计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900061

结合多目标优化算法的模糊聚类有效性指标及应用

崔国楠[1] 王立松[1] 康介祥[2] 高忠杰[2] 王 辉[2] 尹 伟[2]
1 南京航空航天大学计算机科学与技术学院 南京 210000
2 中国航空无线电电子研究院软件部 上海 200233
(xkcaor@163.com)

摘 要 模糊聚类方法可以更有效地对复杂数据集进行分析,由于模糊聚类算法的种类繁多且聚类结果会随着输入的聚类个数的不同而改变,使得模糊聚类算法产生的结果不准确,因此,要获得准确的聚类结果必须确定模糊聚类个数 k。目前已有的研究主要是利用多种模糊聚类有效性指标来确定最优聚类个数 k,但是诸如 SSD,PBM 等模糊聚类指标会随着划分的聚类个数 k 的增加而单调递减,导致聚类个数 k 不准确。为此,文中提出了一种结合多目标优化算法的模糊聚类有效性指标(A Validity Index of Fuzzy Clustering Combined with Multi-objective Optimization Algorithm,OSACF),将模糊聚类度量指标与多目标优化算法(Multi-Objective Optimization Algorithm,MOEA)相结合来解决聚类最优个数 k 的问题。与使用聚类有效性指标不同,OSACF 通过建立聚类个数 k 与聚类度量指标之间的双目标模型并使用 MOEA 优化该双目标模型来确定最优聚类个数 k,避免了聚类有效性指标趋于单调递减的影响。另一方面,OSACF 使用形态形似距离替代传统的欧氏距离度量,避免了聚类形状对计算聚类 k 值的影响。实验结果表明,OSACF 结合 MOEA 得到的最优模糊聚类个数 k 比已有的聚类有效性指标获得的结果更准确。

关键词:聚类有效性指标;模糊聚类;多目标优化算法;模糊聚类个数 k
中图法分类号 TP302

Fuzzy Clustering Validity Index Combined with Multi-objective Optimization Algorithm and Its Application

CUI Guo-nan[1],WANG Li-song[1],KANG Jie-xiang[2],GAO Zhong-jie[2],WANG Hui[2] and YIN Wei[2]
1 School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 210000,China
2 Software Department of China Aeronautical Radio Electronics Research Institute,Shanghai 200233,China

Abstract Fuzzy clustering method can analyze complex data sets more effectively. Because there are many kinds of fuzzy clustering algorithms and the clustering results will change with the number of input clusters,the results of fuzzy clustering algorithm are not accurate,so the number of fuzzy clustering k must be determined in order to obtain certain clustering results. At present,the existing research mainly uses a variety of fuzzy clustering effectiveness indexes to determine the optimal number of clusters k. However,fuzzy clustering indexes such as SSD,PBM will decrease monotonically with the increase of clustering number k,which makes it impossible to determine the optimal number of clusters k. Therefore,this paper proposes a fuzzy clustering validity index (OSACF) combined with a multi-objective optimization algorithm,which combines fuzzy clustering validity with a multi-objective optimization algorithm (MOEA) to solve the optimal number of clusters k problem. Different from using clustering validity index,OSACF establishes a bi-objective model between cluster number k and clustering validity index,and uses MOEA to optimize the bi-objective model to determine the optimal cluster number k,so as to avoid the influence of monotonous decreasing of clustering validity index. On the other hand,OSACF uses morphological similarity distance to replace the traditional Euclidean distance metric,which avoids the influence of cluster shape on the calculation of cluster k. The experimental results show that the optimal fuzzy cluster number k obtained by OSACF combined with MOEA is more accurate than the results obtained by the existing clustering effectiveness indicators.

Keywords Clustering validity index,Fuzzy clustering,Multi-objective optimization algorithm,Number of clusters k

到稿日期:2020-09-08 返修日期:2021-03-22 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:王立松(wangls@nuaa.edu.cn)

