CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 42 卷第 2 期 2020 年 2 月
Vol. 42, No. 2, Feb. 2020

文章编号:1007-130X(2020)02-0317-08

一种基于密度峰值的针对模糊混合数据的聚类算法*

陈奕延[1,2],李　晔[3],李存金[1]

(1.北京理工大学管理与经济学院,北京 100081;2.中国管理科学研究院学术委员会,北京 100036;
3.中国社会科学院大学(研究生院),北京 102488)

摘　要:将 CFSFDP 算法拓展到连续型模糊集和离散型模糊集上,提出了一种针对模糊混合数据的拓展型 CFSFDP 算法,将其命名为 FMD-CFSFDP 算法。FMD-CFSFDP 算法将样本涵盖的经典信息拓展到了模糊集上,利用寻找密度峰值的方法对模糊样本进行聚类,这是一种建立在模糊集上针对模糊混合数据的基于密度的聚类算法。首先简单介绍了 CFSFDP 算法及其改进,给出了"模糊混合数据"的数学概念;然后结合传统模糊欧氏距离的概念,分别提出了误差更小的针对连续型模糊集与离散型模糊集的改进型欧氏距离,在此基础上,依托权值构建了针对混合型模糊数据的整体距离。参考 CFSFDP 算法的聚类步骤给出了 FMD-CFSFDP 算法的聚类步骤。随后,在不同样本量、不同指标数量、不同簇数、不同取数规则的条件下,对算法进行了随机模拟实验并对聚类结果进行了分析。最后分别总结了 FMD-CFSFDP 算法的优缺点,并在此基础上提出了改进方案,为今后深入研究提供了参考。

关键词:模糊混合数据;基于密度峰值的聚类;FMD-CFSFDP 算法;改进型欧氏距离;整体距离

中图分类号:O159;TP301.6　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2020.02.017

A fuzzy mixed data clustering algorithm based on density peaks

CHEN Yi-yan[1,2], LI Ye[3], LI Cun-jin[1]

(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081;
2. Academic Committee, China Academy of Management Science, Beijing 100036;
3. Graduate School, University of Chinese Academy of Social Sciences, Beijing 102488, China)

Abstract: By extending CFSFDP algorithm to continuous fuzzy sets and discrete fuzzy sets, an extended CFSFDP algorithm for fuzzy mixed data is proposed, which is named FMD-CFSFDP algorithm. The FMD-CFSFDP algorithm extends the classical information in the sample to fuzzy sets, and achieves the clustering of fuzzy sets by seeking the density peaks. The proposed FMD-CFSFDP algorithm is a kind of density-based clustering algorithm established on fuzzy set for fuzzy mixed data. Firstly, the CFSFDP algorithm and some of its improvement algorithms are briefly introduced, and the mathematical definition of fuzzy mixed data is given. Secondly, by combining the concept of traditional fuzzy Euclidean distance, the improved Euclidean distance for both continuous and discrete fuzzy sets with smaller error is proposed. On the basis, the weight is introduced to establish the overall distance for fuzzy mixed data. By referring to the clustering steps of the CFSFDP algorithm, the clustering steps of FMD-CFSFDP algorithm are given. Furthermore, under the conditions of different sample size, different index number, different cluster number and different fetching rule, random simulation experiments are carried out on the algorithm and the clustering results are analyzed. Finally, the advantages and disadvantages of

* **收稿日期**:2019-08-05;**修回日期**:2019-10-21
通信作者:李晔(ly1992bigdata@163.com)
通信地址:102488 北京市中国社会科学院大学(研究生院)
Address: Graduate School, University of Chinese Academy of Social Sciences, Beijing 102488, P. R. China

the FMD-CFSFDP algorithm are summarized respectively. On this basis, some improved schemes are proposed, which provides a reference for future in-depth research.

Key words: fuzzy mixed data; density peaks based clustering; FMD-CFSFDP algorithm; improved Euclidean distance; overall distance

1 引言

聚类分析是按照某个特定标准把数据对象划分成子集的过程,每个子集表示一个簇。聚类分析是一种无监督学习过程,其目的是使得簇中的对象彼此相似,但与其它簇对象不相似[1,2]。目前,聚类分析广泛应用于商务智能、生物安全、Web 检索、评价与决策等领域。按照陈彩棠[3]的观点,聚类分析算法可以分为 6 类,包括基于划分[4-6]、层次[7-9]、密度[10]、网格[11,12]、概率模型[13]以及基于约束[14]的聚类算法。这种划分方式并不一定涵盖所有的聚类算法,譬如基于图论[15]的聚类算法,但不论何种算法皆有其各自的特点。

2 相关工作

Rodriguez 等[16]于 2014 年提出了快速搜索和发现密度峰值的聚类 CFSFDP(Clustering by Fast Search and Find of Density Peaks)算法,这是一种基于密度、可自动获得簇的正确个数,并能够解决数据空间分布呈非球形簇的聚类算法。许多学者在 CFSFDP 算法的基础上进行了改进:Wan 等[17]对 CFSFDP 算法中寻找簇中心的决策图方法提出了质疑,提出了一种 Fuzzy CFSFDP 算法,并利用基于流形距离和基于标准差的截断距离对其进行优化;Zhang 等[18]在无线传感网络中将 CFSFDP 算法与层次协议相结合,提出了一种考虑剩余能量的改进型 CFSFDP-E 算法;Qin 等[19]把太赫兹时域光谱与 CFSFDP 算法结合,提出了 PCA-CFSFDP 算法;李晔等[20]提出了针对混合型数据集的 MAO-CFSFDP 算法,并使用该算法解决了实际问题[21],从而验证了该算法的可靠性。

