智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

全局数据驱动的模糊聚类有效性评价指标

唐益明, 刘子龙, 高健玮

引用本文:
唐益明, 刘子龙, 高健玮. 全局数据驱动的模糊聚类有效性评价指标[J]. 智能系统学报, 2026, 21(3): 598-616.
TANG Yiming, LIU Zilong, GAO Jianwei. Global data-driven fuzzy cluster validity index[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 598-616.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202507010

DOI: 10.11992/tis.202507010
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20251222.1754.005

全局数据驱动的模糊聚类有效性评价指标

唐益明, 刘子龙, 高健玮
(合肥工业大学 计算机与信息学院, 安徽 合肥 230601)

摘　要: 现有模糊聚类有效性指标 (cluster validity index, CVI) 在处理具有噪声干扰以及簇规模差异较大的数据集时, 往往难以保持准确的评估性能。为此, 本文提出了一个全局数据驱动指标 (global data driven index, GDD), 该指标针对聚类结果中各簇规模的差异, 设计了一套针对簇内紧致度与簇间分离度的基于簇规模的加权机制, 以此来应对该差异对指标结果的影响。GDD 指标设计时采用簇内紧致度与簇间分离度比值的形式, 对于可能会出现的 CVI 结果根据聚类结果数目单调递增的情况进行遏制; 在考虑模糊隶属度的簇内点的平均距离的基础上, 基于不同规模的簇占数据集中的比重不同, 对于较大的簇给予其更高的权重, 在计算单个簇内紧致度的基础上引入了该簇样本数占总样本的比值, 该加权机制将直接影响该簇紧致度结果对于加和过后的总紧致度的贡献, 从而构建了更客观的簇内紧致度的表达; 综合考虑了所有类中心之间的均值与不同规模的簇占数据集的比重, 对于较大的簇给予更高的权重, 在计算每个聚类中心到聚类中心均值的距离的基础上加入了该簇样本数占总样本的比值, 从而锚定了上述距离结果占所有簇加和的距离结果的比重, 这种加权机制构建了更合理的簇间分离度的表达。实验结果显示, GDD 能够很好地适应各种模糊聚类算法, 而且在面对复杂结构和噪声时表现出较强的鲁棒性。本文提出的 GDD 指标可以在复杂结构与噪声环境下较好地完成对各类模糊聚类算法的评价。

关键词: 聚类算法; 模糊聚类; 聚类有效性指标; 模糊 C 均值算法; 加权紧致度; 加权分离度; 鲁棒性; 抗噪声
中图分类号: TP181; TN99　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)03−0598−19

中文引用格式: 唐益明, 刘子龙, 高健玮. 全局数据驱动的模糊聚类有效性评价指标 [J]. 智能系统学报, 2026, 21(3): 598–616.
英文引用格式: TANG Yiming, LIU Zilong, GAO Jianwei. Global data-driven fuzzy cluster validity index[J]. CAAI transactions on intelligent systems, 2026, 21(3): 598–616.

Global data-driven fuzzy cluster validity index

TANG Yiming, LIU Zilong, GAO Jianwei
(School of Computer Science and Information, Hefei University of Technology, Hefei 230601, China)

Abstract: Existing cluster validity index(CVI) for fuzzy clustering often struggle to maintain accurate evaluation performance when handling datasets containing noise interference or exhibiting significant differences in cluster sizes. To address these limitations, this study proposes a global data-driven(GDD) index. The GDD index incorporates a scale-aware weighting mechanism for intra-cluster compactness and inter-cluster separation to mitigate the adverse impact of imbalanced cluster sizes on validity assessment. First, the GDD index adopts a ratio-based formulation of intra-cluster compactness to inter-cluster separation. This design prevents the undesirable monotonic increase of the index value as the number of clusters grows. Second, to obtain a more objective measure of intra-cluster compactness, the index computes the average distance among data points within each cluster, incorporating fuzzy membership degrees. Crucially, recognizing that clusters of different scales contribute unequally to the overall dataset structure, larger clusters are assigned higher weights. Specifically, the ratio of the number of samples in each cluster to the total number of samples is introduced into the compactness calculation. This weighting scheme directly influences each cluster's contribution to the overall compactness, thereby enhancing representational fairness and accuracy. Third, for inter-cluster separation, the index comprehensively considers both the centroid distribution and the relative sizes of different clusters. Rather than treating all centroids equally, the index assigns higher weights to centroids of larger clusters. When computing the distance from each cluster centroid to the mean of all centroids, the sample-size ratio of the corresponding cluster is incorporated. This adjustment anchors the contribution of each centroid's distance to the total separation measure, resulting in a more reasonable and balanced expression of inter-cluster separation. To evaluate the effectiveness and robustness of the GDD index, extensive experiments were conducted using three representative fuzzy clustering algorithms. Experimental results demonstrate that the GDD index consistently identifies the optimal number of clusters with high accuracy, adapts well across various fuzzy clustering frameworks, and demonstrates strong robustness in challenging scenarios with noise and highly imbalanced cluster sizes. The proposed index provides a more comprehensive and reliable evaluation of fuzzy clustering algorithms in complex, noisy environments.

Keywords: clustering algorithms; fuzzy clustering; clustering validity index; fuzzy C-means algorithms; weighted compactness; weighted separation; robustness; noise robustness

聚类作为一种无监督的机器学习方法[1-2], 其在数据处理和数据分析方向的能力适用于处理不

收稿日期: 2025−07−07.　**网络出版日期**: 2025−12−23.
基金项目: 国家自然科学基金项目 (62576130, 62176083).
通信作者: 高健玮. E-mail: jwgao810@163.com.

同领域和类型的复杂数据[3], 尤其在数据挖掘、模式识别和图像处理等领域得到了广泛的研究和应用。聚类是将一个未标识的数据集划分成不同类簇的过程, 尽可能使相似的数据被划分到同一个簇中, 不相似的数据划分到不同的簇中[4]。聚类在研究和应用过程中有两个重要问题需要考虑: 一是聚类算法, 通过聚类算法来划分数据集, 得到聚类结果; 二是聚类验证[5], 通过评估聚类结果进而评估聚类算法的效果。

随着计算机科学的发展, 尤其是大数据和人工智能的兴起, 聚类算法得到了长足的发展和广泛的应用。经典的 K 均值聚类算法 (K-Means)[6] 和层次聚类方法[7] 最早被提出, 随后催生了基于密度[8-9]、网格和模型等的聚类方法[10-12]。以上这些通常都是硬聚类算法, 在划分过程中要求每个数据点只能划分到一个唯一的簇中。这种非此即彼的划分方式难以适用某些不确定的数据集, 在这种特殊的情况下, 数据点可以同时归属于多个簇, 具有一定的模糊性。这种需求催生了模糊聚类算法。Dunn[13] 于 1973 年提出了模糊 C 均值聚类算法 (fuzzy C-means, FCM), 后续 Bezdek 等[14] 在此基础上发展并推广了 FCM 算法。FCM 是模糊聚类的代表算法, 通过优化隶属度函数来计算每个数据点对于簇的隶属度。此算法成为模糊聚类的重要基础, 在处理不确定性数据方面有了显著进展。但 FCM 也存在缺点, 如对噪声数据敏感。针对这一问题, Krishnapuram 等[15] 提出了可能性 C 均值聚类算法 (possibilistic C-means, PCM), PCM 引入了可能性隶属度, 而非单一的隶属度, 放宽了对隶属度的限制, 赋予噪声点更低的隶属度, 从而增强了聚类的鲁棒性。后续随着模糊聚类应用需求的增加, 研究人员提出了更多种类的模糊聚类算法。例如, Pal 等[16] 提出的模糊可能性 C 均值聚类算法 (fuzzy possibilistic C-means, FPCM)、Antoine 等[17] 提出的可能性模糊 C 均值聚类算法 (possibilistic fuzzy C-means, PFCM)、Zhang 等[18] 提出的基于核的模糊聚类算法 (kernel-based fuzzy C-means, KFCM), KFCM 将核空间的概念引入到模糊聚类算法中, 增强了算法的鲁棒性。

聚类验证是对聚类的结果进行评估, 以验证聚类算法的划分是否正确。聚类验证常用的方法是通过聚类有效性指标 (cluster validity index, CVI)[19] 来对聚类算法划分的结果进行计算, 对比不同划分状态下的计算结果, 选取最优的划分结果。CVI 大体上可以划分为 3 种[20], 即外部有效性[21]、内部有效性[22] 和相对有效性[23], 文中涉及的指标主要是内部有效性指标。内部有效性指标通过聚类结果的内部结构来评估聚类质量, 不依赖于外部标签。常见的内部有效性指标有 Calinski 等[24] 提出的 CH(Calinski-Harabasz) 指标、Davies 等[25] 提出的 DB(Davies-Bouldin) 指标, 还有 Dunn[13] 提出的 Dunn 指标, 但 Dunn 指标对于环状或者线性数据集的适应效果较差。此外还有, Xie 等[26] 的 XBI(Xie-Beni) 指标、Fukuyama 等[27] 提出的 FSI (Fukuyama-Sugeno) 指标、Wu 等[28] 等提出的 WLI(Wu-and-Li) 指标、Liu 等[29] 提出的不平衡指数 (imbalanced index, IMI)、Mittal 等[30] 提出的 SMI(Saraswat-and-Mittal) 指标、Zhu 等[31] 提出的基于方差的聚类有效性 (the variance based clustering validity, VCVI) 指标、Maulikh 等[32] 提出的 MB (Maulik-Bandyopadhyay) 指标和 Tang 等[33] 提出的三重中心关系 (triple center relation, TCR) 指标等, 这些指标都在 CVI 发展进程中起到一定作用。

上文提到的这些 CVI 在聚类质量的评估上起到重要作用, 是聚类分析过程中不可或缺的一部分, 但是它们也存在一些缺点和局限性: 1) 部分 CVI 对于簇的形状较为敏感, 且对于现实数据中常出现的多噪声情况难以有合适的应对方法; 2) 基于 FCM 算法提出的 CVI 会因为 FCM 的错误分类导致出现错误的结果。如 FCM 算法的均匀效应会导致 FCM 算法在处理不平衡数据集时会出现错误结果[34-35]。

考虑以上问题, 本文提出了一种新的 CVI, 称之为全局数据驱动指标 (global data driven index, GDD)。该 CVI 共有 2 个部分组成。GDD 全面考虑了簇内外的数据特征, 同时 2 个部分的组合, 将簇内紧致性和簇间分离性结合起来分析, 从而更好地度量聚类结果的合理性。

针对提出的 CVI, 本文采用 UCI 数据集[36]、人工数据集和 Olivetti Face 数据集共 19 个数据集以及 11 个 CVI 进行了对比实验。为了增加实验的可信度, 分别选取了一些含有噪声的数据集和一些数据分布不均匀的数据集进行对比实验。与此同时, 为了验证 GDD 能否同时适应多种模糊聚类算法, 本文采用了 FCM、PFCM 和 KFCM 这 3 种不同的模糊聚类算法。通过不同的模糊聚类算法进行实验验证, 证明 GDD 有更强的适应力和更广泛的实用价值。此外, 本文论证了 GDD 的收敛性, 以证明新 CVI 在理论上的正确性。

1　部分内部有效性评价指标分析

内部有效性评价指标通过分析聚类结果中簇

的结构来衡量聚类划分的质量，无需使用外部的分类标签来对比判断。具体指标主要关注簇内紧致性（foc）和簇间分离性（fos）两个方面。这里介绍较为经典的 5 种有效性指标。

1) Dunn 指标

Dunn 指标是一种经典且有广泛应用的内部评价指标，其对非球形数据分布有较好的适应性。但是其对噪声敏感且计算较复杂。计算公式为

$$\mathrm{Dunn}^{(+)}=\frac{\min\limits_{1\leqslant i<j\leqslant K}\{\mathrm{dis}(C_i,C_j)\}}{\max\limits_{1\leqslant k\leqslant K}\{\mathrm{diam}(C_k)\}}$$

式中：$\mathrm{dis}(C_i,C_j)=\min\limits_{x_i\in C_i,x_j\in C_j}\{\|x_i-x_j\|\}$，$\mathrm{diam}(C_k)=\max\limits_{x_i,x_j\in C_k}\{\|x_i-x_j\|\}$。$C_i$ 为第 i 个簇。

2) XBI 指标

XBI 指标通过计算簇内部数据点到聚类中心的距离来度量簇内紧致性，计算所有聚类中心之间的距离最小值来度量簇间分离性，考虑因素比较全面。但是其单独选用聚类中心之间的距离最小值来度量簇间分离性，会导致对不规则数据集处理的效果不佳。同时，指标容易随着聚类簇数增加而单调变化。其计算公式为

$$\mathrm{XBI}^{(-)}=\frac{\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N}\mu_{ik}^{m}\|x_i-v_k\|}{N\times\min\limits_{i\neq j}\{\|v_i-v_j\|\}}$$