1 引言

随着互联网技术的发展，聚类在各领域中发挥着重要的作用，它可以主动地对数据点进行分组，使得属于同一个集群的数据点具有极高的相似性，属于不同集群的数据点有着较大的差异性。目前，k-Means，k-Medoids[1-3]等聚类算法在使用前需要确定划分的聚类个数 k 才能对数据集进行聚类划分。由于现实世界中真实数据的复杂性，在没有先验知识的情况下，无法准确地在算法开始前确定最优聚类个数 k，从而影响了聚类算法结果的准确性。针对这一问题，现有研究提出了聚类有效性指标来确定最优聚类 k 值[4-5]。目前，聚类有效性指标主要分为两类：外部有效性指标和内部有效性指标。外部有效性指标指通过将聚类分区结果与事先假定的正确聚类分区进行比较和评估，确定合理的聚类分区[6]。内部有效性指标则指通过检查得到的聚类结果找到最优的聚类分区。通常用衡量聚类内部的紧密度 $comp$ 和分离度 sep 来评估聚类分区 k 的合理性，其中 $comp+sep$ 越小，k 的取值就越合理。借助聚类的有效性指标可以从数据集的信息中更好地分析数据集的结构，从而得到数据集的最优分类数[7]。研究人员提出许多聚类内部有效性指标来对聚类进行验证。文献[8]提出了分块系数(Partition Coeffcient，PC)，文献[9]提出了分块熵(Partition Entropy，PE)，PC 和 PE 指标依赖隶属度矩阵来确定数据集，而未考虑到数据集的集合特征，且它们的数量有减少的趋势；文献[10]提出了 XB 指标，XB 指标评估的是聚类的整体分离度，而忽略了每一个聚类分离度之间的关系；文献[11]从紧密度和分离度角度提出了 FS 指标，然而 FS 指标采用的紧密度和分离度数量级不一致，产生了一定的数据随机性；文献[12]提出了 SC 指标，SC 指标总是为数据结构定义紧密度和分离度，却忽略了每个集群的定义；文献[13-14]提出了 PBM 指数，文献[15]提出了 SSD 指标，当聚类个数 k 的取值范围较大时，PBM 和 SSD 存在着单调递减的趋势；文献[16]提出了 VW 指数，文献[17]提出了 PCAES 指标，但 PCAES 指标缺少对增量的考虑，在处理流数据时，无法很好地监测数据的演化结构。

针对上述聚类有效性指标存在的问题，本文提出了一种结合多目标优化算法的模糊聚类有效性指标 OSACF。该有效性指标的主要思想是提出了一个模糊聚类度量指标与聚类个数 k 的双目标模型，并使用 MOEA 对该双目标模型进行优化。OSACF 可有效解决模糊聚类有效性指标随着 k 值单调递减的问题，并且可以得到最优 k 值。

2 多目标优化

OSACF 有效性指标将度量指标和聚类个数 k 组成一个双目标模型，从而可以利用多目标优化算法对建立的双目标模型进行优化，得到聚类的最优个数 k。

2.1 多目标优化问题

多目标优化问题(Multi-Objective Optimization Problem，MOP)的定义如下：

$$\begin{aligned}&\text{minimize } F(x)=\{f_1(x),f_2(x),\cdots,f_m(x)\}\\&\text{s.t. } \boldsymbol{x}\in R^n\end{aligned}\tag{1}$$

其中，$\boldsymbol{x}=\{x_1,x_2,x_3,\cdots,x_n\}$ 是 n 维欧氏空间中 R^n 的一组具有 n 个变量的解，目标函数 $\boldsymbol{f}(x)$ 是 x 的目标向量，存在 $F:R^n\rightarrow R^m$ 是 n 维向量空间到 m 维目标函数空间的一个映射。设 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别是由 n 个决策变量组成的决策向量。如果解 $\boldsymbol{a}$ 帕累托支配解 $\boldsymbol{b}$，则记作 $\boldsymbol{a}\prec\boldsymbol{b}$，当且仅当 $\forall l=1,2,\cdots,m$ 时目标函数 $f(\boldsymbol{a})_l\leqslant f(\boldsymbol{b})_l$ 并且 $\exists k=1,2,\cdots,m$ 使得 $f(\boldsymbol{a})_k<f(\boldsymbol{b})_k$。设解 $x\in R^n$ 不受任何一个解所支配[18]，则 x 称为一个帕累托最优(Pareto-optimal)。所有帕累托最优解的集合称为帕累托集(PS)，所有帕累托最优目标向量的集合是帕累托前沿(PF)[19]。

2.2 聚类多目标优化算法

在使用 k-Means 或者 k-Medoids 方法形成聚类的过程中，确定 k 值是一个很重要的步骤，其关系到聚类的形成。一些文献已经提出了解决这类问题的方法[18]：例如文献[20]找到了聚类个数 k 与某种度量聚类形成好坏指标之间的关系，并使用遗传进化算法优化得到最优聚类个数，此方法称为先验方法。而另一种方法是先通过实验得到不同的聚类结果，然后根据聚类的有效性指标来评估聚类的结果，此方法称为后验方法。本文将使用先验方法来解决最优 k 值的问题，即找到模糊聚类指标与聚类个数 k 之间的关系，然后通过根据多目标优化算法得到的 PF 来判断其最优解。

3 基于 OSACF 有效性指标的最优 k 值的求解

3.1 建立双目标模型

3.1.1 度量指标的改进

OSACF 有效性指标是将度量指标和聚类个数 k 组成一个双目标模型，因此首先需要一个合适的度量指标，但现有的度量指标不能满足要求，需要改进。

文献[11]提出的 FS 指标结合隶属度 u 以及数据点到数据中心的欧氏距离，虽然能很好地度量聚类内部的紧凑度，但是却忽略了数据的几何特征，从而产生了一定的随机性[21]。为了考虑数据的几何特征，我们需要将 FS 指标中的欧氏距离替换为形态形似距离(Morphology Similarity Distance，MSD)[21]。相比欧氏距离，MSD 考虑到了向量间的形状差异，减少了因几何特征而产生的随机性。

