第 58 卷 第 8 期/2021 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于 Jeffrey 散度相似性度量的加权 FCM 聚类算法

吴辰文，马宁*，蒋雨璠

兰州交通大学电子与信息工程学院，甘肃 兰州 730070

摘要 针对模糊 C 均值(FCM)聚类算法在数据集下聚类效果差的情况，以及基于欧氏距离的相似性度量只考虑数据点之间的局部一致性问题，提出了基于 Jeffery 散度相似性度量加权 FCM 聚类算法(JW-FCM)。引入源于 Jeffery 散度的相似性度量，首先，对于 FCM 算法进行特征加权，对数据的不同特征值赋予适当的权重，再将 Jeffery 散度与加权 FCM 算法进行结合得到 JW-FCM 算法。将 JW-FCM 算法与几种相关算法在人工数据集和 UCI 数据集上进行对比实验，通过实验分析与比较，证明了 JW-FCM 算法具有更好的收敛性、鲁棒性、准确性。实验结果表明，改进算法表现出较好的聚类效果。

关键词 图像处理；聚类算法；加权模糊 C 均值算法；Jeffrey 散度

中图分类号 TP301.6 **文献标志码** A **doi**：10.3788/LOP202158.0810006

Weighted FCM Clustering Algorithm Based on Jeffrey Divergence Similarity Measure

Wu Chenwen, Ma Ning*, Jiang Yufan

School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, China

Abstract In view of the poor clustering effect of the fuzzy C mean (FCM) clustering algorithm under the data set, and the similarity measure based on Euclidean distance only considers the local consistency between data points. This paper presents a weighted FCM clustering algorithm based on Jeffrey divergence similarity measure (JW-FCM), and introduces the similarity measure derived from Jeffery divergence. First, perform feature weighting on the FCM algorithm, assign appropriate weights to different feature values of the data, and then combine the Jeffery divergence with the weighted FCM algorithm to obtain the JW-FCM algorithm. The JW-FCM algorithm is compared with several related algorithms on the artificial data set and UCI data set. Through experimental analysis and comparison, it is proved that the JW-FCM algorithm has better convergence, robustness, and accuracy. The experimental results show that the improved algorithm shows better clustering effect.

Key words image processing; clustering algorithm; weighted fuzzy C means algorithm; Jeffrey divergence

OCIS codes 100.5010; 100.2960

1 引 言

聚类是模式识别领域的一个重要的研究方向[1]，它是基于某些预先固定的相同或不相同的度量来寻找有意义组的方法，在统计数据分析中被广泛应用[2]。聚类在有效的数据分析过程中起着关键作用，它从原始数据集提取必要信息和粒度信息(指信息单元的相对大小或粗糙程度)。与监督学习和判别分析不同，它不涉及标记数据或训练集。在数据挖掘、信息检索、机器学习和计算机视觉等领域，

收稿日期：2020-07-13；**修回日期**：2020-08-11；**录用日期**：2020-09-09

基金项目：国家自然科学基金(61662043，61762057)

* **E-mail**：2996771799@qq.com

聚类分析是一项核心技术。已有大量研究基于不同的假设(如连通性、形心、分布、密度和子空间等)提出了多个聚类算法。由于不同的算法会获得不同的效果,因此很难在实践中确定使用哪种算法[3]。经典算法有 K-means[4]、模糊 C 均值(FCM)[5]、DBSCAN[6]等。

然而,现如今的聚类算法基本上还存在一些问题:1)聚类结果可能是局部最优解,例如 K-means 聚类算法类似于爬山法,会在最近的最优点停止迭代,但是这个最优点不一定是全局的最优点;2)聚类过程中容易受到噪声和环境因素的影响等,因此聚类结果比较依赖于相似性度量。一般来说,在聚类算法中,欧氏相似性度量经常被用来最小化每个点到其最近中心的均方距离。欧氏距离不能总是找到更精确的聚类边界,本研究采用加权的 FCM 聚类方法作为衡量相似性度量方法的基础算法。近年来,研究者们利用非线性相似性度量代替传统的欧氏距离来确定更精确的聚类边界,相对较为熟悉的是三角不等式性质[7]。Banerjee 等[8]在 2005 年使用广义 Bregman 散度作为一种相似性度量方法与 K-means 方法进行结合,改进了传统 K-means 方法的性能。

在目前的模糊聚类算法研究中,FCM 聚类算法有较为完善的理论基础,同时 FCM 算法能计算样本对所有类的隶属度,提供了参考该样本结果可靠性的计算方式。但是 FCM 聚类算法仍然存在如下缺点:1)FCM 算法是一种有效的图像分割算法,但对噪声图像分割效果较差[9];2)加权指数 b 的选取;3)目标函数中相似性度量的定义等。因此不仅需要去探究与解决其聚类结果不稳定性及噪声敏感性,同时还需要提高算法的聚类精确性。

FCM 在欧氏距离和外界因素的影响下,聚类结果将会出现偏差。对于聚类算法而言,度量方式的选择对最终的聚类效果起着至关重要的作用。特征加权是一种接近个体特征最优影响程度的技术。每个特征值的大小都衡量了该特征的重要程度,本文将此特征值称为特征权重,其值在区间[0,1]中。在相应的算法中通过一个学习机制来适应权重,用较小的权重值来表示噪声特征或低质量特征,因此对于差异性的计算贡献不大。如果权重值只有 0 或 1(即完全消除不良特征),则特征加权就减少了特征选择的过程。特征选择可以显著减少学习机制中的计算复杂程度,同时可以尽可能避免重要信息的丢失。又由于 Jeffrey 散度[10]具备较好的数值稳定性和对噪声的鲁棒性,从 Jeffrey 散度出发,本文提出相似性度量的概念,解释了基于 Jeffrey 散度的相似性度量的各种性质,提出基于 Jeffrey 散度的相似性度量的加权 FCM 聚类算法,改进后的加权 FCM 算法的聚类效果有更好的精确性和稳定性。

