基于数据加权策略的模糊聚类改进算法

唐成龙 王石刚 徐 威
(上海交通大学机电设计与知识工程研究所 上海 200240)

摘 要：该文提出了一种数据指数加权的模糊均值聚类策略，引入了指数权因子和影响指数，使得可以在聚类过程中差异化处理各个数据。新策略和现有的 Gustafson-Kessel(G-K)算法相结合，提出了一种新的模糊聚类算法 DWG-K 用于提高聚类质量和挖掘离群点。数据试验表明 DWG-K 在提高聚类质量方面优于现有的 G-K；在离群点挖掘方面，DWG-K 对离群点的判定是全局的，离群点的物理意义清楚，且计算效率明显高于当前广泛采用的基于密度的离群点挖掘算法。

关键词：模糊聚类；数据加权策略；数据加权 G-K；离群点挖掘
中图分类号：TP18 **文献标识码**：A **文章编号**：1009-5896(2010)06-1277-07
DOI: 10.3724/SP.J.1146.2009.00857

Improved Fuzzy Clustering Algorithm Based on Data Weighted Approach

Tang Cheng-long Wang Shi-gang Xu Wei
(*Institute of Mechanical-Electrical Design and Knowledge Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract: A new data exponent weighted fuzzy clustering approach is proposed by introducing a set of exponent weighting factors and influence exponent, the new approach makes it possible to treat the data points discriminatively. The new approach is combined with the existing Gustafson-Kessel (G-K) algorithm and a new algorithm, DWG-K is presented. Numerical experiments show that the DWG-K is better than G-K in improving the quality of clustering, and in the outliers mining, DWG-K detects the outliers with the global view and the physical meaning of outliers is clearer, and moreover, the computational efficiency is significantly higher than the current widely used density-based method.

Key words: Fuzzy clustering; Data weighted approach; Data weighted G-K; Outliers mining

1 引言

人工智能研究涉及到多个子领域[1]。本文讨论了其中两个子领域：模式识别和离群点数据挖掘。在人工智能的应用领域，存在着这样的一种共性需求：对于一个给定的数据集，已知其含有一定数量的离群点，对该数据集的分析同时涉及两个子任务：(1)按照某种给定的标准，识别出给定数量的高质量的模式划分；(2)判定出哪些点是离群点，然后解释离群点的物理含义等。在当前，上述两个任务多数情况下是独立的，前者是模式识别领域中要解决的问题，后者则是离群点挖掘和知识发现领域中要解决的问题。将这两个子问题综合起来，采用一种方法来同时解决，目前国内外的研究还少见报道。本文提出了一种基于数据指数加权的模糊聚类策略。该策略应用于数据集的模糊划分，一方面可以得到高质量的聚类原型，另一方面，也提供了一种可以轻松地判定出数据集中离群点的存在并挖掘其丰富信息的实用手段。

2009-06-05 收到，2010-01-07 改回
国家自然科学基金(50875169)资助课题
通信作者：王石刚 wangshigang@sjtu.eud.cn

2 相关研究工作

2.1 模糊聚类算法

聚类是一种无监督的学习方法，目的是将一个目标数据集划分为若干个类。划分的依据是数据集中数据之间的相似性或相异性。模糊聚类是指数据和原型之间的关系用模糊隶属度来表示。模糊聚类避免了刚性聚类算法中某个数据只能唯一地属于某一类的缺点。

基于均值的模糊聚类算法应用广泛，其目标函数 J_f及约束条件为

$$\left.\begin{aligned} J_f &= \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2} \\ \text{s.t.} \quad & \sum_{i=1}^{c} u_{ij} = 1,\ \sum_{j=1}^{n} u_{ij} > 0 \end{aligned}\right\} \tag{1}$$

式(1)中 u_{ij}为模糊隶属度，d_{ij}为距离，c 表示类的数目，n 为数据集中数据的个数，m 为模糊指数。

基于交互迭代方式求解目标函数最优解涉及到

模糊隶属度 u_{ij} 以及类心原型 $\boldsymbol{V}_i$ 的更新式为

$$u_{ij}=d_{ij}^{-\frac{2}{m-1}}\Bigg/\left(\sum_{i=1}^{c}d_{ij}^{-\frac{2}{m-1}}\right) \tag{2}$$

$$v_i=\sum_{j=1}^{N}{u_{ij}}^m x_j\Bigg/\sum_{j=1}^{N}{u_{ij}}^m \tag{3}$$

上述算法称为概率型模糊均值聚类算法，简称 FCM。在 FCM 基础上，适应于不同的需要，有多种拓展算法。

基于均值的模糊聚类方法是模式识别的一种重要方法，但也存在着明显的不足。近年来对模糊聚类算法的研究，有两个热点，(1)寻求求解目标函数最优值的方法，例如，变尺度混沌求解方法[2]，基于粒子群的优化求解方法[3]等；(2)提出新的目标函数。例如，在图像处理领域有 SFCM 算法[4]，FGFCM 算法[5]。上述研究结果对模糊聚类算法的完善和发展做出了一定的贡献。