设 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 为聚类分析的数据集，$x_i=\{x_{i1},x_{i2},x_{i3},\cdots,x_{im}\}$ 表示 x_i 的 N 个特征。因此，基于 MSD 的紧凑度函数为：

$$Comp(k,U)=\sum_{i=1}^{k}\sum_{j=1}^{n}u_{ij}MSD(x_j,c_i)\tag{2}$$

其中，k 为聚类中心个数，$C=\{c_1,c_2,\cdots,c_k\}$ 为聚类中心的集合，$\boldsymbol{U}$ 为隶属度矩阵，$u_{ij}\in\boldsymbol{U}$ 为第 j 个数据在第 i 个聚类的隶属度系数，$MSD(x_j,c_i)$ 则表示第 j 数据在第 i 个聚类的偏差。

分离度函数检验的是不同聚类之间的关系，利用模糊聚类之间的距离测度可以得到分离度。由于一些模糊聚类之间可能存在重叠点，使用文献[22]提出的分离函数来计算所有模糊聚类之间重叠度的平均值，可以减小聚类间的分离度误差。设 F_p 和 F_q 是属于一个模糊分区 (k,U) 的两个模糊聚类，分离度函数如式(3)－式(6)所示：

$$Sep(k,U)=\frac{2}{k(k-1)}\sum_{p\neq q}^{k}S(F_p,F_q)\tag{3}$$

$$S(F_p,F_q)=\sum_{j=1}^{n}S(x_j:F_p,F_q)h(x_j)\tag{4}$$

$$S(x_j:F_p,F_q)=\min(\boldsymbol{F}_p(x_j),\boldsymbol{F}_q(x_j)) \tag{5}$$

$$\boldsymbol{F}_p(x_j)=\frac{\sum_{j=1}^{n}u_{pj}^m(x_j-c_p)(x_j-c_p)^{\mathrm{T}}}{\sum_{j=1}^{n}u_{pj}^m} \tag{6}$$

其中,$S(\boldsymbol{F}_p,\boldsymbol{F}_q)$且 $p,q\in C$ 为数据集 X 中,模糊聚类分区$\boldsymbol{F}_p$,$\boldsymbol{F}_q$的相似性。$\boldsymbol{F}_p(x_j)$为 x_j 在第 p 个聚类的协方差矩阵,$Sep(k,U)$表示 k 个聚类分区在数据集 X 上的相似性之和,k 值越大,模糊聚类分区相似性越小,$Sep(k,U)$的值就越小。$h(x_j)$表示一种权值,它可以根据模糊聚类之间重叠数据点的共享程度来调整对重叠数据点的强调程度,如式(7)所示:

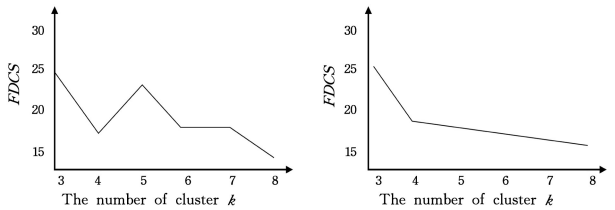
$$h(x_j)=-\sum_{i=1}^{k}u_{pj}(x_j)\log_a u_{qj}(x_j) \tag{7}$$

FDCS 定义为紧密度 $Comp$ 和分离度 Sep 之和,用于度量模糊聚类分区划分情况,FDCS 的值越小,表示模糊聚类内部越紧凑,聚类间的相似度越小,如式(8)所示:

$$FDCS(k,U)=Comp(k,U)+Sep(k,U) \tag{8}$$

3.1.2 FDCS 的转换

经过分析发现,FDCS 并不能直接作为双目标中的目标函数,这是因为只有在正确的聚类中心的情况下 FDCS 的值才会随着聚类 k 值的增大而减小,如图 1 所示。如果在非正确的聚类中心下使用 FDCS 作为目标函数,则可能导致原本正确的 k 值非支配解被控制,从而使得结果陷入局部最优[21]。因此,我们提出以下定理,以保证聚类度量指标在随机聚类中心下单调递减。



(a)Random centroids (b)True centroids

图 1 FDCS 在正确聚类中心和随机聚类中心下不同 k 值时对应的值

Fig. 1 FDCS values with true and rand cluster centroids over different k

定理 1 式(9)是单调递减的。

$$\min F(x)=\{f_1(x)=1-\exp^{-FDCS(k,U)}-k,f_2(x)=k\} \tag{9}$$

证明:

假设随着 k 的增加,$f_1(x)$单调递减,则满足:

$$x_1-x_2>0,f_1(x_1)-f_1(x_2)<0 \tag{9a}$$
$$f_2(x_1)-f_2(x_2)=(k_1-k_2)\geqslant 1$$

$$f_1(x_1)-f_1(x_2)=(k_2-k_1)-(\exp^{-FDCS(k_2,U)}-\exp^{-FDCS(k_1,U)}) \tag{9b}$$