在模糊聚类中,每个聚类都被视为一个模糊集,所有的数据点都属于不同的模糊集,且有相应的隶属度。2004 年,Wang 等[11]使用了 FCM 的全局特征加权方案,为了对特征进行加权,采用了基于学习的方法和进化适应度函数,并且采用梯度下降算法寻找合适的权值。此外,该方法基于数据分布均匀的假设,而在大多数实际数据集中并非如此。2008 年,Hung 等[12]讨论了一种基于 FCM 的全局特征加权图像分割的算法,但是该算法计算的特征权重不适合某些极端情况,换言之,在一些数据集中,特征的权重不能恰当地表示特征的重要性[13]。2013 年,Nazari 等[14]提出自动加权和 FCM 集成的概念。2014 年,Ferreira 等[15]对局部和全局特征加权进行了普遍的研究,利用基本的 FCM 和核距离,讨论了一种利用自适应距离提高聚类质量的聚类方法。2016 年,Saha 等[16]设计了一种具有可分离几何距离的 FCM 算法,并证明了其对噪声特征扰动的鲁棒性。2018 年,林甲祥等[17]针对传统 FCM 算法无法获得令人满意的聚类结果的问题,提出了基于样本与特征双加权的自适应 FCM 聚类算法。根据 Zhou 等[18]在 2018 的文章可知,对这种算法进行实验分析,发现大部分实验中局部自适应距离法优于全局自适应距离法,尽管这种方法可以产生比传统 FCM 更好的结果,但这类算法可能不适用于对大型数据集进行聚类。2020 年,赵战民等[19]针对 FCM 算法对噪声敏感,且不能有效分割具有类大小不均衡特性的图像的问题,提出对类大小不敏感的模糊 C 均值聚类图像分割算法。

2 相关工作

2.1 FCM 聚类算法

1973 年,Dunn[20]引入了 FCM,随后 Bezdek[21]在 1981 年扩展了 FCM。FCM 通过最小化价值函数 $E_f(\boldsymbol{D},K)$找到群,即

$$E_f(K,\boldsymbol{D},\boldsymbol{O})=\sum_{y=1}^{m}\sum_{x=1}^{c}(\lambda_{xy})^f\mid v_y-k_x\mid^2,\ 1\leqslant f<\infty,\quad(1)$$

式中:f 为实数,代表模糊系数。此外,成员等级对指标的影响可以通过 f 来控制。假设增加 f,划分

就会变得更模糊。研究表明，FCM 在 1 和∞之间对 f 的任何值都收敛。K 为样本个数，$K=U_{i=1}^{c}k_i$。λ_{xy} 为样本 k_x 存储于 $\boldsymbol{D}(\boldsymbol{O})_{c\times m}$ 的隶属度，在清晰划分的情况下，$\lambda_{xy}=0$。而当 $\lambda_{xy}=1$ 时，$O_y\in k_x$ 为隶属矩阵元素，能量函数取决于 k_x 和 $\boldsymbol{D}$，$\boldsymbol{D}$ 为样本 k_x 与聚类中心 v_y 的欧氏距离，即

$$\sum_{x=1}^{c}\lambda_{xy}=1,\ y=1,2,\cdots,m,\lambda_{xy}\in[0,1],\ x=1,2,\cdots,c\ \cup\ y=1,2,\cdots,m, \tag{2}$$

$$0<\sum_{y=1}^{m}\lambda_{xy}<c,\ x=1,2,\cdots,c。 \tag{3}$$

FCM 算法采取模糊迭代优化的方法，用隶属度 λ_{xy} 和样本 k_x 方程进行更新，即

$$\lambda_{xy}^{i+1}=\left[\sum_{i=1}^{c}\left(\frac{v_y-k_x^i}{v_y-k_j^{(i)}}\right)\frac{2}{j^{-1}}\right]^{-1}, \tag{4}$$

$$k_x^{(i+1)}=v_y\cdot\sum_{y=1}^{m}[(\lambda_{xy}^{i+1})^f]^{-1}\sum_{y=1}^{m}(\lambda_{xy}^{i+1})^f。 \tag{5}$$

继续更新，直到 $\{|\lambda_{xy}^{i+1}-\lambda_{xy}^{i}|\}<\varepsilon$，其中，$\varepsilon$ 为终止条件，ε 在 0～1 之间，满足 $E_f(\boldsymbol{D},K)$ 局部最小值[22]。

FCM 算法确定聚类中心 v_y 和隶属度 λ_{xy} 的一般步骤如下。

1）用值在 [0,1] 之间的随机数来初始化隶属矩阵 $\boldsymbol{U}$，使其满足(2)式；

2）用(5)式计算聚类中心 v_y，$y=1,2,\cdots,n$；

3）根据(1)式计算价值函数，如果它小于某个已知的阈值，或者它相对于上个价值函数值的改变量小于某一个阈值，那么算法停止；

4）根据(4)式计算新的隶属矩阵 $\boldsymbol{U}$，返回步骤 2。

2.2　加权的 FCM 算法

特征加权是在局部进行的，其方法是根据各个属性对于聚类的贡献程度不同，需要对各个属性赋予相应的权重，从而提高聚类质量。2007 年，韦相等[23]对 FCM 算法进行特征加权，目的是减少迭代次数，提高速度，为了便于对权值进行编码，利用规格化方法进行归一化处理，即

$$\begin{cases}\boldsymbol{w}_{ij}=\dfrac{\boldsymbol{w}'_{ij}}{\boldsymbol{w}'_{\mathrm{sum}}}\\ \boldsymbol{w}'_{\mathrm{sum}}=\sum\limits_{i=1}^{p}\boldsymbol{w}'_{ij}\end{cases},\ 1\leqslant i\leqslant N_{\mathrm{unitynum}}, \tag{6}$$

式中：p 为样本的维数；N_{unitynum} 为遗传算法中每一代群体的个体数；$\boldsymbol{w}_{ij}$ 为 i 行 j 列矩阵；$\boldsymbol{w}'_{ij}$ 为对矩阵进行加权；$\boldsymbol{w}'_{\mathrm{sum}}$ 为对加权矩阵进行求和计算。

2.3　相似性度量方法及其性质

在聚类分析方法中，对于数据集中每个数据对象之间的关系的分析与判断，需要使用一个相似性度量方法进行测量，例如，欧氏距离、马氏距离、指数距离等经常被用到相似性度量方法中。而它们忽略了全局一致性的特征问题，因此在这一部分，给出所使用的相似性度量的定义及各种性质。