虽然 MAO-CFSFDP 算法拓展了数据类型,但它与 CFSFDP 算法及其它改进算法都是建立在经典集合上的聚类算法,而现实生活中的许多对象是不具备严格属性的,无法用“非此即彼”的二值逻辑解释,参考 Zadeh[22]提出的模糊集理论,这些对象是具有模糊概念的事物。目前常用的 PCM[23]、FCM[24]、PFCM[25]等算法依赖统计不确定性理论(概率分布、贝叶斯模型等),将聚类对象与簇之间的隶属关系不确定化,但仍定义在经典集合上,无法解决模糊数据的距离问题。

因此,本文在模糊集理论的基础上,提出了针对由连续型模糊集与离散型模糊集组成的模糊混合数据的聚类算法 FMD-CFSFDP(Fuzzy Mixed Data-Clustering algorithm by Fast Search and Find of Density Peaks)。该算法可满足含有模糊混合数据的样本的聚类需求,继承了 CFSFDP 算法的优点,并且具备 3 项创新:

(1)从理论上将 CFSFDP 算法从经典集扩展到了模糊集上,提出模糊混合数据的概念;

(2)利用连续型模糊集和离散型模糊集,构建了模糊集上针对模糊混合数据的聚类算法;

(3)改进了模糊集上的传统欧氏距离,分别定义了模糊集上针对连续型模糊集和离散型模糊集的改进型欧氏距离,使之相较前者误差减少,令聚类的度量更为合理。

3 模糊混合数据的数学定义

记连续型模糊集为连续型模糊数据,离散型模糊集为离散型模糊数据,假设存在数据集 Θ,若 Θ 中存在 N_1 个连续型模糊数据组成的数据子集 Θ_1,以及 N_2 个离散型模糊数据组成的数据子集 Θ_2,满足:$\Theta_1 \cap \Theta_2 = \varnothing$,$\Theta_1 \cup \Theta_2 = \Theta$,则称数据集 Θ 为模糊混合数据集,简称模糊混合数据。模糊混合数据是由数据形式为连续型模糊数据(连续型模糊集)与离散型模糊数据(离散型模糊集)混合组成的数据集。

4 FMD-CFSFDP 算法步骤

4.1 计算聚类的度量

对于 N 个指标下的 M 个模糊样本,N 个指标中有 N_1 个指标是定量指标,其内容是连续型模糊集,另外 N_2 个指标是定性指标,其内容是离散型模糊集。每个样本有 $N_1 + N_2 = N$ 个模糊集,整个模糊样本集 $\tilde{S} = \{\tilde{s}_i\}_{i=1}^{M}$ 共有 $M \cdot N$ 个模糊集,

这些模糊集构成了模糊混合数据集,简称模糊混合数据。可知,若要求 2 个样本之间的距离,首先需要求 2 个样本在同一指标下的距离。若指标为连续型模糊集,设指标 j 下模糊样本 $\tilde{s}_r$ 对应的模糊集 $C_r^{(j)}$ 的隶属函数为 $C_r^{(j)}(x)$,其中 x 为论域中的元素,另一样本 $\tilde{s}_t$ 在指标 j 下的模糊集 $C_t^{(j)}$ 的隶属函数为 $C_t^{(j)}(x)$,$C_t^{(j)}$ 亦是一个连续型模糊集,若 $X_r^{(j)}=[a_r^{(j)},b_r^{(j)}]\subset R$,$a_r^{(j)}$ 和 $b_r^{(j)}$ 分别是模糊子集 $X_r^{(j)}$ 的上限和下限。$X_t^{(j)}=[a_t^{(j)},b_t^{(j)}]\subset R$,$R$ 为论域,若一共有 N_1 个指标是定量的,即每个模糊样本中存在 N_1 个连续型模糊集,则 2 个模糊样本 $\tilde{s}_r$ 与 $\tilde{s}_t$ 的全部 N_1 个定量指标之间的距离 $L_C(r,t)$ 为:

$$L_C(r,t)=\sum_{j=1}^{N_1}\left[\int_{X_r^{(j)}\cup X_t^{(j)}}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x\right]^{\frac{1}{2}} \tag{1}$$