其中:

$$k_2-k_1\leqslant -1 \text{ and } -1<\exp^{-FDCS(k_2,U)}-\exp^{-FDCS(k_1,U)}<1 \tag{9c}$$

由(9a),(9b),(9c)可得结论:随着 k 的增加,$f_1(x)$保持单调递减。

我们可以将 FDCS 转换为式(9)的形式。由定理 1 可知,此时可以保证 FDCS 在非正确的聚类中心的情况下随着目标函数$f_2(x)$的增加$f_1(x)$单调递减,从而保证了在非正确聚类中心的情况下 FDCS 随着聚类个数 k 的增加而减小。

3.2 双目标模型优化

双目标模型中聚类个数 k 与模糊聚类度量指标是一对冲突函数,本文使用文献[23]中提出的基于分解的多目标优化算法 MaOEAD-2ADV 对该双目标模型进行优化。MaOEAD-2ADV 首先沿着便捷方向向量进行搜索,以实现快速收敛,然后增加方向向量的数量,以逼近更完整的 PF。最后利用 Pareto 优势机制检测各方向向量的有效性,并调整无效方向向量的位置,以更好地拟合不规则 PF 的形状,最终保证了 PF 的均匀性以及规则性。优化流程如算法 1 所示。

算法 1 模糊聚类的 k 值优化

输入:最大迭代次数 Gen;人口数;邻居数;聚类个数 k 的取值范围;当前人口集合

输出:最优 k,即k_{best}

1. 通过初始化操作[P,**DV**,B,z^*,z^{nadir}]=INITILIAZATION 初始化人口 P,初始方向向量 **DV**、邻居指标 B,以及人口 P 中的理想点 z^* 和最低点 z^{nadir}。
2. for t=0 to maxGen do
3. 根据设置好的参数,对集合 P 进行杂交变异操作[Q,z^*]=VARIATION(P,B,z^*,N,m),产生的新子代生成集合 Q,并重新计算新子代的理想点 z^*。
4. 合并集合 P 和 Q,根据理想点和最低点选择 N 个解作为新的集合,即 P=ASSOCIATION_SELECTION(PUQ,**DV**,z^*,z^{nadir},N)。
5. if mod(t,$\emptyset_1$)==0 and K==2 then
6. if $\Delta t<10^{-4}$ then
7. [P,**DV**,B,z^{nadir}]=**DV**_ADJUSTMENT1(P,**DV**,N,T) 调节方向向量的数目。
8. K=N;
9. end
10. end
11. if mod(t,$\emptyset_2$)==0 and==N then
12. [**DV**,B]=DV_ADJUSTMENT2(P,**DV**,T) 调节无效方向向量的位置。
13. end
14. t++
15. end for
16. 根据新集合 P 计算帕累托前沿,即 PF=NONDOMINATED-SELECTION(P);
17. 使用 DB 指标计算 PF 的最优解k_{best}=DB(PF).

在步骤 1 中,首先对相关参数进行初始化操作,如算法 2 所示。随机生成人口集合 P,初始化聚类的关系矩阵 $\boldsymbol{U}$,并沿边界目标方向初始化 m 个初始方向向量 $\boldsymbol{DV}$(由于优化模型为双目标模型,则取 $m=2$)。第 i 个子问题的邻域指数集B_i设为 $\emptyset$,将 z^* 的每个目标初始化为 P 中目标的最小值,将 z^{nadir} 的每个目标设置为$+\infty$并将其作为初始值。

算法 2 初始化

输入:人口数或者为方向向量 **K**;最小聚类数C_{min};最大聚类数C_{max}

输出:初始化的人口集合 P={x^1,…,x^k};方向向量 **DV**={λ^1,…,λ^K};邻居指标 B={B^1,…,B^K};模糊聚类关系矩阵 **U**;理想点 z^*;最低点 z^{nadir}

1. P=$\emptyset$;
2. for i=1 to **K** do
3. 随机生成第 i 个子代x^i=RANDOM(i);
4. 把生成的子代加入 P 集合中 P=P$\cup x^i$;

5. end for
6. for i=C_{min} to C_{max} do
7. 根据生成的 P 构建关系矩阵 **U**,其中$\mathbf{U}_i$代表聚类中心为 i 的关系矩阵$\mathbf{U}_i$=FCM(P,i)。
8. end for
9. for j=1 to m to do
10. $\lambda_j^j=1,\lambda_i^j=0,i=1,\cdots,m,i\neq j$;
11. $B_j=\emptyset$;
12. 计算理想点z^*和最低点z^{nadir}。
13. $z_j^*=\min\limits_{x\in P} f_j(x)$;
14. $z_j^{nadir}=+\infty$;
15. end for

在步骤 3 中,N 个子代从 P 中产生,同时一个空集合 Q 被定义用于新产生的子代根据排序所得到的集合。子代的产生主要根据文献[23]中的 VARIATION 实现。

