智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向模糊C均值算法的MAME聚类有效性指标

唐益明, 陈仁好, 李冰

引用本文:
唐益明,陈仁好,李冰. 面向模糊C均值算法的MAME聚类有效性指标[J]. 智能系统学报, 2023, 18(5): 945-956.
TANG Yiming, CHEN Renhao, LI Bing. A clustering validity index called MAME for the fuzzy c-means algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(5): 945-956.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202212028

DOI: 10.11992/tis.202212028
网络出版地址: https://kns.cnki.net/kcms2/detail/23.1538.TP.20230621.2011.002.html

面向模糊 C 均值算法的 MAME 聚类有效性指标

唐益明，陈仁好，李冰
（合肥工业大学 计算机与信息学院，安徽 合肥 230601）

摘　要：聚类有效性指标可用来评估聚类结果的有效性，并且帮助判别聚类的类别数。现有的面向模糊 C 均值算法的聚类有效性指标存在对于类内紧致性的刻画不太到位、对于类间分离性的度量刻画不够准确的问题。为此，基于类内紧致性和类间分离性两个角度着手设计，提出了一种新的模糊聚类有效性指标——考虑最大值和均值的指标（maximum-mean, MAME）。首先，考虑了整个数据集的综合特征，计算分别分为 K 类和 1 类的情况的比值，提出了一种新的模糊紧致性度量表达式。其次，引入最大聚类中心距离和平均聚类中心距离，提出了一种新的分离性度量方法。最后，从模糊紧致性度量表达式、分离性度量方法出发，提出了 MAME 指标。面向 5 个 UCI 数据集和 6 个人工数据集，和 9 个聚类有效性指标（包括 CH、DB、NPC、PE、FSI、XBI、NPE、WLI 和 *I* 指标）一起进行了对比实验，验证了所提指标的准确性、稳定性，说明了 MAME 指标的鲁棒性较好。

关键词：聚类；模糊聚类；模糊 C 均值；聚类有效性指标；内部指标；外部指标；紧致性；分离性
中图分类号：TP181；TN99　　**文献标志码：**A　　**文章编号：**1673−4785(2023)05−0945−12

中文引用格式：唐益明，陈仁好，李冰．面向模糊 C 均值算法的 MAME 聚类有效性指标 [J]．智能系统学报，2023，18(5)：945–956.
英文引用格式：TANG Yiming, CHEN Renhao, LI Bing. A clustering validity index called MAME for the fuzzy c-means algorithm[J]. CAAI transactions on intelligent systems, 2023, 18(5): 945–956.

A clustering validity index called MAME for the fuzzy c-means algorithm

TANG Yiming, CHEN Renhao, LI Bing
(School of Computer and Information, Hefei University of Technology, Hefei 230601, China)

Abstract: The clustering validity index can be used to evaluate the effectiveness of clustering results and determine the number of clusters. However, existing validity indices for fuzzy c-mean algorithm suffer from the inadequate characterization of intracluster compactness and inaccurate measurement of intercluster separability. To address these issues, we proposed a new fuzzy clustering validity index called maximum-mean (MAME), which considers the maximum and mean values and is designed based on two perspectives, intracluster compactness and intercluster separability. First, considering the comprehensive characteristics of the entire dataset, a new expression of fuzzy compactness measure is put forward by calculating the ratio of cases divided into K clusters and one cluster, respectively. Second, by introducing the maximum and mean distance between cluster centers, a new method is proposed for separability measurement. Finally, the MAME index is put forward on the strength of fuzzy compactness measure expression and the separability measure method. Using five UCI and six artificial datasets, MAME is compared with nine other cluster validity indices, including CH, DB, NPC, PE, FSI, XBI, NPE, WLI, and *I*. The experimental results demonstrate the accuracy and stability of our proposed index, indicating that MAME has good robustness.

Keywords: clustering; fuzzy clustering; fuzzy c-means; clustering validity index; internal criteria; external criteria; compactness; separation

收稿日期：2022−12−18.　**网络出版日期：**2023−06−25.
基金项目：国家重点研发计划项目（2020YFC1523100）；国家自然科学基金项目（62176083，62176084）.
通信作者：唐益明. E-mail: tym608@163.com.

在大数据时代，数据无所不在，如何从海量的数据中挖掘出有价值的信息变成了一个重要的问题[1-4]。日常生活中产生的各种数据无一不蕴含

着各种各样丰富的信息。生产的数据只有经过加工和处理，才能够提炼出真正有价值的信息，其中一个重要的处理机制就是聚类。

聚类将相似性高的数据点划分到同一簇内，相似性低的数据点分离出去。作为一种无监督的机器学习方法[5-8]，在对海量的数据聚类后，将能提取出有价值的信息。按照是否能将数据集中每个样本只划分至一个簇，又可分为硬聚类和模糊聚类。硬聚类的“硬”体现在非 0 即 1。模糊聚类相对于硬聚类而言，其特点体现在“模糊”，通过引入模糊隶属度的概念[9-11]，对某个对象属于某一类的不同程度进行刻画，这使得聚类的结果更加贴近现实意义。其中，最为广泛应用的是模糊 C 均值 (fuzzy C-means, FCM) 算法[11-14]。FCM 算法把聚类过程转化为带约束条件的目标函数优化问题，再通过数学方法求解，最终可以确定聚类结果。

在聚类的研究过程中有两个十分复杂的问题，一是如何划分数据集才能得到最好的聚类结果，二是如何确定该数据集划分的类别数。前者通过聚类算法解决，后者可以通过聚类有效性问题[15-19]来解决。如何确定最佳聚类数是聚类领域的公认难题。虽然聚类的类别数可能由用户的经验或者专家根据领域的知识进行估计得到，但通常我们难以提前得知真实的聚类数。

真实世界中的数据往往复杂且多样，这就要求聚类算法能够准确地根据数据其内部的特征和结构进行聚类。聚类有效性研究就是通过使用聚类有效性指标对聚类结果进行评估，从而分析出聚类的效果。具体而言，在不同聚类数的情况下，运行聚类算法，若得到的结果使得聚类有效性指标函数下取得最优值，则该情况即为最佳聚类数，该划分即为最佳划分。这种研究方法是简洁而有效的。

当前的聚类有效性指标主要涉及 3 类[13-15]，即外部有效性、内部有效性和相对有效性指标。内部有效性指标基于数据集的几何结构信息，从紧致性、分离性、连通性和重叠度等方面对聚类结果加以评价。外部有效性指标通过将聚类结果与外部准则相对比来评估聚类效果。相对有效性指标则根据预先设置的评价标准，对取不同参数的聚类算法进行评估，最终选出较优的参数设置和聚类模式。此外，还有其他类型的评价指标，比如生物类型指标[20]、关联性指标[21]、基于稳定性的指标[22]等，都是为了针对某一特性而研究出来的。