现有模糊聚类算法存在缺点的一个重要原因是，目标函数求最优解的过程中，没有体现出数据集中各个数据的差异性，各个数据点“一视同仁”，即便数据集中明显存在着离群点。在给定初始类心矩阵和类数目之后，聚类结果主要取决于模糊指数 m。而对模数指数 m 的合理取值范围，文献[6]研究认为 m 合理取值在 1.5 和 2。5 之间，且通常取值为 2。当数据集中存在离群点时，模糊聚类的质量差就不难理解了，因为离群点对聚类的质量也作了较大的“贡献”。

2.2 离群点挖掘算法

离群点挖掘是数据挖掘的基本任务之一，当前受到了广泛重视。很多情况下，罕见事件比正常出现事件更令人感兴趣，例如复杂工业生产过程中参数的异常波动等。在离群点挖掘研究领域，离群点自身成为关注焦点。离群点挖掘任务通常描述成“Top-k”原则[7]，即已知数据集中含有 n 个数据点，以及预期的离群点数量 k，发现数据集中显著异常或者不一致的头 k 个数据。

文献[7]将当前广泛采用的离群点挖掘方法总结为 4 类，分别为统计分析、距离、密度、主要特征偏差的方法。现有的离群点挖掘方法主要采用两种策略，(1)将离群点看作是二元性质，即对任一个数据点，判定其是或不是离群点，如统计学、距离的方法；(2)定量分析数据点的离群程度，由此来判定离群点的存在，如基于密度的方法；上述方法中，以基于密度[8,9]和距离[10,11]方法的研究及应用最广。

现有的离群点挖掘算法中，存在三方面的不足：(1)对离群点有用信息的挖掘存在不足，如对离群点仅作二元判定，不能充分挖掘离群点所蕴含的丰富信息。(2)离群点的物理意义难以解释，例如 LOF 算法[8]将离群点定义为数据点在空间的稀疏程度，仅体现了数据的局部信息；(3)计算的方法和效率有待提高，例如目前应用最广的基于密度的方法，计算效率低。

3 基于数据加权的模糊聚类新模型

3.1 模糊聚类算法中的数据加权策略

本节首先给出基于数据指数加权策略的目标函数，然后推导出相关参数更新等式。

3.1.1 目标函数 基于数据加权策略的模糊均值聚类算法的目标函数定义为

$$J_{\mathrm{dwf}}(X,U,C)=\sum_{j=1}^{n}e^{t_j}\sum_{i=1}^{c}u_{ij}^m d_{ij}^2 \tag{4}$$

式(4)中，J_{dwf} 为基于数据加权策略的模糊聚类目标函数，和式(1)相比，J_{dwf} 中引入了一组新参数 e^{t_j}，称为指数权因子，其中 t 称为权因子。显然，t 不论取何值，指数权因子均大于零。式(4)中的其它符号的含义同式(1)，模糊隶属度的约束条件也同式(1)。

分析式(4)，目标函数 J_{dwf} 中引入了 n 个新变量，即 n 个指数权因子，此时对目标函数求最优解需要补充约束条件。本文给出的补充约束条件为所有的指数权因子乘积约定为 1，即

$$\prod_{j=1}^{n}e^{t_j}=1 \tag{5}$$

在对式(4)求最优解时，也采用交互迭代方法。特别地，新策略中引入的指数权因子，也是迭代更新的，指数权因子的更新迭代式须事先给出。

式(5)中约束条件是 n 个指数权因子的乘积为 1，在实际使用中不方便。式(6)给出了和式(5)完全等价的约束条件，即所有权因子的和为 0。

$$\prod_{j=1}^{n}e^{t_j}=1\Leftrightarrow\sum_{j=1}^{n}t_j=0 \tag{6}$$

对目标函数求最优解的数学方法中，拉格朗日方法是一种常用方法。综合式(1)，式(4)和式(6)，有约束条件的目标函数最优解求解式可以写为

$$J_{\mathrm{dwf},\phi_1,\phi_2}=\sum_{j=1}^{n}e^{t_j}\sum_{i=1}^{c}u_{ij}^m d_{ij}^2+\phi_1\left(\sum_{j=1}^{n}t_j-0\right)+\phi_2\left(\sum_{i=1}^{c}t_j-1\right) \tag{7}$$