上述方法是将 N_1 个指标下的连续型模糊集对应的欧氏距离相加,可对此进行改进以便减少系统性误差。设 $X_r^{(j)}=[a_r^{(j)},b_r^{(j)}]\subset R$,$X_t^{(j)}=[a_t^{(j)},b_t^{(j)}]\subset R$,$X_r^{(j)}$ 与 $X_t^{(j)}$ 均为有界区间,另设有界区间 $X_{rt}^{(j)}$,满足 $X_{rt}^{(j)}=[a_{rt}^{(j)},b_{rt}^{(j)}]$。其中 $a_{rt}^{(j)}=a_r^{(j)}\wedge a_t^{(j)}$,$b_{rt}^{(j)}=b_r^{(j)}\vee b_t^{(j)}$,对于 N_1 个指标,可以确定一个区间 $X_{rt}=[a_{rt},b_{rt}]$,满足条件:$X_{rt}=[\bigwedge_{j=1}^{N_1}a_{rt}^{(j)},\bigvee_{j=1}^{N_1}b_{rt}^{(j)}]$,称区间 X_{rt} 为 2 个样本之间的最大公共积分域。另外,若存在 P 个指标($1\leqslant P\leqslant N_1$)的模糊集的积分区域 $X_P^{(j)}$ 为双侧无界区间($-\infty,+\infty$)或单侧无界区间,此时确定公共积分域的计算方法与 $X_P^{(j)}$ 为有界区间时一样,也是寻找最大公共积分域,可认为无界区间是有界区间的无限拓展。为涵盖所有情况,本文还是用 X_{rt} 表示 2 个模糊样本 $\tilde{s}_r$ 与 $\tilde{s}_t$ 的最大公共积分区域,则此时定义 N_1 个连续型模糊集上 2 个模糊样本 $\tilde{s}_r$ 与 $\tilde{s}_t$ 的改进型距离 $L_C(r,t)$ 为:

$$L_C(r,t)=\left[\int_{X_{rt}}\sum_{j=1}^{N_1}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x\right]^{\frac{1}{2}} \tag{2}$$

对于式(1),设其整体误差为 $e_{\sum}(r,t)$,系统误差为 $e_{\sum}^{(S)}(r,t)$,随机误差为 $e_{\sum}^{(R)}(r,t)$,有:

$$e_{\sum}(r,t)=e_{\sum}^{(S)}(r,t)+e_{\sum}^{(R)}(r,t) \tag{3}$$

对于式(2),设其整体误差为 $e_{IE}(r,t)$,系统误差为 $e_{IE}^{(S)}(r,t)$,随机误差为 $e_{IE}^{(R)}(r,t)$,有:

$$e_{IE}(r,t)=e_{IE}^{(S)}(r,t)+e_{IE}^{(R)}(r,t) \tag{4}$$

由于随机误差是不能人为控制的,假设 $e_{\sum}^{(R)}(r,t)=e_{IE}^{(R)}(r,t)$,$\int_{X_r^{(j)}\cup X_t^{(j)}}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x$ 的误差为 $e_j^{(S)}(r,t)$,因为 $\forall X_{rt}^{(j)}\subseteq X_{rt}$,则对于第 j 个指标,其公共积分域为 $X_{rt}^{(j)}$,显然,根据定积分的性质,对于单个模糊样本下的模糊集 $C_r^{(j)}$ 或 $C_t^{(j)}$ 而言,非积分域的部分:

$$C_r^{(j)}(x)=0,$$
$$\int_{X_t^{(j)}}|C_r^{(j)}(x)|^2\mathrm{d}x=0$$

或

$$C_t^{(j)}(x)=0,$$
$$\int_{X_r^{(j)}}|C_t^{(j)}(x)|^2\mathrm{d}x=0$$

即对于 $C_r^{(j)}$ 与 $C_t^{(j)}$,非公共积分域的积分:

$$\int_{(X_{rt}-X_{rt}^{(j)})}\sum_{j=1}^{N_1}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x\equiv 0$$

其中,$X_{rt}-X_{rt}^{(j)}$ 表示差集,即 X_{rt} 中除 $X_{rt}^{(j)}$ 以外的积分域,即对于单独指标而言,它的误差只存在于有定义的积分域部分,没有定义的非积分域部分的所有积分均为 0,则不存在误差,因此有:

$$e\left\{\int_{X_{rt}}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x\right\}=$$
$$e\left\{\int_{X_r^{(j)}\cup X_t^{(j)}}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x\right\}=e_j^{(S)}(r,t) \tag{5}$$

则式(1)的系统误差为:

$$e_{\sum}^{(S)}(r,t)=\sum_{j=1}^{N}[e_j^{(S)}(r,t)]^{\frac{1}{2}} \tag{6}$$

根据定积分的性质,$\int_{X_{r,t}}\sum_{j=1}^{N_1}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x=\sum_{j=1}^{N_1}\int_{X_{r,t}}|C_r^{(j)}(x)-C_t^{(j)}(x)|^2\mathrm{d}x$,则式(2)的系统误差为:

$$e_{IE}^{(S)}(r,t)=\left[\sum_{j=1}^{N}e_j^{(S)}(r,t)\right]^{\frac{1}{2}} \tag{7}$$

显然 $e_{\sum}^{(S)}(r,t)\geqslant e_{IE}^{(S)}(r,t)$,由于 $e_{\sum}^{(R)}(r,t)=e_{IE}^{(R)}(r,t)$,则有 $e_{\sum}(r,t)\geqslant e_{IE}(r,t)$,相比之下,使用式(2)定义的改进型欧氏距离的误差比使用式(1)的更小。