学者们在内部有效性指标的设计中投入了很多精力，现存的内部有效性指标的设计主要分为以下几类。第一类是基于几何结构信息，如 Calinski 提出的 CH 指标[23]、Davies 提出的 DB 指标[24]等。CH 指标用类内离差矩阵来度量紧密度，用类间离差矩阵来度量分离度。DB 指标用类内样本点到其聚类中心的距离来度量类内紧致性，用聚类中心之间的距离来度量类间分离性。第二类基于隶属度，如 Bezdek 提出的用于模糊聚类的有效性指标，分离系数 PC[25] 和分离熵 PE[26] 指标，以及 Roubens 提出的标准分离系数 NPC 和 NPE[27] 指标等。PC 和 PE 指标考察的维度是隶属度信息，并没有考虑到样本的结构信息，且 PC 指标还有一个缺点，就是单调变化。为了克服这个缺点，NPC 指标应运而生，但其也没有对样本信息进行全面的考量。此外，还有一些指标基于数据集的结构信息和隶属度，比如 Xie 提出的 XBI[28] 指标、Fukuyama 提出的 FS[29] 指标。XBI 指标是一种比值型的指标，但指标的性能不稳定。FS 指标是一种求和型指标，这类指标和 DB 指标原理相同。在 XBI 指标的基础上，通过在其中引入聚类中心之间的中值距离，Wu 等提出了 WLI 指标[30]，在实际中也有不错的表现。后来还有学者提出了 I 指标[31]，其由 3 部分信息构成，也取得了不俗的效果。

目前来看，现有的面向模糊 C 均值算法的内部有效性指标尚存在一些比较典型的问题：

1) 对于类内紧致性的刻画不太到位。现有的大多数指标都是用类内距离表示簇内的紧致性。类内距离越小则认为类内紧致性更好。考虑到模糊聚类的特点，其实这个是不够充分的。大多指标对此都处理得较为简单。这方面 WLI 考虑得稍好，但也没有考虑整个数据集的综合特征。

2) 对于类间分离性的度量刻画不够准确。大多数有效性指标的处理机制过于简单、粗糙，比如 XBI 采用聚类中心之间的最小距离来刻画类间分离性，FS 和 VCVI 采用聚类中心和平均聚类中心的差值再求和来刻画。而这种对类间分离性度量的刻画方式并不准确，需要更为综合性地考虑。

基于上述原因，本文提出一种新的聚类有效性指标，即考虑最大值和均值的指标（maximum-mean, MAME）。该指标从类内紧致性和类间分离性两个角度着手设计。首先，考虑了整个数据集的综合特征，计算分别分为 K 类和 1 类的情况的比值，提出了一个新的模糊紧致性度量表达

式。其次, 引入最大聚类中心距离和平均聚类中心距离, 提出了一个新的分离性度量方法。最后, 从模糊紧致性和分离性度量方法出发, 提出了 MAME 指标。通过大量的实验, 在多个数据集上均验证了所提指标较以往的聚类有效性指标有明显的性能改善, 特别是面对复杂多样的数据集时, 也能表现良好。这进一步证明了新提出的聚类有效性指标, 即 MAME 指标的合理性和有效性。

1　相关工作

近年来, 大量的研究致力于设计有效的模糊聚类有效性指标。它们的目的是通过对聚类结果进行有效性评估从而能够更好地进行聚类。这里介绍 3 种比较典型的指标。

XBI 指标[28] 是目前应用较为广泛的指标之一。XBI 指标将类内紧致性与类间分离性的比值作为其结果, 计算方式简单, 且准确率较高。其使用数据簇内的数据点到其聚类中心的距离之和来度量类内紧致性, 类间分离性由 N 倍的最小数据簇之间的距离来表示。具体公式为

$$X_{\mathrm{BI}^{(-)}}=\frac{\sum_{k=1}^{K}\sum_{i=1}^{N}\mu_{ik}^{m}d(x_i,v_k)}{N\times\wedge_{i\neq j}d(v_i,v_j)} \tag{1}$$

式中: K是数据集的类别数; N是数据集中数据样本的个数; m表示模糊加权系数; x_j表示数据集中第j个样本点; v_i和v_j分别表示数据集的第i个和第j个聚类中心; $\wedge$表示对数据集取最小值; v_k表示数据集的第k个聚类中心; u_{ik}表示第i个样本点属于第k个聚类中心的隶属度。XBI 指标不仅考虑了数据集内部的几何结构信息, 还考虑了模糊隶属度信息, 能较为全面的评价一个聚类算法的优劣。但 XBI 指标有个缺点, 当划分的类的数量不断增加时, XBI 指标函数的值会不断减小, 当 K无限接近于N时, 其值会无限趋向于 0, 在这种情况下, 该指标就不能有效评价。

WLI 指标[30] 的具体定义为

$$W_{\mathrm{LI}}^{(-)}=\frac{\sum_{k=1}^{K}\left(\frac{\sum_{i=1}^{N}\mu_{ik}^{2}\|x_i-v_k\|^2}{\sum_{i=1}^{N}\mu_{ik}}\right)}{\wedge_{i\neq j}\{\|v_i-v_j\|^2\}+\triangleleft_{i\neq j}\{\|v_i-v_j\|^2\}} \tag{2}$$

其中, $\triangleleft$表示数据集取平均值, 其他符号(比如 K、N等)的定义和前面一样。并且,

$$\triangleleft_{i\neq j}\left\{\left\|v_i-v_j\right\|^2\right\} \tag{3}$$

表示两聚类中心间的距离的中值, 而

$$\wedge_{i\neq j}\left\{\left\|v_i-v_j\right\|^2\right\} \tag{4}$$

表示最小的聚类中心间的距离值。为了面对簇中心分布密集的数据集时也有不错的表现, WLI 指标引入了中值距离, 但其对于含有噪声的数据集表现依然不佳。

I指标[31] 的具体定义为

$$I^{(+)}=\left(\frac{1}{K}\times\frac{E_1}{E_K}\times D_K\right)^{p} \tag{5}$$

其中, K是数据集的类别数, 且

$$E_K=\sum_{k=1}^{K}\sum_{i=1}^{N}\mu_{ik}\|x_i-v_k\| \tag{6}$$

p是小于 1 的任意实数。该指标由 3 部分构成, 彼此之间相互竞争和制衡, 但其十分依赖于初始值的设定, 导致结果具有很大的不确定性。

根据以上对于指标的简要介绍, 这里分析其存在的问题。XBI 指标对聚类划分的质量进行了详细评价, 但是当聚类数 K 非常大并趋向数据样本总数 N 时, 效果不理想。WLI 指标加入了中值距离, 可以防止聚类中心分布较近时指标值大的情况, 但是对于噪声点适应性不佳。I 指标中隶属度、聚类中心、两个集群之间最大分离度三者相互竞争, 但是对于密集型簇的效果不明显, 且依赖于初始值的设定。

表 1 是各类指标的优缺点汇总。总体而言, 对于引言部分提及的 2 个问题(即对于类内紧致性的刻画不太到位、对于类间分离性的度量刻画不够准确), 这些指标都难以有效地解决。

表 1　各类指标的优缺点对比

Table 1　Advantages and disadvantages of various indicators

指标	提出年份	优点	缺点
XBI	1991	对聚类划分的质量进行了详细评价	当聚类数K非常大并趋向数据样本总数N时, 指标的值会和聚类数成反比
WLI	2015	加入了中值距离, 能够防止聚类中心分布很近的时候, 指标函数值很大的情况	对有大量数据噪声点的数据集来说, 此指标的性能不佳
I	2002	隶属度、聚类中心、两个集群之间最大分离度三者相互竞争	只考虑到聚类最大分离度, 对密集型簇的分类评价效果不明显, 且非常依赖于参数初始值的设定

2　新的聚类有效性指标

以往的模糊聚类有效性指标或多或少地会存在一些问题，并不能有效地对聚类结果进行评估，从而难以有效地指导聚类。例如，CH 指标和 DB 指标，主要的评判依据是数据集的结构信息，并没有考虑模糊隶属度，虽然也可以应用于模糊聚类算法中，但其准确性以及使用范围都大打折扣。这两个指标在一般数据集中表现良好，但是当遇到较复杂的数据集，比如噪声点较多或者是数据簇相互之间重叠度较大的时候，得不到较好的结果。