在步骤 4 中,将父种群 P 和种群 Q 结合,调用文献[23]提供的关联选择算法,从合并种群中选择 N 个解,作为新的种群 P。

在步骤 5—步骤 10 中,当第 t 代沿两个不同方向的向量减小到小于某一个阈值,即 10^{-4}时,说明所有子问题已经很好地进行了收敛,并通过 MaOEAD-2ADV 的 DV_ADJUSTMENT1[23]对方向向量的数量进行调整,即将方向矢量个数调整为 K,其中 K 为人口数。

在步骤 11—步骤 17 中,方向向量个数由 $K=2$ 扩展到 $K=N$ 后,在多目标优化问题中可能存在许多无效的方向向量。通过调用 MaOEAD-2ADV 的 DV_ADJUSTMENT2 来检测和调整这些无效的方向矢量。首先检测有效和无效的方向矢量。由于每个非支配解都与一个方向向量相关联,如果方向向量不包含相关的非支配解,则该方向向量覆盖的子区域很可能不包含 Pareto 最优解,那么该方向向量就被视为无效,否则被视为有效。

最后在步骤 14—步骤 15 中,通过 MOEA 得到关于 k 值的 PF 后,使用 Davies-Bouldin(DB) 指数[24](在本文中,我们仍然使用 MSD 替换 DB 指数中的欧式距离),如式(10)、式(11)所示。通过聚类内散射与聚类间分离之和的比值来得到最优的 k 值,即 DB 指数越小则所划分的聚类个数 k 越好,DB 在规定范围内得到的最小值即为最优 k 值。

$$DB=\frac{1}{k}\sum_{i=1}^{k}R_i \tag{10}$$

$$R_i=\max_{j,j\neq i}\left\{\frac{S_i+S_j}{d_{ij}}\right\} \tag{11}$$

其中,$S_i=\frac{1}{V_i}\sum_{x\in V_i}MSD(x,c)$为第 i 个聚类内的散射值,V_i为该聚类的数据点数量,c 为该聚类的聚类中心,$d_{ij}=MSD(c_i,c_j)$为两个聚类中心的形态相似距离。

4 实验及其结果分析

4.1 实验环境和实验参数设置

OSACF 实验环境如下:CPU 2.3 GHz 双核 Intel Corei5;内存 16 GB 2133 MHz LPDDR3;硬盘 SSD 512GB;操作系统 macOS Catalina 10.15.6。OSACF 的实验参数设置与 MaOEAD-2ADV 非常相似,其中差分进化(DE)中的 $\delta=0.9$,$CR=1.0$,$F=0.5$,$\eta=20$,$pm=1/n$。初始人口大小 P 设置为 80,最大迭代数设置为 100,邻居尺寸设置为 15,网格参数 K 设置为 80。在初始化过程中,理想点由初始种群获得,而最低点由簇的个数决定。

4.2 实验准备

实验将在 6 个人工数据集以及 6 个 UCI 真实数据集上把 OSACF 与别的聚类有效性指标(PC,PE,XB,PCAES,SC,VW)得到的最优 k 值进行对比测试。这 6 个人工数据集分别为:Data_3,Data_3Noise,Data_4,Data_5,Data_6 和 Data_4X,如图 2 所示,其中名称中的数字表示数据集的集群个数。表 1 列出了对上述所介绍数据集的简要说明,其中 Data _4X 考虑到了很强的重叠性,Data_3Noise 中存在噪声点。上述人工数据集与文献[25-26]使用的数据集相似,在大多数文献中 Iris 都被认为只有两个聚类,并且认为 $k=2$ 时聚类数量达到最优,但仍然有些聚类算法能产生 3 个聚类,因此总体来说 $k=2$ 或者 3 都被认为是合理的。