若指标为离散型模糊集,设指标 k 下模糊样本 r 的取值 x 的隶属度为 $D_r^{(k)}(x)$,另一样本 t 在指标 k 下的信息亦是一个离散型模糊集。设取值 x 的隶属度为 $D_t^{(k)}(x)$,若离散型模糊集 $D_r^{(k)}$ 与 $D_t^{(k)}$ 拥有相同的元素,其组成的集合记为 $\{x_g^{(k)}\}_{g=1}^{h(k)}$,

$h(k)$ 表示指标 k 下离散模糊集 $D_r^{(k)}$ 与 $D_t^{(k)}$ 的元素上限数。若一共有 N_2 个指标是定性的，即每个模糊样本中存在 N_2 个离散型模糊集，参照连续型模糊集的改进型欧氏距离，则其改进距离为：

$$L_{\mathrm{D}}(r,t)=\left(\sum_{k=1}^{N_2}\sum_{g=1}^{h(k)}\left|D_r^{(k)}(x_g^{(k)})-D_t^{(k)}(x_g^{(k)})\right|^2\right)^{\frac{1}{2}} \tag{8}$$

同样可以证明式(8)的误差要小于传统欧氏距离的，称式(8)为离散型模糊集的改进型欧氏距离。参考 Huang[26] 提出的 K-prototypes 算法中的相似性度量，设 2 个模糊样本 $\tilde{s}_r$ 与 $\tilde{s}_t$ 的整体距离为 $L(r,t)$，则有：

$$L(r,t)=L_{\mathrm{C}}(r,t)+L_{\mathrm{D}}(r,t) \tag{9}$$

4.2 其余聚类步骤

当得到任意 2 个模糊样本 $\tilde{s}_r$ 与 $\tilde{s}_t$ 的整体距离 $L(r,t)$ 后，将其作为聚类的度量，参考 MAO-CFSFDP 算法的步骤，其算法流程图如图 1 所示。

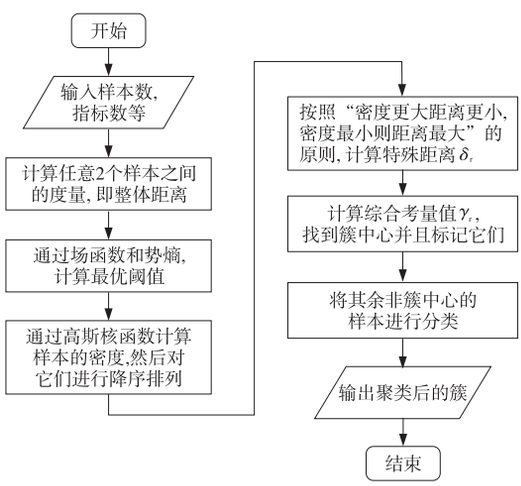


Figure 1 Flow chart of FMD-CFSFDP algorithm

图 1 FMD-CFSFDP 算法流程图

FMD-CFSFDP 算法是顺序结构，所以其最大时间复杂度是 $O(N\cdot M^2)$，而 CFSFDP 算法的复杂度是 $O(M^2)$[27]，显然，由于数据形式和度量都变得复杂，所以 FMD-CFSFDP 算法的复杂度要高于 CFSFDP 算法的。

5 随机模拟实验

本节进行 2 组随机模拟实验，每组进行 25 次，一共 50 次。第 1 组模拟实验一共包含 200 个模糊样本，即 $M=200$，共有 3 个指标，$N=3$，设论域 $X=\mathbf{R}$，3 个指标中有 2 个是 $\mathbf{R}$ 上的连续型模糊子集，$N_1=2$，记为指标 A 和指标 B，1 个是 $\mathbf{R}$ 上的离散型模糊子集(文中简称离散型模糊集)，$N_2=1$，记为指标 C。事先人为将这些模糊样本分成 2 个簇，$C=2$。前 100 个样本为一个簇，设 2 个连续型模糊子集均为正态模糊集，其中第 1 个指标 A 对应的隶属函数的参数 μ,σ 分别来自 2 个正态分布 $N(155,2.3^2)$，$N(1.6,0.26^2)$(这里的分布可以是任意的，仅代表 1 种取数规则)，第 2 个指标 B 对应的隶属函数的参数 μ,σ 分别来自 2 个正态分布 $N(41,2.3^2)$，$N(2.1,0.43^2)$，第 3 个指标 C 是离散型模糊集，任意指定 x_g 与 g 构成的集合 $\{x_g\}_{g=1}^5$，设 $\{x_g\}_{g=1}^5=\{x_g=g\}_{g=1}^5$，$x_g$ 对离散型模糊集 $\widetilde{C}$ 的隶属度由均匀分布 $U(0,1)$ 生成；后 100 个样本为另一个簇，其中第 1 个指标 A 对应的隶属函数的参数 μ,σ 分别来自 2 个正态分布 $N(163,3.1^2)$，$N(2.3,0.25^2)$，第 2 个指标 B 对应的隶属函数的参数 μ,σ 分别来自 2 个正态分布 $N(30,3.4^2)$，$N(3.1,0.52^2)$，第 3 个指标 C 对应的 5 个元素($\{x_g\}_{g=1}^5=\{1,2,3,4,5\}$)对离散型模糊集 $\widetilde{C}$ 的隶属度由均匀分布 $U(0,1)$ 生成。对聚类效果的检验与评价可引用 Al-Shammary 等[28]提出的聚类正确率。定义算法 f 在模糊样本集上的聚类正确率如式(10)所示：

$$ac\_rate(D/f)=\frac{\sum_{i=1}^{K}corr\_c_i}{|D|} \tag{10}$$

其中，K 表示该样本集真实的簇的个数，$corr\_c_i$ 表示第 i 个簇中被正确聚类的模糊样本个数，$|D|$ 为模糊样本个数。$corr\_c_i$ 越大，则说明聚类效果越好。计算第 1 组随机模拟实验的聚类正确率和最优阈值 L^*，如表 1 所示。平均聚类正确率 $MC=63.38\%$，聚类正确率的标准差 $SD=6.3628$。