式中：N 为样本总数，K 为聚类数。这里的距离通常采用欧氏距离。

3) WLI 指标

WLI 指标在度量簇内紧致性的公式中加入了模糊隶属度的和，更好地度量簇内数据的紧致性，同时采用聚类中心之间距离的最小值和中值来度量簇间分离性，对分离性的度量更加全面。其计算公式为

$$\mathrm{WLI}^{(-)}=\frac{\sum\limits_{k=1}^{K}\left(\dfrac{\sum\limits_{i=1}^{N}\mu_{ik}^{2}\|x_i-v_k\|^2}{\sum\limits_{i=1}^{N}\mu_{ik}}\right)}{\min\limits_{i\neq j}\{\|v_i-v_j\|^2\}+\mathrm{median}\limits_{i\neq j}\{\|v_i-v_j\|^2\}}$$

式中 min 与 median 为簇中心之间的距离的最小值与中位数值。通常采用欧氏距离进行计算。

4) IMI 指标

IMI 指标在簇间分离性的度量上加入了不同簇之间的数据点数量比值，考虑了不同簇之间的不平衡比，对分离性的刻画更加准确。其公式为

$$\mathrm{IMI}^{(-)}=\frac{\sum\limits_{k=1}^{K}\dfrac{\sum\limits_{i=1}^{N}u_{ik}^{q}\|x_i-v_k\|}{\sum\limits_{i=1}^{N}u_{ik}}}{\min\limits_{i\neq j}\{\delta_{ij}\|v_i-v_j\|^2\}+\mathrm{median}\limits_{i\neq j}\{\delta_{ij}\|v_i-v_j\|^2\}}$$

式中：$\delta_{ij}=f_i/f_j, f_i>f_j$，$f_i=\sum\limits_{n=1}^{N}u_{ni}$，可以看出 δ_{lj} 是簇 C_i 和簇 C_j 的不平衡比。

5) TCR 指标

TCR 指标通过计算聚类中心之间距离最小值和均值，以及聚类中心方差来度量簇间分离性，三重因素相互作用，互相平衡，更好地刻画了分离性。其计算公式为

$$\mathrm{TCR}^{(-)}=\frac{\sum\limits_{k=1}^{K}\dfrac{\sum\limits_{i=1}^{N}u_{ik}^{2}\|x_i-v_k\|^2}{\sum\limits_{i=1}^{N}\max\limits_{1\leqslant k\leqslant K}u_{ik}}}{\dfrac{N}{K-1}\left(\min\limits_{i\neq j}\{\|v_i-v_j\|^2\}\times\mathrm{mean}\limits_{i\neq j}\{\|v_i-v_j\|^2\}\times\sum\limits_{k=1}^{K}\|v_k-\bar{v}\|^2\right)}$$

式中：N 为样本数，K 为聚类数，min 和 mean 分别为簇中心之间距离的最小值与均值。通常采用欧氏距离进行计算。

2　全局数据驱动指标

GDD 指标主要由簇内紧致性和簇间分离性两个部分组成。在参考 XBI 指标的整体结构基础上，GDD 指标对簇内紧致性和簇间分离性的表达进行一些改进。

2.1　簇内紧致性表达

GDD 指标的簇内紧致性的表达分为两个部分。第一部分 foc 为簇内所有数据点之间的平均距离，在此基础上，考虑到模糊聚类中的模糊隶属度，在计算数据点之间的欧氏距离时加上两点对当前簇的模糊隶属度，定义的公式为

$$f_{\mathrm{oc1}}=\frac{\sum\limits_{1\leqslant i<j\leqslant n_k}u_{ik}u_{jk}\|x_i-x_j\|}{\dfrac{n_k(n_k-1)}{2}}$$

式中：n_k 为第 k 个簇的数据点数目，u_{ik} 为在第 k 个簇中第 i 个样本的模糊隶属度，x_i，x_j 分别为该簇中两个互不相同的样本。距离计算采用欧氏距离。

第二部分 foc 为单个簇内部数据点的标准差, $\bar{x}$表示簇内样本的样本均值, 表达式为

$$f_{\mathrm{oc2}}=\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\|x_i-\bar{x}\|^2}$$

将两个部分相加得到对单个簇内部紧致性的表达式, 同时考虑到不同的簇所含数据点数目的不同, 其在整个数据集中所占的比重不同。所以每个簇对整体数据紧凑程度的贡献不应相同, 所占比重高的簇应当在整个数据集紧凑程度表达上有更高的权重。因此整体紧致性表达式由每个簇内部紧致性表达式加权相加得来, 权重为当前簇的数据点数目占总体数据点数目的比值。最终表达式为

$$f_{\mathrm{oc}}=\sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{\sum_{1\leqslant i<j\leqslant n_k}u_{ik}u_{jk}\|x_i-x_j\|}{\frac{n_k(n_k-1)}{2}}+\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\|x_i-\bar{x}\|^2}\right)$$

使用加权距离和来衡量数据紧凑程度能更好地反映簇内的数据几何结构, 也能更好应对形状各异的数据集。此外标准差的加入能增强 GDD 指标对簇内数据点紧致性和离散程度的评估能力。两个部分的结合使得 GDD 指标对簇内紧致性的评估更加全面。

2.2 簇间分离性表达

GDD 指标的簇间分离性的表达分为两个部分。第一部分 fos 以所有聚类中心到聚类中心均值距离之和为基底, 同时考虑到不同聚类中心所在簇所含数据点数目的不同, 其在整个数据集中所占的比重不同。所以每个聚类中心所在簇对整体数据的贡献不应相同, 所占比重高的簇应当在整个数据集上有更高的权重。因此, 在每个聚类中心到聚类中心均值的距离上加上对应的权重。权重为当前聚类中心所在簇的数据点数目占总体数据点数目的比值。最终表达式为

$$f_{\mathrm{os1}}=\sum_{k=1}^{K}\left(\frac{n_k}{N}\|v_k-\bar{v}\|\right)$$

式中: $\bar{v}$为所有簇中心的均值; N 为样本总数。同样采用欧氏距离进行计算。

第二部分 fos 为所有聚类中心之间距离的均值, 表达式为

$$f_{\mathrm{os2}}=\operatorname*{mean}_{i\neq j}\{\|v_i-v_j\|\}$$

将两个部分相加即得到 GDD 指标对于簇间分离性的刻画, 表达式为

$$f_{\mathrm{os}}=\sum_{k=1}^{K}\left(\frac{n_k}{N}\|v_k-\bar{v}\|\right)+\operatorname*{mean}_{i\neq j}\{\|v_i-v_j\|\}$$

关于簇间分离性刻画的第一部分, 加权距离和在整体上对所有聚类中心与中心点的离散程度和分离性做评估, 权重的引入进一步考虑了不同大小的簇对整体数据分离程度的影响。关于第二部分, 聚类中心之间距离均值从另一个方面对所有聚类中心之间的分离程度做全面的评估。两者的互相结合使得 GDD 指标对簇间分离性的刻画更加全面深刻。

综上所述, GDD 指标由两个部分组成: 第一部分 foc 用来刻画簇内数据点之间的紧致性; 第二部分 fos 用于刻画聚类中心之间分离性, 也就是簇间分离性。

$$\mathrm{GDD}(K)^{(-)}=\frac{f_{\mathrm{oc}}}{f_{\mathrm{os}}}=\frac{\sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{\sum_{1\leqslant i<j\leqslant n_k}u_{ik}u_{jk}\|x_i-x_j\|}{\frac{n_k(n_k-1)}{2}}+\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\|x_i-\bar{x}\|^2}\right)}{\sum_{k=1}^{K}\left(\frac{n_k}{N}\|v_k-\bar{v}\|\right)+\operatorname*{mean}_{i\neq j}\{\|v_i-v_j\|\}}\tag{1}$$

2.3 CVI 有效性证明

下面将对 GDD 指标的有效性做出证明, 其中用到了 Dunn 指标作为参照。其公式为

$$\mathrm{Dunn}^{(+)}=\frac{\min\limits_{1\leqslant i<j\leqslant K}\{\mathrm{dis}(C_i,C_j)\}}{\max\limits_{1\leqslant k\leqslant K}\{\mathrm{diam}(C_k)\}}$$

其中$\mathrm{dis}(C_i,C_j)$和$\mathrm{diam}(C_k)$定义分别为

$$\mathrm{dis}(C_i,C_j)=\min_{x_i\in C_i,x_j\in C_j}\{\|x_i-x_j\|\}$$

$$\mathrm{diam}(C_k)=\max_{x_i,x_j\in C_k}\{\|x_i-x_j\|\}$$

Dunn 指标的论文中证明当 Dunn>1 时, Dunn 指标会指出正确的聚类数目并且 Dunn 指标的数值越大标识当前的聚类效果越好。分析 GDD 指标的构成, 猜测当在正确的聚类划分结构下, Dunn 指标变得足够大, GDD 指标会变得足够小。接下来将对这一猜测进行证明。

假设在聚类数目为 K 时, 对数据集进行划分得到划分结果。分别使用 GDD 指标和 Dunn 指标对结果进行评估, 得到评价数值 $\mathrm{GDD}(K)$ 和 $\mathrm{Dunn}(K)$。

定理 1 设$k\in\{2,3,\cdots,N-1\}$, 且$\max\limits_{1\leqslant k\leqslant K}\{\mathrm{dia}^2(C_k)\}\geqslant 1$成立。同时, $\mu_{ik}(1\leqslant i\leqslant N,1\leqslant k\leqslant K)$是模糊隶属度, 而$\omega_{ik}(1\leqslant i\leqslant N,1\leqslant k\leqslant K)$是相应硬划分的隶属度,

其公式为

$$\omega_{ik}=\begin{cases}1, & k=\underset{1\leqslant k^*\leqslant K}{\operatorname{argmax}}\{\mu_{ik^*}\}\\ 0, & \text{其他}\end{cases}$$

则可得

$$\mathrm{GDD}(K)\leqslant\frac{2}{\mathrm{Dunn}(K)}\tag{2}$$

证明　设 K 类划分为数据集 X 的优化划分，其中 $X=\{x_i|1\leqslant i\leqslant N\}$，聚类中心为 $v_k(1\leqslant k\leqslant K)$，隶属度为 $\mu_{ik}(1\leqslant i\leqslant N,1\leqslant k\leqslant K)$。$n_k$ 为第 k 个簇内部数据点的数目。推导 foc 和 $\mathrm{diam}(C_k)$ 的大小关系可得公式为

$$\mathrm{diam}(C_k)=\max_{x_i,x_j\in C_k}\{\|x_i-x_j\|\}\geqslant\frac{\sum\limits_{1\leqslant i<j\leqslant n_k}\|x_i-x_j\|}{\frac{n_k(n_k-1)}{2}}=f_{\mathrm{oc1}}\tag{3}$$

$$\mathrm{diam}(C_k)=\max_{x_i,x_j\in C_k}\{\|x_i-x_j\|\}\geqslant\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\|x_i-\bar{x}\|^2}=f_{\mathrm{oc2}}\tag{4}$$

结合式 (3) 和式 (4)，且聚类结果中所有的簇都满足式 (3) 和式 (4)，所以对式 (3) 和式 (4) 的加权和也满足。推出下式：

$$\begin{aligned}f_{\mathrm{oc}}=&\sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{\sum\limits_{1\leqslant i<j\leqslant n_k}u_{ik}u_{jk}\|x_i-x_j\|}{\frac{n_k(n_k-1)}{2}}+\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\|x_i-\bar{x}\|}^{\,2}\right)\\&\leqslant 2\max\nolimits_{1\leqslant k\leqslant K}\{\mathrm{diam}(C_k)\}\end{aligned}\tag{5}$$

同时，由于

$$f_{\mathrm{os}}\geqslant\underset{1\leqslant i<j\leqslant K}{\mathrm{mean}}\left\{\left||v_i-v_j|\right|^2\right\}\geqslant\mathrm{dis}(C_i,C_j)\geqslant\min_{1\leqslant i<j\leqslant K}\mathrm{dis}(C_i,C_j)\}\tag{6}$$

结合式 (5)、(6) 以及 $\max\limits_{1\leqslant k\leqslant K}\{\mathrm{dia}^2(C_k)\}\geqslant 1$，可推出：

$$\mathrm{GDD}(K)=\frac{f_{\mathrm{oc}}}{f_{\mathrm{os}}}\leqslant\frac{2\max\limits_{1\leqslant k\leqslant K}\{\mathrm{diam}(C_k)\}}{\min\limits_{1\leqslant i<j\leqslant K}\{\mathrm{dis}(C_i,C_j)\}}\leqslant\frac{2}{\mathrm{Dunn}(K)}$$

至此证明完毕。

从定理 1 的式 (2) 可知，在正确的聚类划分结构下，Dunn 指标变得足够大，GDD 指标会变得足够小，说明对应的聚类划分效果是非常理想的。

2.4　CVI 时间复杂度分析

时间复杂度是衡量一个 CVI 面对由聚类算法划分后的数据集的处理效率的重要参照。一个优秀的 CVI 往往在拥有优秀识别正确率的基础上尽可能的达到更小的时间复杂度。时间复杂度因为有对 CVI 的实用性评估的功效，需要对其进行分析。