2.4　Jeffrey 散度

在传统的相似性度量方法中聚类结果质量会受到加权 FCM 算法的影响，因此使用根据 Kullback-Leibler 散度改进之后的 Jeffery 散度来改进算法，这方种法具有数值稳定性和对噪声鲁棒性，对于方差的不稳定性的影响较小[24]。它的定义为

$$J(\boldsymbol{Y}|\boldsymbol{X})=\int\left[\log_2\frac{f_X(x)}{m(x)}f_X(x)+\log_2\frac{f_Y(x)}{m(x)}f_Y(x)\right]\mathrm{d}x, \tag{7}$$

式中：$m(x)$ 为符合条件的概率密度函数，$m(x)=[f_X(x)+f_Y(x)]/2$；假设随机变量 $\boldsymbol{Z}$ 的概率密度函数为 $m(x)$，即 $f_Z(x)=m(x)$。

定义 1　估计 Jeffrey 散度 J_n，它是在一组 $n\times n$ 的正定矩阵上定义的，即

$$\partial(\boldsymbol{C},\boldsymbol{D})=(\boldsymbol{C}-\boldsymbol{D})\log_2\frac{\boldsymbol{C}}{\boldsymbol{D}}, \tag{8}$$

式中：$\boldsymbol{C}=|\boldsymbol{C}|$；$\boldsymbol{D}=|\boldsymbol{D}|$；$\partial(\boldsymbol{C},\boldsymbol{D})$ 为对矩阵 $\boldsymbol{C}$ 和 $\boldsymbol{D}$ 求偏导。

定义 2　任何两个数据点 $c,d\in\mathbf{R}_+^n$（正实数），之间的相似性可以表示为映射 $S:\mathbf{R}_+^n\times\mathbf{R}_+^n\rightarrow\mathbf{R}_+\cup\{0\}$，可以定义为

$$\boldsymbol{Z}(c,d)=\partial[\psi(c),\psi(d)], \tag{9}$$

式中：$\psi(\boldsymbol{O})=\mathrm{diag}(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_m)$，$\boldsymbol{O}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_m)\in\mathbf{R}_+^n$ 为实正向量。

所提出的相似性度量包括如下一些属性。

命题 1

$$\boldsymbol{Z}(c,d)=\boldsymbol{Z}(d,c), \tag{10}$$

式中：c 和 d 为两个数据点，为 $\boldsymbol{Z}(c,d)=\partial[\psi(c),\psi(d)]$ 的映射。

命题 2

$$\boldsymbol{Z}(c,d)\geqslant\boldsymbol{0}\ \&\ \boldsymbol{Z}(c,d)=\boldsymbol{0},\ \text{if } c=d。 \tag{11}$$

定理 1　相似性度量不是 Bregman 散度。

定理 2　$\boldsymbol{Z}(e\circ c,e\circ d)=e\boldsymbol{Z}(c,d)$ 表示 $e\in\mathbf{R}_+^n$，其中，$e\circ c$ 为 e 和 c 之间的 Hadamard 积。

定理 3　相似性度量是 f-散度。

3　算法改进

3.1　将 Jeffery 散度相似性度量与加权 FCM 相结合

使用 Jeffery 散度的相似性度量的加权 FCM 通

过以下公式求解来实现分组。

$$\min_{K=(k_1,k_2,\cdots,k_c)\in\mathbf{R}^{n\times c}\atop \boldsymbol{D}\in N} E_f(K,\boldsymbol{D};\boldsymbol{O})=\sum_{y=1}^{n}\sum_{x=1}^{c}(\lambda_{xy})^f\boldsymbol{Z}(o_y,k_x),1\leqslant s<\infty,\quad(12)$$

式中：o_y 为隶属矩阵的元素到样本 k_x 的映射，即

$$N=\left\{\boldsymbol{D}=[\lambda_{xy}]_{\substack{x=1,2m,\cdots,c\\ y=1,2,\cdots,n}}\middle|\lambda_{xy}\in[0,1],\sum_{x=1}^{c}\lambda_{xy}=1,\sum_{y=1}^{n}\lambda_{xy}>0\right\}。\quad(13)$$

(12)式精确解不存在，文献[25]中存在着一种交替优化方法如下所示。

定理 4　假设 $\tau_y=\{x\mid x\in[1,c],o_y=k_x^i\}$，下式是(4)式的替代式，也是证明算法的收敛性的必要条件，和(5)式是相同的。在交替优化算法中，证明 E_f 的收敛性。

$$\lambda_{xy}^{i+1}=\begin{cases}\sum_{i=1}^{c}\left[\dfrac{\boldsymbol{Z}(o_y,k_x^{(i)})}{\boldsymbol{Z}(o_y,k_j^{(i)})}\right], & \tau_y=0\\ |\tau_y|^{-1}, & \tau_y\neq0\ \text{and}\ x\in\tau_y\\ 0, & \tau_y\neq0\ \text{and}\ x\notin\tau_y\end{cases},\quad(14)$$

式中的 FCM 准则可用定理 5 中的优化无约束 FCM 准则来表述。

定理 5　简化 FCM 标准出现的公式与文献[25]相似，即

$$\min_{K\in\mathbf{R}^{n\times c}}E'_f(K;\boldsymbol{O})=\sum_{y=1}^{n}\left[\sum_{x=1}^{c}\boldsymbol{Z}(o_y,k_x)^{\frac{2}{1-f}}\right]^{1-f}。\quad(15)$$

收敛加权 FCM：假设 $f(\gamma_1,\gamma_2,\cdots,\gamma_c)=(\sum_{x=1}^{c}\gamma_x^s)^{\frac{1}{s}}$，$z=(1-f)^{-1}<0$，则(15)式表示为

$$E'_f(k_1,k_2,\cdots,k_c;\boldsymbol{O})=\sum_{y=1}^{n}f(\gamma_{k_1},\gamma_{k_2},\cdots,\gamma_{k_x})|_{\gamma_{xy}}=z(o_y,k_x)^2。\quad(16)$$

