = y_1, y_2, \dots, y_n$ , 目标点维度 $V$ , 聚类簇数 $K$ ;

初始化设置: 权重系数 $\alpha$ 、距离系数 $\beta$ 、权重选择参数 $a$ 、随机聚类中心 $c_1, c_2, \dots, c_K$ , 初始权重 $\omega = 1/V$ , 最大迭代次数 $X_{\max}$ , 收敛值 $\varepsilon$ ;

输出: 目标分配矩阵 $R_{ik}$ , 最终聚类中心 $c_k (k = 1, 2, \dots, K)$ .

Step 1: Repeat

Step 2: for  $i = 1$  to  $N$

Step 3: for  $k = 1$  to  $K$

$$\text{Step 4: } R_{ik} = \begin{cases} 1, & \arg \min \sum_{M=1}^V (W_{kv} \pm a)^\alpha d_{ik}, \\ 0, & \text{else.} \end{cases}$$

Step 5: end

Step 6: end

Step 7: for  $k = 1$  to  $K$

$$\text{Step 8: } c_{kv} = (y_{iv} \cdot R_{ik}) / \sum_{i \in N} R_{ik}$$

Step 9: end

Step 10: for  $k = 1$  to  $K$

Step 11: for  $v = 1$  to  $M$

Step 12: if  $W_{kv}^{(x-1)} = 0$

Step 13: set  $W_{kv}^{(x)} = 0$

$$\text{Step 14: else } W_{kv}^{(x)} = \frac{Ma+1}{\left(\sum_{u=1}^V \frac{D_{kv\beta}}{D_{ku\beta}}\right)^{1/\alpha-1}} - a, \\ D_{kv\beta} = \sum_{i \in S_k} d_{ik}.$$

Step 15: end

Step 16: end

Step 17: until  $C^{(x)} - C^{(x-1)} < \xi$  or  $x \geq X_{\max}$ ,

Step 18: return  $R_{ik}$  and  $C_k$

### 4 实验评估

实验评估主要考察算法在下述方面的聚类有效性:

1) 对于添加了噪声维度的不同数据集, 考察FSK-means与K-means算法的聚类恢复效果;

2) 对于相同的无噪声数据集, 比较FSK-means算法与K-means、WK-means、iWMK-means算法的聚类恢复情况;

3) 对相同的无噪声数据集, 考察不同 $\alpha$ 、 $\beta$ 参数取值下的算法聚类恢复效果, 寻找最优参数取值.

在算法复杂度方面, WK-means、iWMK-means、FSK-means算法在循环嵌套和迭代方法上具有一致

性,因此本节实验评估不予特别考虑.

4.1 数据集选择

实验评估所用的数据集为UCI数据库中7组真实数据集和5组人工随机生成的高斯数据集,其中真实数据集Wine、Iris、Heart disease、glass数据集已经被文献[4,7]验证过,另外选取Libras movement、gesture phase、MFCCs高维数据集对聚类有效性的2)和3)进行验证. 在添加噪声方面,每组真实数据集添加均值为0、方差为0.1的1~5列不等的噪声维度. 人工数据集方面,每组数据集目标数为500,维度为5,均值为0,方差为0.3,聚类数为5,各数据集目标数与维数情况如表2所示.

表2 各数据集目标数与维数

| 数据集             | 目标数   | 真实维度 | 噪声维度 |
|-----------------|-------|------|------|
| Wine            | 178   | 13   | 3    |
| Iris            | 150   | 4    | 1    |
| Heart disease   | 303   | 75   | 5    |
| glass           | 214   | 10   | 2    |
| Libras movement | 360   | 91   | 5    |
| gesture phase   | 9 900 | 50   | 4    |
| MFCCs           | 7 196 | 22   | 3    |
| GM data 1~5     | 500   | —    | 5    |

4.2 标准化初始聚类中心

在对每组数据集进行聚类分析前,需要初始化聚类中心. 为了更真实地体现算法的聚类效果,排除初始聚类中心不同对算法性能造成的影响. 本文在实验评估时均使用*iK*-means方法选取的初始聚类中心<sup>[17]</sup>,即后续实验比较的是*iK*-means、*iWK*-means、*iMWK*-means、*iFSK*-means算法使用相同聚类中心恢复得到的正确率.

4.3 有无噪声情况下的比较

运用*iK*-means算法和*iFSK*-means算法对是否添加了噪声维度的真实数据集进行聚类,得到结果如表3所示.