分析式 (1)，将 GDD 分成以下部分进行时间复杂度的讨论。

1）计算全局质心。对于全局质心，有

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

对此，需要遍历数据集中所有共 n 个点，每个点 d 维，所以该部分的时间复杂度为 $O(nd)$。

2）计算模糊加权平均距离。对于每个簇 j 分别进行处理。每个簇的簇大小 $|C_j|=n_j\leqslant\tau$，其中 τ 为所有簇中最大的样本数量，所需计算的所有数据点对共 $O(\tau^2)$ 对，每对计算距离 $||x_i-x_j||$ 的复杂度为 $O(d)$。每对需乘其隶属度 $O(1)$。因此该部分的总计算量为 $\sum\limits_{j=1}^{k}O(\tau^2d)$，考虑其最坏时间复杂度为 $O(K\tau^2d)$，其中 K 为簇的数量。

3）计算簇内标准差之和。计算每个簇的质心 $\overline{x_j}$ 所需的时间复杂度为 $O(\tau d)$，接着计算 $\sum\limits_{x_i\in C_j}||x_i-\overline{x_j}||^2$，其时间复杂度为 $O(\tau d)$，这只是单个簇的复杂度，该部分总的时间复杂度为 $\sum\limits_{j}O(\tau d)=O(nd)$。

4）计算加权类间距离。由于在以上部分已经计算了全局质心 $\bar{x}$ 和每个簇的质心 $\overline{x_j}$，因此该部分计算 $||\overline{x_j}-\bar{x}||^2$ 的时间复杂度为 $O(d)$，这同样是单个簇的复杂度，一共有 K 个簇，从而类间距离部分总共 $O(Kd)$，又因为需要遍历数据集内所有共 n 个点，这部分需要 $O(nd)$ 因此该部分的总复杂度为 $O(Kd+nd)$。

5）计算簇中心间的平均距离。单个簇中心对的距离 $||\overline{x_i}-\overline{x_j}||$ 的时间复杂度为 $O(d)$，而该簇中心对一共有 K^2 对，因此该部分的总复杂度为 $O(K^2d)$。

至此，可以得到 GDD 指标的时间复杂度为 $O(K\tau^2d+nd+Kd+nd+K^2d)$，而当聚类正确的情况下聚类数 K 应当远小于样本数 n，因此 GDD 的时间复杂度最终记为 $O(K\tau^2d+nd)$。

表 1 中给出了文中参与实验的 12 个指标的时间复杂度。

表 1　12 个指标的时间复杂度

Table 1　Time complexity of 12 indicators

序号	CVI	时间复杂度
1	CH	$O(K\tau d)$
2	Dunn	$O(K^2\tau^2d)$
3	DB	$O(K\tau^2d)$
4	MB	$O(Knd)$
5	IMI	$O(Kn(d+K))$
6	XBI	$O(Knd)$
7	VCVI	$O(K\tau d)$

续表 1

序号	CVI	时间复杂度
8	FSI	$O(Knd)$
9	WLI	$O(Knd)$
10	SMI	$O(K^2\tau^2d+Knd)$
11	TCR	$O(Kn(d+K))$
12	GDD	$O(K\tau^2d+nd)$

从表 1 中可以看出, GDD 的时间复杂度为 $O(K\tau^2d+nd)$, 属于高开销指标, 同为高开销指标的还有 Dunn 指标与 SMI 指标, 而其余指标在时间复杂度上均优于上述三者, 这意味着 GDD 指标在实际使用时可能需要更多的时间来处理数据并得出结果。

3　GDD 指标的仿真与对比实验

本文选取 3 个模糊聚类算法、11 个对比 CVI 以及 19 个数据集来做对比实验。3 个模糊聚类算法分别是 FCM 算法、PFCM 算法和 KFCM 算法; 11 个对比 CVI 分别是 CH、Dunn、DB、MB、IMI、XBI、VCVI、FSI、WLI、SMI 和 TCR 指标; 19 个数据集分别是 9 个 UCI 数据集、9 个人造数据集和 Olivetti Face 数据集。实验的软件环境如下: 操作系统是 Windows 11OS, 编程软件是专业数学分析软件。实验的硬件环境如下: CPU 是 Intel(R) Core(TM)i9-12900 KF, 显卡是 NVIDIA GeForce RTX3090, 内存是 64 GB。

3.1　数据集与对比指标

在本实验中, 主要采用 3 类数据集进行研究, 分别为 UCI 数据集、人工合成数据集以及 Olivetti Face 数据集。本实验选取 UCI 数据集中的 9 个数据集, 分别是 Zoo、Hayes-Roth、Iris、Glass、Dermatology、Breast Cancer、Balance Scale、Libras、Letter 数据集。其中 Zoo 数据集来自动物数据, 数据维度为 16 维, 共有 101 个样本点, 正确的聚类数目为 7 类; Hayes-Roth 数据集来自社科研究, 数据维度为 4 维, 共有 132 个样本点, 正确的聚类数目为 3 类; Iris 数据集来自不同品种的鸢尾花数据, 数据维度为 4 维, 共有 150 个样本点, 正确的聚类数目为 3 类; Glass 数据集来自具有不同元素含量的玻璃的数据, 数据维度为 9 维, 共有 214 个样本点, 正确的聚类数目为 6 类; Dermatology 数据集来自一种疾病的医疗数据, 数据维度为 34 维, 共有 366 个样本点, 正确的聚类数目为 6 类; Breast Cancer 数据集来自肿瘤数据, 用于预测肿瘤的性质, 数据维度为 30 维, 共有 569 个样本点, 正确的聚类数目为 2 类; Balance Scale 数据集来自模拟心理学实验结果, 数据维度为 4 维, 共有 625 个样本点, 正确的聚类数目为 3 类; Libras 数据集来自人手部动作数据, 数据维度为 90 维, 共有 360 个样本点, 正确的聚类数目为 15 类; Letter 数据集来自字符图像数据, 数据维度为 16 维, 共有 20 000 个样本点, 正确的聚类数目为 26 类。为了避免实验的数据集的类型单一, 本实验同样选取部分人造数据集进行实验来增强实验的可信性。人工数据集作为评估聚类算法在不同结构、密度、维度和噪声条件下性能的工具, 其允许研究者精确控制簇的结构、形状、密度、重叠度及噪声水平的特性, 能够针对不同的被测对象设计对应的测试内容, 从而系统地评估被测对象的鲁棒性、可拓展性和对特定模式的识别能力。人工数据集常常通过高斯分布采样或者数据点, 再通过高斯噪声引入坐标扰动, 通过不同的设计与配比得到满足要求的数据集合。其中包含有 9 个人造数据集, 分别为 Data_60、Data_150、Circle、Jain、X8D5K、Data_77、E6、Dim_128 和 Dim_256 数据集。Data_60 数据集的数据维度为 2 维, 共有 60 个样本点, 正确的聚类数目为 3 类; Data_150 数据集的数据维度为 2 维, 共有 150 个样本点, 正确的聚类数目为 3 类; Circle 数据集的数据维度为 2 维, 共有 150 个样本点, 正确的聚类数目为 2 类; Jain 数据集的数据维度为 2 维, 共有 373 个样本点, 正确的聚类数目为 2 类; X8D5K 数据集的数据维度为 8 维, 共有 1 000 个样本点, 正确的聚类数目为 5 类; Data_77 数据集的数据维度为 2 维, 共有 1000 个样本点, 正确的聚类数目为 7 类; E6 数据集的数据维度为 2 维, 共有 8 537 个样本点, 正确的聚类数目为 4 类; Dim_128 数据集的数据维度为 128 维, 共有 1 024 个样本点, 正确的聚类数目为 16 类; Dim_256 数据集的数据维度为 256 维, 共有 1024 个样本点, 正确的聚类数目为 16 类。

同时, 本实验选取 Olivetti Face 数据集, 来验证实验在图像数据上的准确性。实验前对 Olivetti Face 数据集进行切割处理, 处理后的数据集包含 40 类人脸图像, 每类 10 张图像, 10 张图像是同一个人在不同状态下的面部图像, 每张图像大小为 4096 个像素点。

图 1 给出了部分上述数据集在正确聚类数目下的分布情况。其中维度大于 2 的数据集采用 TSNE 方法降维到 3 维后给出。Olivetti Face 数据集的缩略图如图 2 所示。

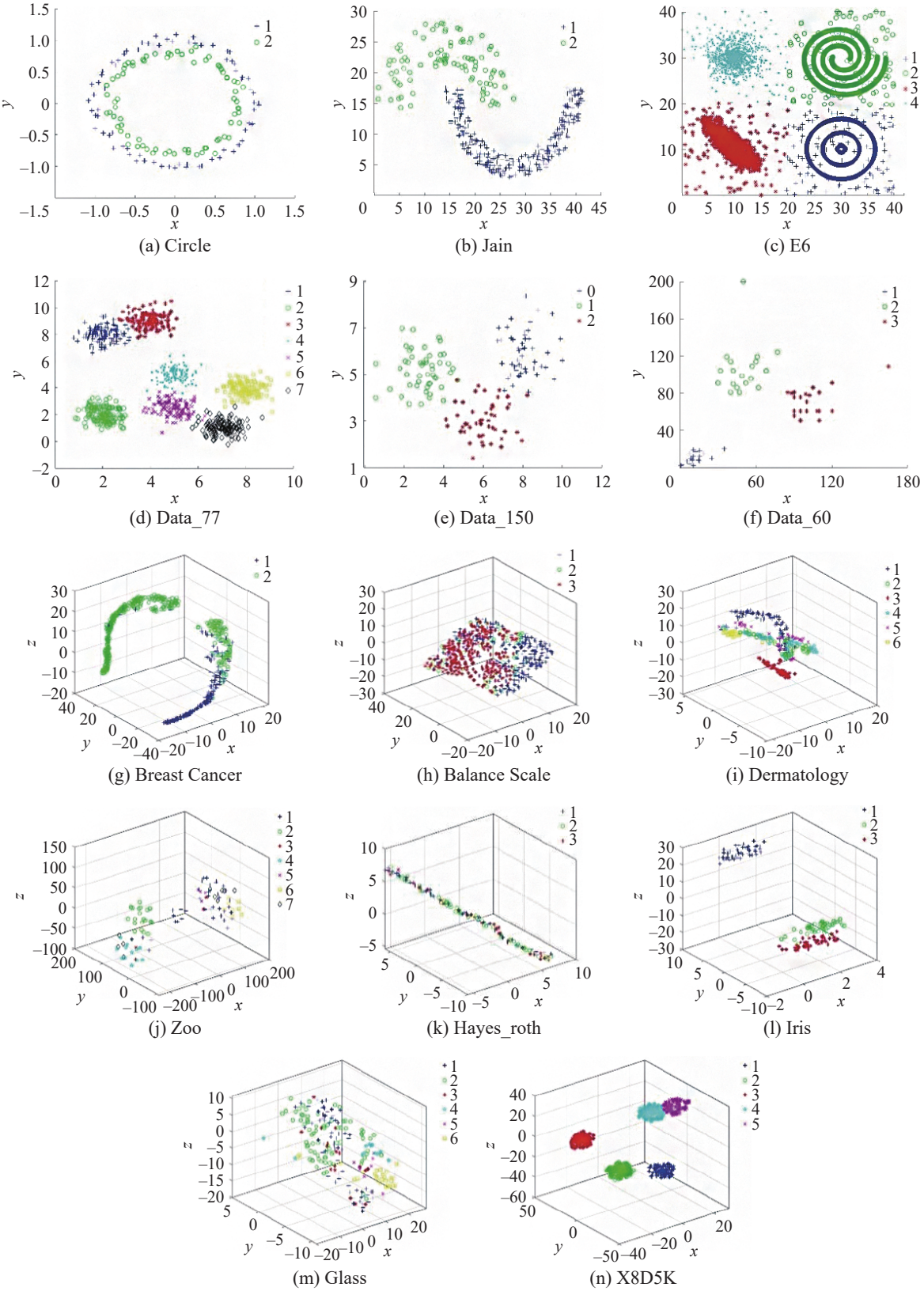


图 1　数据集的分布

Fig. 1　Distribution of datasets

本实验选取另外 11 个 CVI 作为对比指标，分别是 CH、Dunn、DB、MB、IMI、XBI、VCVI、FSI、WLI、SMI 和 TCR 指标。通过与各种不同 CVI 进行对比实验，来验证提出的 GDD 指标具有更强的适应性和更高的准确性。

3.2　对比实验

本实验采用 FCM 算法、PFCM 算法和 KFCM 算法对 CVI 的准确性进行验证。在正确聚类数目为

2~10 的数据集上, 每个数据集进行 10 轮实验, 每轮实验的 K 值从 2 变到 10; 在正确聚类数目为 2~30 的数据集上, 每个数据集进行 30 轮实验, 每轮实验的 K 值从 2 变到 30。实验结束后, 收集所有实验数据, 并根据不同 CVI 最优值的分布统计各轮实验中 CVI 对应的最优聚类数目。表 2 给出了 UCI 数据集在上述 3 种算法下的实验结果, 其中的数据表示各轮实验中最优聚类数目及其出现的频次。例如, 数据“$2^{7}3^{3}$”表示在 10 轮实验中, 聚类数目为 2 的最优值出现了 7 次, 而聚类数目为 3 的最优值出现了 3 次。