引理 1　由(15)式定义，E'_f 的主项为

$$\mathrm{maj}^lE'_f=E'_f(k_1^i,k_2^i,\cdots,k_c^i;\boldsymbol{O})+\sum_{y=1}^{c}\sum_{x=1}^{c}\frac{\mathrm{d}f}{\mathrm{d}\gamma_{xy}}|_{(i)}[\boldsymbol{Z}(o_y,k_x)^2-\boldsymbol{Z}(o_y,k_x^i)^2],$$

$$[E'_f(k_1,k_2,\cdots,k_c;\boldsymbol{O})]\leqslant\mathrm{maj}^lE'_f,\quad(17)$$

式中：导数取 $k_1^i,k_2^i,\cdots,k_c^i$。

定理 6　(交替优化的最速下降法)如果用优化原理调整步长(17)式，则序列 k_x^{i+1} 以等式形式出现在交替优化算法中，(14)式和(5)式以及应用在(15)式的最速下降算法的序列是相同的。

推论 1(优化加权 FCM 的全局收敛性)约化加权 FCM 算法在(12)式中说明它收敛到一个鞍点。

推论 2(优化加权 FCM 的局部收敛性)如果 $K=(k_1^*,k_2^*,\cdots,k_c^*)$ 有 $\boldsymbol{Z}(k_1^*,k_2^*,\cdots,k_c^*)$ 是 E'_f 的局部最小，因此假设起点 $K^{(0)}=(k_1^{(0)},k_2^{(0)},\cdots,k_c^{(0)})$ 在邻域中选取，则加权 FCM 算法收敛到 $K^*=(k_1^*,k_2^*,\cdots,k_c^*)$。

推论 3(优化加权 FCM 的收敛速度)FCM 在非奇异局部极小值附近与正定 Hessian 矩阵线性重合。

3.2　改进算法步骤

基于 Jeffrey 散度相似性度量的加权 FCM 算法的步骤如下。

1)经过对数据集进行预处理获得聚类数目 k，将聚类数目 k 用作起始聚类中心的代表点，经过初步划分后形成簇；

2)依据预处理获得的数据集，对样本进行初始划分，得到权重向量；

3)将权重向量引入 FCM 算法，由(6)式对 FCM 算法进行特征加权，初始化隶属度矩阵 $\boldsymbol{u}_{ij}$，设置迭代精度参数 ε，迭代次数 $t=0$，模糊加权指数 $m=2$；

4)根据新的聚类中心代入 Jeffrey 散度相似性度量中计算新的距离，重新计算隶属度和聚类中心，进行迭代循环，当聚类中心的结果不再发生变化时，算法结束。

4　实验结果与分析

4.1　实验介绍

实验环境：Windows 7 操作系统，CPU i5-2410M 2.30 GHz，内存为 4 GB，编程环境为 MALAB R2016b。为了对改进算法进行评估和分析，使用 UCI 数据集和人工数据集进行实验，实验数据集如表 1 所示。其中 UCI 数据集包括 Wine、Thyroid，人工数据集包括 S1、Isquare2、D31、Spiral。对比算法包括：FCM、K-means、GFSFDP 和 JW-FCM 算法。其中：K-means 算法采用欧氏距离建立相似度矩阵，是一种只适用于凸数据的聚类算法。密度峰值聚类(DPC)算法是一种能够快速搜索及查询并且利用决策图来确定聚类中心的算法[26]。

表 1　实验数据集属性

Table 1 Experimental data set properties

Data set	Sample number	Attributes	Number of categories
Wine	178	13	3
Thyroid	215	5	3
D31	3100	2	31
S1	5000	2	15
Isquare2	1741	2	6
Spiral	312	2	3

4.2　评价指标

本文使用聚类准确率(ACC)[27]、调整兰德系数(ARI)[28]以及鲁棒性(Entropy)对改进算法的聚类效果进行比较与分析。

改进算法的准确率使用 ACC 来进行评价，即

$$A_{\mathrm{ACC}}=\frac{\sum_{i=0}^{n}\delta\left[\hat{c}_i,\mathrm{map}(c_i)\right]}{n},\tag{18}$$

式中：c_i 为所提改进算法的类标签；$\hat{c}_i$ 为数据的类标签；$\delta(x,y)$为函数；map(x)为映射函数对所获取的中心和真实的中心进行映射。

调整 ARI 为

$$A_{\mathrm{ARI}}=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)},\tag{19}$$

式中：a 为归属于隶属矩阵 $\boldsymbol{U}$ 的同类且属于 $\boldsymbol{V}$ 的同类的数据的对数目；b 为归属于 $\boldsymbol{U}$ 的同类但属于 $\boldsymbol{V}$ 的不同类的数据的对数目；c 为归属于 $\boldsymbol{U}$ 的不同类而归属于 $\boldsymbol{V}$ 的同类的数据的对数目；d 为归属于 $\boldsymbol{U}$ 的不同类且归属于 $\boldsymbol{V}$ 的不同类的数据的对数目。ARI 越近似 1 说明聚类效果越好，越临近 0 说明聚类效果越差。

Entropy 即熵值，在 1854 年由 Clausius[29]提出，表示一个系统的内在混乱程度。Entropy 的取值在[0,1]间，取值越小表示算法中的混乱程度越低，聚类效果越好。

4.3　实验与分析

在本节中进行了多组实验，并将本文所提基于 Jeffery 散度的相似性度量加权 FCM 聚类算法与相关算法(FCM、K-means 和 DPC 算法)进行了对比与分析。所用到的数据集主要包括：D31、Isquare2、S1、Spiral 4 个人工数据集以及 Wine 和 Thyroid 2 个 UCI 数据集。其中 D31、Isquare2、S1、Spiral 均为二维数据集，D31 的数据样本数量较多，Wine 和 Thyroid 均为多维数据集。

本组实验分别在 D31、Isquare2、S1 、Spiral 4 个人工数据集以及 Wine 和 Thyroid 2 个 UCI 数据集上进行，实验数据集属性如表 1 所示，实验收敛性结果如图 1 所示。从图 1 可以看出每个算法的收敛效率的情况，K-means 和 DPC 收敛性较弱，FCM 较