又例如 PE 指标，其仅仅考虑了模糊聚类的隶属度信息，该指标会随着聚类数的变化而单调变化，虽指标形式简单，易于计算，但是数据集只要稍微复杂一点，就不能达到理想的效果。作为改进，NPC 与 NPE 指标在一定程度上缓解了 PC 与 PE 指标的单调性问题，但是最终的聚类评价效果还是不理想。FS、XBI、WLI 等 3 种指标都是基于数据集的内部几何结构信息与隶属度信息，相对于其他指标而言，更为全面和综合，但他们计算量较大，运算较为复杂。I指标中的 3 个因子之间能够相互协调和制衡，但是它过度地依赖于初始值地设定，导致结果具有很大的不确定性。

所以这里，我们提出了一个新的聚类有效性评价指标，即 MAME 指标。该指标基于类内紧致性和类间分离性两个方面，又综合考虑了以往指标存在的问题，尽可能地简化指标的运算复杂度，使得指标即使处理复杂数据集时也能得到较满意的效果，能够更加清晰、准确地评价聚类结果。

2.1　紧致性度量分析

紧致性用来衡量类中每个数据样本之间的紧密程度，一个好的划分就要求类内的数据点尽可能地紧密，而类间的数据点尽可能分离。

在参考了 WLI、I、XBI 指标中关于紧致性的度量之后，此处给出新的模糊紧致性度量表达式为

$$j_{zx}(K)=\frac{E_K}{E_1} \tag{7}$$

其中，

$$E_K=\sum_{k=1}^{K}\frac{\sum_{i=1}^{N}\mu_{ik}^{m}\|x_i-v_k\|^2}{\sum_{i=1}^{N}\mu_{ik}} \tag{8}$$

其中，$\sum_{i=1}^{N}\mu_{ik}$是模糊基数。E_1是E_K当$k=1$时的值（即整个数据集分为 1 类的情形），反应了整个数据集的综合特征。此外，

$$\sum_{i=1}^{N}\mu_{ik}^{m}\|x_i-v_k\|^2 \tag{9}$$

称为类内平方误差和。这里实际上考虑了分别分为 K 类和 1 类的情况的比值（即E_K和E_1的比值）。

一般情况下，类内误差平方和通常会随着 K 的增加而减小。即当类别划分数增加时，每类对应的类内平方误差和就越小。模糊紧致性衡量的是一个数据簇内部的数据样本之间的紧凑程度。数据簇内部的数据样本之间越紧凑，就说明该划分聚类效果好，所以，一般来讲，j_{zx} 的值越小越好。

2.2　分离性度量分析

分离性用来衡量每个划分之间的分离程度，两个聚类之间的分离性越大说明划分效果越好。因此，为了提高聚类有效性指标的性能，就需要重新设计一种新的分离性度量表达式，以便更精准地度量类之间的分离性。

我们提出得到新的分离性度量方法为

$$f_{lx}(K)=\frac{\vee_{i\neq j}\left\|v_i-v_j\right\|^2+\exists_{i\neq j}\left\|v_i-v_j\right\|^2}{K} \tag{10}$$

式中：N 是数据集中数据样本的个数，v_i表示第 i 个聚类中心，同理v_j表示第 j 个聚类中心；$\vee$表示对数据集取最大值，$\exists$表示对数据集取均值。其中，K是为了防止聚类中心之间的最小距离和平均距离太小而导致的指标值过大的情况。

新的分离性度量表达式不仅考虑了数据簇中心之间的最大距离，还引入了数据簇中心之间的平均距离，使得新提出的聚类有效性指标的性能更加的准确和稳定。

2.3　MAME 函数表达式

根据前两节提出的新的模糊紧致性度量表达式和分离性度量方法，本节提出一种新的聚类有效性指标，即 MAME 指标。新的聚类有效性指标从分离性和紧致性两个角度上评估聚类结果的有效性，具体公式如下：

$$M_{AME}(K)=\left[\frac{f_{lx}(K)}{j_{zx}(K)}\right]^p \tag{11}$$

或者，可以重写为

$$M_{AME}(K)=\left[\frac{\vee_{i\neq j}\left\|v_i-v_j\right\|^2+\exists_{i\neq j}\left\|v_i-v_j\right\|^2}{K\times E_K/E_1}\right]^p \tag{12}$$

式中：p是一个不小于 1 的任意实值，这里p取 2。紧致性衡量的是数据簇内部的的紧致程度，即一个簇中所有数据样本的分布是否比较紧密，越紧密则说明该划分效果较好；分离性衡量的是数据集中不同簇之间的分布情况，两个簇之间相隔的

越远, 则越能说明聚类的结果较优。本文中提出的新的聚类有效性指标表现为簇间分离性和簇内紧致性的比值, 由此分析, 当聚类数目确定时, MAME 指标的值越大, 则说明聚类的效果越好。

2.4 MAME 计算算法

如下给出 MAME 的计算算法。其过程基本为: 首先进行 FCM 算法的迭代, 然后进行 MAME 公式的计算, 并且发现最大值对应的聚类数, 即为本算法对应的最优聚类数。

算法 1　有效性指标 MAME 的计算算法

输入　最大迭代次数M, 迭代停止的误差ε, 隶属度矩阵$\boldsymbol{U}=[u_{ij}]$,最小聚类数$K_{\min}$和最大聚类数$K_{\max}$。

输出　每种聚类数所对应的 MAME 指标值。

1)设定M、ε、最大聚类数$K_{\max}$。初始化隶属度矩阵$\boldsymbol{U}=[u_{ij}]$(注意需要满足$\sum_{i=1}^{c}u_{ij}=1$), 令初始迭代次数$k=0$, $K_{\min}=2$, $K=K_{\min}$, $m=2$。

2)更新隶属度矩阵 $\boldsymbol{U}$:

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{K}\left(\frac{x_i-v_k}{x_i-v_j}\right)^{\frac{2}{m-1}}} \tag{13}$$

3)更新聚类中心 V:

$$v_k=\frac{\sum_{i=1}^{N}\mu_{ik}^{m}x_i}{\sum_{i=1}^{N}\mu_{ik}^{m}} \tag{14}$$

4)令$k=k+1$。

5)如果$\left\|V^{(k+1)}-V^{(k)}\right\|\geqslant\varepsilon$且$k<M$, 返回 2)。否则, 继续。

6)用式(7)计算紧致性。

7)用式 (10) 计算分离性。

8)利用式 (11) 得到 MAME 的值。

9)$K=K+1$。

10)如果$K\leqslant K_{\max}$, 返回 2)。否则, 继续。

11)找出 $M_{\mathrm{AME}}(K)$ 的最大值, 该值对应的 K 即为最优划分数。

12)结束。

3 仿真实验与分析

为了证明新提出的聚类有效性指标 MAME 指标的合理性和有效性, 采用模糊 C 均值算法 FCM 来进行实现验证。首先在不同的 K 值下运行 FCM 算法, 然后使用指标逐个检验聚类结果, 最终选取最优指标值所对应的 K 值即为聚类的最优划分数。在此实验中, 使用的环境为 Intel(R) Core(TM)i5-4200 U CPU @ 1.60 GHz 以及 RAM 4.00 GB 和 Windows 7 旗舰版, 编程软件采用 VC++6.0。