表 2 第一部分是 FCM 算法下 CVI 的实验结果。以其中的 Dermatology 数据集为例, Dermatology 数据集的正确聚类数目为 6 类, MB 指标 10 轮的结果中 7 轮是 4 类为最优值, 2 轮是 5 类为最优值, 1 轮是 6 类为最优值, 其他 CVI 不再赘述。统计出现次数最多的最优值聚类数目作为该指标的评价结果, MB 指标结果为 4 类, WLI 指标结果为 5 类, GDD 指标结果为 6 类。可以看出, MB 指标和 WLI 指标的结果不正确, GDD 指标正确得出了 Dermatology 数据集的聚类数目。

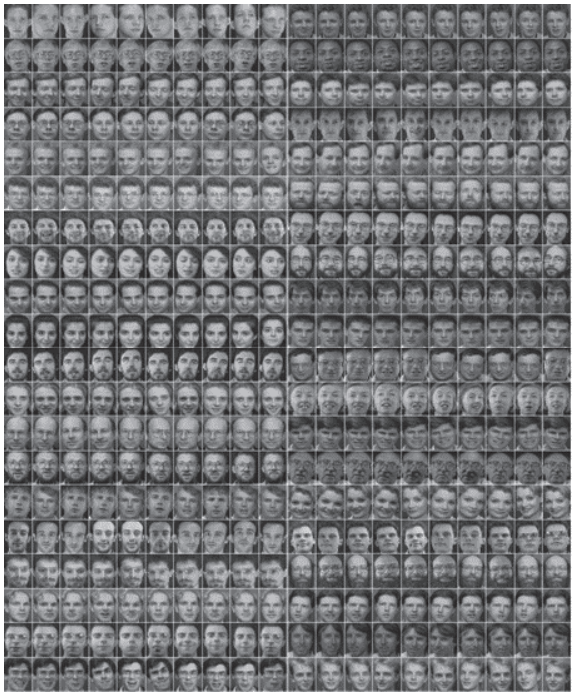


图 2　Olivetti Face 数据集
Fig. 2　Olivetti Face dataset

表 2　UCI 数据集的实验结果
Table 2　Experimental results of UCI dataset

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Zoo	2^{10}	2^{10}	2^{10}	7^{10}	$5^{2}7^{8}$	2^{10}	$2^{8}3^{1}4^{1}$	3^{10}	$2^{3}7^{7}$	7^{10}	7^{10}	7^{10}
	Hayes-Roth	2^{10}	$9^{3}10^{7}$	3^{10}	3^{10}	$2^{3}3^{7}$	$2^{1}3^{7}4^{2}$	$2^{1}3^{5}4^{4}$	$3^{2}4^{8}$	3^{10}	3^{10}	$2^{3}3^{6}4^{1}$	3^{10}
	Iris	2^{10}	$3^{9}5^{1}$	3^{10}	3^{10}	$3^{8}5^{2}$	3^{10}	$3^{6}4^{1}5^{3}$	10^{10}	3^{10}	3^{10}	$2^{1}3^{9}$	3^{10}
	Glass	2^{10}	$2^{6}3^{4}$	3^{10}	2^{10}	$2^{2}5^{2}6^{6}$	$5^{3}6^{7}$	2^{10}	$3^{8}5^{1}6^{1}$	$4^{7}5^{1}6^{2}$	$5^{9}6^{1}$	$5^{2}6^{8}$	6^{10}
	Dermatology	$2^{8}3^{2}$	$2^{2}3^{2}5^{6}$	2^{10}	$4^{7}5^{2}6^{1}$	$6^{7}7^{3}$	$6^{7}7^{3}$	2^{10}	$2^{2}3^{8}$	$4^{1}5^{5}6^{4}$	$2^{7}6^{3}$	$4^{3}6^{7}$	$4^{2}6^{8}$
	Breast Cancer	2^{10}	$2^{8}6^{2}$	2^{10}	2^{10}	2^{10}	$2^{8}3^{2}$	$2^{8}3^{2}$	2^{10}	$2^{8}3^{1}5^{1}$	2^{10}	2^{10}	2^{10}
	Balance Scale	$2^{8}3^{1}5^{1}$	$3^{2}4^{8}$	$2^{7}3^{3}$	$2^{8}5^{2}$	$2^{3}3^{2}4^{5}$	$2^{1}3^{3}4^{6}$	$3^{8}6^{2}$	$2^{5}3^{2}5^{3}$	$3^{1}5^{7}6^{2}$	$2^{7}3^{2}5^{1}$	$2^{2}4^{7}6^{1}$	$4^{1}5^{9}$
	Libras	$2^{22}3^{3}$ $5^{4}7^{1}$	$2^{15}3^{12}$ 4^{3}	$2^{15}3^{5}$ 4^{10}	$18^{22}20^{8}$	$14^{4}15^{20}$ 16^{6}	$13^{9}14^{15}$ 15^{6}	$2^{25}3^{3}$ 6^{2}	30^{30}	$14^{7}15^{17}$ 17^{6}	$2^{16}10^{6}$ $11^{4}15^{4}$	$10^{3}13^{2}$ $14^{20}15^{5}$	$11^{4}14^{2}$ $15^{22}16^{2}$
	Letter	$2^{20}3^{6}$ $4^{2}5^{2}$	$2^{6}3^{17}4^{7}$	2^{30}	$2^{10}3^{18}$ 5^{2}	$26^{10}27^{18}$ 30^{2}	$20^{5}22^{15}$ $23^{4}25^{6}$	$2^{19}5^{3}$ $10^{3}14^{2}$ 15^{3}	$28^{25}29^{1}$ 30^{4}	$24^{2}25^{18}$ $26^{4}27^{6}$	2^{30}	$22^{1}26^{25}$ $27^{2}30^{2}$	$22^{2}26^{23}$ 27^{5}
PFCM	Zoo	2^{10}	2^{10}	2^{10}	7^{10}	$5^{1}6^{1}7^{8}$	2^{10}	$2^{8}4^{2}$	3^{10}	$2^{3}7^{7}$	$4^{2}7^{8}$	7^{10}	7^{10}
	Hayes-Roth	3^{10}	$2^{1}9^{2}10^{7}$	3^{10}	3^{10}	$2^{6}3^{3}5^{1}$	$3^{8}4^{2}$	$3^{7}4^{3}$	$3^{5}4^{3}7^{2}$	$3^{9}4^{1}$	3^{10}	$2^{2}3^{6}4^{2}$	$3^{9}4^{1}$
	Iris	2^{10}	$3^{9}5^{1}$	3^{10}	$2^{2}3^{8}$	$3^{8}4^{1}5^{1}$	3^{10}	$3^{6}4^{2}5^{2}$	10^{10}	3^{10}	3^{10}	3^{10}	3^{10}
	Glass	2^{10}	$2^{7}3^{3}$	$2^{1}3^{8}5^{1}$	2^{10}	$5^{3}6^{7}$	$4^{1}5^{1}6^{8}$	2^{10}	3^{10}	$4^{3}6^{7}$	$4^{1}5^{8}6^{1}$	$5^{2}6^{8}$	6^{10}
	Dermatology	$2^{8}3^{2}$	$2^{2}3^{1}5^{7}$	$2^{8}5^{2}$	$4^{4}5^{1}6^{5}$	$4^{1}6^{7}7^{2}$	$6^{7}7^{3}$	2^{10}	$2^{2}3^{8}$	$5^{7}6^{3}$	$2^{7}6^{3}$	$4^{2}5^{1}6^{7}$	$4^{1}6^{9}$
	Breast Cancer	2^{10}	$2^{8}6^{2}$	2^{10}	$2^{9}4^{1}$	2^{10}	2^{10}	2^{10}	2^{10}	$2^{8}3^{2}$	2^{10}	2^{10}	2^{10}
	Balance Scale	$2^{8}3^{2}$	$2^{1}3^{2}4^{7}$	$2^{7}3^{3}$	$2^{8}4^{1}5^{1}$	$2^{1}3^{3}4^{6}$	$3^{3}4^{7}$	$3^{8}6^{2}$	$2^{7}5^{3}$	$5^{8}6^{2}$	$2^{7}3^{2}5^{1}$	$2^{2}4^{8}$	5^{10}
	Libras	$2^{23}3^{3}$ 5^{4}	$2^{15}3^{10}$ $4^{3}6^{2}$	$2^{15}3^{5}$ $4^{9}7^{1}$	$18^{21}20^{6}$ 24^{3}	$14^{2}15^{21}$ $16^{6}20^{1}$	$13^{6}14^{15}$ $15^{6}17^{3}$	$2^{24}3^{3}$ $6^{2}10^{1}$	30^{30}	$14^{4}15^{17}$ $17^{6}20^{3}$	$2^{16}10^{10}$ 15^{4}	$10^{3}14^{23}$ 15^{4}	$11^{4}15^{25}$ 16^{1}
	Letter	$2^{22}3^{6}$ 4^{2}	$2^{6}3^{17}$ $4^{5}8^{2}$	2^{30}	$2^{10}3^{18}$ $5^{1}6^{1}$	$26^{10}27^{20}$	$22^{15}23^{9}$ 25^{6}	$2^{23}10^{3}$ $14^{2}15^{2}$	$28^{25}29^{2}$ 30^{3}	$25^{18}26^{6}$ 27^{6}	2^{30}	$26^{25}27^{3}$ 30^{2}	$22^{2}26^{20}$ $27^{5}28^{3}$

续表 2

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
KFCM	Zoo	2^{10}	2^{10}	2^{10}	$2^{1}7^{9}$	$4^{1}5^{1}7^{8}$	2^{10}	$2^{9}4^{1}$	$3^{7}5^{3}$	$2^{2}4^{1}7^{7}$	$5^{1}7^{9}$	7^{10}	7^{10}
	Hayes-Roth	3^{10}	$2^{1}10^{9}$	3^{10}	$2^{2}3^{8}$	2^{10}	3^{10}	$2^{1}3^{8}4^{1}$	$2^{1}3^{7}4^{3}$	3^{10}	$3^{7}6^{3}$	$3^{6}4^{2}7^{2}$	3^{10}
	Iris	2^{10}	$3^{9}4^{1}$	3^{10}	$2^{2}3^{8}$	$2^{1}3^{9}$	$3^{8}5^{2}$	$3^{7}5^{2}8^{1}$	10^{10}	3^{10}	3^{10}	3^{10}	3^{10}
	Glass	2^{10}	$2^{7}3^{2}6^{1}$	3^{10}	2^{10}	$5^{2}6^{8}$	$2^{1}5^{1}6^{8}$	2^{10}	3^{10}	$2^{1}4^{6}6^{3}$	$2^{2}5^{7}6^{1}$	$3^{1}5^{2}6^{7}$	$4^{2}6^{8}$
	Dermatology	$2^{8}3^{2}$	$4^{2}5^{8}$	2^{10}	$4^{4}6^{6}$	$2^{1}6^{7}7^{2}$	$6^{8}7^{2}$	2^{10}	$3^{8}5^{2}$	$3^{2}5^{7}6^{1}$	$2^{7}5^{2}6^{1}$	$4^{2}6^{8}$	$6^{9}7^{1}$
	Breast Cancer	2^{10}	$2^{8}3^{1}6^{1}$	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	$2^{8}3^{2}$	2^{10}	2^{10}	2^{10}
	Balance Scale	$2^{8}3^{2}$	$2^{1}4^{9}$	$2^{5}3^{3}4^{2}$	$2^{5}3^{3}4^{2}$	$2^{1}3^{4}4^{5}$	$2^{2}3^{3}4^{5}$	$3^{7}6^{2}7^{1}$	2^{10}	$5^{8}6^{2}$	$2^{7}5^{1}6^{2}$	$2^{1}4^{6}5^{3}$	$5^{9}6^{1}$
	Libras	$2^{22}3^{2}$ $5^{4}10^{2}$	$2^{15}3^{9}$ $4^{4}7^{2}$	$2^{15}3^{3}$ $4^{10}5^{2}$	$15^{3}18^{21}$ $20^{5}24^{1}$	$14^{4}15^{22}$ 16^{4}	$13^{7}14^{17}$ 15^{6}	$2^{21}3^{4}$ 6^{5}	30^{30}	$15^{17}17^{6}$ 20^{7}	$2^{16}3^{1}$ $10^{9}15^{4}$	$10^{3}13^{2}$ $14^{21}15^{4}$	$14^{2}15^{22}$ 16^{6}
	Letter	$2^{19}3^{4}$ $4^{2}5^{2}7^{3}$	$2^{6}3^{19}$ 4^{5}	2^{30}	$2^{10}3^{18}$ 5^{2}	$24^{5}26$ $^{10}27^{15}$	$20^{4}22^{15}$ $23^{5}25^{6}$	$2^{21}4^{2}10^{3}$ $14^{2}15^{2}$	$28^{24}29^{4}$ 30^{2}	$25^{18}26^{6}$ $27^{3}30^{3}$	2^{30}	$20^{4}22^{2}$ $26^{21}27^{3}$	$22^{2}26^{24}$ $27^{3}29^{1}$