Table 1 25 results of the 1st random simulation

表 1 第 1 组随机模拟 25 次的实验结果

实验数	最优阈值 L^*	聚类正确率/%	实验数	最优阈值 L^*	聚类正确率/%
1	0.643 2	60.5	14	0.660 3	74.5
2	0.673 9	64.0	15	0.639 1	58.0
3	0.643 7	61.5	16	0.646 9	62.0
4	0.646 7	53.0	17	0.648 2	57.5
5	0.656 7	60.0	18	0.612 9	61.0
6	0.626 5	72.0	19	0.672 3	54.5
7	0.646 9	63.5	20	0.641 2	57.5
8	0.647 4	71.0	21	0.626 5	72.0
9	0.606 3	57.0	22	0.646 9	63.5
10	0.668 9	60.5	23	0.647 4	71.0
11	0.659 7	69.5	24	0.606 3	57.0
12	0.646 3	71.5	25	0.668 9	60.5
13	0.646 3	71.5			

显然,第 1 组 25 次随机模拟实验的平均聚类正确率为 63.38%,聚类的标准差为 6.362 8,各组实验的离散型模糊集中的元素各自相同,区分度仅体现在隶属度上,而隶属度的区分度亦不高,所以在绘图中可暂时忽略离散型指标 C,使用第 1 个指标 A 下的正态模糊集的均值 $\mu^{(A)}$ 作为横轴坐标,第 2 个指标 B 下的正态模糊集的均值 $\mu^{(B)}$ 作为纵轴坐标绘制二维的聚类效果图,不同簇用不同颜色表示[29],绘制第 1 组随机模拟中的聚类正确率最高的第 17 次实验的黑白聚类效果图,如图 2 所示。

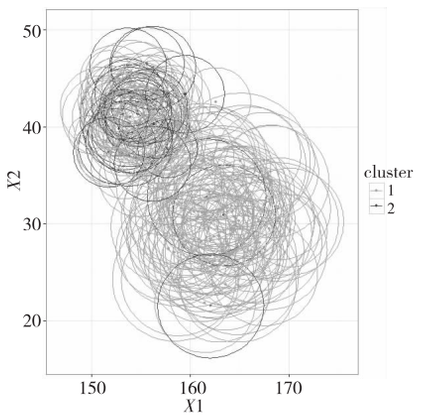


Figure 2 Clustering effect diagram of the 17th experiment in the 1st random simulation

图 2 第 1 组第 17 次实验的聚类效果图

第 2 组模拟一共包含 300 个模糊样本,$M=300$,共有 3 个指标,$N=3$,设论域 $X=\mathbf{R}$,3 个指标中有 2 个是 $\mathbf{R}$ 上的连续型模糊子集,连续型指标个数 $N_2=2$,与上组模拟一样,依然记 2 个对应连续型模糊集的指标为 A 和 B,1 个对应 $\mathbf{R}$ 上的离散型模糊子集(文中简称离散型模糊集)的指标为 C,离散型指标个数 $N_2=1$。这些模糊样本事先人为地分成 3 个簇,$C=3$。前 100 个样本为第 1 个簇,设 2 个连续型模糊子集均为正态模糊集,其中第 1 个指标 A 对应的隶属函数的参数 μ,σ 分别来自正态分布 $N(5.3,0.32^2)$ 与指数分布 $Exp(11)$,第 2 个指标 B 对应的隶属函数的参数 μ,σ 分别来自 2 个均匀分布 $U(5,7)$,$U(0.12,0.34)$,第 3 个指标 C 是离散型模糊集,任意指定 x_g 与 g 构成的集合 $\{x_g\}_{g=1}^4$,设 $\{x_g\}_{g=1}^4=\{x_g=g\}_{g=1}^4$,$x_g$ 对离散型模糊集 C 的隶属度由均匀分布 $U(0,1)$ 生成;中间 100 个样本为第 2 个簇,其中第 1 个指标 A 对应的隶属函数的参数 μ,σ 分别来自正态分布 $N(12,0.23^2)$ 与指数分布 $Exp(12)$,第 2 个指标 B 对应的隶属函数的参数 μ,σ 分别来自 2 个均匀分布 $U(14,20)$,$U(1.01,1.34)$,第 3 个指标 C 对应的 4 个元素($\{x_g\}_{g=1}^4=\{1,2,3,4\}$)对离散型模糊集的隶属度由均匀分布 $U(0,1)$ 生成;后 100 个样本为第 3 个簇,其中第 1 个指标 A 对应的隶属函数的参数 μ,σ 分别来自正态分布 $N(14.6,0.15^2)$ 与指数分布 $Exp(9)$,第 2 个指标 B 对应的隶属函数的参数 μ,σ 分别来自 2 个均匀分布 $U(9,12)$,$U(0.37,1.05)$,第 3 个指标 C 对应的 4 个元素($\{x_g\}_{g=1}^4=\{1,2,3,4\}$)对离散型模糊集的隶属度由均匀分布 $U(0,1)$ 生成。同理,利用式(10)计算第 2 组随机模拟实验的聚类正确率和最优阈值 L^*,如表 2 所示。平均聚类正确率 $MC=44.64$,聚类正确率的标准差 $SD=3.6092$。