3.1 数据集与对比指标

本次实验挑选了 11 个数据集, 即 Flame、Banknote、Habe、Jain、WDBC 和 AD1、AD2、AD3、AD4、Data-E6、Data-Fc1。前 5 个数据集为 UCI 数据集[32], 来自真实世界的真实数据; 后 6 个数据集是人工数据集(来自于文献 [30]), 分布较复杂且含有很多噪声点, 可以从不同角度对指标评估。UCI 数据集中, Flame 有 240 个数据样本, 每个数据样本具有 2 个属性, 共有 2 类; Banknote 是从纸币鉴别过程中提取出来的数据集, 数据量为 1 372, 共 4 个属性, 分为 2 类; Habe 的数据量为 306, 有 2 个属性, 分为 2 类; Jain 的数据量为 373, 共 2 个属性, 分为 2 类; WDBC 描述的是乳腺癌诊断的信息, 数据量为 569, 共有 30 个属性, 分为 2 类。在人工数据集中, 下述 4 个数据集的数据点均有 2 个属性, 其中, AD1 的数据量为 800, 分为 4 类; AD2 的数据量为 300, 分为 3 类; AD3 的数据量为 300, 分为 3 类; AD4 的数据量为 450, 分为 3 类; 而 Data-E6 含有 8 537 个数据样本, 具有 2 个属性, 分为 4 类; Data-Fc1 含有 1053 个数据样本, 具有 2 个属性, 分为 5 类。

在这些数据集上对新提出的指标进行实验并与多种有效性指标进行比较。采用了 9 个具有代表性的指标, 分别是 $\mathrm{CH}^{(+)}$ 指标、$\mathrm{DB}^{(-)}$ 指标、$\mathrm{PE}^{(-)}$ 指标、$\mathrm{NPC}^{(+)}$ 指标、$\mathrm{NPE}^{(-)}$ 指标、$\mathrm{FS}^{(-)}$ 指标、$\mathrm{XBI}^{(-)}$ 指标、I 指标 (+)、$\mathrm{WLI}^{(-)}$ 指标。这里 (+) 表示该指标的值越大越好, 即最优值对应最大的值。同时 (−) 表示指标的值越小越好, 即最优值对应最小的值。

3.2 面向 UCI 数据集的实验

由于 FCM 聚类算法的初始聚类中心是随机初始化的, 因此可能每次得到的聚类结果都不一样, 进而导致对于聚类数 K, 聚类有效性指标函数的最优值可能也不一样。所以为了保证实验结果的稳定性, 将每个实验在不同的 K 值下重复运行 10 次。每一轮的每一个聚类有效性指标都会记录一个最大或者最小值, 该最大或是最小值所对应的 K 值即为最优划分数。每一轮结束后, 都会统计出一个最优值, 10 次实验结束后, 同一个评价指标会得到 10 个极值所对应的 K 值, 用 K^{*}表示其统计结果。若同一个指标的 10 轮结果中极值所对应的 K 值都相同, 则说明此指标的稳定性较强, 正确率较高。

在 UCI 数据集上实验的结果如表 2 所示, 其中, X^{Y}表示 K=X 的值出现了 Y 次。

表 2　UCI 数据集上实验的结果

Table 2　Results of experiments on UCI datasets

数据集	$PE^{(-)}$	$NPE^{(-)}$	$NPC^{(+)}$	$FSI^{(-)}$	$XBI^{(-)}$	$CHI^{(+)}$	$DBI^{(-)}$	$WLI^{(-)}$	$I^{(+)}$	$MAME^{(+)}$
Flame	$2^9 4^1$	2^{10}	$4^9 2^1$	$8^7 2^3$	4^{10}	4^{10}	4^{10}	$4^2 5^8$	$2^9 5^1$	2^{10}
Banknote	2^{10}	2^{10}	2^{10}	2^{10}	$9^8 2^2$	2^{10}	$2^2 4^8$	2^{10}	$3^7 4^3$	2^{10}
Habe	2^{10}	2^{10}	$2^8 3^2$	$7^8 2^2$	$4^9 5^1$	4^{10}	$4^6 3^4$	2^{10}	3^{10}	2^{10}
Jain	$2^8 3^2$	2^{10}	$5^9 2^1$	$8^6 2^4$	$6^6 5^4$	10^{10}	8^{10}	5^{10}	$2^2 3^8$	2^{10}
WDBC	2^{10}	2^{10}	2^{10}	9^{10}	$10^8 6^2$	$9^6 2^4$	2^{10}	2^{10}	$4^9 2^1$	$2^8 3^2$

以 Flame 为例，其数据量为 240，具有 4 个属性，共分为 2 类。由表 2 中数据可知，运行了 10 次后，PE 指标的结果中仅出现了一次 K 为 4 是最优值的情况，其余的 9 次得到的最优值均为 2；NPE 指标的结果中，最优值均是 2；NPC 指标的结果中，只有一次划分为 2 类，其余均划分为 4 类；FSI 指标的结果中，有 3 次的划分结果为 2，7 次的划分结果为 8；XBI 指标的结果中，最优值都是 4；CHI 指标的结果中，最优值都是 4 类；DB 指标的结果中，均划分为 4 类；WLI 指标的结果中，有 2 次得到的最优值是 4，8 次划分为 5 类；I 指标的结果中，9 次划分为 2 类，一次划分为 5 类；在 MAME 指标的结果中，最优值均为 2。最终选取出现次数最多的最优值所对应的 K 值作为评价指标最终的结果，见表 3。其中，结果右上角带“*”的表示本次结果不是最优划分。Best 列表示数据集的真实聚类数。

表 3　UCI 数据集上指标得到的最优值

Table 3　Optimal value of the index on UCI datasets

数据集	最优值	$PE^{(-)}$	$NPE^{(-)}$	$NPC^{(+)}$	$FSI^{(-)}$	$XBI^{(-)}$	$CHI^{(+)}$	$DBI^{(-)}$	$WLI^{(-)}$	$I^{(+)}$	$MAME^{(+)}$
Flame	2	2	2	4^*	8^*	4^*	4^*	4^*	5^*	2	2
Banknote	2	2	2	2	9^*	9^*	2	4^*	2	3^*	2
Habe	2	2	2	2	7^*	4^*	4^*	4^*	2	3^*	2
Jain	2	2	2	5^*	8^*	6^*	10^*	8^*	5^*	3^*	2
WDBC	2	2	2	2	9^*	10^*	9^*	2	2	4^*	2

由表 3 中数据可知，PE 指标、NPE 指标和 MAME 指标在数据集 Flame、Banknote、Habe、Jain 和 WDBC 上均得到了正确的聚类结果；而 NPC 指标只在数据集 Banknote、Habe 和 WDBC 上得到了正确的结果；FSI 指标和 XBI 指标在数据集 Flame、Banknote、Habe、Jain 和 WDBC 上划分错误；CH 指标仅在数据集 Banknote 上实现了正确划分，在其余数据集上均判断错误；DBI 指标仅在数据集 WDBC 上实现了正确划分，而在其余数据集上均判断错误；WLI 指标则在数据集 Banknote、Habe 和 WDBC 上实现了正确划分，而在其余 2 个数据集上判断错误；I 指标仅在数据集 Flame 上得到了正确划分，而在其余数据集上均判断错误。由此可见，新提出的 MAME 指标的正确率为 100%，而其他的指标在面对稍微复杂的数据集时，就会出现准确率不高、且不稳定的缺点。