统计表 2 中的数据，选取实验结果中出现次数最多的聚类数目作为最终的指标评价结果，形成表 3。例如 FCM 算法下，Dermatology 数据集 Dunn 指标的实验结果是“$4^{7}5^{2}6^{1}$”，比较不同聚类数目出现的次数，得出最终的聚类数目为 4 类。

表 3　UCI 数据集的 Eff 统计结果

Table 3　Eff statistical results of UCI dataset

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Zoo	2	2	2	**7**	**7**	2	2	3	**7**	**7**	**7**	**7**
	Hayes-Roth	2	10	**3**	**3**	**3**	**3**	**3**	4	**3**	**3**	**3**	**3**
	Iris	2	**3**	**3**	**3**	**3**	**3**	**3**	10	**3**	**3**	**3**	**3**
	Glass	2	2	3	2	**6**	**6**	2	3	4	5	**6**	**6**
	Dermatology	2	5	2	4	**6**	**6**	2	3	5	2	**6**	**6**
	Breast Cancer	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Balance Scale	2	4	2	2	4	4	**3**	2	5	2	4	5
	Libras	2	2	2	18	**15**	14	2	30	**15**	2	14	**15**
	Letter	2	3	2	3	27	22	2	28	25	2	**26**	**26**
	Eff	0.11	0.22	0.33	0.44	0.78	0.56	0.44	0.11	0.56	0.44	0.78	0.89
PFCM	Zoo	2	2	2	**7**	**7**	2	2	3	**7**	**7**	**7**	**7**
	Hayes-Roth	**3**	10	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Iris	2	**3**	**3**	**3**	**3**	**3**	**3**	10	**3**	**3**	**3**	**3**
	Glass	2	2	3	2	**6**	**6**	2	3	**6**	5	**6**	**6**
	Dermatology	2	5	2	**6**	**6**	**6**	2	3	5	2	**6**	**6**
	Breast Cancer	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Balance Scale	2	4	2	2	4	4	**3**	2	5	2	4	5
	Libras	2	2	2	18	**15**	14	2	30	**15**	2	14	**15**
	Letter	2	3	2	3	27	22	2	28	25	2	**26**	**26**
	Eff	0.22	0.22	0.33	0.44	0.67	0.56	0.44	0.22	0.67	0.44	0.78	0.89
KFCM	Zoo	2	2	2	**7**	**7**	2	2	3	**7**	**7**	**7**	**7**
	Hayes-Roth	**3**	10	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Iris	2	**3**	**3**	**3**	**3**	**3**	**3**	10	**3**	**3**	**3**	**3**
	Glass	2	2	3	2	**6**	**6**	2	3	4	5	**6**	**6**
	Dermatology	2	5	2	**6**	**6**	**6**	2	3	5	2	**6**	**6**
	Breast Cancer	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Balance Scale	2	4	2	2	4	4	**3**	2	5	2	4	5
	Libras	2	2	2	18	**15**	14	2	30	**15**	2	14	**15**
	Letter	2	3	2	3	27	22	2	28	25	2	**26**	**26**
	Eff	0.22	0.22	0.33	0.56	0.67	0.56	0.44	0.22	0.56	0.44	0.78	0.89

注：加粗代表该数值和正确的聚类数目一致。

此时需要借助评价因子来对 CVI 的性能进行评价。选取的评价因子为有效性比率 (effectiveness ratio, Eff)[37], 其值为 CVI 得到正确聚类数目结果的数据集个数与总的数据集个数的比值。Eff 的计算公式为

$$L_{\mathrm{Eff}}=\frac{1}{D}\sum_{d=1}^{D}\theta,\theta=\begin{cases}1, & k\text{是正确聚类数}\\ 0, & \text{其他}\end{cases}$$

分析表 3 的统计数据可以看出, 本文提出的 GDD 指标在 UCI 数据集上的实验结果中相较于其他 CVI 准确率更高。GDD 指标在 FCM 算法、PFCM 算法和 KFCM 算法上均是有 1 个数据集的结果错误, 准确率达到 89%。其他准确率较高的指标, 例如 IMI 指标和 TCR 指标, 其在 FCM 算法上的准确率大致在 78% 左右, 在 PFCM 算法和 KFCM 算法上的准确率稍低一点在 70% 左右。剩余的指标准确率大多都在 60% 以下。从实验结果可以分析出, 本文提出的 GDD 指标在 3 个算法上的准确率均高于其他 CVI, 说明 GDD 指标对于不同算法的适应力均优于其他 CVI。同时对比不同 CVI 在不同算法下的准确率, 例如 WLI 指标在 FCM 算法上的准确率为 56%, 而在 PFCM 算法下的准确率为 67%, 这在一定程度上说明不同的 CVI 有着更契合 CVI 本身的算法。

为了更精细地确定一个 CVI 的评价效果和稳定程度, 在此引入另外一个评价因子来对实验结果进行评价。引入的评价因子为有效性偏差 (Bias), 其值为实验所得的聚类数目和正确聚类数目之间差值的绝对值。Bias 的计算公式为

$$L_{\mathrm{Bias}}=\sum_{d=1}^{D}|k-k_{\mathrm{true}}|$$

应用 Bias 评价因子后得到表 4。

表 4　UCI 数据集的 Bias 统计结果

Table 4　Bias statistical results of UCI dataset

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Zoo	5	5	5	0	0	5	5	4	0	0	0	0
	Hayes-Roth	1	7	0	0	0	0	0	1	0	0	0	0
	Iris	1	0	0	0	0	0	0	7	0	0	0	0
	Glass	4	4	3	4	0	0	4	3	2	1	0	0
	Dermatology	4	1	4	2	0	0	4	3	1	4	0	0
	Breast Cancer	0	0	0	0	0	0	0	0	0	0	0	0
	Balance Scale	1	1	1	1	1	1	0	1	2	1	1	2
	Libras	13	13	13	3	0	1	13	15	0	13	1	0
	Letter	24	23	24	23	1	4	24	2	1	24	0	0
	Bias	53	54	50	33	2	11	50	36	6	43	2	2
PFCM	Zoo	5	5	5	0	0	5	5	4	0	0	0	0
	Hayes-Roth	0	7	0	0	1	0	0	0	0	0	0	0
	Iris	1	0	0	0	0	0	0	7	0	0	0	0
	Glass	4	4	3	4	0	0	4	3	0	1	0	0
	Dermatology	4	1	4	0	0	0	4	3	1	4	0	0
	Breast Cancer	0	0	0	0	0	0	0	0	0	0	0	0
	Balance Scale	1	1	1	1	1	1	0	1	2	1	1	2
	Libras	13	13	13	3	0	1	13	15	0	13	1	0
	Letter	24	23	24	23	1	4	24	2	1	24	0	0
	Bias	52	54	50	31	3	11	50	35	4	43	2	2
KFCM	Zoo	5	5	5	0	0	5	5	4	0	0	0	0
	Hayes-Roth	0	7	0	0	1	0	0	0	0	0	0	0
	Iris	1	0	0	0	0	0	0	7	0	0	0	0
	Glass	4	4	3	4	0	0	4	3	2	1	0	0
	Dermatology	4	1	4	0	0	0	4	3	1	4	0	0
	Breast Cancer	0	0	0	0	0	0	0	0	0	0	0	0
	Balance Scale	1	1	1	1	1	1	0	1	2	1	1	2
	Libras	13	13	13	3	0	1	13	15	0	13	1	0
	Letter	24	23	24	23	1	4	24	2	1	24	0	0
	Bias	52	54	50	31	3	11	50	35	6	43	2	2

分析表 4 中的数据，以 KFCM 算法下的运行结果为例，Bias 值比较低的有 IMI 指标、WLI 指标、TCR 指标和 GDD 指标，其中 GDD 指标的 Bias 值最低，足以说明本文提出的 GDD 指标在 UCI 数据集上有更好的评价效果和更稳定的评价能力，即使在 GDD 指标给出错误评价结果的情况下，得到的错误聚类数目和正确的聚类数目也不会相差太多。相反，其他 CVI，如 CH 指标、Dunn 指标和 DB 指标，这些 CVI 的 Bias 值都在 50 上下，相较于其他 CVI 高出很多。这说明 CH 指标、Dunn 指标和 DB 指标这些 CVI 在错误评价后得到的聚类数目会和正确的聚类数目相差较大，其稳定性较差。在其他几个算法下，GDD 的 Bias 值也是最低的。

同时，为了更直观地体现不同 CVI 在不同数据集下的评价结果和正确值之间的差距波动，选取 FCM 算法下实验结果绘制图 3，其中红色点画线是当前数据集的正确聚类数。分析图 3 中的折线图可以发现，本文提出的 GDD 指标是 12 个 CVI 中比较稳定的指标，在绝大多数数据集上评价结果都是正确的，在评价错误的数据集 Balance Scale 上，GDD 指标的结果也和正确的聚类数目相差不多。相较于其他 CVI，例如 CH 指标、Dunn 指标和 FSI 指标，这些 CVI 在大部分数据集上评价结果都和正确的聚类数目相差较大。

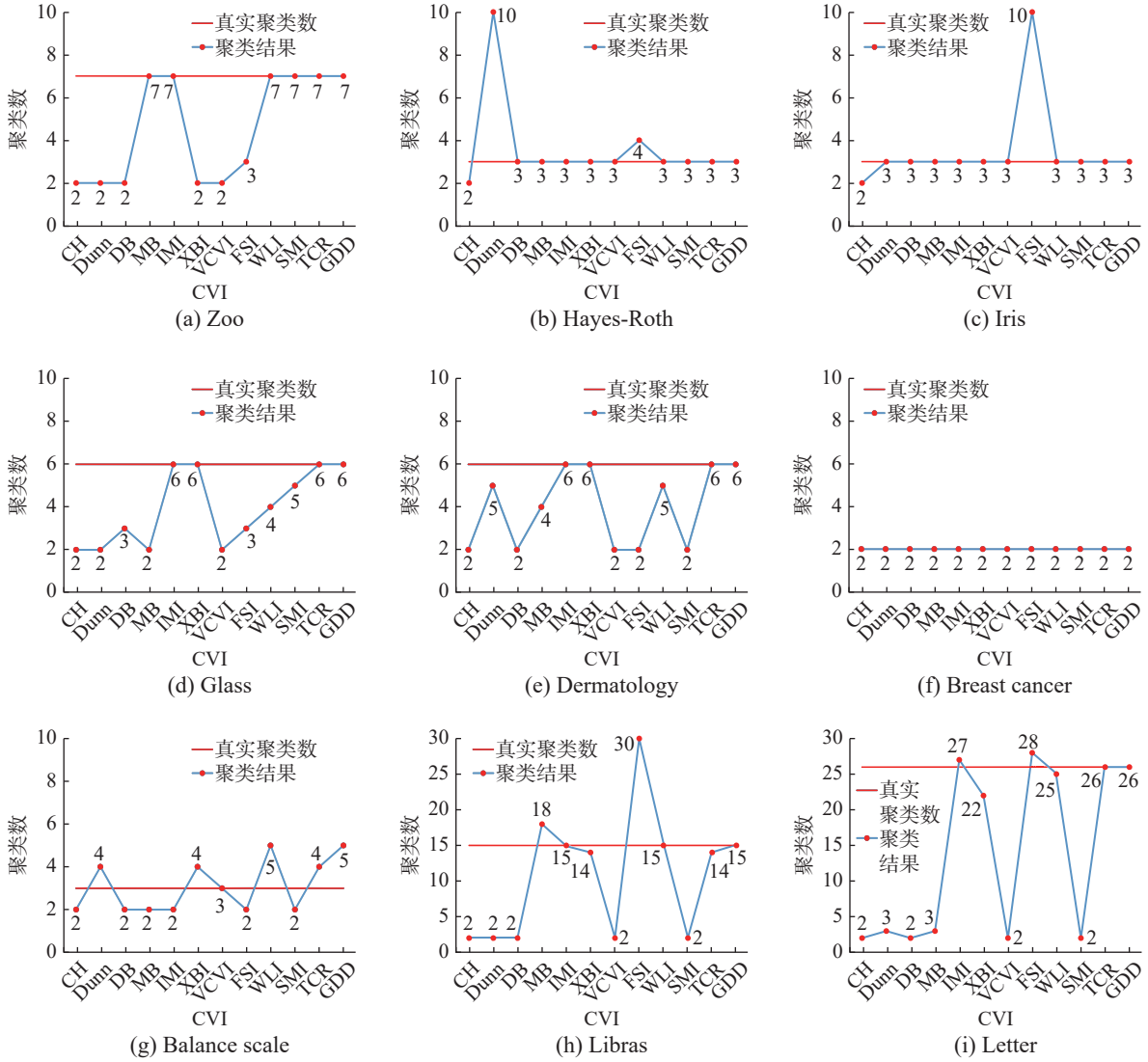


图 3　FCM 算法下 UCI 数据集的统计结果

Fig. 3　Statistical results of UCI dataset under the FCM algorithm

图 4 给出了 12 个指标在 FCM 算法下的 Dermatology 数据集的计算结果数值变化，其中红色“·”形状标记表示当前聚类数目下的 CVI 取最优值。这里，FSI 本应为值越小越好，在图 4 中将其纵坐标翻转后标记为最大值为最优聚类数。通过分析图 3、图 4，可以发现 IMI 指标、XBI 指标、TCR 指标和 GDD 指标在当前数据集上的聚类结果是正确的，其中正确的聚类数目为 6 类。然而分析