Table 2 25 results of the 2nd random simulation

表 2 第 2 组随机模拟 25 次实验结果

实验数	最优阈值 L^*	聚类正确率/%	实验数	最优阈值 L^*	聚类正确率/%
1	0.486 7	42.33	14	0.505 4	46.00
2	0.498 4	40.67	15	0.483 1	46.00
3	0.483 5	40.00	16	0.483 6	52.00
4	0.481 3	39.67	17	0.500 5	41.00
5	0.503 2	47.00	18	0.499 4	44.33
6	0.496 9	52.00	19	0.506 5	45.67
7	0.497 7	52.00	20	0.460 7	44.00
8	0.499 8	42.33	21	0.481 6	44.00
9	0.482 5	45.33	22	0.484 5	43.67
10	0.502 0	46.33	23	0.496 0	49.33
11	0.499 8	43.33	24	0.476 2	41.33
12	0.426 7	43.33	25	0.501 2	42.00
13	0.502 0	42.33			

显然,第 2 组 25 次随机模拟的平均聚类正确率下降为 44.64%,聚类的标准差为 $SD=3.609\ 2$,说明相比第 1 组随机模拟,第 2 组随机模拟的实验结果较第 1 组有所提升,但聚类正确率有所下降,聚类正确率最高的是第 6、第 7 和第 16 次实验,均为 52.00%,与绘制图 2 采用的方法一样,暂时忽略离散型指标 C 的影响,使用第 1 个指标 A 下的正态模糊集的均值 $\mu^{(A)}$ 作为横轴坐标,第 2 个指标 B 下的正态模糊集的均值 $\mu^{(B)}$ 作为纵轴坐标绘制二维的聚类效果图,不同的簇使用不同颜色表示[29],绘制出第 6 次实验的黑白聚类效果图,如图 3 所示。

显然,从彩图[29]可以看出,代表 3 个簇的彩色团块中夹杂着不同的颜色,说明第 2 组的聚类的效果不如第 1 组第 17 次实验理想。另外,将表 1 与

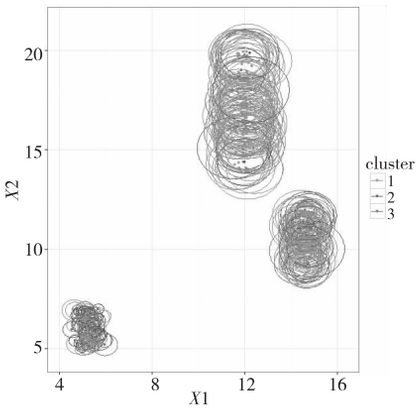
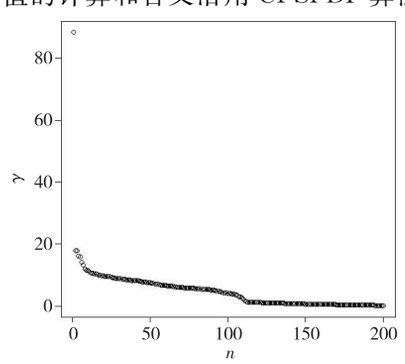


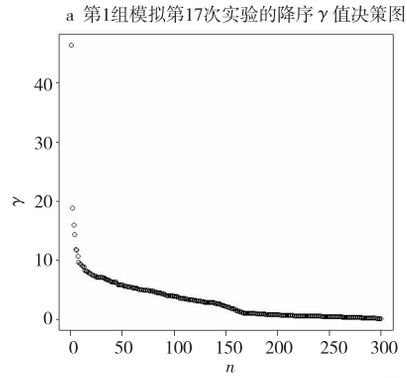
Figure 3 Clustering effect diagram of the 6th experiment of the 2nd random simulation

图 3 第 2 组第 6 次实验的聚类效果图

表 2 中的聚类正确率与参考文献[16,20]比较,显然可以发现 FMD-CFSFDP 算法的聚类正确率没有 CFSFDP 和 MAO-CFSFDP 的高。这是因为样本每一个指标下的模糊集中对应的每种状态(元素)都被当成数值参与运算,对于任意连续型模糊集而言则均有无数个元素参与运算,故聚类正确率会较前 2 者偏低。分别画出第 1 组随机模拟中第 17 次实验,以及第 2 组随机模拟中第 6 次实验下 γ 值降序排列后的决策图,如图 4 所示,由于非簇中心的 γ 会比较平滑,故可以利用跳跃点判断簇中心个数,γ 值的计算和含义沿用 CFSFDP 算法。