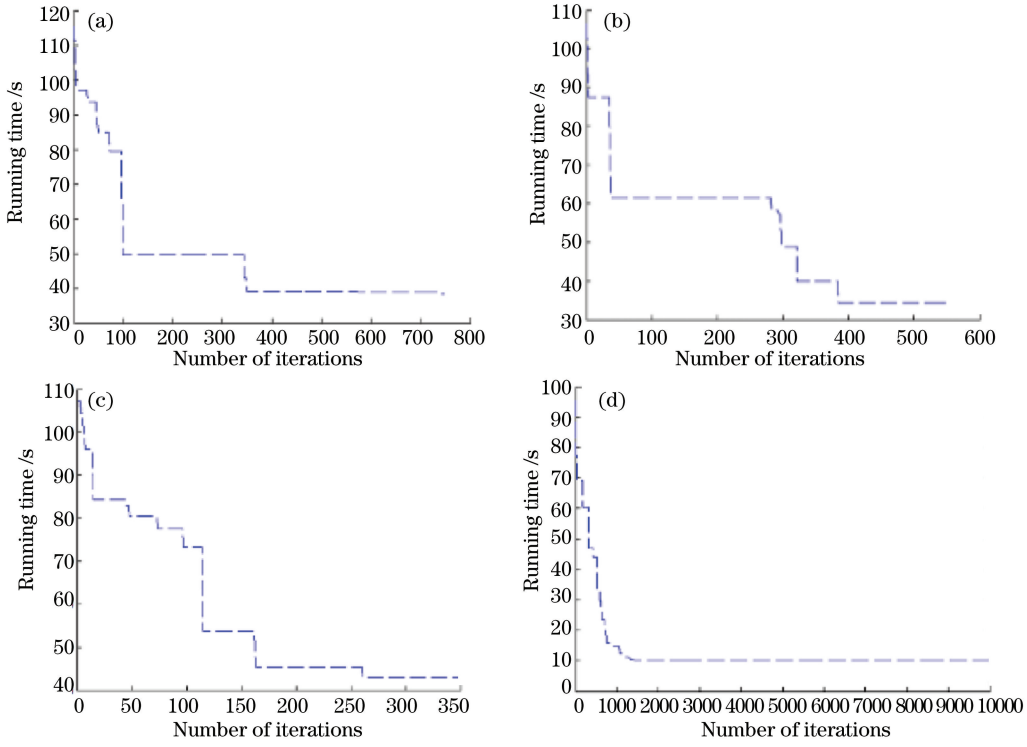

图 1　四种聚类算法的收敛性分析。(a) FCM；(b) K-means；(c) DPC；(d) JW-FCM

Fig. 1 Convergence analysis of four clustering algorithms. (a) FCM; (b) K-means; (c) DPC; (d) JW-FCM

好，JW-FCM 收敛性相对其他几个算法收敛性最好。聚类效果图如图 2～4 所示。从图 2 可以看出，FCM、K-means 和 DPC 算法在 Spiral 数据集上的聚类效果没有正确地划分类别，而 JW-FCM 算法能将数据正确地划分类别。在图 3 中 FCM、K-means 和 DPC 算法在 S1 数据集上的聚类效果依然没有正确地聚类，而 JW-FCM 算法同样能将此数据正确地划分类别。在图 4 中，对于 ISquare2 数据集，JW-FCM 算法聚类效果是比较好的，FCM 算法仅次之，而 K-means 、DPC 算法的聚类效果是比较差的。

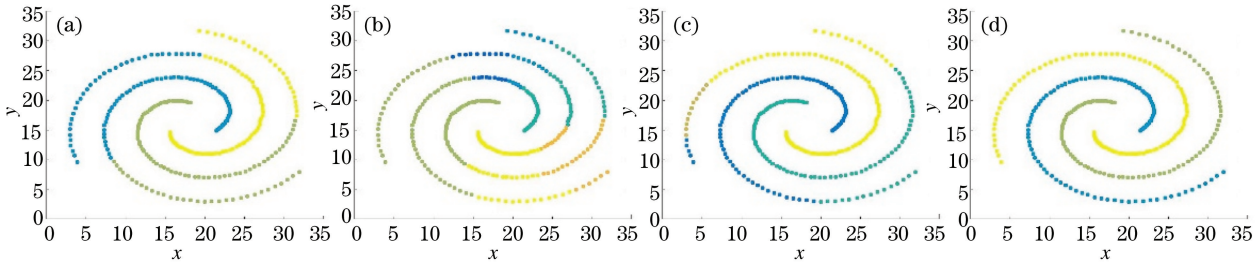

图 2　四种聚类算法对数据集 Spiral 聚类结果。(a) FCM；(b) K-means；(c) DPC；(d) JW-FCM

Fig. 2 Clustering results of four clustering algorithms on Spiral data set. (a) FCM; (b) K-means; (c) DPC; (d) JW-FCM

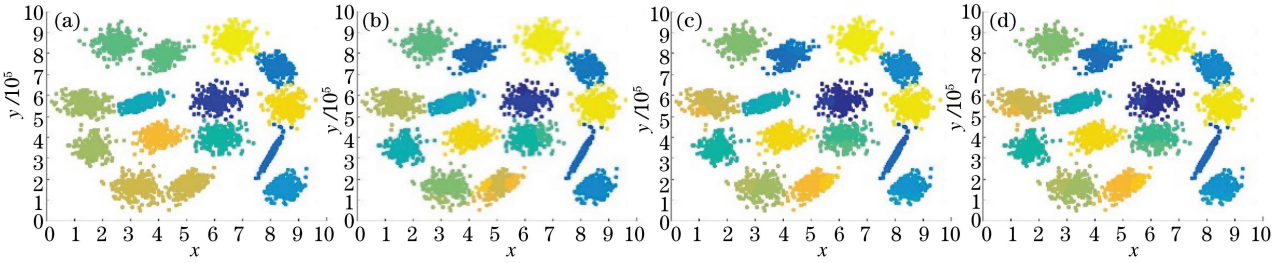

图 3　四种聚类算法对数据集 S1 聚类的结果。(a) FCM；(b) K-means；(c) DPC；(d) JW-FCM

Fig. 3 Clustering results of four clustering algorithms on S1 data set. (a) FCM; (b) K-means; (c) DPC; (d) JW-FCM