数值曲线的走势, 可以发现 CH 指标和 VCVI 指标的数值曲线随聚类数目 K 值的变化呈单调变化趋势, 这表明这些 CVI 在当前数据集上的评价结果失去实际意义, 难以有效反映聚类的真实质量。

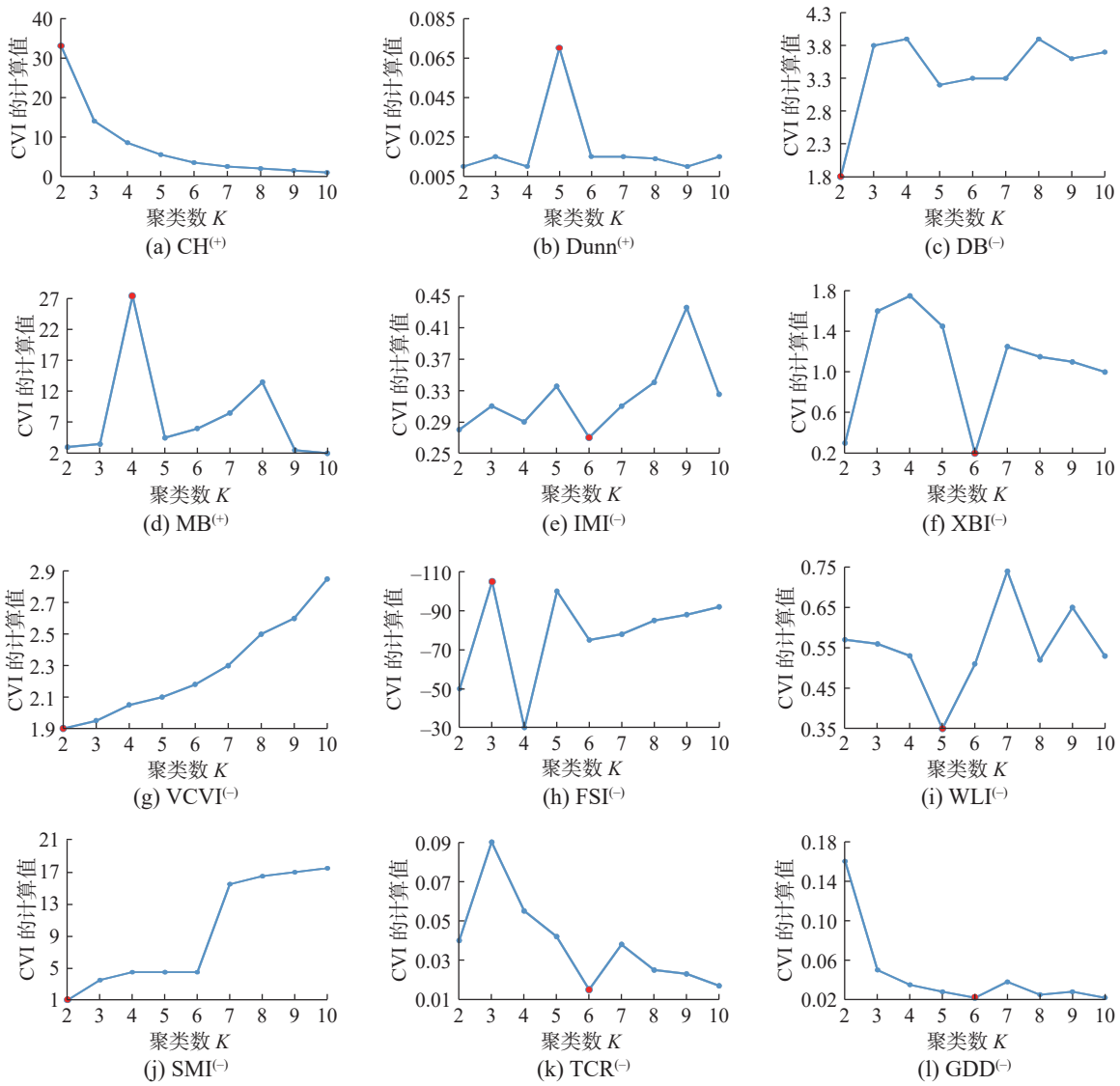


图 4　FCM 算法下 Dermatology 数据集运行结果

Fig. 4　Results of Dermatology dataset under FCM algorithm

表 5 中间部分是 PFCM 算法下 CVI 的实验结果。以其中的 Data77 数据集为例, Data77 数据集的正确聚类数目为 7 类, IMI 指标 10 轮的结果中 5 轮是 5 类为最优值, 4 轮是 7 类为最优值, 1 轮是 2 类为最优值, DB 指标 10 轮的结果中 8 轮是 3 类为最优值, 1 轮是 2 类为最优值, 1 轮是 4 类为最优值, 新提出的 GDD 指标 10 轮的结果中 7 轮是 7 类为最优值, 3 轮是 5 类为最优值, 其他 CVI 不再赘述。统计出现次数最多的最优值聚类数目作为该指标的评价结果, IMI 指标结果为 5 类, DB 指标结果为 3 类, GDD 指标结果为 7 类。可以看出, IMI 指标和 DB 指标的结果都不正确, GDD 指标正确得出了 Data77 数据集的聚类数目。统计表 5 中的数据, 选取实验结果中出现次数最多的聚类数目作为最终的指标评价结果。

表 5　人造数据集的实验结果

Table 5　Experimental results of artificial dataset

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Data_60	$3^{9}4^{1}$	$3^{7}4^{3}$	3^{10}	3^{10}	$3^{8}4^{2}$	3^{10}	3^{10}	3^{10}	$3^{9}4^{1}$	3^{10}	$3^{8}4^{2}$	3^{10}
	Data_150	3^{10}	$2^{1}3^{9}$	$2^{2}3^{7}5^{1}$	3^{10}	$2^{1}3^{9}$	3^{10}	$2^{1}3^{8}4^{1}$	3^{10}	3^{10}	3^{10}	$2^{1}3^{9}$	3^{10}
	Circle	$2^{2}3^{7}4^{1}$	$2^{2}3^{6}4^{2}$	$2^{6}3^{4}$	3^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}

续表 5

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Jain	$2^2 3^8$	$2^3 3^7$	$2^3 3^6 5^1$	2^{10}	2^{10}	$2^4 3^3 4^3$	$2^9 3^1$	$3^2 4^8$	$2^8 4^2$	2^{10}	$2^7 4^3$	$2^7 3^1 4^2$
	X8D5K	2^{10}	$4^1 5^8 7^1$	$2^9 5^1$	3^{10}	$4^2 5^8$	$3^5 4^2 5^3$	$2^1 4^6 5^3$	4^{10}	5^{10}	$5^7 6^1 7^2$	$3^1 5^9$	5^{10}
	Data_77	$2^9 4^1$	2^{10}	$2^1 3^7 4^2$	$6^8 7^2$	$4^1 5^6 7^3$	$4^2 5^8$	5^{10}	5^{10}	$4^3 5^6 7^1$	$2^2 4^3 5^5$	$5^5 6^3 7^2$	$5^2 6^1 7^7$
	E6	2^{10}	$2^6 5^4$	$4^2 5^8$	2^{10}	$2^4 5^6$	4^{10}	$2^1 3^3 4^6$	2^{10}	$2^2 4^8$	4^{10}	$4^9 5^1$	4^{10}
	Dim_128	$2^{24} 4^3$ 6^3	$2^{22} 3^3$ $5^4 7^1$	$2^{20} 3^3$ 4^7	$2^{17} 3^{10}$ 5^3	$9^1 15^9$ $16^{16} 17^4$	$14^{17} 15^5$ $16^7 20^1$	$2^{15} 3^9$ $4^2 5^4$	$24^1 27^3$ $28^9 30^{17}$	$10^1 14^{17}$ $15^5 16^7$	$2^{29} 4^1$	$9^1 11^2 13^2$ $15^7 16^{18}$	$10^3 12^2$ $15^5 16^{20}$
	Dim_256	$2^{20} 3^2$ $4^5 6^3$	$2^{19} 3^7$ $4^2 6^2$	$2^{19} 3^5$ 6^6	$2^7 3^{12}$ $5^6 10^5$	$10^2 15^2$ $16^{19} 17^7$	$12^3 14^5$ $15^{18} 16^4$	$2^6 3^{14}$ $5^8 10^2$	$28^9 29^4$ 30^{17}	$14^7 15^{17}$ $16^3 17^3$	2^{30}	$9^1 14^2$ $15^{14} 16^{13}$	$10^3 15^2$ $16^{21} 17^4$
PFCM	Data_60	3^{10}	$3^7 4^2 6^1$	3^{10}	3^{10}	$3^9 5^1$	3^{10}	$2^1 3^7 4^2$	3^{10}	$2^1 3^9$	3^{10}	$2^2 3^8$	$3^9 4^1$
	Data_150	3^{10}	$2^2 3^8$	$3^9 4^1$	3^{10}	3^{10}	3^{10}	3^{10}	$2^2 3^8$	3^{10}	3^{10}	3^{10}	3^{10}
	Circle	$2^4 3^6$	$2^4 3^3 4^3$	$2^6 3^2 4^2$	3^{10}	2^{10}	2^{10}	2^{10}	2^{10}	$2^9 3^1$	2^{10}	2^{10}	2^{10}
	Jain	2^{10}	$2^2 3^8$	$2^3 3^6 5^1$	2^{10}	2^{10}	$2^4 3^3 4^3$	2^{10}	$2^2 4^8$	$2^8 3^1 4^1$	2^{10}	2^{10}	$2^8 3^1 4^1$
	X8D5K	2^{10}	$2^1 4^1 5^7 7^1$	2^{10}	3^{10}	$4^2 5^8$	$3^7 5^3$	4^{10}	$4^7 5^3$	5^{10}	5^{10}	$3^2 5^8$	5^{10}
	Data_77	$2^9 4^1$	2^{10}	$2^1 3^8 4^1$	$5^2 6^7 7^1$	$2^1 5^5 7^4$	$4^3 5^7$	$5^4 6^3 7^3$	5^{10}	$3^1 4^2 5^7$	$2^1 4^4 5^5$	$5^5 6^1 7^4$	$4^1 5^1 7^8$
	E6	2^{10}	$2^6 4^1 5^3$	$2^1 4^1 5^8$	2^{10}	$2^4 5^6$	4^{10}	$4^6 5^4$	2^{10}	$2^2 4^8$	4^{10}	$2^1 4^6 5^3$	4^{10}
	Dim_128	$2^{21} 3^4$ $4^3 6^2$	$2^{23} 4^3$ 5^4	$2^{20} 3^3$ $4^4 9^3$	$2^{14} 3^{11} 5^2$ $9^2 11^1$	$10^2 15^8$ $16^{17} 17^3$	$9^6 14^9$ $15^5 16^{10}$	$2^{14} 3^9$ $4^2 5^5$	$24^1 27^2$ $28^{11} 30^{16}$	$14^{10} 15^4$ $16^{15} 20^1$	2^{30}	$13^2 15^{11}$ $16^{16} 20^1$	$15^6 16^{20}$ 17^4
	Dim_256	$2^{25} 4^5$	$2^{21} 3^5$ $4^2 9^2$	2^{30}	$2^{14} 3^9$ $5^6 7^1$	$15^2 16^{20}$ 17^8	$10^2 14^5$ $15^{20} 16^3$	$2^3 3^{16}$ $5^8 6^3$	$26^2 28^7 30^{21}$	$14^{18} 15^5 16^2$ $17^2 20^3$	2^{30}	$15^{18} 16^{12}$	$10^1 14^2$ $16^{21} 17^6$
KFCM	Data_60	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}
	Data_150	3^{10}	$3^7 4^3$	3^{10}	3^{10}	3^{10}	3^{10}	$2^1 3^9$	3^{10}	3^{10}	3^{10}	3^{10}	3^{10}
	Circle	$2^3 3^7$	$3^6 4^4$	$2^7 3^3$	3^{10}	$2^8 4^2$	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}
	Jain	$2^2 3^7 4^1$	$3^7 4^2 5^1$	$2^2 3^6 4^2$	2^{10}	2^{10}	$2^4 3^3 4^3$	2^{10}	$2^1 3^1 4^8$	$2^8 3^1 4^1$	2^{10}	$2^7 4^3$	$2^3 3^4 4^3$
	X8D5K	$2^9 3^1$	$4^2 5^8$	2^{10}	$2^4 3^6$	$4^6 5^4$	$2^1 3^8 4^1$	4^{10}	4^{10}	$4^2 5^8$	5^{10}	$3^4 5^6$	5^{10}
	Data_77	$2^8 5^2$	2^{10}	$3^5 4^2 5^3$	$5^2 6^8$	$4^1 5^6 7^3$	$4^2 5^7 6^1$	5^{10}	5^{10}	$3^1 4^3 5^6$	$4^3 5^5 7^1$	$5^4 6^3 7^3$	$4^1 5^2 7^7$
	E6	2^{10}	$2^6 5^4$	$4^2 5^8$	2^{10}	$2^3 4^1 5^6$	4^{10}	$3^3 4^6 5^1$	2^{10}	$2^2 4^8$	4^{10}	$2^1 4^9$	4^{10}
	Dim_128	$2^{25} 3^1$ $4^2 6^2$	$2^{21} 3^4$ $5^3 7^2$	2^{30}	$2^{16} 3^9$ $5^2 7^3$	$10^2 15^8$ $16^{16} 17^4$	$10^1 14^{17}$ $15^5 16^7$	$2^4 3^{16}$ $4^5 5^3 7^2$	$24^5 25^{10}$ 30^{15}	$14^{20} 15^5$ 17^5	2^{30}	$14^1 15^{14}$ 16^{15}	$14^4 15^2$ 16^{24}
	Dim_256	$2^{22} 4^5$ 6^3	$2^{22} 3^5$ $4^2 6^1$	2^{30}	$2^9 3^{15}$ $5^5 7^1$	$14^2 16^{20}$ 17^8	$14^2 15^{22}$ $16^5 18^1$	$2^5 3^{15} 5^{10}$	$25^{12} 30^{18}$	$14^{14} 16^6$ 17^{10}	$2^{25} 3^5$	$14^2 15^{16}$ 16^{12}	$15^2 16^{22}$ 17^6

