

引用格式:唐益明,宋小成,任福继,等. 面向权重的可能性 MKFC 算法[J]. 微电子学与计算机,2020,37(12):6-11. [TANG Y M, SONG X C, REN F J, et al. Weight-oriented possibilistic MKFC algorithm[J]. Microelectronics & Computer, 2020, 37(12): 6-11.]

# 面向权重的可能性 MKFC 算法

唐益明, 宋小成, 任福继, 丰刚永

(合肥工业大学 计算机与信息学院, 安徽 合肥 230601)

**摘 要:** 基于多核的模糊聚类(MKFC)是当前聚类领域的最新热点,但是其很难通过人工方式确定在组合中核函数权重大小,并更好地调整所使用的不同内核函数的权重比. 为了解决以上问题,设计了自动称量 MKFC 算法的提案. 首先,给出了该算法目标函数的核心公式及其内在思想;其次,给出了 Mercer 核函数及多核距离的计算方法;然后给出了核函数权重的自动计算模式,由此形成了所提算法的流程;最后,在 8 个 UCI 数据库中进行测试,凭借与 5 种算法的分析对比,发现所提算法是较为理想的聚类方法.

**关键词:** 聚类;模糊 C 均值;可能性聚类;核函数

**中图分类号:** TP18

**文献标识码:** A

**文章编号:** 1000—7180(2020)12—0006—06

## Weight-oriented possibilistic MKFC algorithm

TANG Yi-ming, SONG Xiao-cheng, REN Fu-ji, FENG Gang-yong

(School of Computer and Information, Hefei University of Technology, Hefei 230601, China)

**Abstract:** Multi-kernel fuzzy clustering (MKFC) is a hot topic in the field of clustering. However, it is difficult to manually determine the weight of kernel function in the combination, and difficult to allocate the weight of kernel function ideally. Focusing on such problem, a weight-oriented possibilistic MKFC algorithm was put forward, which is based on multi-kernel mechanism, possibilistic clustering and automatic weight calculation. Firstly, the core formula of the objective function and its inner idea were given. Secondly, the calculation methods of the Mercer kernel function and multi-kernel distance were given. Then the automatic calculation mode of kernel function weight was given and the flow of the proposed algorithm was formed. Finally, the proposed algorithm is tested in 8 UCI databases, and compared with the 5 algorithms. It is found that the proposed algorithm is an ideal clustering method.

**Key words:** clustering; Fuzzy C-means; possibilistic clustering; kernel function

### 1 引言

模糊聚类在聚类分析领域中是一个非常重要的研究方向<sup>[1]</sup>,它主要是在模糊集理论的基础上,建立起一个非常实用的算法系统. 到目前为止已经存在数千种聚类算法,在各种聚类算法中,提出的时间更快,并且最常用的通用算法是模糊 C 平均算法.

FCM 算法从本质上来看是通过迭代使得对目标函数达到最小值,得到位置最优的聚类中心,进而获得聚类的结果. 然而,令人遗憾的是,FCM 算法的聚类结果非常容易受到噪声点和离群点的干扰和影响,这主要是由算法中对于隶属度的约束造成的. 因此,为克服这一弱点,使用概率值<sup>[2]</sup>来代替原始算法的成员,并构造了概率 PCM 算法,其在处理存在噪

收稿日期: 2019—12—02; 修回日期: 2020—02—10

基金项目: 国家自然科学基金(61673156, 61432004, U1613217, 61672202), 安徽省自然科学基金(1408085MKL15, 1508085QF129), 中国博士后科学基金(2014T70585), 国家“八六三”高技术研究发展计划基金(2012AA011103)

声和离群点时可以表现出更好的抗干扰性. 然后, PCM 却常存在聚类中心重叠的弱点<sup>[3]</sup>.

为此,文献[4]重新设计了新的可能性聚类算法,可解决原来中心重叠的弱点. 但是,该算法鲁棒性不行,受参数干扰太大. 文献[5]综合了 FCM 和 PCM 的内在机理,建立了 PFCM. PFCM 既对数据中可能存在的噪声表现出了较好的抗噪性,同时也避免了原先 PCM 算法会出现的聚类中心重叠的问题. 然而 PFCM 算法的不足之处在于,对于算法中的参数设置不同的值会导致算法的运行结果产生巨大的差异.