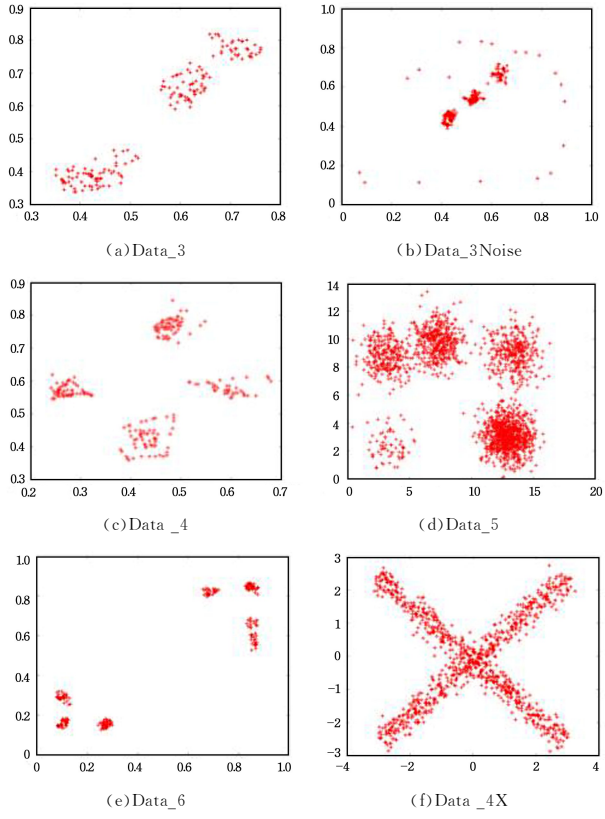


图 2 人工数据集聚类

Fig. 2 Artificial datasets of cluster

表 1 人工数据集

Table 1 Artificial datasets

data set	data number	dimension	cluster of number
Data_3	100	2	3
Data_4	200	2	4
Data_3Noise	130	2	3
Data_5	2000	2	4 or 5
Data_6	300	2	6
Data_4X	1000	2	4
Wine	178	13	3
WBCD	569	30	2
BLD	345	7	2
Iris	150	4	2 or 3
Breast Cancer	699	9	2
Pima	768	8	2

4.3 实验结果

表 2 列出了不同算法从人工数据集和真实数据集计算得到的最优聚类数，k_{opt} 表示预期的最优聚类个数。

表 2 不同算法在 $k\in[2,18]$ 时的最优聚类个数

Table 2 Optimal clustering number of different algorithms in $k\in[2,18]$

data set	k_{opt}	PC	PE	XB	PCAES	SC	VW	OSACF
Data_3	3	2	2	4	3	3	3	3
Data_3Noise	3	2	2	4	3	4	3	3
Data_4	4	4	2	4	4	4	4	4
Data_5	4 or 5	4	6	4	4	5	5	5
Data_6	6	6	6	4	6	4	6	6
Data_4X	4	2	2	9	2	3	4	4
WBCD	2	2	2	2	2	2	2	2
BLD	2	2	2	2	2	4	2	2
Wine	3	2	2	3	3	3	3	3
Iris	2 or 3	2	2	2	2	3	2	2
Breast Cancer	2	2	2	2	10	2	2	2
Pima	2	2	2	10	9	2	2	2

结合表 2 的结果，在聚类分离效果较为明显的 Data_3，Data_4，Data_6 数据集中，大多数有效性指标能得到预期的结果。PCAES，VW，OSACF 由于分配了归一化的分配系数和分离指数，能更好地识别以当前对象为聚类中心时是否具有成为良好聚类的能力，因此在数据集 Data_3Noise 上它们对噪声具有较强的鲁棒性。对于 Data_5 数据集，数据集的部分区域的点存在重叠，部分有效性指标如 PC 未使用聚类的分离度和紧密度考虑聚类划分，XB 和 PCAES 使用分离度评估聚类内对象分布状况，因此建议的最优 k 值为 4，而 OSACF，VW，SC 则对每一个聚类的分离度和紧密度进行综合评估，建议最优 k 为 5，说明在 Data_5 数据集下 OSACF 的结果更准确。在带有连接点且存在点重叠的 Data_4X 中，OSACF 由于结合了多目标优化算法，避免了有效性指标的单调递减的趋势，且使用 MSD 代替欧氏距离避免了聚类形状带来的影响，因此仍然能找到与预期相符的最优 k 值。另一方面，在真实数据集 BLD，WBCD，Wine，Breast Cancer 和 Pima 中，几乎所有有效性指标都能计算出与预期结果相同的最优 k 值。尽管在 Iris 数据集中不同算法得到的结果为 $k=2$ 或 $k=3$，OSACF 得到的最优 k 值也在这个合理范围内。根据上述与 PC，PE，XB，PCAES，SC，VW 所产生的结果进行比较，我们认为 OSACF 得到的最优 k 值是有效的。

为了进一步研究 OSACF 结果的准确性，我们根据 OSACF 以及 6 种有效性指标，将计算得到的 k_{opt} 作为 FCM 算法的输入值，在 Data_4X 数据集上得到的相应的聚类结果如图 3 所示。可以看到，OSACF 和 VW 根据相应的 k_{opt} 在 Data_4X 上得到的聚类结果符合实验预期，而其余指标由于 k_{opt} 不准确导致 FCM 产生的聚类结果不符合预期。综合上述实验结果，相比别的有效性指标，OSACF 得到的 k_{opt} 更准确，说明 OSACF 在不同数据集中得到的 k_{opt} 是有效的。