分析表3所示的结果可以得出: 1) 添加了噪声的数据集在运行*iK*-means算法时聚类正确率下降很大,即不能区分噪声维度和真实维度;而*iFSK*-means算法在特征选择时可去除噪声维度的干扰,减小噪声对聚类结果的影响. 2) Heart disease、Libras movement、gesture phase、MFCCs相对Wine、Iris、glass是高维数据集,*iK*-means算法对噪声没有辨别,因此在各个数据集添加噪声时均下降很快,*iFSK*-means算法对噪声维度的去除,在高维数据时比低维度数据更有效.

表3 有无噪声情况下数据集聚类准确率

| 数据集             | <i>iK</i> -means |      | <i>iFSK</i> -means( $\alpha = 1.2, \beta = 1.8$ ) |      |
|-----------------|------------------|------|---------------------------------------------------|------|
|                 | Contains         | No   | Contains                                          | No   |
| Wine            | 53.4             | 81.4 | 64.6                                              | 83.7 |
| Iris            | 46.7             | 84.0 | 79.3                                              | 90.0 |
| Heart disease   | 63.0             | 69.3 | 73.3                                              | 91.4 |
| glass           | 49.5             | 68.2 | 61.2                                              | 96.2 |
| Libras movement | 41.9             | 76.9 | 74.7                                              | 85.6 |
| gesture phase   | 56.5             | 71.3 | 70.1                                              | 81.4 |
| MFCCs           | 29.5             | 59.9 | 62.8                                              | 79.8 |

4.4 特征赋权算法的比较

特征赋权的算法,在*iWK*-means、*iMWK*-means、*iFSK*-means算法中因权重系数 $\alpha$ 与距离系数 $\beta$ 的取值差异而得到不同的聚类结果,*iWK*-means、*iMWK*-means算法参数值选择参考文献[4, 6-7]的不同数据集的最优取值实验,*iFSK*-means在 $a = 1/N$ 时不同 $\alpha$ 、 $\beta$ 系数取值下的聚类结果如表4~表10所示.

从实验呈现的结果看,在权重系数和距离系数确定的情况下,其算法恢复正确率为

$$\begin{aligned} &\text{Rate}(\text{iMWK-means}) > \text{Rate}(\text{iWK-means}) > \\ &\text{Rate}(\text{iK-means}). \end{aligned}$$

表4 Iris数据集聚类结果

| 算法                 | 系数取值        |            | 准确率/% |
|--------------------|-------------|------------|-------|
|                    | 权重 $\alpha$ | 距离 $\beta$ |       |
| <i>iK</i> -means   | —           | 2          | 84.0  |
| <i>iWK</i> -means  | 1.1         | 2          | 88.7  |
| <i>iMWK</i> -means | 1.2         | 1.2        | 94.7  |
| <i>iFSK</i> -means | 1.2         | 1.8        | 90.0  |
|                    | 2.2         | 1.8        | 88.7  |
|                    | 1.6         | 2.5        | 93.0  |
|                    | 1.6         | 1.2        | 73.3  |

表5 Wine数据集聚类结果

| 算法                 | 系数取值        |            | 准确率/% |
|--------------------|-------------|------------|-------|
|                    | 权重 $\alpha$ | 距离 $\beta$ |       |
| <i>iK</i> -means   | —           | 2          | 81.4  |
| <i>iWK</i> -means  | 1.1         | 2          | 81.4  |
| <i>iMWK</i> -means | 1.2         | 1.2        | 84.8  |
| <i>iFSK</i> -means | 1.2         | 1.8        | 83.7  |
|                    | 2.2         | 1.8        | 80.9  |
|                    | 1.6         | 2.5        | 84.8  |
|                    | 1.6         | 1.2        | 81.4  |

表 6 Heart disease 数据集聚类结果

| 算法         | 系数取值        |            | 准确率 / % |
|------------|-------------|------------|---------|
|            | 权重 $\alpha$ | 距离 $\beta$ |         |
| iK-means   | —           | 2          | 69.3    |
| iWK-means  | 1.1         | 2          | 80.4    |
| iMWK-means | 1.2         | 1.2        | 84.1    |
| iFSK-means | 1.2         | 1.8        | 91.4    |
|            | 2.2         | 1.8        | 78.5    |
|            | 1.6         | 2.5        | 84.5    |
|            | 1.6         | 1.2        | 78.2    |