其中 ϕ_1 和 ϕ_2 为拉格朗日乘法算子，分别对应于指数权因子和模糊隶属度的约束条件。

式(7)中分别对 t_j 和 u_{ij} 求偏导数，有

$$\left.\begin{aligned}\frac{\partial J_{\mathrm{dwf},\phi_1,\phi_2}}{\partial t_j}&=e^{t_j}\sum_{i=1}^{c}u_{ij}^m d_{ij}^2+\phi_1\\ \frac{\partial J_{\mathrm{dwf},\phi_1,\phi_2}}{\partial u_{ij}}&=me^{t_j}u_{ij}^{m-1}d_{ij}^2+\phi_2\end{aligned}\right\}\tag{8}$$

令上述两个偏导数均取值为 0，得到两个乘法算子的表达式为

$$\phi_1=-e^{t_j}\sum_{i=1}^{c}u_{ij}^m d_{ij}^2\tag{9}$$

$$\phi_2=-me^{t_j}u_{ij}^{m-1}d_{ij}^2\tag{10}$$

下文分别推导指数权因子，模糊隶属度和类心矩阵的更新表达式。

3.1.2 指数权因子更新迭代等式　对式(9)进行变换，求出指数权因子的表达式为

$$e^{t_j}=\frac{-\phi_1}{\sum_{i=1}^{c}u_{ij}^m d_{ij}^2}\tag{11}$$

将 n 个指数权因子相乘，有下式

$$\prod_{j=1}^{n}e^{t_j}=\prod_{j=1}^{n}\left[\frac{-\phi_1}{\sum_{i=1}^{c}u_{ij}^m d_{ij}^2}\right]\tag{12}$$

考虑到指数权因子的乘积为 1 这一约束条件，乘法算子 ϕ_1 为

$$-\phi_1=\prod_{j=1}^{n}\left[\left(\sum_{i=1}^{c}u_{ij}^m d_{ij}^2\right)^{\frac{1}{N}}\right]\tag{13}$$

将式(13)代入到式(11)，求得数据因子的更新迭代式为

$$e^{t_j}=\frac{\prod_{j=1}^{n}\left[\left(\sum_{i=1}^{c}u_{ij}^m d_{ij}^2\right)^{1/n}\right]}{\sum_{i=1}^{c}u_{ij}^m d_{ij}^2}\tag{14}$$

3.1.3 模糊隶属度更新等式　对式(10)进行变换，求出模糊隶属度 u_{ij} 为

$$u_{ij}=\left(\frac{\phi_2}{-me^{t_j}}\right)^{\frac{1}{m-1}}d_{ij}^{-\frac{2}{m-1}}\tag{15}$$

对于数据 $\boldsymbol{x}_j$，考虑到模糊隶属度的约束条件，有式(16)。

$$\sum_{i=1}^{c}u_{ij}=\left(\frac{\phi_2}{-me^{t_j}}\right)^{\frac{1}{m-1}}\sum_{i=1}^{c}d_{ij}^{-\frac{2}{m-1}}\tag{16}$$

注意到式(16)的左边为 1，继续变换可得

$$\left(\frac{\phi_2}{-me^{t_j}}\right)^{1/(m-1)}=\frac{1}{\sum_{i=1}^{c}d_{ij}^{-2/(m-1)}}\tag{17}$$

由式(15)和式(17)，得到模糊隶属度 u_{ij} 的表达式为

$$u_{ij}=\frac{d_{ij}^{-\frac{2}{m-1}}}{\sum_{i=1}^{c}d_{ij}^{-\frac{2}{m-1}}}\tag{18}$$

比较式(18)和式(2)发现，新策略下模糊隶属度的更新等式和常规模糊聚类算法下模糊隶属度更新等式相同，这表明，在新数据加权策略下，并没有改变模糊隶属度的定义。

3.1.4 类心矩阵更新等式　在新策略中，类心矩阵的迭代式为

$$\overline{v}_i=\frac{\sum_{j=1}^{n}\overline{u}_{ij}^{\,m}x_j}{\sum_{j=1}^{N}\overline{u}_{ij}^{\,m}}\tag{19}$$