和 UCI 数据集类似，同样在统计结果上使用 Eff 和 Bias 两个评价因子来对结果进行统计分析，形成 Eff 统计表和 Bias 统计表如表 6、表 7 所示。分析表 6 的统计数据可以看出，本文提出的 GDD 指标在人造数据集上的实验结果中相较于其他 CVI 准确率更高。GDD 指标在 FCM 算法、PFCM 算法上对所有人造数据集的评价结果都是正确的，在 KFCM 算法上对 Jain 数据集的评价结果错误，准确率分别达到 100%、100% 和 89%。其他准确率较高的指标，例如 IMI 指标和 TCR 指标，其在 FCM 算法和 PFCM 算法上的准确率在 78% 左右，在 KFCM 算法上的准确率稍低一点，在 70% 左右。WLI 指标和 SMI 指标的准确率稍低一点，在 67% 左右。剩余的指标准确率大多都在 60% 以下。从实验结果可以分析出，新提出的 GDD 指标在 3 个算法上的准确率均高于其他 CVI，说明 GDD 指标对于不同算法的适应力均优于其他 CVI。同时对比不同 CVI 在不同算法下的准确率，例如 XBI 指标在 FCM 算法和 KFCM 算法上的准确率为 56%，而在 PFCM 算法下的准确率为 67%，这在一定程度上说明不同的 CVI 有着更契合 CVI 本身的算法。

表 6　人造数据集的 Eff 统计结果

Table 6　Eff statistical results of artificial dataset

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Data_60	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Data_150	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**

续表 6

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Circle	3	3	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Jain	3	3	3	**2**	**2**	**2**	**2**	4	**2**	**2**	**2**	**2**
	X8D5K	2	**5**	2	3	**5**	3	4	4	**5**	**5**	**5**	**5**
	Data_77	2	2	3	6	5	5	5	5	5	5	5	**7**
	E6	2	2	5	2	5	**4**	**4**	2	**4**	**4**	**4**	**4**
	Dim_128	2	2	2	2	**16**	14	2	30	14	2	**16**	**16**
	Dim_256	2	2	2	3	**16**	15	3	30	15	2	15	**16**
	Eff	0.22	0.33	0.33	0.33	0.78	0.56	0.56	0.33	0.67	0.67	0.78	1
PFCM	Data_60	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Data_150	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Circle	3	**2**	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Jain	**2**	3	3	**2**	**2**	**2**	**2**	4	**2**	**2**	**2**	**2**
	X8D5K	2	**5**	2	3	**5**	3	4	4	**5**	**5**	**5**	**5**
	Data_77	2	2	3	6	5	5	5	5	5	5	5	**7**
	E6	2	2	5	2	5	**4**	**4**	2	**4**	**4**	**4**	**4**
	Dim_128	2	2	[2]	2	**16**	**16**	2	30	**16**	2	**16**	**16**
	Dim_256	2	2	2	2	**16**	15	3	30	14	2	15	**16**
	Eff	0.33	0.44	0.33	0.33	0.78	0.67	0.56	0.33	0.78	0.67	0.78	1
KFCM	Data_60	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Data_150	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	Circle	3	3	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
	Jain	3	3	3	**2**	**2**	**2**	**2**	4	**2**	**2**	**2**	3
	X8D5K	2	**5**	2	3	4	3	4	4	**5**	**5**	**5**	**5**
	Data_77	2	2	3	6	5	5	5	5	5	5	5	**7**
	E6	2	2	5	2	5	**4**	**4**	2	**4**	**4**	**4**	**4**
	Dim_128	2	2	2	2	**16**	14	3	30	14	2	**16**	**16**
	Dim_256	2	2	2	3	**16**	15	3	30	14	2	15	**16**
	Eff	0.22	0.33	0.33	0.33	0.67	0.56	0.56	0.33	0.67	0.67	0.78	0.89

注: 加粗代表该数值和正确的聚类数目一致。

表 7　人造数据集的 Bias 统计结果

Table 7　Bias statistical results of artificial dataset

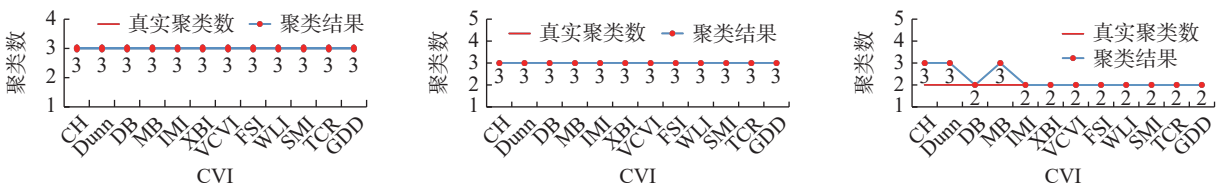
算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Data_60	0	0	0	0	0	0	0	0	0	0	0	0
	Data_150	0	0	0	0	0	0	0	0	0	0	0	0
	Circle	1	1	0	1	0	0	0	0	0	0	0	0
	Jain	1	1	1	0	0	0	0	2	0	0	0	0
	X8D5K	3	0	3	2	0	2	1	1	0	0	0	0
	Data_77	5	5	4	1	2	2	2	2	2	2	2	0
	E6	2	2	1	2	1	0	0	2	0	0	0	0
	Dim_128	14	14	14	14	0	2	14	14	2	14	0	0
	Dim_256	14	14	14	13	0	1	13	14	1	14	1	0
	Bias	40	37	37	33	3	7	30	35	5	30	3	0

续表 7

算法	数据集	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
PFCM	Data_60	0	0	0	0	0	0	0	0	0	0	0	0
	Data_150	0	0	0	0	0	0	0	0	0	0	0	0
	Circle	1	0	0	1	0	0	0	0	0	0	0	0
	Jain	0	1	1	0	0	0	0	2	0	0	0	0
	X8D5K	3	0	3	2	0	2	1	1	0	0	0	0
	Data_77	5	5	4	1	2	2	2	2	2	2	2	0
	E6	2	2	1	2	1	0	0	2	0	0	0	0
	Dim_128	14	14	14	14	0	0	14	14	0	14	0	0
	Dim_256	14	14	14	14	0	1	13	14	2	14	1	0
	Bias	39	36	37	34	3	5	30	35	4	30	3	0
KFCM	Data_60	0	0	0	0	0	0	0	0	0	0	0	0
	Data_150	0	0	0	0	0	0	0	0	0	0	0	0
	Circle	1	1	0	1	0	0	0	0	0	0	0	0
	Jain	1	1	1	0	0	0	0	2	0	0	0	1
	X8D5K	3	0	3	2	1	2	1	1	0	0	0	0
	Data_77	5	5	4	1	2	2	2	2	2	2	2	0
	E6	2	2	1	2	1	0	0	2	0	0	0	0
	Dim_128	14	14	14	14	0	2	13	14	2	14	0	0
	Dim_256	14	14	14	13	0	1	13	14	2	14	1	0
	Bias	40	37	37	33	4	7	29	35	6	30	3	1

分析表 7 中的数据，以 FCM 算法下的运行结果为例，Bias 值比较低的有 IMI 指标、WLI 指标、TCR 指标和 GDD 指标，其中 GDD 指标的 Bias 值最低，足以说明本文提出的 GDD 指标在人造数据集上有更好的评价效果和更稳定的评价能力，即使在 GDD 指标给出错误的评价结果的情况下，得到的错误聚类数目和正确的聚类数目也不会相差太多。相反，其他 CVI 例如 CH 指标、Dunn 指标和 DB 指标，这些 CVI 的 Bias 值都在 40 上下，相较于其他 CVI 高出很多。这说明 CH 指标、Dunn 指标和 DB 指标这些 CVI 在错误评价后得到的聚类数目会和正确的聚类数目相差较大，其稳定性较差。

同时，为了更直观地体现不同 CVI 在不同数据集下的评价结果和正确值之间的差距波动，选取 FCM 算法下实验结果绘制图 5，其中红色点画线是当前数据集的正确聚类数。分析图 5 中的折线图，可以发现，本文提出的 GDD 指标是 12 个 CVI 中比较稳定的指标，在绝大多数数据集上评价结果都是正确的，在 KFCM 算法下评价错误的数据集 Jain 上，GDD 指标的结果也和正确的聚类数目相差不多。如 CH 指标、Dunn 指标和 FSI 指标，这些 CVI 在大部分数据集上评价结果都和正确的聚类数目相差较大。



(a) Data_60　(b) Data_150　(c) Circle

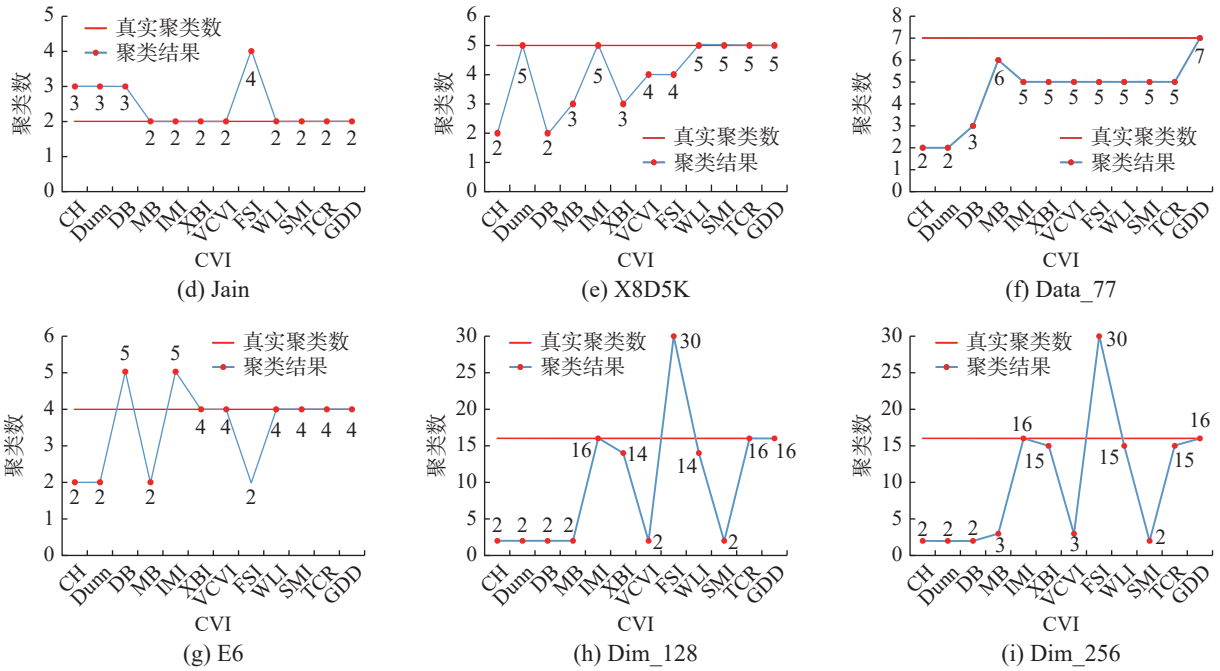


图 5　FCM 算法下人造数据集的统计结果

Fig. 5　Statistical results of artificial dataset under FCM algorithm

接下来使用 Olivetti Face 数据集进行实验来检验本文提出的 GDD 指标面对图像类数据集的效果, 验证其能否适应不同类型的数据。表 8 是 Olivetti Face 数据集在 FCM、PFCM 和 KFCM 这 3 种算法下的运行结果。统计表 8 中的数据, 选取实验结果中出现次数最多的聚类数目作为最终的指标评价结果, 同时应用 Eff 和 Bias 评价因子以形成表 9。