每个指标的 K 从 2 ~ 10 得到的结果在 Flame 数据集上的变化情况见图 1。图 1 中，横坐标表示的是聚类数 K 值，纵坐标表示对应的评价指标。星标记的是每个指标函数值的极值所对应的最优分类数。由图可知，只有 NPE、PE、I 和 MAME 指标在 K 为 2 时的极值是正确的，其他的指标均有偏差。可以看出，各指标的最优值出现的 K 值不同，且每个指标算法收敛时所取的最优值的收敛方向不同，评价函数在取不同的 K 值时，都会对应一个值。各个指标选取结果中的最大或最小值来作为这个聚类评价函数的最优值的情况各不相同。例如，CH 指标、NPC 指标、I 指标均是取评价函数结果中最大值作为最优值；其余指标均是取评价函数结果中最小值为最优值。而新提出的 MAME 指标则是取评价函数结果中最大值为最优值。

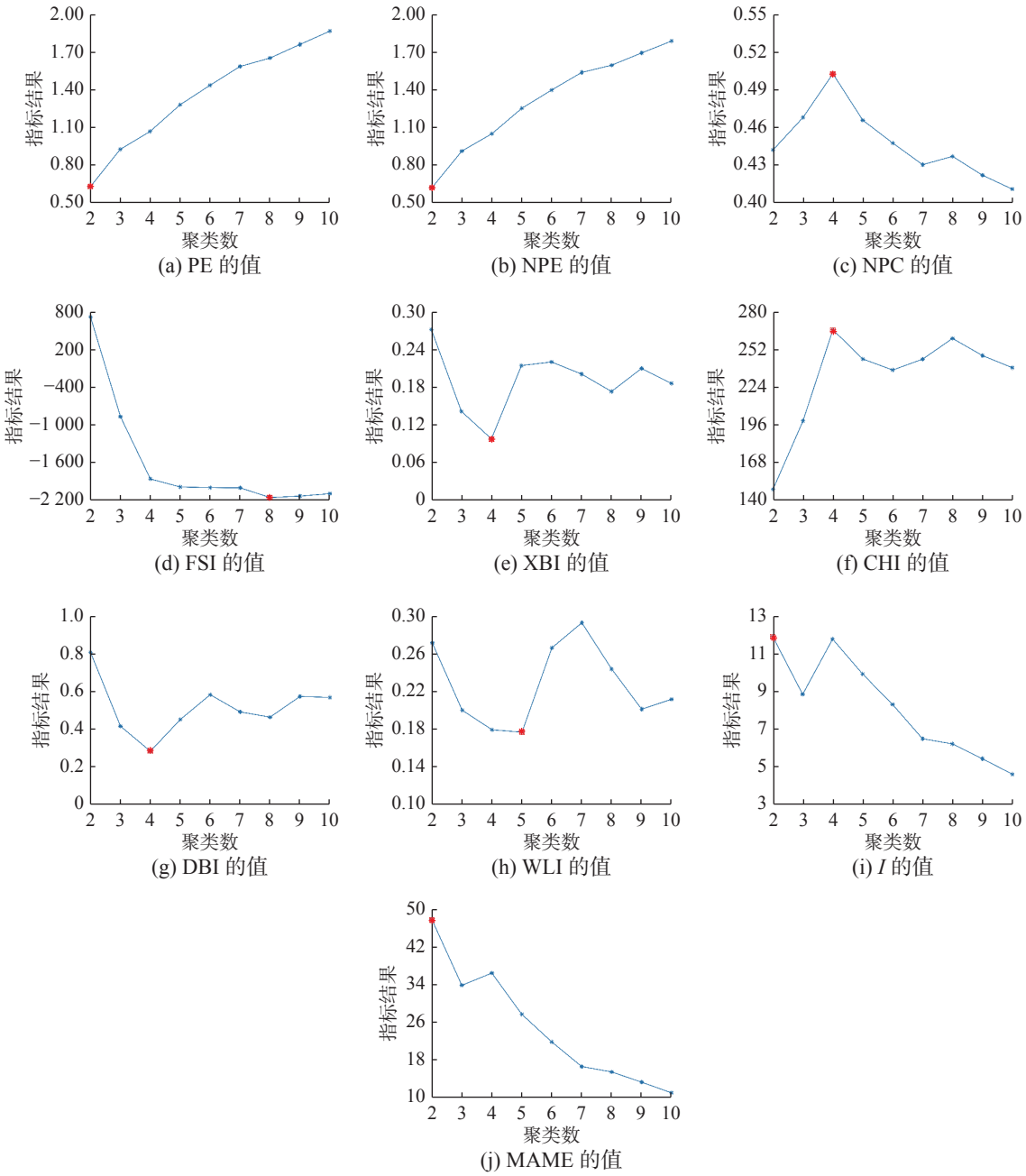


图 1　各指标在 Flame 数据集上的结果对比

Fig. 1　Comparison of results for each metric on the Flame dataset

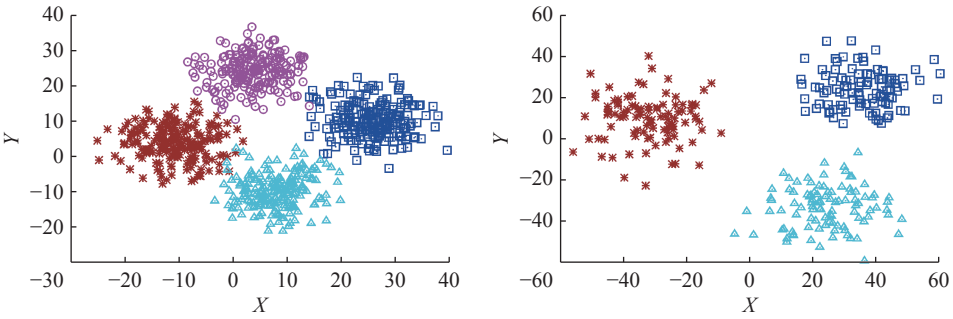
图 1 和表 2~3 中的数据均清晰地显示了新提出的聚类有效性指标对于 UCI 真实数据集表现出了较高的准确率和稳定性。

3.3　面向人工数据集的实验

图 2 给出了 4 个人工数据集 AD1~AD4[28] 的分布情况，图 3 给出了 2 个人工数据集 E6 和 Fc1 的分布情况，人工数据集 AD1~AD4 以及 E6、Fc1 上的实验结果如表 4 所示，X^Y表示 $K=X$ 的值出现了 Y 次。

表 4 是人工数据集 AD1~AD4、E6 和 Fc1 的运行结果。Best 列表示数据集的真实聚类数。以 AD1 数据集为例，AD1 的数据量为 800，具有 2 个属性，共分为 4 类。由表 4 中数据可知，运行了 10 次实验后，PE 指标的结果中，得到的最优值均为 2 类；NPE 指标的结果中，最优值都是 2；NPC 指标的结果中，有 9 次得到了正确划分，剩余 1 次的最优值为 2；FS 指标的结果中，有 7 次实现了正确划分，而 3 次得到了最优值是 2 类；XBI 指标的结果中，最优值都是 4 类；CHI 指标的结果中，最优值都是 4；DBI 指标的结果中，最优值均为 4 类；WLI 指标的结果中，最优值均为 4 类；I 指标的结果中，有 9 次的最优值是 4，1 次最优值是 3 类；而

MAME 指标的 10 次结果中，得到的最优值都是 4 类。最终选取出现次数最多的最优值所对应的 K 值作为评价指标最终的结果，结果见表 5。其中，结果右上角带“*”的表示本次结果不是最优划分。