式中 $\overline{u}_{ij}$ 称为加权的模糊隶属度，其计算公式为

$$\overline{u}_{ij}^m=e^{t_j/s}u_{ij}^m\tag{20}$$

特别地，式(20)中引入了一个常数 s，称为影响指数。影响指数 s 对类心矩阵的影响是显性的，其对模糊隶属度、距离测度乃至目标函数的寻优过程也是有显著影响的，只不过这种影响是隐形的。s 的取值在第 5 节中详细讨论。

考虑到 $\overline{u}_{ij}$ 的表达式(20)，类心迭代的更新可改为

$$\overline{v}_i=\frac{\sum_{j=1}^{n}e^{t_j/s}u_{ij}^m x_j}{\sum_{j=1}^{N}e^{t_j/s}u_{ij}^m}\tag{21}$$

将式(21)和式(3)相比较发现，现有的模糊聚类下，类心迭代值是模糊隶属度、数据自身的综合结果。而在数据加权模糊聚类策略下，除了模糊隶属度和数据自身外，类心迭代等式还和模糊指数因子有关，即增加了一个手段，这说明获得更好质量的聚类结果是完全可能的。

3.2 DWG-K 算法

G-K 是 FCM 的一种拓展算法，其距离测度采用了马氏距离。计算马氏距离时，采用了模糊的协方差矩阵。马氏距离定义式和类内数据模糊协方差矩阵的定义式详细见文献[12]。G-K 较之采用欧氏距离测度的聚类算法，更适用于变量间存在相关性的数据集的聚类分析，从数据集中挖掘到更多的有效信息。

将数据加权策略和 G-K 算法相结合，提出了一种新的基于数据加权模糊聚类算法，简称为 DWG-K。

4 试验

本节通过数据试验评估 DWG-K 的性能。试验包括两部分，试验 1 为 DWG-K 下聚类质量试验，比较对象是 G-K。试验 2 为 DWG-K 对离群点的判定能力，比较对象是 LOF。

4.1 试验数据集

本文试验的第 1 个数据集为 IRIS 数据集。IRIS 集是一个真实数据集，其包括 3 个分组，其中第 1 组和后面两组是线性分离的，后面两组之间存在一定的重叠，因此适用 G-K 进行聚类操作。试验部分的第 2 个数据集来自作者的一个研究项目。在极薄带钢的平整轧制过程中，采用了智能控制策略。其中的一个任务是对平整后带钢的平坦度实时测量值的模式进行识别。在实际生产线中以 1 s 为周期采集了连续 3 卷带钢共 1154 组平坦度测量值，并将测量值转换为 4 阶勒让德多项式的 1–4 次系数，由此得到的数据集，称为 FL4C。FL4C 数据集中 4 个系数存在较强的相关性，G-K 适用于对该数据集进行聚类分析。

试验 1 采用了 IRIS 和 FLC4，由 DWG-K 和 G-K，比较了两种算法下的聚类质量。试验 2 采用了 FL4C 为试验对象，这是因为 IRIS 是一个不含离群点的数据集，而 FL4C 明显含有离群点。

4.2 试验 1 聚类质量分析

试验 1 的主要目的是评估聚类的质量和算法的计算效率。

4.2.1 聚类质量评价函数 本文采用紧致度、分离度和紧致度和分离度的比率 3 个指标来评价聚类质量。紧致度 Cmp 和分离度 Spt 的定义式分别为

$$\mathrm{Cmp}=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\left(x_j-\overline{v}_i\right)^2 \tag{22}$$

$$\mathrm{Spt}=\min_{i\neq k}\left\|\overline{v}_i-\overline{v}_k\right\|^2 \tag{23}$$

Cmp 是模糊隶属度、数据集自身、类心矩阵和模糊指数的函数。Cmp 体现了数据划分为 c 类之后数据的紧致性，即类内数据量化的相似程度。Spt 是类心矩阵的函数，体现了数据划分为 c 类之后数据的分离性，即类和类之间尽可能地不相似。本文将 Cmp 和 Spt 以及二者的比率 Cmp/Spt 作为对 DWG-K 和 G-K 在给定类数目 c 之后聚类质量的综合评价指标。

在 DWG-K 和 G-K 下，评价标准具体是，当 c 和 m 一致时，两种算法下，Cmp 小，Spt 大以及 Cmp/Spt 小的情况下，算法更优，当几个指标的判断出现不一致时，以 Cmp/Spt 作为最终评价指标。