伴随着聚类算法的不断完善和发展,将数据从原始特征空间映射至核空间进行处理成为一种流行的趋势. 由此,文献[6]建立了 KFCM 算法. 在此之后,Genon 以统计学为出发点,提出了一种新的以核为基础的机器学习算法<sup>[7]</sup>. Tzortzis 以及 Likas 认为可以通过一种确定增量的算法来消除以往的算法中存在的对于初始化敏感的问题<sup>[8]</sup>. 但是,这种使用核距离的聚类算法其自身也存在较为显著的问题,即更擅长于处理球状的数据,但是在处理具有其他形态和性质的数据时很难得到良好的运行效果.

在算法的实际使用过程中,对于不同种类的核函数其适合处理的数据的形态是不同的. 仅仅使用某一种单一核函数不足以应对各种不同类型的数据. 为了解决此问题,一些学者试图引入多核的概念<sup>[9]</sup>,并且在此基础上构造出使用多个不同核函数的模糊聚类算法. 基于多核的概念<sup>[10]</sup>是由 Lanckriet 第一次提出,并且引出了后来的支持向量机的凸优化问题. 考虑到算法的效率,Bach 提出了一种优化算法,该算法可基于几个内核功能来最小化排序和效率问题<sup>[11]</sup>. 后来通过半无限线性规划解决了这个问题在文献[12]中,介绍了一种使用多个核函数的图像分割策略.

文献[13]提出了一种多核模糊聚类方法 MKFC,该方法可以非常灵活地选择内核功能,同时可以从不同角度增加信息源. MKFC 虽然解决了单一核函数在运用的过程中无法应对不同类型的数据的问题,然而基于多核的模糊聚类存在着显而易见的缺点,即很难通过自动设定的模式确定在组合中不同的核函数的权重的大小,难以对核函数权重进行较为理想的细腻分配<sup>[14]</sup>.

为了避免以往算法中出现的问题,在这篇文章中我们设计出一种新的面向权重的可能性 MKFC 算法. 该算法兼容了 FCM 算法和 PCM 算法的特

点,与此同时不再像 FCM 算法那样容易受到噪声点的干扰以及像 PCM 算法那样产生集群中心重叠的问题. 本文提出的算法在使用多核概念的基础上进一步整合了似然值的概念. 这样,获得的聚类结果更加准确. 同时,所提出的算法可以自适应地确定不同内核的权重值. 此外,这里还建立了权重的自动计算模式,以使内核空间中的不同数据点属于每个内核功能集群中心的权重.

## 2 面向权重的可能性 MKFC 算法

### 2.1 目标函数的核心公式

多核模糊聚类的思想是把许多个满足 Mercer 核条件的  $\chi_k$  进行加权和形成总的  $\chi$ ,以此来增强模糊聚类算法在处理不同类型数据集时的适应能力,其表达式为(其中  $\beta_k \geq 0$ ):

$$\chi(x_i, x_j) = \sum_{k=1}^K \beta_k \chi_k(x_i, x_j) \quad (1)$$

在最初的 FCM 算法中,数据样本的成员资格必须满足总数 1 的限制. 这种限制使算法在运行的过程中容易受到噪声点和离群点的干扰. 因此,为了有效地避免这种情况,在这里提出的新算法中,取消了这一约束条件,使用可能性值来发挥隶属度值的作用,进而设计出一种权重导向的可能性 MKFC 算法. 与此同时,在该算法中通过惩罚因子,来对算法的计算进行优化处理,这种做法增强了算法的性能.

这里提出的目标函数如下:

$$J = \sum_{i=1}^c \sum_{j=1}^n (u_{ij} + t_{ij})^m D_{ij}^2 + \sum_{i=1}^c r_i \sum_{j=1}^n (\partial_j - t_{ij})^\eta \quad (2)$$

式中, $c$  代表要聚类的簇数; $n$  代表要处理的数据集中的数据总量; $u_{ij}$  表示第  $j$  个数据点相对于第  $i$  个簇的隶属度的值; $m$  为模糊系数,它的取值区间为  $[1.5, 2.5]$ ,在一般情况下值取 2; $\eta$  一般取和  $m$  一样的值. $t_{ij}$  表示第  $j$  个点相对于第  $i$  个簇的典型值; $r_i$  为惩罚因子; $D_{ij}^2$  为第  $j$  个点关于第  $i$  类的多核距离.