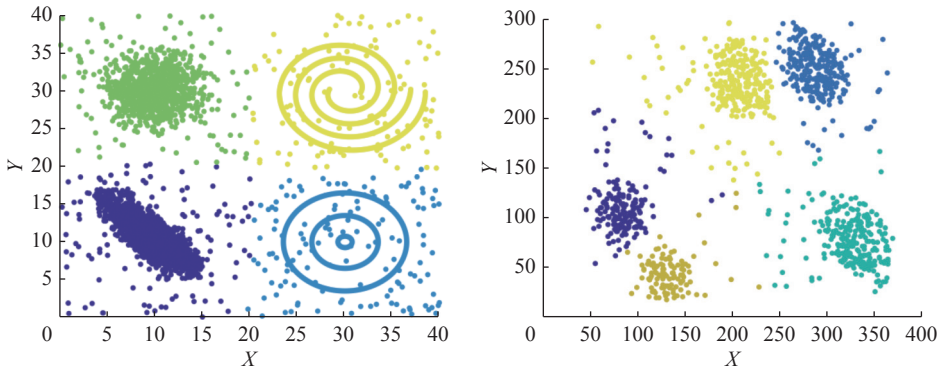


(a) AD1 数据集　(b) AD2 数据集

(c) AD3 数据集　(d) AD4 数据集

图 2　4 个人工数据集

Fig. 2　Four artificial datasets



(a) E6 数据集　(b) Fc1 数据集

图 3　人造数据集 E6 和 Fc1 的分布

Fig. 3　The distribution of artificial datasets E6 and Fc1

表 4　人工数据集上实验的结果

Table 4　Results of experiments on artificial datasets

数据集	最优值	PE$^{(-)}$	NPE$^{(-)}$	NPC$^{(+)}$	FSI$^{(-)}$	XBI$^{(-)}$	CHI$^{(+)}$	DBI$^{(-)}$	WLI$^{(-)}$	$I^{(+)}$	MAME$^{(+)}$
AD1	4	2^{10}	2^{10}	$4^{9}2^{1}$	$4^{7}2^{3}$	4^{10}	4^{10}	4^{10}	4^{10}	$4^{9}3^{1}$	4^{10}
AD2	3	3^{10}	$3^{9}2^{1}$	3^{10}	$5^{9}4^{1}$	3^{10}	3^{10}	$2^{1}3^{9}$	3^{10}	$3^{7}4^{3}$	3^{10}
AD3	3	$3^{1}2^{9}$	2^{10}	$2^{8}3^{2}$	$9^{8}2^{2}$	3^{10}	$8^{9}2^{1}$	3^{10}	3^{10}	3^{10}	3^{10}
AD4	3	3^{10}	3^{10}	$3^{9}2^{1}$	5^{10}	3^{10}	5^{10}	3^{10}	3^{10}	5^{10}	3^{10}
E6	4	$4^{8}2^{2}$	$4^{7}2^{3}$	$4^{7}2^{3}$	$5^{8}6^{2}$	$4^{8}2^{2}$	$4^{6}2^{4}$	$4^{7}2^{3}$	$4^{9}2^{1}$	$4^{9}6^{1}$	4^{10}
Fc1	5	3^{10}	2^{10}	$2^{6}4^{4}$	$5^{8}4^{2}$	$5^{8}6^{2}$	2^{10}	$3^{7}2^{3}$	$5^{9}3^{1}$	5^{10}	5^{10}

表 5　人工数据集上指标得到的最优值

Table 5　Optimal values of the indexes on artificial datasets

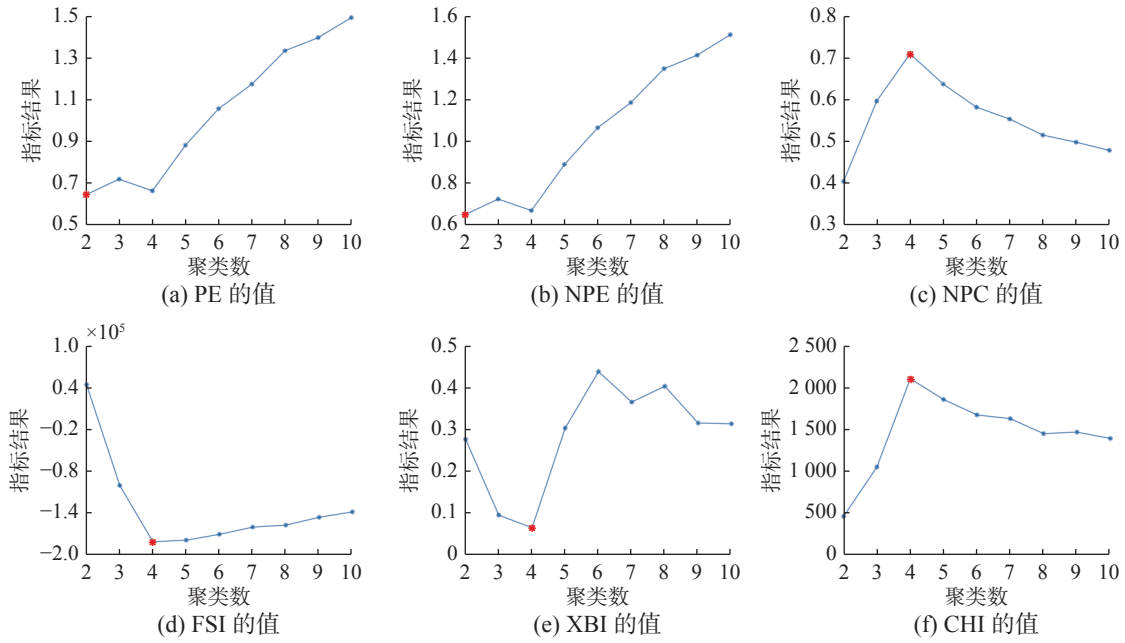
数据集	最优值	PE$^{(-)}$	NPE$^{(-)}$	NPC$^{(+)}$	FSI$^{(-)}$	XBI$^{(-)}$	CHI$^{(+)}$	DBI$^{(-)}$	WLI$^{(-)}$	$I^{(+)}$	MAME$^{(+)}$
AD1	4	2*	2*	4	4	4	4	4	4	4	4
AD2	3	3	3	3	5*	3	3	3	3	3	3
AD3	3	2*	2*	2*	9*	3	8*	3	3	3	3
AD4	3	3	3	3	5*	3	5*	3	3	5*	3
E6	4	4	4	4	5*	4	4	4	4	4	4
Fc1	5	3*	2*	2*	5	5	2*	3*	5	5	5

由表 5 中可知, XBI 指标、WLI 指标和 MAME 指标在数据集 AD1、AD2、AD3、AD4、E6 和 Fc1 上全都实现了正确划分; DBI 指标在数据集 AD1、AD2、AD3、AD4 和 E6 上实现了正确划分, 仅在 Fc1 上没有实现正确划分; I 指标在数据集 AD1、AD2、AD3、E6 和 Fc1 上实现了正确划分, 仅在 AD4 上没有实现正确划分; NPC 指标在数据集 AD1、AD2、AD4 和 E6 上实现了正确划分, 而在其余数据集上划分错误; PE 指标和 NPE 指标在数据集 AD2、AD4 和 E6 上实现了正确划分, 其余数据集上判断错误; CHI 指标仅在数据集 AD1、AD2 和 E6 上划分正确, 其余数据集上均划分错误; FS 指标仅在数据集 AD1 和 Fc1 上实现了正确划分, 而在其余数据集上均判断错误。观察以上结果可知, 有些指标在面对稍微复杂的数据集时, 就会出现准确率不高且不稳定的缺点, 效果有点差强人意。