4.2.2 聚类结果及分析

(1) 数据集聚类 IRIS 数据集在 DWG-K、G-K 下聚类结果的 3 个评价指标值见表 1。数据试验中 c 分别为 2，3，4，模糊指数 m 的取值均为 2，目标函数停止迭代的条件是 ε =0.0001。两种算法下给出的初始类心矩阵均由 FCM 算法生成。另外，DWG-K 所需要的 s 的取值见表 1。s 的取值方法见第 5 节分析。

表 1 IRIS 在 G-K 和 DWG-K 下的 3 种评价指标值

c	G-K			DWG-K			
	Cmp	Spt	Cmp /Spt	Cmp	Spt	Cmp /Spt	s
2	52.730	12.563	4.197	53.350	13.566	3.932	1.5
3	31.592	1.103	28.622	31.913	1.540	20.716	1.5
4	22.799	0.671	33.946	24.655	0.940	26.228	1.0

FL4C 在 DWG-K、G-K 下聚类结果的 3 个评价指标值见表 2。在该试验中，m 取值为 2，ε =0.001。c 的取值则考虑了实际应用，即在带钢的平坦度控制系统中，待识别的模式数目在 8–16 中优选。试验中，令 c 分别取值 8 至 16，在此条件下分别计算了 DWG-K 和 G-K 下聚类结果的评价指标值。

表 2 FL4C 在 G-K 和 DWG-K 下的 3 种评价指标值

c	G-K			DWG-K			
	Cmp	Spt	Cmp/Spt	Cmp	Spt	Cmp/Spt	s
8	757.796	0.778	973.003	757.414	1.159	653.169	−2.5
9	667.607	0.636	1048.604	669.205	1.100	607.950	−2.5
10	596.984	0.597	999.698	602.371	0.663	908.286	−2.5
11	539.884	0.533	1012.325	547.288	1.179	464.017	−2.5
12	494.274	0.674	732.650	498.429	1.491	334.260	−2.5
13	455.827	0.758	601.138	457.176	1.355	337.319	−2.5
14	421.344	0.854	493.178	424.327	1.204	352.191	−2.5
15	391.131	0.150	2602.361	396.416	0.694	570.429	−2.5
16	364.313	0.581	626.179	366.483	1.230	297.783	−2.5

(2)聚类质量分析　对于IRIS数据集，见表1，和 FL4C 数据集，见表 2，均有以下结论：不论 c 为何值，DWG-K 下的 Cmp 值，均要比 G-K 对应的 Cmp 值要略大，说明 G-K 下的紧致性要略好，DWG-K 下的 Spt 值，均要比 G-K 下对应的 Spt 要大，说明 DWG-K 下的分离性要好。DWG-K 下的 Cmp/Spt 值，均要比 G-K 下对应的的 Cmp/Spt 要小，说明从紧致性和分离性的综合效果来看，DWG-K 的质量更好。

4.2.3 计算效率　算法的计算效率通常采用迭代次数和计算过程所需要的时间两个指标来衡量。此处对两个数据集的试验，采用了以上两个计算效率指标。其中计算各进行了 5 次，并给出了 5 次计算的时间平均值，IRSI 数据集见表 3。FL4C 数据集见表 4。

表 3 IRIS 数据集在 G-K 及 DWG-K 下的计算效率

c	G-K		DWG-K		
	迭代次数	平均时间(s)	迭代次数	平均时间(s)	s
2	31	0.3244	35	0.1410	1.5
3	75	0.2497	59	0.1767	1.5
4	114	0.2704	36	0.2119	1.0

表 4 FL4C 数据集在 G-K 及 DWG-K 下的计算效率

c	G-K		DWG-K		
	迭代次数	平均时间(s)	迭代次数	平均时间(s)	s
8	122	1.5216	87	0.9670	−2.5
9	241	3.1260	78	0.8853	−2.5
10	53	0.7087	112	1.2111	−2.5
11	83	1.2562	120	1.3895	−2.5
12	100	1.7380	282	3.5912	−2.5
13	89	1.6499	388	5.1032	−2.5
14	80	1.6965	135	2.0657	−2.5
15	258	5.2468	26	0.4230	−2.5
16	233	4.8162	70	1.2042	−2.5

对于 IRIS 集，3 种给定类心数目下，DWG-K 算法下所需要的平均时间均小于G-K算法所需要的时间。对于 FL4C 数据集，c=8、9、15、16 时 DWG-K 所需时间要短，c=10、11、12 、13、14 时 DWG-K 所需时间要长。