此外,为了降低数据中可能存在的噪声点和离群点对聚类结果的影响,对于任意数据都建立  $\partial_j$ ,以此展示该点的价值. 通过这种做法,大大减轻了数据中离群点和噪声点对于迭代结果的干扰. 这里,使用符号  $\varphi$  来表示某种非线性映射函数,  $\varphi: R^L \rightarrow H, x \rightarrow \varphi(x)$ . 其中,  $x \in R^L, H$  为推演出的高维空间. 假定所需要处理数据的集合为  $x = \{x_1, x_2, \dots, x_n\}$ ,则样本权重  $\partial_j$  的计算方式如下:

$$d_j = \sum_{k=1}^n \exp(-\vartheta \|x_j - x_k\|^2) \quad (3)$$

式中,  $\vartheta$  是一个大于 0 的实数;  $\|x_j - x_k\|$  反映了欧式距离的值. 对于其中的某一个数据点来看, 该点与数据集中其他的数据点的距离越小, 其样本权重越大; 反之则越小.

一般而言, 离群点或者噪声点与其他数据点的距离相对较远, 这样一来其权重就会比较小, 对聚类的结果的影响也就会变小. 这样通过这个参数设置可以有效降低离群点或者噪声点的影响.

在计算过程中, 如果聚类中心位于原始特征空间中, 通过模糊聚类算法中的计算方式, 可以方便地计算出聚类中心. 凭借内在的推演机制, 获得迭代的模式:

$$\varphi(\hat{v}_i) = \frac{\sum_{j=1}^n (u_{ij} + t_{ij})^m \varphi(x_j)}{\sum_{j=1}^n (u_{ij} + t_{ij})^m} \quad (4)$$

凭借对  $J_\lambda$  求偏导, 能建立  $u_{ij}$  的计算模型. 这里  $\sum_{i=1}^c \sum_{j=1}^n u_{ij} = n$  必须要满足. 通过与 PCM 类似的方式, 也可以得到典型值  $t_{ij}$  的迭代目标公式, 隶属度和典型值的迭代目标公式分别如下所示:

$$u_{ij} = \frac{1}{\sum_{s=1}^c \left[ \frac{\sum_{l=1}^L [1 - \chi(x_j, \hat{v}_i)]}{\sum_{l=1}^L [1 - \chi(x_j, \hat{v}_s)]} \right]^{\frac{1}{m-1}}} \quad (5)$$

$$t_{ij} = \frac{1}{1 + \left\{ \frac{2b \sum_{l=1}^L [1 - \chi(x_j, \hat{v}_i)]}{\sigma^2} \right\}^{\frac{1}{\eta-1}}} \quad (6)$$

式中,  $r_i$  为惩罚因子, 它的表达式如下:

$$r_i = p \frac{\sum_{j=1}^n u_{ij}^m d(x_j, v_i)}{\sum_{j=1}^n u_{ij}^m} \quad (7)$$

## 2.2 多核距离的计算方法

根据 Mercer 等人提出的核函数理论, 如果函数  $\chi: X \times X \rightarrow R$  为一个正定核, 当且仅当  $\chi$  是正定的且满足方程:

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j \chi(x_i, x_j) \geq 0 \quad (8)$$

对任意的  $n \geq 2$  成立, 其中  $c_r \in R, r = 1, \dots, n$ . 同时 Mercer 核可以表示成为内积的形式, 如下所示:

$$\chi(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (9)$$

核距离根本意义和原始特征空间中的欧氏距离从本质上看是不同的. 距离  $D_{ij}$  的公式可以表示如下:

$$D_{ij} = (\varphi(x_j) - \varphi(\hat{v}_i))^T (\varphi(x_j) - \varphi(\hat{v}_i)) = \chi(x_j, x_j) - 2\chi(x_j, \hat{v}_i) + \chi(\hat{v}_i, \hat{v}_i) \quad (10)$$