a 第1组模拟第17次实验的降序 γ 值决策图



b 第2组模拟第6次实验的降序 γ 值决策图

Figure 4 The descending γ decision diagram in two different experiments

图 4 2 次不同实验的降序 γ 值决策图

由 γ 的情况以及聚类结果可知,第 1 组模拟的第 17 次实验中,人工划分的簇数是 2 个,根据图 4a 中所示,其拥有 2 个跳跃点,故自动识别出的簇数也是 2 个;而第 2 组模拟的第 6 次实验中,人工划分的簇数是 3 个,但从图 4b 中可以看出,其自动识别出的簇数为 6。显然,FMD-CFSFDP 算法利用 γ 值的跳跃点来自动识别簇数是不稳定的。因为不论连续型模糊集还是离散型模糊集,计算其相应的改进型欧氏距离中使用的隶属度的取值是在[0,1],因此算出的改进型欧氏距离较小,导致整体距离 L、最优阈值 L^*、密度 ρ、特殊距离 δ 与综合考量值 γ 也较小,反映在图像中的区分度较低,因此单纯通过视觉识别 γ 值跳跃点就变得比较困难。

6 结束语

FMD-CFSFDP 算法可满足模糊混合数据的聚类需求,在模糊集上继承了 CFSFDP 算法的大多数优点,本文的主要创新之处在于 FMD-CFSFDP 算法把 CFSFDP 算法从经典集扩展到了模糊集上,同时也吸收了 MAO-CFSFDP 算法的优势,赋予“模糊混合数据”数学涵义,改进了作为度量的传统模糊欧氏距离,使改进后的改进型欧氏距离具有更小的误差,可以提高聚类精度。

然而,纵有上述创新,FMD-CFSFDP 算法仍存在以下 3 个缺点:

(1)FMD-CFSFDP 算法中的模糊样本涵盖的信息是模糊集,但模糊样本与簇之间的隶属关系依然使用了硬划分而未能考虑模糊的特性,虽然模糊数学上的许多计算,包括模糊贴近度、模糊度、模糊距离等都是把模糊量转化为经典量,最终计算结果也都是经典数值,这在模糊数学上是合理的。但是,从模糊集到经典集的转化过程中往往会损失一些信息,特别是对于模糊样本的划分,如果采用硬划分则会造成聚类正确率在一定程度上的下降。

(2)虽然使用了误差较传统欧氏距离更小的改进型欧氏距离,并利用权值对其进行了加权处理,从而得到整体距离,但由于其权值是固定的,无法自适应调整,这无疑会削弱整体距离的区分度,从而导致聚类正确率相比 CFSFDP 算法有所下降。

(3)FMD-CFSFDP 算法未能解决 CFSFDP 算法中利用视觉识别跳跃点寻找簇数的方法的缺陷,这在一定程度上是由于度量的计算使用了隶属度,从而导致最后综合考量值的区分度过低,无法利用

视觉有效地寻找跳跃点。

针对上述缺点,未来可对 FMD-CFSFDP 算法做如下拓展改进:

(1)将模糊样本与簇之间的隶属关系也定义在模糊集上,从而使样本信息和隶属关系均建立在模糊集上,这或许可以减少聚类划分造成的信息损失,提高聚类正确率;

(2)放弃使用隶属度进行计算的模糊距离及其相关一切改进,寻找新的可以体现模糊数据属性的计算公式作为度量;

(3)放弃利用视觉识别几何图像中 γ 值的特征寻找簇数的方式,可以研究一套量化的数学模型来自动寻找簇的个数,这样即便发生前述缺点(2)和(3)中的情况,微小的差异也可以被数学模型轻易识别出来,从而提高了聚类的区分度。

综上,FMD-CFSFDP 算法虽然存在不足之处,但它的提出可为进一步研究模糊集上的聚类算法提供参考。

参考文献:

[1] Han J W, Kamber M, Pei J. Data mining: Concept and technique(Third Edition) [M]. Fan Ming, Meng Xiao-feng, translation. Beijing: China Machine Press, 2012. (in Chinese)

[2] Bishop C. Neural networks for pattern recognition [M]. New York: Oxford University Press, 1995.

[3] Chen Cai-tang. Research on K-modes clustering algorithm of dissimilarity measure [D]. Taiyuan: Taiyuan University of Technology, 2012. (in Chinese)

[4] Macqueen J. On convergence of K-means and partitions with minimum average variance [J]. Annals of Mathematical Statistics, 1965, 36(3): 1084.

[5] Chaturvedi A, Green P E, Carroll J D. K-modes clustering [J]. Journal of Classification, 2001, 18(1): 35-55.

[6] Karbach W, Voss A, Schuckey R, et al. Model-K: Prototyping of the knowledge level [C]// Proc of the 1st International Conference on Knowledge Modeling & Expertise Transfer, 1991: 195-208.

[7] Ma Ru-ning, Wang Xiu-li, Ding Jun-di. Multilevel core-sets based aggregation clustering algorithm [J]. Journal of Software, 2013, 24(3): 490-506. (in Chinese)

[8] Wilcox H, Nichol R C, Zhao Gong-bo, et al. Simulation tests of galaxy cluster constraints on chameleon gravity [J]. Monthly Notices of the Royal Astronomical Society, 2016, 462(1): 715-725.