此处对两种算法的计算效率进行的试验研究，重点并不在于对所需时间的严格对比，而是以试验的方式说明，新算法中引入的指数权因子并没有大大地延长计算所需要的时间。实际上，在 DWG-K 下，因为每一步均需要计算指数权因子，程序迭代一次所需要的时间比 G-K 要长，但又正是因为引入了指数权因子，且在合适的 s 的作用下，DWG-K 算法加快了收敛速度，使得迭代总次数减少，二者综合下，使得 DWG-K 具有和 G-K 差别不大的计算效率，这正是此处试验想要表达的思想。

4.3 试验 2　离群点挖掘

4.3.1 基于 DWG-K 的离群点挖掘方法　基于数据加权的模糊聚类策略一方面可以获得更好的聚类质量，另一个目的是将该策略用于挖掘数据集中的离群点。离群点的挖掘中一个基本问题是什么样的点被定义为离群点。基于数据加权策略的方法用于挖掘数据集中的离群点时，将离群点定义为数据点和聚类得到的类心的模糊距离关系。

注意到公式(15)，$\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}$ 表达了 $\boldsymbol{x}_j$ 到类心矩阵的一种模糊距离测度，这个值越大，说明 $\boldsymbol{x}_j$ 是离群点的可能性越大，将所有的数据的模糊距离测度按照从大到小的次序重新排列，最大的 k 个模糊聚类对应的数据被判定为离群点。变换式(23)得

$$e^{t_j}\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}=\prod_{j=1}^{n}\left[\left(\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}\right)^{1/n}\right] \tag{24}$$

等式右侧在目标函数收敛到最优值之后，此时是一个常数。也即对于数据集中任何一个数据 $\boldsymbol{x}_j$，e^{t_j} 和 $\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}$ 的乘积是相等的。显然，最大的 k 个模糊距离测度对应了 k 个最小的模糊指数因子。因此对离群点的判定可以依据模糊指数因子的取值来确定。定义数据 $\boldsymbol{x}_j$ 的离群度 $O_{\mathrm{dwf}}(\boldsymbol{x}_j)$ 如下：

$$O_{\mathrm{dwf}}\left(\boldsymbol{x}_j\right)=\frac{\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}}{\max\left[\sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}\right]}=\frac{\min\left(e^{t_j}\right)}{e^{t_j}} \tag{25}$$

$O_{\mathrm{dwf}}(\boldsymbol{x}_j)$ 是一个归一化值，取值范围为 $(0,\ 1]$。DWG-K 评定离群点如下，将所有数据点的离群度数值由大到小排序，前 k 个离群度所对应的数据点判定为离群点，即最大的 k 个离群度取值所对应的数据点为离群点。显然，$\boldsymbol{x}_j$ 的离群度表达的是数据 $\boldsymbol{x}_j$ 和全体类心距离的全局关系。

4.3.2 LOF 算法　离群点挖掘中，最著名和实用的方法是基于密度的方法，如 LOF 算法[7,8]。LOF 首先计算数据集中所有数据之间的距离，进而计算每个数据的可达密度，最后通过局部的偏离度来判断离群点的存在。

数据点的局部离群点因子表征了数据的离群程度。由此定义，如果数据 $\boldsymbol{x}_j$ 不是局部离群点，则

LOF($\boldsymbol{x}_j$)值接近于 1 。LOF($\boldsymbol{x}_j$)越大，则 $\boldsymbol{x}_j$是离群点的可能性越大。因此，由 LOF($\boldsymbol{x}_j$)的大小来判定数据集中离群点的存在，即前 k 个最大的 LOF 值为离群点。

从 LOF 的计算过程来看，有两个明显的缺点，(1)是计算量大，计算每一个数据点的相关参数值时，都需要对整个数据库扫描一次，下面的数据试验验证了这一点。(2)局部离群度反映了数据集的微观特性，且是从数学上解释的，离群点的物理意义不明确。

4.3.3 离群点挖掘结果对比 以 FL4C 数据集为研究对象，采用 DWG-K 对该数据集进行聚类分析，算法中 c=12，m=2，s=−2.5。并假定数据集中含有 10%的离群点，即 1154×10%=115 个离群点。按此条件对该数据集进行聚类分析，并将得到的各个数据的离群度绘制于图 1(a)中。图 1(a)中包含的信息如下：首先每个数据点，不论是不是离群点，均归属于某一个“类”。其次，离群点可以清楚地判定出来，图 1(a)中点划线为离群点判别线，在该线之上有 115 个点被判定为离群点。也即点划线的纵坐标为离群度从大到小的排序中第 116 个值。