采用  $\hat{v}_i$  来刻画核函数空间中的聚类中心, 其模式为:

$$k(x_j, \hat{v}_i) = \varphi(x_j) \varphi(\hat{v}_i) = \frac{\varphi(x_j) \varphi(x_i) \sum_{j=1}^N (u_{ij} + t_{ij})^m}{\sum_{j=1}^N (u_{ij} + t_{ij})^m} = \frac{\chi(x_i, x_j) \sum_{j=1}^N (u_{ij} + t_{ij})^m}{\sum_{j=1}^N (u_{ij} + t_{ij})^m}. \quad (11)$$

因此, 第  $i$  个聚类中心到其自己的距离, 如果采用高斯核函数可以表示为如下所示:

$$\chi(\hat{v}_i, \hat{v}_i) = \varphi(\hat{v}_i) \varphi(\hat{v}_i) = \frac{\chi(x_i, x_i) \sum_{j=1}^N (u_{ij} + t_{ij})^m \sum_{j=1}^N (u_{ij} + t_{ij})^m}{\sum_{j=1}^N (u_{ij} + t_{ij})^m \sum_{j=1}^N (u_{ij} + t_{ij})^m} \quad (12)$$

在具体的计算过程中, 不同的数据点在核空间中乘积的表达式是:  $\chi_l(x_i, x_j) = \varphi_l(x_i)^T \varphi_l(x_j)$ . 在本文中, 所使用的核函数主要是高斯核函数:

$$\chi_l(x_u, x_{j_l}) = \exp(-\|x_u - x_{j_l}\|^2 / \sigma) \quad (13)$$

## 2.3 权重的自动计算模式

任何 Mercer 条件的函数都可以用作内核函数. 所谓 Mercer 条件具体来说就是: 若存在一个函数  $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ , 同时对于任意一个平方积函数  $f(x)$  都有  $\int_{x \times x} k(x, z) f(x) f(z) dx dz \geq 0$ , 则可以说其满足 Mercer 条件. 为辨析多核函数是否符合 Mercer 条件, 可以引入新参数  $\varphi$ :

$$\varphi = \sum_{l=1}^L w_l \varphi_l(x), \quad w_l \geq 0 \quad (14)$$

在引入权重指数之后, 基于核的距离可以表示为如下公式:

$$D_{ij}^2 = \sum_{l=1}^L w_l^2 \chi_l(x_i, x_i) - 2 \sum_{j=1}^n \sum_{l=1}^L w_l^2 \chi_l(x_i, v_j) + \sum_{j=1}^n \sum_{j'=1}^n \sum_{l=1}^L w_l^2 \chi_l(v_j, v_{j'}). \quad (15)$$

固定隶属度后,我们对公式进行简化如下:令

$$\theta_l = \sum_{i=1}^c \sum_{j=1}^n \chi_l(\varphi(x_i), \varphi(x_j)) - \frac{\sum_{l=1}^L (u_{ij} + t_{ij})^m \varphi(x_{il})}{\sum_{l=1}^L (u_{ij} + t_{ij})^m} + \left[ \frac{\sum_{l=1}^L (u_{ij} + t_{ij})^m \varphi(x_{il})}{\sum_{l=1}^L (u_{ij} + t_{ij})^m} \right]^2 \quad (16)$$

化简后得到

$$J(\omega) = \sum_{l=1}^L \sum_{i=1}^c \sum_{j=1}^n (u_{ij} + t_{ij})^m \theta_l \omega_l^2 + \sigma^2 \sum_{i=1}^c \sum_{j=1}^n (\partial_i - t_{ij})^\eta. \quad (17)$$

同时权重的函数必须满足如下条件限制:

$$J_\lambda(\omega, \lambda) = \sum_{l=1}^L \sum_{i=1}^c \sum_{j=1}^n (u_{ij} + t_{ij})^m \theta_l \omega_l^2 + \sigma^2 \sum_{i=1}^c \sum_{j=1}^n (\partial_i - t_{ij})^\eta - 2\lambda (\sum_{k=1}^c \omega_k - 1). \quad (18)$$

这里进一步进行转换:

$$J_\lambda(\omega, \lambda) = \sum_{l=1}^L \sum_{i=1}^c \sum_{j=1}^n (u_{ij} + t_{ij})^m \theta_l \omega_l^2 + \sigma^2 \sum_{i=1}^c \sum_{j=1}^n (\partial_i - t_{ij})^\eta - 2\lambda (\sum_{k=1}^c \omega_k - 1). \quad (19)$$