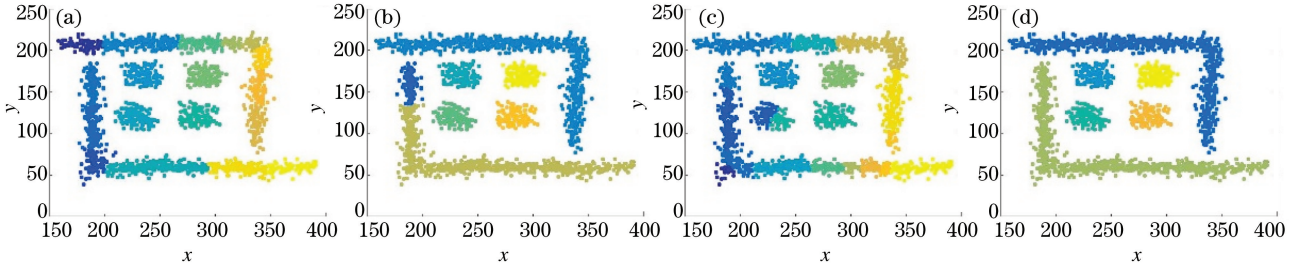

图 4　四种聚类算法对数据集 ISquare2 聚类的结果。(a) FCM；(b) K-means；(c) DPC；(d) JW-FCM

Fig. 4 Clustering results of four clustering algorithms on ISquare2 data set. (a) FCM; (b) K-means; (c) DPC; (d) JW-FCM

通过对图 2～4 三组实验的可视化对比实验分析可知：JW-FCM 算法比 K-means、FCM 和 DPC 算法更擅长对复杂形状的数据进行聚类。以上四种聚类算法在人工数据集上的性能对比如表 2 所示，从表 2 可以看出，JW-FCM 算法在 Spiral、S1、ISquare2 数据集上的聚类指标值都是 1，在 wine 数据集的聚类指标值 JW-FCM 算法均大于其他算法，聚类效果最好。根据表中聚类指标值来看，JW-FCM 算法聚类性能最优，FCM 次之，而 K-means 和 DPC 相对来说效果一般。加入不同比例的噪声数据后，FCM 和 K-means 的 Entropy 变化较大，DPC 和 JW-FCM 算法的 Entropy 指标变化较为接近。综上，JW-FCM 算法在保持较高的准确性的同时，还具有较好的鲁棒性。由此可见，用 Jeffrey 散度相似性度量代替欧氏距离很大程度上提高了加权 FCM 算法的聚类性能，与此同时，提高了聚类效果的稳定性和准确性，此外为了更直观地展示对比效果，将各个数据进行求均值，如表 2 最后一行 Mean 所示，ACC 对比图如图 5 所示，ARI 对比图如图 6 所示。

表 2　四种聚类算法在数据集上的性能对比

Table 2 Performance comparison of four clustering algorithms on data set

Data set	ACC				ARI				Entropy			
	FCM	*K*-means	DPC	JW-FCM	FCM	*K*-means	DPC	JW-FCM	FCM	*K*-means	DPC	JW-FCM
Wine	0.408	0.456	0.441	0.898	0.007	0.006	0.012	0.854	0.150	0.250	0.160	0.100
Thyroid	0.615	0.653	0.524	0.750	0.168	0.196	0.075	0.698	0.600	0.780	0.470	0.420
D31	0.875	0.358	0.846	0.993	0.654	0.324	0.756	0.882	0.320	0.320	0.230	0.180
S1	0.657	0.685	0.876	1.000	0.598	0.698	0.897	1.000	0.760	0.890	0.760	0.630
Isquare2	0.976	0.764	0.985	1.000	0.975	0.708	0.968	1.000	0.080	0.270	0.130	0.060
Spiral	0.356	0.405	0.529	1.000	0.003	0.011	0.276	1.000	0.960	0.960	0.660	0.460
Mean	0.648	0.554	0.700	0.940	0.401	0.324	0.497	0.906	0.478	0.578	0.402	0.308

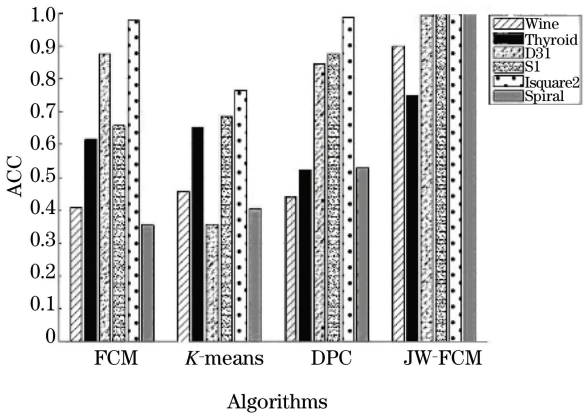

图 5　ACC 指标对比图

Fig. 5 Comparison of ACC indicators

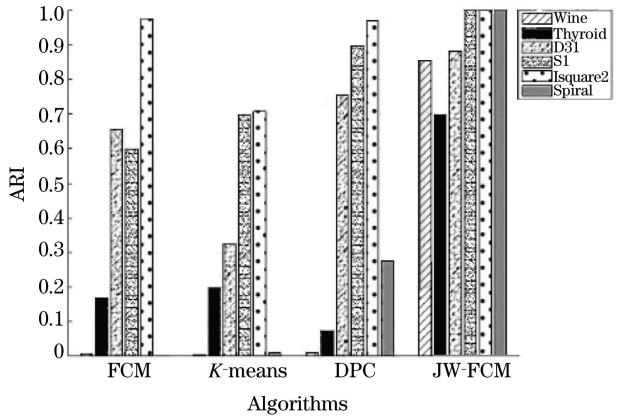

图 6　ARI 指标对比图

Fig. 6 Comparison of ARI indicators

比较了收敛性、准确率、聚类效果之后，在实验中，将噪声数据按照 0.1、0.2、0.3、0.4、0.5 的比例加入 6 个数据集中分别聚类，实验结果如图 7～12 所示，可以看出，在 D31、Isquare2、S1、Spiral 、Wine 和 Thyroid 六个数据集下，JW-FCM 的 Entropy 值最低，混乱程度越低，聚类效果最好，证明了本文算法具有较好的噪声鲁棒性。