每个指标的 K 从 2~10 得到的结果在 AD1 数据集上的变化情况见图 4。图 4 中, 横坐标表示的是聚类数 K 值, 纵坐标表示对应的评价指标。星标记的是每个指标函数值的极值所对应的最优分类数。除 PE、NPE 指标在 K 为 4 时得到的极值是错误的外, 其他的指标都得到了正确的分类结果。可以看出, 新提出的聚类有效性指标, 即 MAME 指标在 6 个人工数据集中都取得了正确的分类结果。

观察表 4 和表 5 中结果可知, 新提出的 MAME 聚类有效性指标对于人工数据集确实有较高的有准确率和稳定性。

此外, 我们还将 MAME 指标分别与其他指标的实验结果进行了克鲁斯卡尔−沃利斯检验 (Kruskal-Wallis test)。该指标用于检验 MAME 与其他聚类有效性指标是否有显著性差异, 若克鲁斯卡尔−沃利斯检验的最终值小于 0.05, 则说明两指标间存在显著性差异。表 6 列出了 MAME 指标与其他指标之间克鲁斯卡尔−沃利斯检验的最终结果。例如, MAME 与 FSI 指标间的结果值为 $1.653\,1\times10^{-4}$, 说明两指标间的差异是显著的。总体而言, MAME 指标与其他指标之间存在着差异。



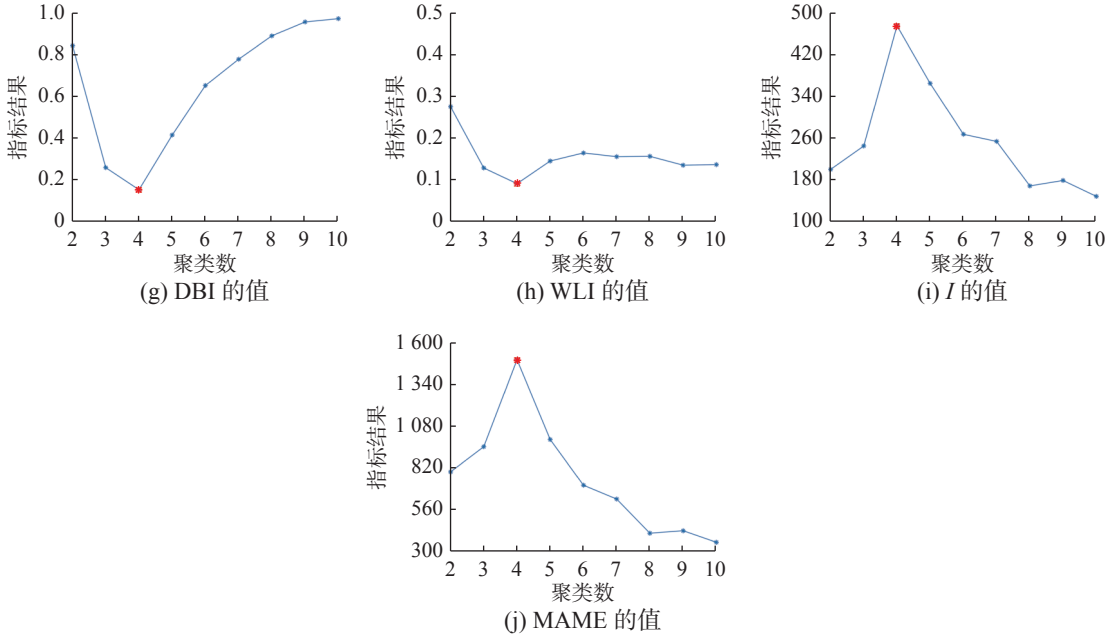


图 4　各指标在 AD1 数据集上的结果对比

Fig. 4　Comparison of the results of each index on AD1 dataset

表 6　MAME 和其他指标之间的 Kruskal-Wallis 检验结果

Table 6　Final values of Kruskal-Wallis tests between MAME and every other CVI.

指标	$PE^{(-)}$	$NPE^{(-)}$	$NPC^{(+)}$	$FSI^{(-)}$	$XBI^{(-)}$	$CHI^{(+)}$	$DBI^{(-)}$	$WLI^{(-)}$	$I^{(+)}$
检验结果	0.292 0	0.169 2	0.861 4	$1.653\,1\times10^{-4}$	0.009 4	0.039 1	0.099 8	0.273 7	0.123 3

4　结束语

近些年来，随着聚类技术的不断发展，各种聚类算法层出不穷，如何评价聚类结果的有效性变成了一个重要的问题。聚类有效性问题就是用来评估聚类结果以及帮助判别聚类的类别数的。但现有的聚类有效性指标对于类内紧致性的刻画不太到位、对于类间分离性的度量刻画不够准确。

为了解决以上的问题，使聚类有效性问题能更加有效地指导聚类过程，本文提出了一个新的模糊聚类有效性指标，即 MAME 指标。针对现有指标对类内紧致性刻画不到位的问题，在考虑了整个数据集的综合特征的基础上，计算分别分为 K 类和 1 类的情况的比值，提出了一个新的模糊紧致性度量表达式。对于类间分离性度量不准确问题，引入最大聚类中心距离和平均聚类中心距离，提出了一个新的分离性度量方法。最后，基于类内紧致性和类间分离性表达式提出了 MAME 指标。

为了证明新指标的可行性与准确性，在多个真实数据集和人工数据集上进行了实验。其结果均验证了所提指标较以往的聚类有效性指标有明显的性能改善，且对不同类型数据集的适应能力较强，表明新的聚类有效性指标可以更准确地指导聚类过程从而得到最优结果。即使面对复杂多样和噪声点较多的数据集时也能得到较满意的效果，能够更加清晰、准确地评价聚类结果，并且适应的范围也比较广泛，准确性较高。这进一步证明了新提出的聚类有效性指标，即 MAME 指标的合理性和有效性。

本文的创新性体现在以下 3 个方面：1）考虑了分别分为 K 类和 1 类的情况的比值，提出了一个新的模糊紧致性度量表达式。2）引入最大聚类中心距离和平均聚类中心距离，提出了一个新的分离性度量方法。3）从模糊紧致性度量表达式、分离性度量方法出发，提出了一个面向模糊聚类的新的聚类有效性指标 MAME。

未来，将基于本文提出的 MAME 指标去设计一种新的模糊聚类算法，进一步提升其在发现最优聚类数方面的能力。进一步将研究适用于多种聚类算法[33-34] 的聚类有效性指标。

参考文献：

[1]　TANG Yiming, PEDRYCZ W. Oscillation-bound estimation of perturbations under bandler-kohout subproduct[J].

IEEE transactions on cybernetics, 2022, 52(7): 6269–6282.

[2] TANG Yiming, PEDRYCZ W, REN Fuji. Granular symmetric implicational method[J]. IEEE transactions on emerging topics in computational intelligence, 2022, 6(3): 710–723.

[3] TANG Yiming, LI Lin, LIU Xiaoping. State-of-the-art development of complex systems and their simulation methods[J]. Complex system modeling and simulation, 2021, 1(4): 271–290.

[4] 唐益明, 刘晓平. 二值命题逻辑的无损求解 [J]. 计算机学报, 2013, 36(5): 1097–1114.
TANG Yiming, LIU Xiaoping. Lossless solving of two-valued propositional logic[J]. Chinese journal of computers, 2013, 36(5): 1097–1114.