对权重  $\omega_k$  求偏积分,有如下表示:

$$\frac{\partial J_\lambda}{\partial \omega_k} = 2 \sum_{i=1}^c \sum_{j=1}^n \sum_{l=1}^L (u_{ij} + t_{ij})^m \theta_l \omega_l - 2\lambda = 0 \quad (20)$$

因为  $\sum_{l=1}^L \omega_l = 1$ , 及  $H = (au_{ij}^m + bt_{ij}^\eta)$ , 则原式可

表示为:

$$\sum_{l=1}^L \omega_l = (\frac{1}{\theta_1 H} + \frac{1}{\theta_2 H} + \dots + \frac{1}{\theta_m H}) \lambda = 1 \quad (21)$$

再然后,可得到  $\lambda$  的计算方式如下:

$$\lambda = \frac{1}{\frac{1}{\theta_1 H} + \frac{1}{\theta_2 H} + \dots + \frac{1}{\theta_m H}} \quad (22)$$

不同核函数的权重计算公式如下所示:

$$\omega_l = \frac{\frac{1}{\theta_l H}}{\frac{1}{\theta_1 H} + \frac{1}{\theta_2 H} + \dots + \frac{1}{\theta_L H}} \quad (23)$$

至此,我们引入了可变的动态权重以表示不同核函数的变化. 内核函数的权重必须满足非负条件.

## 2.4 所提算法的流程

根据上述内容,我们可以获得面向权重的似然 MKFC 算法. 其运行流程如下所示:

算法:面向权重的可能性 MKFC 算法

输入:数据集包含  $n$  个点,簇的数目设为  $c$ , 设置模糊系数  $m > 1$ , 且  $n = \{x_i\}_{i=1}^n$ , 选定核函数分别为  $\{\chi_l\}_{l=1}^L, \{u_{ij}\}_{i=1, j=1}^{c, n}, \{\omega_l\}_{l=1}^L$ .

输出:  $\omega_l, u_{ij}, v_i$  及聚类结果.

Step 1: 初始化  $t_{ij}^{(0)}$ , 迭代次数以  $t$  来刻画, 其中最大的取  $t_{\max}$ ; 并初始化  $t = 1$ .

Step 2: 由式(4) 推演得到  $v_k^{(t)}$ .

Step 3: 由式(5) 推演得到  $u_{ij}^{(t)}$ .

Step 4: 由式(6) 推演得到  $t_{ij}^{(t)}$ .

Step 5: 根据公式(23) 更新权重  $\omega_l$ .

Step 6:  $t = t + 1$ . 如果条件  $|v_i^{(t)} - v_i^{(t-1)}| < \epsilon$  或  $t > t_{\max}$  成立, 执行 step7. 否则, 返回 Step 2.

Step 7: 输出各个参数以及最终的结果. 结束.

## 3 实验与分析

在这一节当中,为了探究我们提出的算法的具体性能,我们设计了一些实验. 这里采用了 5 个算法与新算法进行对比,对比算法分别为 FCM、PCM、PFCM、KFCM、MKFC.

UCI 数据库是由加利福尼亚大学建立的数据库,构建该数据库的目的即是为了对提出的各种不同类型的机器学习算法的性能进行评估. 目前,该数据库中存在着已经存在着数千种不同的数据集,并且实际的数据集的数量还在不断的增长. 我们在实验中所使用的数据皆来源于 UCI 数据库. 可以在表 1 中查看每个数据集的详细信息.

为了对 UCI 数据集的情况进行进一步的了解,在这里通过 Iris 数据库为例进行详细的描述. Iris 数据集涉及 150 个具体的数据,属性有 4 个,具体为 sepal length(花萼长度), sepal width(花萼宽度), petal length(花瓣长度), petal width(花瓣宽度). 根据数据点的特征,150 个数据点一共可以分为三类,代表了三种不同种类的鸢尾花: Iris Setosa, Iris Versicolour, Iris Virginica. 我们将 Iris 数据集通过降维的方式映射三维空间中(原来是 4 维的),如图 1 所示,可以直观地找到三维数据集的模式.