图 1(b)为 LOF 算法下对 FL4C 的局部离群点进行的判定。点划线的解释同前，点划线之上的 115 个数据判定为离群点。对比图 1(a)和 1(b)，可以看出，对数据集中离群点的判定，两种方法下判定的结果具有类似的趋势，当然也存在着一定的区别，区别主要是两种算法下对离群度不同定义造成的。例如，LOF 下，离群度被定义为数据在高维空间中的稀疏程度。越稀疏，表明在数据的维数球域内数据点很少，据此判定离群点的存在，这是一种微观局部的判定方法。而在 DWG-K 下，离群度表达为某个数据点和类心坐标的模糊平方距离之和的大小，此时对离群点的判定是一种全局判定，对离群点信息的挖掘也更全面，即不但知道哪些点是离群点，而且知道是隶属于哪个“类”的离群点，这是 LOF 所不具有的优点。

由于对离群点的定义不同，两种算法其实并不能简单地评价其优劣，要根据具体的应用环境来选择。但是，对于这种情况：对数据集进行聚类，同时需要判定离群点存在的情况下，显然 DWG-K 的优势是 LOF 所不具有的。

4.3.4 计算效率 为了比较 DWG-K 下和 LOF 的计算效率，做了数据试验。试验用的计算机 CPU 主频 2.4 GHz，内存为 1 GB，试验的参数同 4.3.3 节说明。试验重复了 5 次，DWG-K 平均用时 3.2160 s，LOF 用时 12.2977 s，表明 DWG-K 对离群点的挖掘计算所需要的时间仅相当于 LOF 的四分之一左右，具有明显优势。

5 指数权因子和影响指数的讨论

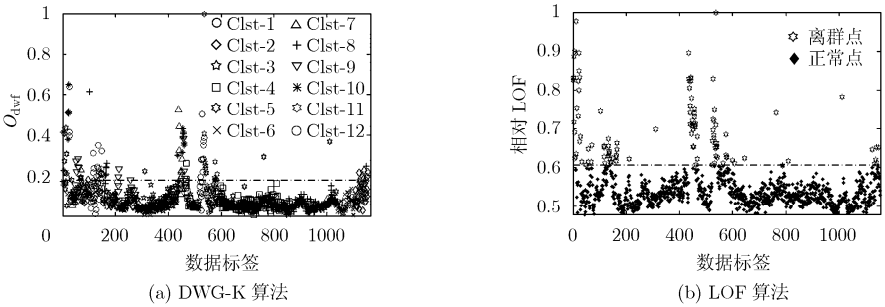
5.1 指数权因子

数据加权策略中引入指数权因子，算法的精妙之处在于提出了所有指数权因子的乘积为 1 这样的一个约束条件，由此实现了“差异化”对待数据的这一设想。另外，特别指出的是，数据加权算法下，迭代更新的是指数权因子这一整体。并不用关注权因子 t。 在算法的实现中，t 的作用是产生一组符合约束条件的指数权因子。从这个角度看，数据加权策略并非一定要采用指数权函数的形式。数据权函数也可以是其它表达式，只要这些权函数的乘积约定为 1。

5.2 影响指数

在数据加权模糊聚类算法中，影响指数 s 是一个十分重要的参数，其作用和角色类似于模糊聚类算法中模糊指数 m。影响指数 s 的作用和取值方法目前还没有严格的数学验证和证明。在实际应用中，往往通过“试算法”，即通过选择一系列有代表性的影响指数 s 来试算，从中找到一个最优的 s。

本文中对 s 的研究主要依据数据试验来进行。类似于对 m 的约定，本文提出 s 的取值范围为 $|s|\geqslant 1$，在这个范围内选择较优的 s 值。下文以 IRIS 为研究对象，以 Cmp/Spt 为评价指标，在不同的 c



(a) DWG-K 算法　　(b) LOF 算法

图 1 DWG-K 和 LOF 算法下 FL4C 数据集中离群点挖掘

的作用下，计算 Cmp/Spt 值，小 Cmp/Spt 的值所对应的 s 为优。表 5 为 IRIS 数据集在不同 c 和 s 条件下，Cmp/Spt 的计算值。