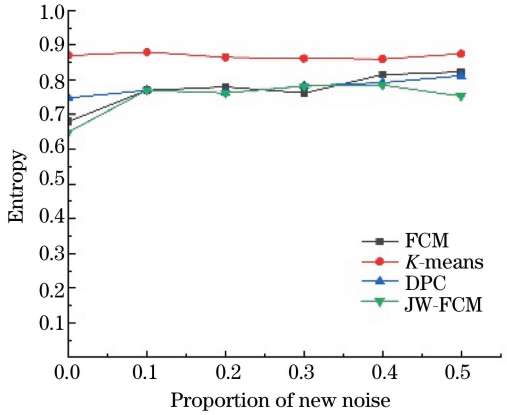

图 7　在数据集 Wine 上的鲁棒性对比

Fig. 7 Robustness comparison on Wine data set

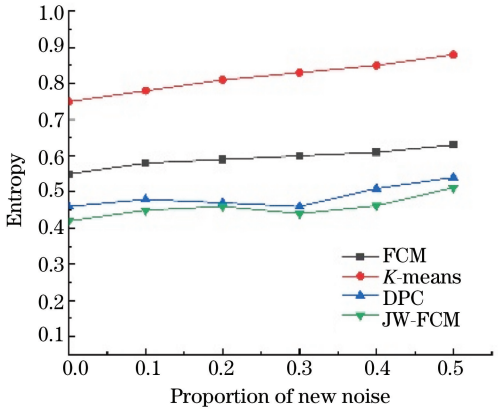

图 8　在数据集 Thyroid 上的鲁棒性对比

Fig. 8 Robustness comparison on Thyroid data set

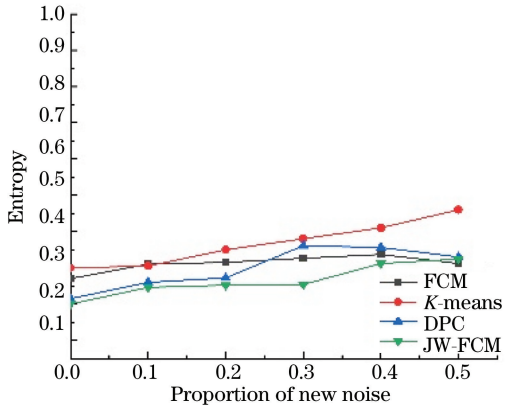

图 9　在数据集 D31 上的鲁棒性对比

Fig. 9 Robustness comparison on D31 data set

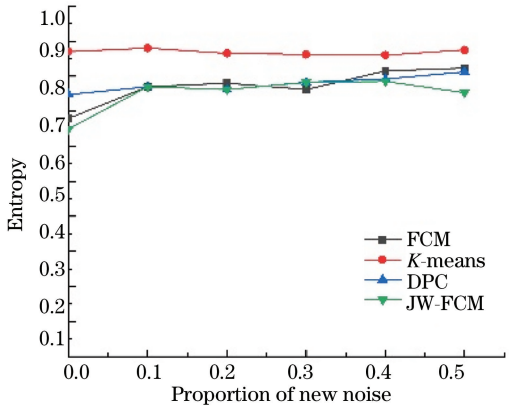

图 10　在数据集 S1 上的鲁棒性对比

Fig. 10 Robustness comparison on S1 data set

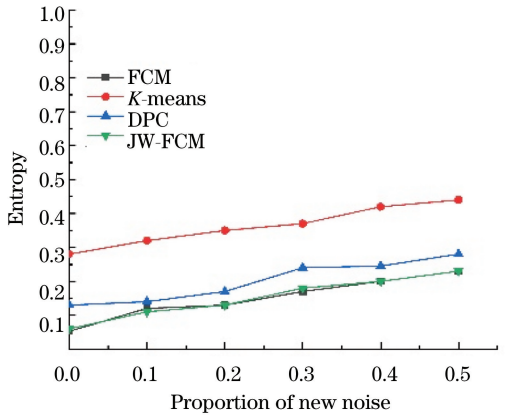

图 11　在数据集 Isquare2 上的鲁棒性对比

Fig. 11 Robustness comparison on Isquare2 data set

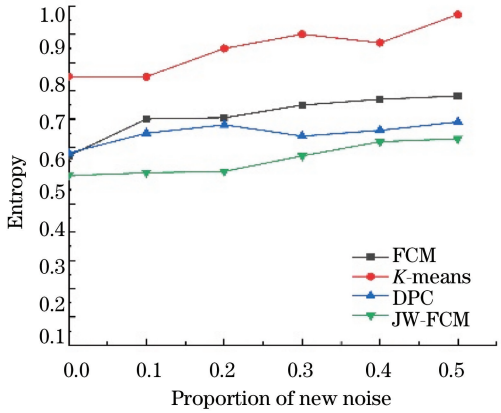

图 12　在数据集 Spiral 上的鲁棒性对比

Fig. 12 Robustness comparison on Spiral data set

5　结　　论

传统的 FCM 算法在使用欧氏距离相似性度量中，只考虑数据点之间的局部一致性问题，所以聚类效果差，对此本文将 FCM 算法进行特征加权，引入遗传算法加快算法的收敛速度，再结合 Jeffrey 散度对算法进行改进，在加权 FCM 算法中使用 Jeffrey 相似性度量代替了欧氏距离，所提 JW-FCM 算法保证了局部极小。通过将 FCM、*K*-means、DPC 算法和本文 JW-FCM 算法在 UCI 数据集及人工数据集中进行实验比较及分析，验证了改进算法具有较好的聚类效果，较原始算法有所提升，并且具备高效性、鲁棒性及准确性。但是该算法在处理大规模数据时，其时间复杂度相对较高，同时复杂性度量的研究也是算法下一步的主要研究目标。

参　考　文　献

[1] Zhang J, Fu J P, Li X H. Low-rank regularized heterogeneous tensor decomposition algorithm for subspace clustering [J]. Laser & Optoelectronics Progress, 2018, 55(7): 071003.
张静, 付建鹏, 李新慧. 基于低秩正则化异构张量分解的子空间聚类算法[J]. 激光与光电子学进展, 2018, 55(7): 071003.

[2] Pei X B, Wu T, Chen C B. Automated graph regularized projective nonnegative matrix factorization for document clustering [J]. IEEE Transactions on Cybernetics, 2014, 44(10): 1821-1831.