表 1 实验用到的数据库

| 数据集种类                | 数据集<br>样本数 | 样本属<br>性个数 | 样本聚<br>类个数 |
|----------------------|------------|------------|------------|
| Iris                 | 150        | 3          | 3          |
| Glass identification | 214        | 9          | 6          |
| Ecoli                | 336        | 7          | 18         |
| Yeast                | 1 484      | 8          | 10         |
| balance-scale        | 625        | 4          | 3          |
| Wine                 | 178        | 13         | 3          |
| Variant              | 160        | 10         | 8          |
| Pima                 | 768        | 8          | 2          |

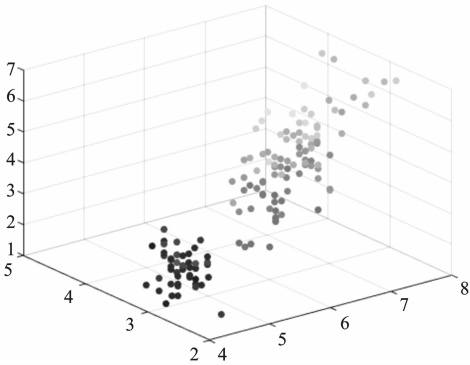


图 1 Iris 数据集分布图

通过实验,我们测试了该算法的性能效果. 在实验中,我们计算了算法结果中正确分类的数据点的数量,以计算聚类的准确性并评估通过聚类获得的结果的准确性. 在具体的操作中,需要执行如下的步骤:

- (1)对于实验中使用的数据集,必须首先删除数据中包含的标签信息.
- (2)随后,采用六种算法形成各自的聚类格局.
- (3)最后我们需要统计运行的结果中被正确分类的数据点的数量,用其除以总的样本量即可得到聚类精确度的值. 准确度的大小反映聚类性能的好坏.

对于在实验的过程中各个参数以及超参数的配置,我们将各个算法的最大迭代次数的值设置为 100,将停止界限取为  $1 \times 10^{-5}$ ,模糊系数  $m$  的值设置为 1.08,计算初始  $k$  值设置为 6.2. 为了选择内核功能,本文使用了几种高斯内核功能的组合. 实验所得到的结果数据如表 2 所示.

从表 2 的结果数据中我们可以看出,与其他对比算法相比较,我们所提出的面向权重的可能性 MKFC 算法在所有的实验数据集上运行所得到的

聚类精度都是最高的,均在 0.9 以上.

通过数据集 wine 作为例子来看,wine 为葡萄酒数据集,其来源于三种不同种类的葡萄酒. 数据集中一共存在 178 条记录,每条记录有 13 个特征属性,分别为葡萄酒的 13 中化学成分. 我们需要根据着 13 个特征属性判断数据点分别属于哪一种葡萄酒. 尽管数据集 wine 中存在较多的属性特征,并且数据的分布也比较复杂,但是在实验的过程中,我们提出的算法依然获得了非常理想的结果,其聚类精度达到了 0.925. 相比之下,其他比较算法在葡萄酒数据集上实现的最高聚类准确性仅为 0.841. 这说明所提算法的性能比以往算法有了很大提升.

从实验的结果可以看出,由于本文提出的面向权重的可能性 MKFC 算法在计算的过程中使用了可能性值,使得新算法可以很好的抵抗数据中的噪声点和离群点的影响;而且,映射到更高维内核空间的聚类数据不会引起一致的聚类问题. 所以新算法比其他算法可以生成更为优异的聚类结果.

表 2 聚类精度的实验结果

|                      | FCM   | PCM   | PFCM  | KFCM  | MKFC  | 所提<br>算法 |
|----------------------|-------|-------|-------|-------|-------|----------|
| Iris                 | 0.877 | 0.668 | 0.808 | 0.914 | 0.932 | 0.959    |
| Glass identification | 0.813 | 0.739 | 0.846 | 0.888 | 0.901 | 0.953    |
| Ecoli                | 0.775 | 0.751 | 0.818 | 0.869 | 0.909 | 0.968    |
| Seeds_dataset        | 0.889 | 0.779 | 0.902 | 0.811 | 0.907 | 0.963    |
| Wine                 | 0.708 | 0.409 | 0.688 | 0.777 | 0.841 | 0.925    |
| Variant              | 0.842 | 0.653 | 0.764 | 0.882 | 0.901 | 0.971    |
| Pima                 | 0.582 | 0.355 | 0.754 | 0.773 | 0.831 | 0.934    |
| Yeast                | 0.807 | 0.738 | 0.855 | 0.880 | 0.892 | 0.979    |

4 结束语

本文提出了一种针对 MKFC 的加权算法. 从 UCI 数据库中选择了八个数据集用于实验,从结果中可以看出,提出的面向权重的可能性 MKFC 算法可以通过不同权重的变化,调整其他内核功能在集群过程中的作用,最终获得更好的集群结果.