表 7 glass 数据集聚类结果

| 算法         | 系数取值        |            | 准确率 / % |
|------------|-------------|------------|---------|
|            | 权重 $\alpha$ | 距离 $\beta$ |         |
| iK-means   | —           | 2          | 68.2    |
| iWK-means  | 1.1         | 2          | 88.3    |
| iMWK-means | 1.2         | 1.2        | 94.4    |
| iFSK-means | 1.2         | 1.8        | 96.2    |
|            | 2.2         | 1.8        | 79.9    |
|            | 1.6         | 2.5        | 91.1    |
|            | 1.6         | 1.2        | 88.3    |

表 8 Libras movement 数据集聚类结果

| 算法         | 系数取值        |            | 准确率 / % |
|------------|-------------|------------|---------|
|            | 权重 $\alpha$ | 距离 $\beta$ |         |
| iK-means   | —           | 2          | 76.9    |
| iWKK-means | 1.1         | 2          | 82.5    |
| iMWK-means | 1.2         | 1.2        | 86.4    |
| iFSK-means | 1.2         | 1.8        | 85.6    |
|            | 2.2         | 1.8        | 84.4    |
|            | 1.6         | 2.5        | 85.6    |
|            | 1.6         | 1.2        | 85.6    |

表 9 gesture phase 数据集聚类结果

| 算法         | 系数取值        |            | 准确率 / % |
|------------|-------------|------------|---------|
|            | 权重 $\alpha$ | 距离 $\beta$ |         |
| iK-means   | —           | 2          | 71.3    |
| iWKK-means | 1.1         | 2          | 75.9    |
| iMWK-means | 1.2         | 1.2        | 80.9    |
| iFSK-means | 1.2         | 1.8        | 81.4    |
|            | 2.2         | 1.8        | 72.5    |
|            | 1.6         | 2.5        | 82.6    |
|            | 1.6         | 1.2        | 77.0    |

表 10 MFCCs 数据集聚类结果

| 算法         | 系数取值        |            | 准确率 / % |
|------------|-------------|------------|---------|
|            | 权重 $\alpha$ | 距离 $\beta$ |         |
| iK-means   | —           | 2          | 59.9    |
| iWK-means  | 1.1         | 2          | 73.4    |
| iMWK-means | 1.2         | 1.2        | 80.6    |
| iFSK-means | 1.2         | 1.8        | 80.8    |
|            | 2.2         | 1.8        | 78.6    |
|            | 1.6         | 2.5        | 79.8    |
|            | 1.6         | 1.2        | 77.2    |

上述规律可以体现出特征赋权的  $K$ -means 算法相比于传统  $K$ -means 算法的聚类更有效性,尤其是在高维数据集 Heart disease、Libras movement、gesture phase、MFCCs 中,聚类性能提高得更为明显. 其次,对比 iFSK-means 与 iMWK-means 算法的恢复准确率,如表 11 所示.

表 11 iMWK-means 与 iFSK-means 恢复准确率比较

| 数据集             | iMWK-means | iFSK-means |
|-----------------|------------|------------|
| Wine            | high       | —          |
| Iris            | high       | —          |
| Heart disease   | —          | high       |
| glass           | —          | high       |
| Libras movement | high       | —          |
| gesture phase   | —          | high       |
| MFCCs           | —          | high       |

在表 11 中:MFCCs 数据集通过 iFSK-means 算法所得结果的提高并不是十分明显,原因是 iFSK-means 的参数取值与 iMWK-means 的相近,其中  $\alpha$  值均取 1.2,  $\beta$  值分别取 1.2 和 1.8,因此在特征维度的选择和去除上并不能取得明显效果. gesture phase 数据集正确率由 iMWK-means 的 80.9 % 提高到 iFSK-means 的 81.4%,仅就正确率变化而言,聚类效果提升并不明显,但由于数据集本身数据量大,数据恢复的正确个数实际增加了 168 个点,因此可以认为 iFSK-means 算法是能够在特征赋权前对特征进行有效选择和去除的. Heart disease 数据集恢复率提高相对多,但实际目标点数并不多,且在该数据集中数值差异较大,维度  $V$  大、聚类簇数  $K$  小,在不同类簇中很多维数据是相同的,因此 iFSK-means 在实际选择时可以较好地对其进行去除,从而留下在不同簇差异较大的特征进行聚类恢复. Glass 数据集的数据差异也很大,目标点较少,因此相对恢复的目标点数提高也并不是十分明显.

就 iMWK-means 算法和 iFSK-means 算法恢复

效果而言,当维度较大且数据点差异明显时,特征选择和去除的方法更能取得良好效果,实际原理就是在特征赋权之前对其进行判定选择.而在低维数据或数据差异不大时,特征选择会起到相反的效果.