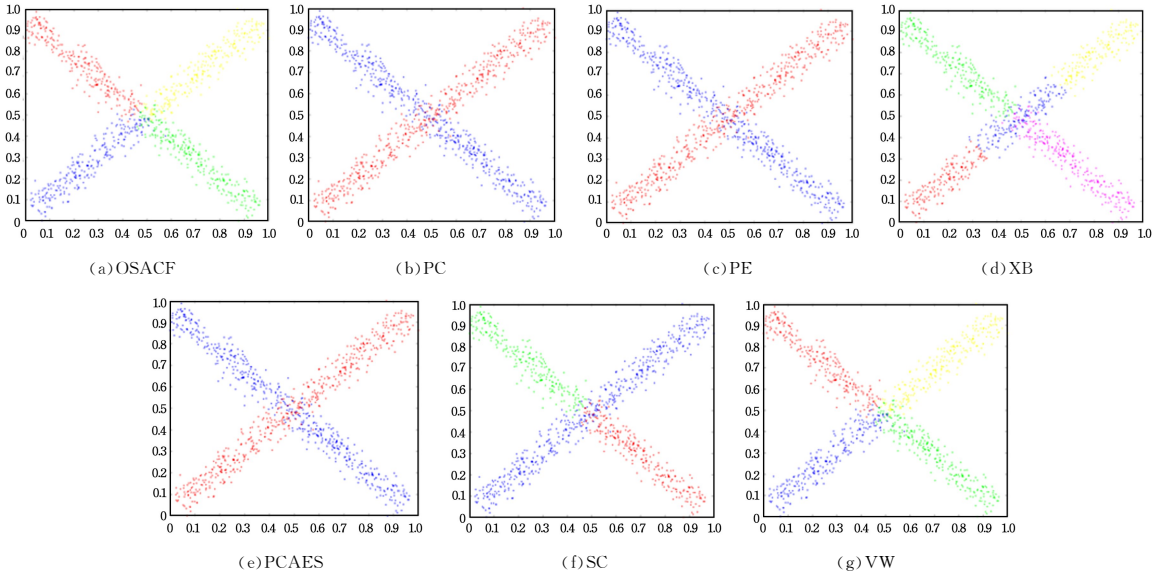


图 3 不同聚类有效性指标集合 FCM 算法在 Data_4X 数据集上的聚类结果

Fig. 3 Clustering results of different clustering indexes combined with FCM algorithm on Data_4X

4.4 OSACF 的应用实例

本节将 OSACF 得到的聚类最优个数 k 与 FCM 相结合运用到直升机试飞数据的分析中，通过分析直升机试飞数据来预测直升机的飞行状态。测试数据集记录的是直升机在空中旋转时的相关参数，该数据集包含 31333 条测试数据，每条数据包含 4 个属性：发动机转速、气压高度、油量、大气压力。我们设置聚类个数 k 的最大取值范围 $k_{max}=8,9,10,11$。根据测试结果得出，当 $k_{max}=8$ 时，$k_{opt}=2$，当 $k_{max}=9$ 时，$k_{opt}=6$，当 $k_{max}=10$ 时，$k_{opt}=4$，当 $k_{max}=11$ 时，$k_{opt}=6$。根据不同的 k_{max} 得到聚类结果显示，将数据分为 4 组，不同聚类之间具有很大的差异。如果数据被划分为更多的聚类，例如 $k_{max}=11,12,13$ 等，一些聚类会呈现出相似的模式（例如 $k_{max}=11$ 的 k_{opt} 与 $k_{max}=9$ 的 k_{opt} 相似），并可能包含在之前的 4 个聚类中，因此取 $k_{max}=10$，$k_{opt}=4$ 最为合理，其通过 OSACF 得到 $k_{opt}=4$ 为当前聚类的最优 k 值，如图 4 所示。接下来根据最优个数 $k_{opt}=4$ 结合 FCM 算法得到最终结果，如表 3 所列。由表 3 可知，在第一个聚类中，在气压高度较低和大气压力数值较大的状态下，或者在第 3 个聚类中发动机转速较快的情况下，耗

油量都会有所上升，从而可以推断出在直升机从悬停状态变为旋转状态或者直升机正处于旋转状态时，气压高度的变化和大气压力的变化会导致耗油量发生变化。在第二个聚类中，在发动机转速较慢、气压高度和大气压强较小的情况下，耗油量数值较低，可以推断出直升机可能从旋转状态变为悬停状态，从而导致耗油量相对较少。

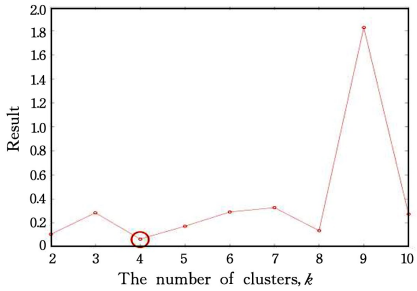


图 4 直升机试飞数据集的最优 k 值

Fig. 4 Optimal k value of flight test dataset

表 3 直升机试飞数据聚类划分

Table 3 Clustering of flight test data

engine speed	Barometric altitude	Oil volume	Atmospheric pressure
5280.10	−58.056	24.456	964.744
2039.93	−56.337	23.750	964.325
5514.77	−32.697	24.514	964.710
0	−53.754	24.407	0

结束语 本文提出了一种 OSACF 有效性指标，它结合多目标优化算法解决了模糊聚类有效性指标随着聚类个数 k 增加而减小的问题。与一般有效性指标不同，本文将聚类任务转换为一个双模模型，其中两个目标分别为聚类度量指标以及聚类个数 k，并使用多目标优化算法得到该双目标模型的 PF，最后通过 DB 指数分析 PF 得到的最优 k 值，使其在不同类型数据集下仍能准确地计算出聚类个数 k。与其他 6 种不同的聚类有效性指标的比较结果表明，OSACF 对本研究中大多数数据集都是有效的，并且可以应用到预测飞机的飞行状态中。然而，目前 OSACF 只解决了聚类双目标领域中的问题，暂时还未考虑到聚类三目标和多目标问题，在未来的研究中可考虑将 OSACF 指标应用到更多的模糊聚类的相关领域中。

参 考 文 献

[1] BEZDEK J C, EHRLICH R, FULL W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2/3): 191-203.