由表 5 可知，对于 c=2 和 3，最佳的 s 为 1.5，而对于 c= 4 ，最佳的 s 为 1。这解释了在第 4 节的数据试验中，s 这样取值的原因。

表 5 IRIS 集不同 c、s 的作用下 Cmp/Spt 的值

s	c	Cmp/Spt	c	Cmp/Spt	c	Cmp/Spt
1	2	5.0052	3	133.9787	**4**	**26.2288**
1.25	2	3.9732	3	44.6703	4	64.5486
1.5	**2**	**3.9326**	**3**	**20.7165**	4	36.5406
2	2	3.9514	3	21.6992	4	35.9867
2.5	2	3.9842	3	22.2743	4	35.7922
3	2	4.0126	3	22.7031	4	35.6009
4	2	4.0539	3	23.5547	4	35.2347
5	2	4.081	3	24.2227	4	34.9717

6 结束语

本文提出了一种新的基于数据加权的模糊均值聚类方法，新方法的创新性主要体现在“数据加权”和“权的模糊化”，优点在于可以获得更好的聚类质量、可以轻松地挖掘出数据集中离群点的信息。特别是对一个数据集的分析同时涉及这两方面需求时，新算法显示出其独有的优势。

本文所提出的聚类模型和方法是对现有模糊聚类算法的重要发展，可以和现有的多种模糊聚类算法相结合，开发出更高效的有针对性的算法。本文的后续工作包括研究更广义的权函数，提高算法的鲁棒性等。特别地，对于影响指数 s 的选取，本文是通过数据试验的方式来“试选”，文章没有给出严格的数学上的分析和证明，这是今后工作要解决的问题之一。

参 考 文 献

[1] 蔡自兴, 徐光佑著. 人工智能及其应用. 第三版, 北京: 清华大学出版社, 2004: 10–23.
Cai Zi-xing and Xu Guang-you. Artifcial Intelligence: Principles and Applications Third Edition, Beijing: Tsinghua Press, 2004: 10–23.

[2] Li Chao-shun, Zhou Jian-zhong, and Li Qing-qing. A fuzzy clustering algorithm based on mutative scale chaos optimization. Advances in Neural Networks. ISNN 2008, Berlin/Heidelberg: Springer. 2008, 5264: 259–267.

[3] Runkler T A and Katz C. Fuzzy clustering by particle swarm optimization. Proceedings of 2006 IEEE International Conference on Fuzzy Systems. Vancouver, BC, 2006: 601–608.

[4] Chuang Keh-shih, Tzeng Hong-long, and Chen Sharon. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*. 2006, 30(1): 9–15.

[5] Cai Wei-ling, Chen Song-can, and Zhang Dao-qiang. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition*, 2007, 40(3): 825–838.

[6] Pal N R and Bezdek J C. On cluster validity for the Fuzzy c-means Model. *IEEE Transactions on Fuzzy Systems*. 1995, 3(3): 370–378.

[7] Kamber M and Han Jia-wei. Data Mining: Concepts and Techniques. 2rd edition. Singapore: Elsevier Press. 2005: 295–300.

[8] Breunig M M, Kriegel Hans-peter, and Raymond T N, *et al*.. LOF: Identifying density-based local outliers. Proceedings of ACM SIGMOD International Conference on Management of Data, Dallas, Texas: ACM Press. 2000, 29: 93–104.

[9] Cao Hui, Si Gang-quan, Zhu Wen-zhi, and Zhang Yan-bin. Enhancing effectiveness of density-based outlier mining. International Symposiums on Information processing, Moscow, May 23–25, 2008.

[10] Ghoting A, Parthasarathy S, and Otey M E. Fast mining of distance-based outliers in high-dimensional dataset. *Data Mining Knowledge Discovery,* 2008, 16(3): 349–364.

[11] Weng Xiao-qing and Shen Jun-yi. Detecting outlier samples in multivariate time series dataset. *Knowledge-Based Systems,* 2008, 21(8): 807–812.

[12] Gustafson E E and Kessel W C. Fuzzy clustering with a fuzzy covariance matrix. Proceedings of IEEE Conference on Decision Control. San Diego, Californian, Piscataway, NJ. 1979: 761–766.

唐成龙：　男，1971 年生，博士生，高级工程师，研究方向为轧钢理论、人工智能、数据挖掘.

王石刚：　男，1957 年生，博士，教授，博士生导师，研究方向为复杂机电系统设计、机器视觉、人工智能.

徐　威：　男，1969 年生，博士，研究方向为机器人、微电子装备、人工智能.