4.5 最佳参数取值

iFSK-means 算法共给出  $a$ 、 $\alpha$ 、 $\beta$  三个参数取值,实验中均取  $a = 1/N$ ,下面主要针对权重和距离系数的不同取值进行实验和讨论,实验结果如表 12 所示.

表 12 不同  $\alpha$ 、 $\beta$  系数下 iFSK-means 聚类实验结果 %

| 数据集                 | iFSK-means 算法准确率            |                             |                             |                             |
|---------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|                     | $\alpha = 1.2, \beta = 1.8$ | $\alpha = 2.2, \beta = 1.8$ | $\alpha = 1.6, \beta = 2.5$ | $\alpha = 1.6, \beta = 1.2$ |
| Wine                | 83.7                        | 80.9                        | <b>84.8</b>                 | 81.4                        |
| Iris                | 90.0                        | 88.7                        | <b>93.0</b>                 | 87.3                        |
| Heart disease       | <b>91.4</b>                 | 78.5                        | 84.5                        | 78.2                        |
| glass               | <b>96.2</b>                 | 79.9                        | 91.1                        | 88.3                        |
| Libras movement     | <b>85.6</b>                 | 84.4                        | <b>85.6</b>                 | <b>85.6</b>                 |
| gesture phase       | 81.4                        | 72.5                        | <b>82.6</b>                 | 77.0                        |
| MFCCs               | <b>80.8</b>                 | 78.6                        | 79.8                        | 77.2                        |
| GM data 1 ~ 5(mean) | 80.3                        | 76.2                        | <b>78.7</b>                 | 74.9                        |

在表(12)中,GM给出的是5个数据集聚类恢复率的平均值,可以看到系数的最佳取值集中在 $\alpha = 1.2, \beta = 1.8$ 和 $\alpha = 1.6, \beta = 2.5$ 两组取值之间,因为真实数据集的差异较大,不能够有效辨别系数的最佳取值,所以给出GM数据集的聚类恢复结果.从随机数据的恢复结果可以看出, $\alpha = 1.2, \beta = 1.8$ 情况略显优势,但在实际问题中还需要根据数据集形状和数据特点进行参数选择.其次,权重调节参数 $a$ 在同组实验中的取值没有变过,主要是依据不同数据集的维度而固定设置的,该参数是否对聚类结果产生较大影响,需要进行后续验证.

5 结 论

特征赋权的  $K$ -means 算法是一种有效处理混合噪声维度、不规则形状类簇、高维数据集聚类的有效方法.本文提出了基于特征选择和去除的改进  $K$ -means 算法,通过与 WK-means、iMWK-means 算法的聚类恢复性进行比较,探讨了带噪声情况下的聚类效果;对比了同一数据集下不同特征赋权算法的聚类结果,分析了在特征赋权之前对其进行选择和去除的必要性;给出了 FSK-means 的最佳参数选择值.实验结果表明了该算法在处理高维数据和特征差异度较大时聚类性能的优势.但是,本文算法还有需要探讨和改进的方面:

- 1) 在处理数据的分类属性特征时,并没有给出具体方法,算法实验前只是进行了人工筛选.
- 2) 在参数  $a$  的取值上没有进行探讨,没有验证其他给定值能否进行有效的特征选择.另外,使用

( $w \pm a$ ) 的方法对特征进行选择和去除,使得算法略显粗糙和简单,需要改进后实现对特征权重更加灵活地处理.

3) 改进算法对簇间散度信息利用不充分,对噪声维度和数据不规则维度的分辨力有限.

参考文献(References)

[1] Bai L, Cheng X Q, Liang J Y, et al. Fast density clustering strategies based on the  $k$ -means algorithm[J]. Pattern Recognition, 2017, 71(3): 375-386.

[2] Jiang X P, Li C H, Sun J. A modified  $K$ -means clustering for mining of multimedia databases based on dimensionality reduction and similarity measures[J]. Cluster Computing, 2017, 20(10): 1-8.

[3] 黄月, 吴成东, 张云洲, 等. 基于  $K$  均值聚类的二进制传感器网络多目标定位方法[J]. 控制与决策, 2013, 28(10): 1497-1501.

(Huang Y, Wu C D, Zhang Y Z, et al. Multi-objective localization method based on  $K$ -means clustering in binary sensor network[J]. Control and Decision, 2013, 28(10): 1497-1501.)