[5] 唐益明, 张征, 芦启明. 分段二次方转换函数驱动的高斯核模糊 C 均值聚类 [J]. 山东大学学报(理学版), 2020, 55(3): 107–112,120.
TANG Yiming, ZHANG Zheng, LU Qiming. Gaussian kernel fuzzy C-means clustering driven by piecewise quadratic transfer function[J]. Journal of Shandong university (natural science), 2020, 55(3): 107–112,120.

[6] 唐益明, 赵跟陆, 任福继, 等. 图像分割的 EMKPFC 算法 [J]. 南京大学学报(自然科学), 2017, 53(3): 569–578.
TANG Yiming, ZHAO Genlu, REN Fuji, et al. The EMKPFC algorithm of image segmentation[J]. Journal of Nanjing university (natural sciences), 2017, 53(3): 569–578.

[7] 黄河燕, 曹朝, 冯冲. 大数据情报分析发展机遇及其挑战 [J]. 智能系统学报, 2016, 11(6): 719–727.
HUANG Heyan, CAO Zhao, FENG Chong. Opportunities and challenges of big data intelligence analysis[J]. CAAI transactions on intelligent systems, 2016, 11(6): 719–727.

[8] ZHANG Changqing, FU Huazhu, HU Qinghua, et al. Generalized latent multi-view subspace clustering[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(1): 86–99.

[9] ZHU Shuwei, XU Lihong, GOODMAN E D. Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering[J]. IEEE transactions on cybernetics, 2022, 52(9): 9846–9860.

[10] 贾鹤鸣, 张棕淇, 姜子超, 等. 基于混合身份搜索黏菌优化的模糊 C-均值聚类算法 [J]. 智能系统学报, 2022, 17(5): 999–1011.
JIA Heming, ZHANG Zongqi, JIANG Zichao, et al. An optimization fuzzy C-means clustering algorithm based on the hybrid identity search and slime mold algorithms[J]. CAAI transactions on intelligent systems, 2022, 17(5): 999–1011.

[11] HU Xianghui, TANG Yiming, PEDRYCZ W, et al. Fuzzy clustering with knowledge extraction and granulation[J]. IEEE transactions on fuzzy systems, 2023, 31(4): 1098–1112.

[12] TANG Yiming, HU Xianghui, PEDRYCZ W, et al. Possibilistic fuzzy clustering with high-density viewpoint[J]. Neurocomputing, 2019, 329: 407–423.

[13] TANG Yiming, REN Fuji, PEDRYCZ W. Fuzzy C-Means clustering through SSIM and patch for image segmentation[J]. Applied soft computing, 2020, 87: 105928.

[14] TANG Yiming, PAN Zhifu, PEDRYCZ W, et al. Viewpoint-based kernel fuzzy clustering with weight information granules[J]. IEEE transactions on emerging topics in computational intelligence, 2023, 7(2): 342–356.

[15] OZKAN I, TÜRKŞEN I B. MiniMax ε-stable cluster validity index for Type-2 fuzziness[J]. Information sciences, 2012, 184(1): 64–74.

[16] CAMPBELL T, KULIS B, HOW J. Dynamic clustering algorithms via small-variance analysis of Markov chain mixture models[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(6): 1338–1352.

[17] 谢娟英, 周颖, 王明钊, 等. 聚类有效性评价新指标 [J]. 智能系统学报, 2017, 12(6): 873–882.
XIE Juanying, ZHOU Ying, WANG Mingzhao, et al. New criteria for evaluating the validity of clustering[J]. CAAI transactions on intelligent systems, 2017, 12(6): 873–882.

[18] BAI Liang, LIANG Jiye, CAO Fuyuan. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters[J]. Information fusion, 2020, 61: 36–47.

[19] LI Xiangtao, WONG K C. Evolutionary multiobjective clustering and its applications to patient stratification[J]. IEEE transactions on cybernetics, 2019, 49(5): 1680–1693.

[20] HANDL J, KNOWLES J, KELL D B. Computational cluster validation in post-genomic data analysis[J]. Bioinformatics, 2005, 21(15): 3201–3212.

[21] POPESCU M, KELLER J M, BEZDEK J C, et al. Correlation cluster validity[C]//2011 IEEE International Conference on Systems, Man, and Cybernetics. Anchorage: IEEE, 2011: 2531−2536.

[22] CHEN Na, XU Zeshui, XIA Meimei. Hierarchical hesitant fuzzy K-means clustering algorithm[J]. Applied mathematics-a journal of Chinese universities, 2014, 29(1): 1–17.

[23] CALIŃSKI T, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in statistics, 1974, 3(1): 1–27.

[24] DAVIES D L, BOULDIN D W. A cluster separation

measure[J]. IEEE transactions on pattern analysis and machine intelligence, 1979, PAMI-1(2): 224–227.

[25] BEZDEK J C. Numerical taxonomy with fuzzy sets[J]. Journal of mathematical biology, 1974, 1(1): 57–71.

[26] BEZDEK J C. Cluster validity with fuzzy sets[J]. Journal of cybernetics, 1973, 3(3): 58–73.

[27] ROUBENS M. Pattern classification problems and fuzzy sets[J]. Fuzzy sets and systems, 1978, 1(4): 239–253.

[28] 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.

[29] FUKUYAMA Y, SUGENO M. A new method of choosing the number of clusters for the fuzzy c-means method[C]// Proc 5th Fuzzy Syst Symp. Kobe: IEEE, 1989: 247−250.

[30] WU C H, OUYANG Chensen, CHEN Liwen, et al. A new fuzzy clustering validity index with a Median factor for centroid-based clustering[J]. IEEE transactions on fuzzy systems, 2015, 23(3): 701–718.

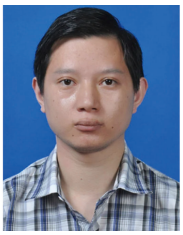
[31] MAULIK U, BANDYOPADHYAY S. Performance evaluation of some clustering algorithms and validity indices[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(12): 1650–1654.

[32] ASUNCION A, NEWMAN D J. UCI Machine Learning Repository [EB/OL]. (2001-1-15)[2020-01-01]. http://archive.ics.usi.edu/ml/datasets.html, 2007.

[33] 王跃, 杨燕, 王红军. 一种基于少量标签的改进迁移模糊聚类 [J]. 智能系统学报, 2016, 11(3): 310–317.
WANG Yue, YANG Yan, WANG Hongjun. An improved transfer fuzzy clustering with few labels[J]. CAAI transactions on intelligent systems, 2016, 11(3): 310–317.

[34] 谷文祥, 郭丽萍, 殷明浩. 模糊 c-均值算法和万有引力算法求解模糊聚类问题 [J]. 智能系统学报, 2011, 6(6): 520–525.
GU Wenxiang, GUO Liping, YIN Minghao. A solution for a fuzzy clustering problem by applying fuzzy c-means algorithm and gravitational search algorithm[J]. CAAI transactions on intelligent systems, 2011, 6(6): 520–525.

作者简介：



唐益明，教授，博士，主要研究方向为聚类、模糊逻辑与系统、情感计算、图像处理、计算机辅助设计。主持国家自然科学基金项目 3 项。发表（或录用）学术论文 80 余篇，授权发明专利 5 项。

陈仁好，硕士研究生，主要研究方向为聚类、粒度计算、情感计算。

李冰，硕士研究生，主要研究方向为聚类、聚类有效性指标、情感计算。