[2] ZHANG P Z, ZHANG H Y. A Review of Features and Labels Dimensionality Reduction Methods of Multi Label Data[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2020, 37(5): 23-29.

[3] WANG Z H, WANG S Y, DU H. Improved Fuzzy C-means Clustering Algorithm Based on Density-Sensitive Distance[J]. Computer Engineering, 2021, 47(5): 88-96, 103.

[4] GAN G, MA C, WU J. Data clustering: theory, algorithms, and applications[M]. Society for Industrial and Applied Mathematics, 2020.

[5] MATHER P, TSO B. Classification methods for remotely sensed data[M]. CRC Press, 2016.

[6] CUI H, ZHANG K, FANG Y, et al. A clustering validity index based on pairing frequency[J]. IEEE Access, 2017, 5: 24884-24894.

[7] VAIDYA J, SHAFIQ B, BASU A, et al. Differentially private naive bayes classification[C]// 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE, 2013, 1: 571-576.

[8] BEZDEK J C. Numerical taxonomy with fuzzy sets[J]. Journal of Mathematical Biology, 1974, 1(1): 57-71.

[9] BEZDEK J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics, 1973, 3: 58-73.

[10] XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.

[11] FUKUYAMA Y. A new method of choosing the number of clusters for the fuzzy c-mean method[C]// Proc. 5th Fuzzy Syst. Symp., 1989. 1989: 247-250.

[12] ZAHID N, LIMOURI M, ESSAID A. A new cluster-validity for fuzzy clustering[J]. Pattern Recognition, 1999, 32(7): 1089-1097.

[13] PAKHIRA M K, BANDYOPADHYAY S, MAULIK U. Validity index for crisp and fuzzy clusters[J]. Pattern Recognition, 2004, 37(3): 487-501.

[14] PAKHIRA M K, BANDYOPADHYAY S, MAULIK U. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification[J]. Fuzzy Sets and Systems, 2005, 155(2): 191-214.

[15] WANG R, LAI S, WU G, et al. Multi-clustering via evolutionary multi-objective optimization[J]. Information Sciences, 2018, 450: 128-140.

[16] ZHANG Y, WANG W, ZHANG X, et al. A cluster validity index for fuzzy clustering[J]. Information Sciences, 2008, 178(4): 1205-1218.

[17] WU K L, YANG M S. A cluster validity index for fuzzy clustering[J]. Pattern Recognition Letters, 2005, 26(9): 1275-1291.

[18] MIRJALILI S, JANGIR P, SAREMI S. Multi-objective ant lion optimizer: a multi-objective optimization algorithm for solving engineering problems[J]. Applied Intelligence, 2017, 46(1): 79-95.

[19] YANG X S. Nature-inspired optimization algorithms[M]. Academic Press, 2020.

[20] GUO Y, WENG G. K-means++ clustering-based active contour model for fast image segmentation[J]. Journal of Electronic Imaging, 2018, 27(6): 063013.

[21] LI Z, YUAN J, ZHANG W. Fuzzy C-mean algorithm with mor-

phology similarity distance[C]// 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2009,3:90-94.

[22] YANG S,LI K,LIANG Z,et al. A novel cluster validity index for fuzzy c-means algorithm[J]. Soft Computing,2018,22(6):1921-1931.

[23] CAI X,MEI Z,FAN Z. A decomposition-based many-objective evolutionary algorithm with two types of adjustments for direction vectors [J]. IEEE Transactions on Cybernetics, 2017, 48(8):2335-2348.

[24] WANG L,CUI G,ZHOU Q,et al. A multi-clustering method based on evolutionary multiobjective optimization with grid decomposition[J]. Swarm and Evolutionary Computation, 2020, 55:100691.

[25] REZAEE B. A cluster validity index for fuzzy clustering[J]. Fuzzy Sets and Systems,2010,161(23):3014-3025.

[26] DAVIES D L,BOULDIN D W. A cluster separation measure [J]. IEEE transactions on Pattern Analysis and Machine Intelligence,1979(2):224-227.

CUI Guo-nan, born in 1996, postgraduate. His main research interests include multi-object optimization method and data mining.

WANG Li-song, born in 1969, associate professor, is a member of China Computer Federation. His main research interests include avionics safety analysis, data management in distributed environments, formal methods and model-based safety analysis, and wireless sensor network.