[4] Chan E Y, Ching W K, Ng M K, et al. An optimization algorithm for clustering using weighted dissimilarity measures[J]. Pattern Recognition, 2004, 37(5): 943-952.

[5] Huang J Z, Ng M K, Rong H, et al. Automated variable weighting in  $k$ -means type clustering[J]. IEEE Trans on Pattern Analysis and Machine Learning, 2005, 27(5): 657-668.

[6] Chen E Y, Ye Y, Xu X, et al. A feature group weighting method for subspace clustering of high-dimensional data[J]. Pattern Recognition, 2012, 45(1): 434-446.

- [7] Amorim R C, Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in  $K$ -means clustering[J]. Pattern Recognition 2012, 45(3): 1061-1075.
- [8] Amorim R C D, Komisarczuk P. On initializations for the minkowski weighted  $k$ -means[C]. Int Conf on Advances in Intelligent Data Analysis. Helsinki: IEEE, 2012: 45-55.
- [9] Amorim R C D, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors[J]. Information Science, 2015, 324: 126-245.
- [10] Amorim R C D, Mirkin B. Selecting the Minkowski exponent for intelligent  $K$ -means with feature weighting[M]. Clusters, Orders, Trees: Methods and Applications, Optimization and its Applications. Berlin: Springer, 2014: 103-117.
- [11] Tsai C Y, Chui C C. Developing a feature weight self-adjustment mechanism for a  $K$ -means clustering algorithm[J]. Computational Statistics Data Analysis, 2008, 52(10): 4685-4672.
- [12] Chen X, Ye X, Xu X, et al. A feature group weighting method for subspace clustering of high-dimensional data[J]. Pattern Recognition, 2012, 45(1): 434-446.
- [13] Ji J C, Bai T, Zhou C G, et al. An improved  $K$ -prototypes clustering algorithm for mixed numeric and categorical data[J]. Neurocomputing, 2013, 120(22): 590-596.
- [14] 陈爱国, 王士同. 基于多代表点的大规模数据模糊聚类算法[J]. 控制与决策, 2016, 31(12): 2122-2130.  
(Chen A G, Wang S T. Fuzzy clustering algorithm based on multiple medoids for large-scale data[J]. Control and Decision, 2016, 31 (12): 2122-2130.)
- [15] 李向丽, 耿鹏, 邱保志. 混合属性数据集的聚类边界检测技术[J]. 控制与决策, 2015, 30(1): 171-175.  
(Li X L, Geng P, Qiu B Z. Clustering boundary detection technology for mixed attribute data set[J]. Control and Decision, 2015, 30 (1): 171-175.)
- [16] Anaraki F P, Becker S. Preconditioned data sparsification for big data with applications to PCA and  $K$ -means[J]. IEEE Trans on Information Theory, 2017, 63(5): 2954-2974.
- [17] Chiang M M, Mirkin B. Intelligent choice of the number of clusters in  $k$ -means clustering: An experimental study with different cluster spreads[J]. J of Classification, 2010, 27(1): 1-38.
- [18] 李武, 赵娇燕, 严太山. 基于平均差异度优选初始聚类中心的改进  $K$ -均值聚类算法[J]. 控制与决策, 2017, 32 (4): 759-762.  
(Li W, Zhao J Y, Yan T S. Improved  $K$ -means clustering algorithm optimizing initial clustering centers based on average difference degree[J]. Control and Decision, 2017, 32 (4): 759-762.)
- [19] 王莉, 周献中, 沈捷. 一种改进的粗  $K$  均值聚类算法[J]. 控制与决策, 2012, 27 (11): 1711-1719.  
(Wang L, Zhou X Z, Shen J. An improved rough  $K$ -means clustering algorithm[J]. Control and Decision, 2012, 27(11): 1711-1719.)

## 作者简介

杨华晖 (1992—), 男, 博士生, 从事装备测试及数据挖掘的研究, E-mail: yanghuahui1991@163.com;

孟晨 (1963—), 男, 教授, 博士生导师, 从事装备保障网络化、装备全寿命周期管理等研究, E-mail: mengchen63@163.com;

王成 (1980—), 男, 讲师, 博士, 从事装备全生命周期管理、测试系统软件体系的研究, E-mail: 32626364@qq.com;

姚远志 (1988—), 男, 工程师, 博士, 从事复杂装备维修策略、装备 MRO 关键技术的研究, E-mail: yaoyunzhi@vip.qq.com.

(责任编辑: 闫 妍)