所提出的新算法吸纳了以往各算法的优点,避免了 FCM 算法对离群值更敏感,并且在 PCM 算法聚类时易于生成一致的聚类问题. 除此以外,将数据从原始的低纬度空间转换至高维空间的做法可以显著提高算法的鲁棒性.

在未来的研究中,将本文提出的算法与模糊推理进行融合<sup>[15]</sup>;另外,从粒度计算的方向出发,进一步对粒度模糊聚类算法进行开发.

## 参考文献:

- [1] TANG Y M, HU X H, PEDRYCZW, et al. Possibilistic fuzzy clustering with high-density viewpoint [J]. *Neurocomputing*, 2019,329(1): 407-423.
- [2] KELLERJM, KRISHNAPURAMR. A possibilistic approach to clustering [J]. *IEEE Transaction on Fuzzy Systems*, 1993,1(2):98-110.
- [3] KOUTROUMBASKD, XENAKISD, RONTOGIAN-NIS A A. On the Convergence of the Sparse Possibilistic C-Means Algorithm [J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(1): 324-337.
- [4] ZHANG J S, LEUNGYW. Improved Possibilistic C-Means Clustering Algorithms [J]. *IEEE Transactions on Fuzzy Systems*, 2004, 12(2):209-217.
- [5] PALN R, PAL K, JAMES M, et al. Possibilistic Fuzzy c-Means Clustering Algorithm [J]. *IEEE Transaction on Fuzzy Systems*, 2005, 13(4):107-116.
- [6] YANG M S, TSAI H S. A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction [J]. *Pattern Recognition Letters*, 2008, 29 ( 12 ): 1713-1725.
- [7] GONEN M, ALPAYDINE. Localized multiple kernel learning [C]. In: *Proc. 25th Annu. Int. Conf. Mach. Learning*, 2008:352 - 359.
- [8] TZORTZISG, LIKASA. The global kernel k-means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks* [J]. *IEEE Transactions on Neural Networks*, 2009,20(7):1181-1194.
- [9] NGUYENDD, NGOLT, PHAM L T, et al. Towards hybrid clustering approach to data classification: Multiple kernels based interval-valued fuzzy C-means algorithms [J]. *Fuzzy Sets and Systems*, 2015, 27(9): 17-39.
- [10] LANCKRIETGR, DEBT, CRISTIANINI N, et al. A statistical framework for genomic data fusion [J]. *Bioinformatics*, 2004,20(16):2626-2635.
- [11] BACHFR, LANCKRIETRG, JORDAN I. Multiple kernel learning, conic duality, and the SMO algorithm [C]. *International Conference of DBLP*. 2008, 6-13.
- [12] ZHAO B, KWOKJ, ZHANG C. Multiple kernel clustering [C]. In: *Proc. 9th SIAM Int. Conf. Data Mining*, 2009:638-649.
- [13] HUANG H C, CHUANG Y Y. Multiple Kernel Fuzzy Clustering [J]. *IEEE Transactions on Fuzzy Systems*, 2012,20(1):120-134.
- [14] CHEN S C, ZHANG D Q. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure [J]. *IEEE Transactions on System Man and Cybernetics - Part B*, 2004, 34(4):1907-1916.
- [15] TANG Y M, PEDRYCZW. On the  $\alpha(u,v)$ -symmetric implicational method for R- and (S, N)-implications [J]. *International Journal of Approximate Reasoning*, 2018,92(1): 212-231.

## 作者简介:

唐益明 男,(1982-),博士,副研究员.研究方向为聚类、模糊理论、图像处理. E-mail:tym608@163.com

宋小成 男,(1994-),硕士研究生.研究方向为聚类算法.

任福继 男,(1959-),博士,教授.研究方向为自然语言处理、模糊推理与聚类、情感计算与情感机器人.

丰刚永 男,(1991-),硕士研究生.研究方向为聚类算法.