表 8　Olivetti Face 数据集的实验结果

Table 8　Experimental results of Olivetti Face dataset

算法	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	30^{30}	$2^{17}3^{5}$ $4^{5}6^{3}$	2^{30}	$2^{23}3^{2}$ $4^{1}6^{2}9^{2}$	$9^{5}10^{23}$ 12^{2}	$25^{6}29^{5}$ 30^{19}	$2^{1}3^{4}4^{25}$	30^{30}	$5^{2}8^{3}$ $9^{5}10^{20}$	2^{30}	$10^{20}11^{10}$	$8^{5}9^{3}$ 10^{22}
PFCM	30^{30}	$2^{21}3^{1}$ $4^{5}5^{3}$	2^{30}	$2^{21}5^{3}$ $9^{4}10^{2}$	$2^{1}8^{2}$ $9^{4}10^{23}$	$22^{3}25^{4}$ $28^{4}30^{19}$	$3^{5}4^{25}$	30^{30}	$7^{3}9^{1}$ $10^{23}11^{3}$	2^{30}	$8^{2}10^{19}$ $11^{5}14^{4}$	$5^{1}9^{4}$ $10^{21}12^{4}$
KFCM	$25^{1}29^{1}$ 30^{28}	$2^{10}3^{16}$ $5^{3}7^{1}$	2^{30}	$2^{25}4^{5}$	$8^{2}9^{4}$ $10^{22}20^{2}$	$22^{4}29^{5}$ 30^{21}	$3^{2}4^{21}$ $5^{5}10^{2}$	30^{30}	$9^{5}10^{22}11^{3}$	2^{30}	$9^{3}10^{22}$ $11^{3}14^{2}$	$9^{5}10^{19}$ $11^{2}13^{4}$

表 9　Olivetti Face 数据集的统计结果

Table 9　Statistical results of Olivetti Face dataset

算法	评价因子	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	Value	30	2	2	2	**10**	30	4	30	**10**	2	**10**	**10**
	Eff	0	0	0	0	1	0	0	0	1	0	1	1
	Bias	20	8	8	8	0	20	6	20	0	8	0	0
PFCM	Value	30	2	2	2	**10**	30	4	30	**10**	2	**10**	**10**
	Eff	0	0	0	0	1	0	0	0	1	0	1	1
	Bias	20	8	8	8	0	20	6	20	0	8	0	0
KFCM	Value	30	3	2	2	**10**	30	4	30	**10**	2	**10**	**10**
	Eff	0	0	0	0	1	0	0	0	1	0	1	1
	Bias	20	7	8	8	0	20	6	20	0	8	0	0

注: 加粗代表该数值和正确的聚类数目一致。

评估一个 CVI 的优劣不仅需要各种类型、各种结构的数据集进行实验，同样也需要检测 CVI 在含有噪声的数据集上的效果。一个稳定的 CVI 应该能适应添加有不同程度噪声数据的数据集，并且能得出正确的结果。接下来选取 X8D5K 数据集，X8D5K 数据集的数据维度为 8 维，共有 1 000 个样本点，正确的聚类数目为 5 类。向其中加入不同程度的噪声数据，分别为 2%、4%、6%、8% 和 10%。在这 5 种噪声下的数据集上进行实验，分别使用 FCM 算法、PFCM 算法和 KFCM 算法。统计实验后的结果如表 10 所示。

表 10　噪声数据集的统计结果

Table 10　Statistical results of noisy dataset

算法	噪声比例/%	CH	Dunn	DB	MB	IMI	XBI	VCVI	FSI	WLI	SMI	TCR	GDD
FCM	0	2	**5**	2	3	**5**	3	4	4	**5**	**5**	**5**	**5**
	2	3	**5**	2	3	**5**	4	**5**	4	**5**	4	**5**	**5**
	4	2	**5**	2	3	4	4	4	4	**5**	4	**5**	**5**
	6	2	4	3	3	3	3	3	6	4	4	**5**	**5**
	8	2	4	3	3	3	4	3	6	3	3	6	**5**
	10	2	3	3	3	2	3	3	**5**	2	3	4	4
PFCM	0	2	**5**	2	3	**5**	3	4	4	**5**	**5**	**5**	**5**
	2	2	**5**	2	3	**5**	3	4	3	**5**	4	**5**	**5**
	4	3	**5**	3	3	**5**	4	3	4	**5**	**5**	**5**	**5**
	6	3	4	3	3	4	**5**	3	3	4	7	4	**5**
	8	2	3	4	3	3	**5**	3	2	3	4	3	**5**
	10	2	3	**5**	3	3	4	2	2	3	4	3	**5**
KFCM	0	2	**5**	2	3	4	3	4	4	**5**	**5**	**5**	**5**
	2	2	**5**	**5**	3	**5**	3	4	3	**5**	**5**	**5**	**5**
	4	2	4	**5**	3	4	4	4	4	5	4	**5**	**5**
	6	2	3	4	3	4	3	3	3	4	**5**	4	**5**
	8	2	4	3	3	3	3	3	3	3	4	4	**5**
	10	2	2	3	4	3	2	4	3	3	3	3	6

注：加粗代表该数值和正确的聚类数目一致。

4　结束语

1) 本文提出了一个新的模糊聚类有效性评价指标 GDD。GDD 指标采用簇内紧致度与簇间分离度比值的形式抑制了聚类结果数目单调递增的情况，同时在簇内紧致度中在计算单个簇紧致度的基础上引入了该簇样本占总样本的比值，构建了更客观的紧致度表达；在簇间分离度中在计算聚类中心到聚类中心均值的基础上增加了该簇样本站总样本的比值，构建了更合理的簇间分离度表达。

2) 在复杂度分析中，虽然 GDD 指标拥有与 Dunn 和 SMI 指标相同的时间复杂度，但是其拥有更加优异的性能，并且性能同样优于其他时间复杂度更低的指标，说明了 GDD 在能接受的时间复杂度下达到了比较指标中最好的性能。

3) 在面对 3 个模糊聚类算法、11 个对比 CVI 的实验中，GDD 指标在 UCI 数据集中均优于其他指标，在人造数据集的实验中 GDD 仅在 KFCM 算法下的 Jain 数据集出现了错误，在人脸数据集 Olivetti Face 中 GDD 同样得到了较为优秀的效果。

4) 在噪声实验中，FCM 算法下噪声为 8% 时只有 GDD 指标得到了正确结果，而噪声为 10% 时只有 FSI 得到了正确结果。PFCM 算法下当噪声为 10% 时只有 DB 和 GDD 能够得到正确结果。KFCM 算法下当噪声为 8% 时只有 GDD 得到了正确结果，而噪声为 10% 时则没有指标得到正确聚类数。

综上，GDD 指标具有全新的指标设计思路、可接受的时间复杂度与较好的实验表现，在噪声

实验中体现了其较强的鲁棒性。在未来研究中，将试图结合模糊逻辑的思想来进行聚类有效性评价指标的设计与应用。

参考文献:

[1] TANG Yiming, PAN Zhifu, HU Xianghui, et al. Knowledge-induced multiple kernel fuzzy clustering[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(12): 14838−14855.

[2] 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.

[3] TANG Yiming, REN Fuji, PEDRYCZ W. Fuzzy C-Means clustering through SSIM and patch for image segmentation[J]. Applied soft computing, 2020, 87: 1−16.

[4] TANG Yiming, WU Wenbin, PEDRYCZ W, et al. Clustering interval and triangular granular data: modeling, execution, and assessment[J]. IEEE transactions on neural networks and learning systems, 2025, 36(6): 10000−10014.

[5] TANG Yiming, LI Bing, PEDRYCZ W, et al. A clustering validity index with multi-granularity fusion for multiple fuzzy clustering algorithms[J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(10): 8379−8396.

[6] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California, 1965.

[7] ROSS H H, SOKAL R R, SNEATH P H A, et al. Principles of numerical taxonomy[J]. Systematic zoology, 1964, 13(2): 106.

[8] CAMPELLO R J G B, MOULAVI D, SANDER J. Density-based clustering based on hierarchical density estimates[C]//The 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Gold Coast: PAKDD, 2013.

[9] 吕莉, 陈威, 肖人彬, 等. 面向密度分布不均数据的加权逆近邻密度峰值聚类算法[J]. 智能系统学报, 2024, 19(1): 165−175.
LYU Li, CHEN Wei, XIAO Renbin, et al. Density peak clustering algorithm based on weighted reverse nearest-neighbor for uneven density datasets[J]. CAAI transactions on intelligent systems, 2024, 19(1): 165−175.

[10] ZHAO Yanchang, SONG Junde. GDILC: a grid-based density-isoline clustering algorithm[C]//Proceedings of the International Conference on Information Technology and Intelligent Network. Beijing: IEEE, 2001.

[11] VIJAY R K, NANDA S J, SHARMA A. A spatio-temporal binary grid-based clustering model for seismicity analysis[J]. Pattern analysis and applications, 2024, 27(1): 14.

[12] 孙林, 梁娜, 徐久成. 基于邻域互信息与 K-means 特征聚类的特征选择[J]. 智能系统学报, 2024, 19(4): 983−996.
SUN Lin, LIANG Na, XU Jiucheng. Feature selection using neighborhood mutual information and feature clustering with K-means[J]. CAAI transactions on intelligent systems, 2024, 19(4): 983−996.

[13] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of cybernetics, 1973, 3(3): 32−57.

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

[15] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE transactions on fuzzy systems, 1993, 1(2): 98−110.

[16] PAL N R, PAL K, BEZDEK J C. A mixed C-means clustering model[C]//Proceedings of 6th International Fuzzy Systems Conference. Barcelona: IEEE, 1997.

[17] ANTOINE V, GUERRERO J A, ROMERO G. Possibilistic fuzzy c-means with partial supervision[J]. Fuzzy sets and systems, 2022, 449: 162−186.

[18] ZHANG Daoqiang, CHEN Songcan, PAN Zhisong, et al. Kernel-based fuzzy clustering incorporating spatial constraints for image segmentation[C]//Proceedings of the 2003 International Conference on Machine Learning and Cybernetics. Washington: IEEE, 2003.

[19] PUNIT R, ZAHRA G, BEZDEK J C, et al. Approximating Dunn's cluster validity indices for partitions of big data[J]. IEEE transactions on cybernetics, 2018, 49(5): 1629−1641.

[20] HASSAN B A, TAYFOR N B, HASSAN A A, et al. From a-to-z review of clustering validation indices[J]. Neurocomputing, 2024, 601: 128−198.

[21] FUKUI K, NUMAO M. Neighborhood-based smoothing of external cluster validity measures[C]//Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2012: 354−365.

[22] 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.

[23] VENDRAMIN L, CAMPELLO R J G B, HRUSCHKA E

R. Relative clustering validity criteria: a comparative overview[J]. Statistical analysis & data mining, 2010, 3(4): 209−235.

[24] CALINSKI R B, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in statistics, 1974, 3(1): 1−27.

[25] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE transactions on pattern analysis and machine intelligence, 1979, 2(1): 224−227.

[26] 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.

[27] FUKUYAMA Y, SUGENO M. A new method of choosing the number of clusters for the fuzzy c-means method[C]// Proceedings of the 5th Fuzzy Systems Symposium. Tokyo: Japan Society for Fuzzy Theory and Systems, 1989.

[28] WU Chihhung, 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.

[29] LIU Yun, JIANG Yanfang, HOU Tao, et al. A new robust fuzzy clustering validity index for imbalanced data sets[J]. Information sciences, 2021, 547: 579−591.

[30] MITTAL H, SARASWAT M. A new fuzzy cluster validity index for hyperellipsoid or hyperspherical shape close clusters with distant centroids[J]. IEEE transactions on fuzzy systems, 2020, 29(11): 3249−3258.

[31] ZHU Erzhuo, MA Zhujuan, LI Xuejun, et al. An effective partitional clustering algorithm based on new clustering validity index[J]. Applied soft computing, 2018, 71: 608−621.

[32] MAULIKH 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.

[33] TANG Yiming, HUANG Jiajia, PEDRYCZ W, et al. A fuzzy clustering validity index induced by triple center relation[J]. IEEE transactions on cybernetics, 2023, 53(8): 5024−5036.

[34] LIU Yun, HOU Tao, LIU Fu, et al. Improving fuzzy c-means method for unbalanced dataset[J]. Electronics letters, 2015, 51(23): 1880−1881.

[35] ZHOU Kaile, YANG Shanlin. Exploring the uniform effect of FCM clustering: a data distribution perspective[J]. Knowledge-based systems, 2016, 96: 76−83.

[36] DUA D, GRAFF C. UCI machine learning repository [EB/OL]. (2017−01−01)[2025−03−20]. http://archive.ics.uci.edu/ml.

[37] SALEM S A, NANDI A K. Development of assessment criteria for clustering algorithms[J]. Pattern analysis and applications, 2009, 12(1): 79−98.

作者简介：

唐益明，教授，博士，主要研究方向为聚类、模糊逻辑与推理、情感计算和图像处理。主持国家自然科学基金项目 4 项。发表学术论文 100 余篇，获国家发明专利授权 8 项。E-mail：tym608@163.com。

刘子龙，硕士研究生，主要研究方向为聚类和聚类有效性指标。E-mail：2024170934@mail.hfut.edu.cn。

高健玮，博士研究生，主要研究方向为聚类、粒计算和模糊推理。E-mail：jwgao810@163.com。

[责任编辑：刘冰洁]