[3] Tong X J, Wu Y C. An improved spectral ensemble clustering algorithm in data mining [J]. Journal of Terahertz Science and Electronic Information Technology, 2020, 18(3): 497-503.
童绪军, 吴义春. 数据挖掘中一种改进的谱组合聚类算法[J]. 太赫兹科学与电子信息学报, 2020, 18(3): 497-503.

[4] Wang Q, Wang C, Feng Z Y, et al. Review of *K*-means clustering algorithm [J]. Electronic Design Engineering, 2012, 20(7): 21-24.
王千, 王成, 冯振元, 等. *K*-means 聚类算法研究综述[J]. 电子设计工程, 2012, 20(7): 21-24.

[5] Hathaway R J, Bezdek J C, Hu Y K. Generalized fuzzy c-means clustering strategies using Lp norm

distances[J]. IEEE Transactions on Fuzzy Systems, 2000, 8(5): 576-582.

[6] Zhou S G, Zhou A Y, Cao J. A data-partitioning-based DBSCAN algorithm[J]. Journal of Computer Research and Development, 2000, 37(10): 1153-1159.
周水庚, 周傲英, 曹晶. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展, 2000, 37(10): 1153-1159.

[7] Chakraborty S, Das S. *K*-means clustering with a new divergence-based distance metric: convergence and performance analysis[J]. Pattern Recognition Letters, 2017, 100: 67-73.

[8] Banerjee A, Merugu S, Dhillon I, et al. Clustering with bregman divergences[C]// Proceedings of the 2004 SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2004.

[9] Zhu Z L, Wang J F. Image segmentation based on adaptive fuzzy C-means and post processing correction[J]. Laser & Optoelectronics Progress, 2018, 55(1): 011004.
朱占龙, 王军芬. 基于自适应模糊 C 均值与后处理的图像分割算法[J]. 激光与光电子学进展, 2018, 55(1): 011004.

[10] Puzicha J, Hofmann T, Buhmann J M. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 17-19, 1997, San Juan, Puerto Rico, USA. New York: IEEE Press, 1997: 267-272.

[11] Wang X Z, Wang Y D, Wang L J. Improving fuzzy c-means clustering based on feature-weight learning[J]. Pattern Recognition Letters, 2004, 25(10): 1123-1132.

[12] Hung W L, Yang M S, Chen D H. Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation[J]. Pattern Recognition Letters, 2008, 29(9): 1317-1325.

[13] Xing H J, Wang X Z, Ha M H. A comparative experimental study of feature-weight learning approaches[C]//2011 IEEE International Conference on Systems, Man, and Cybernetics, October 9-12, 2011, Anchorage, AK, USA. New York: IEEE Press, 2011: 3500-3505.

[14] Nazari M, Shanbehzadeh J, Sarrafzadeh A. Fuzzy c-means based on automated variable feature weighting[C]// Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2013, March 13-15, 2013, Hong Kong.

[15] Ferreira M R P, de Carvalho F D A T. Kernel fuzzy c-means with automatic variable weighting[J]. Fuzzy Sets and Systems, 2014, 237: 1-46.

[16] Saha A, Das S. Geometric divergence based fuzzy clustering with strong resilience to noise features[J]. Pattern Recognition Letters, 2016, 79: 60-67.

[17] Lin J X, Wu L P, Wu J W, et al. Adaptive FCM clustering algorithm based on sample and feature weights[J]. Journal of Natural Science of Heilongjiang University, 2018, 35(2): 244-252.
林甲祥, 吴丽萍, 巫建伟, 等. 基于样本与特征双加权的自适应 FCM 聚类算法[J]. 黑龙江大学自然科学学报, 2018, 35(2): 244-252.

[18] Zhou Z P, Zhu S W. Kernel-based multiobjective clustering algorithm with automatic attribute weighting[J]. Soft Computing, 2018, 22(11): 3685-3709.

[19] Zhao Z M, Zhu Z L, Liu Y J, et al. Fuzzy C-means clustering algorithm for image segmentation insensitive to cluster size[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021001.
赵战民, 朱占龙, 刘永军, 等. 对类大小不敏感的图像分割模糊 C 均值聚类方法[J]. 激光与光电子学进展, 2020, 57(2): 021001.

[20] 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.

[21] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. Boston: Springer, 1983, 442:1-13.

[22] Seal A, Karlekar A, Krejcar O, et al. Fuzzy c-means clustering using Jeffreys-divergence based similarity measure[J]. Applied Soft Computing, 2020, 88: 106016.

[23] Wei X, Tang X H. Advanced dimension power FCM arithmetic with genetics[J]. Journal of Honghe University, 2007, 5(2): 39-42.
韦相, 汤兴华. 一种改进了的基于遗传算法的维特征加权改进 FCM 算法[J]. 红河学院学报, 2007, 5(2): 39-42.

[24] Zheng J, You H J. Change detection with SAR images based on radon transform and Jeffrey divergence[J]. Journal of Radars, 2012, 1(2): 182-189.
郑瑾, 尤红建. 基于 Radon 变换和 Jeffrey 散度的 SAR 图像变化检测方法[J]. 雷达学报, 2012, 1(2): 182-189.

[25] Groll L, Jakel J. A new convergence proof of fuzzy c-

means[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(5): 717-720.

[26] Chen Y W, Shen L L, Zhong C M, et al. Survey on density peak clustering algorithm[J]. Journal of Computer Research and Development, 2020, 57(2): 378-394.
陈叶旺, 申莲莲, 钟才明, 等. 密度峰值聚类算法综述[J]. 计算机研究与发展, 2020, 57(2): 378-394.

[27] Shang F H, Jiao L C, Shi J R, et al. Fast affinity propagation clustering: a multilevel approach[J]. Pattern Recognition, 2012, 45(1): 474-486.

[28] Vinh N X, Epps J, Bailey J. Bibliometrics: information theoretic measures for clusterings comparison[C]// The International Conference on Machine Learning, July 11-14, 2010, Qingdao, China. New York: ACM, 2010: 2837-2854.

[29] Clausius R. Ueber eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie[J]. Annalen Der Physik Und Chemie, 1854, 169(12): 481-506.