[9] Lorbeer B, Kosareva A, Beva B, et al. Variations on the clustering algorithm BIRCH [J]. Big Data Research, 2018, 11: 44-53.

[10] Bryant A, Cios K. RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1109-1121.

[11] Dat N D, Phu V N, Tran V T N, et al. Sting algorithm used English sentiment classification in a parallel environment [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2017, 31(7): 1750021.

[12] Sheikholeslami G, Chatterjee S, Zhang A D. WaveCluster: A wavelet-based clustering approach for spatial data in very large databases [J]. The VLDB Journal — the International Journal on Very Large Data Bases, 2000, 8(3-4): 289-304.

[13] Kohonen T. Self-organized formation of topologically correct feature maps [J]. Biological Cybernetics, 1982, 43(1): 59-69.

[14] Zhang X P, Wu J J, Si H F, et al. Spatial clustering with obstacles constraints by ant colony optimization and quantum particle swarm optimization [C]// Proc of 2009 International Conference on Artificial Intelligence and Computational Intelligence, 2009: 154-158.

[15] Bu Z, Gao G L, Li H J, et al. CAMAS: A cluster-aware multiagent system for attributed graph clustering [J]. Information Fusion, 2017, 37: 10-21.

[16] Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

[17] Wan M, Yin S Q, Tan T, et al. Optimized fuzzy clustering by fast search and find of density peaks [C]// Proc of the 2018 IEEE 3rd International Conference on Cloud Computing and Big data Analysis(ICCCBDA), 2018: 83-87.

[18] Zhang Y M, Liu M D, Liu Q W. An energy-balanced clustering protocol based on an improved CFSFDP algorithm for wireless sensor networks [J]. Sensors, 2018, 18(3): doi: 10.3390/s18030881.

[19] Qin B Y, Li Z, Luo Z H, et al. Terahertz time-domain spectroscopy combined with PCA-CFSFDP applied for pesticide detection [J]. Optical and Quantum Electronics, 2017, 49(7): 1-12.

[20] Li Ye, Chen Yi-yan, Zhang Shu-fen. Design of mixed data clustering algorithm based on density peak [J]. Journal of Computer Applications, 2018, 38(2): 483-490. (in Chinese)

[21] Chen Y Y, Li Y. Analysis of the causes of cancer death: A cluster study of 51 constituencies based on US presidential elections[J]. Basic & Clinical Pharmacology & Toxicology, 2018, 123(SI, Supplements): 75-76.

[22] Zadeh L A. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338-353.

[23] Barni M, Cappellini V, Mecocci A. A possibilistic approach to clustering-comments [J]. IEEE Transactions on Fuzzy Systems, 1996, 4(3): 393-396.

[24] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy C-means clustering-algorithm [J]. Computers & Geosciences, 1984, 10(2-3): 191-203.

[25] Pal N R, Pal K, Keller J M, et al. A possibilistic fuzzy c-means clustering algorithm [J]. Journal of Cybernetics,

1974,3(3):32-57.

[26] Huang Z. Clustering large data sets with mixed numeric and categorical values [C]// Proc of the 1st Pacific-Asia Knowledge Discovery and Data Mining Conference,1997:39-54.

[27] Zhang Wen-kai. Research on density-based hierarchical clustering algorithm [D]. Hefei: University of Science and Technology of China,2015. (in Chinese)

[28] Al-Shammary D,Khalil I,Tari Z,et al. Fractal self-similarity measurements based clustering technique for SOAP web messages [J]. Journal of Parallel and Distributed Computing,2013,73(5):664-676.

[29] Chen Yi-yan. Effect color picture of clusering [EB/OL]. [2019-12-26]. https://blog.csdn.net/Dr_Chenyiyan/article/details/103715064. (in Chinese)

附中文参考文献:

[1] Han J W, Kamber M, Per J. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2012.

[3] 陈彩棠. 相异度量的 K-modes 聚类算法研究[D]. 太原:太原理工大学,2012.

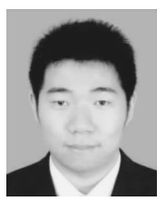
[7] 马儒宁,王秀丽,丁军娣. 多层核心集凝聚算法[J]. 软件学报,2013,24(3):490-506.

[20] 李晔,陈奕延,张淑芬. 基于密度峰值的混合型数据聚类算法设计[J]. 计算机应用,2018,38(2):483-490.

[27] 张文开. 基于密度的层次聚类算法研究[D]. 合肥:中国科学技术大学,2015.

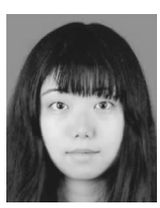
[29] 陈奕延. 彩色聚类效果图[EB/OL]. [2019-12-26]. https://blog.csdn.net/Dr_Chenyiyan/article/details/103715064.

作者简介:



陈奕延(1986-),男,北京人,博士,高级工程师,CCF 会员(49827M),研究方向为复杂不确定性科学决策和聚类分析。**E-mail**:townjam_sovietnia@163.com

CHEN Yi-yan, born in 1986, PhD, senior engineer, CCF member(49827M), his research interests include complex uncertainty scientific decision-making, and clustering analysis.



李晔(1992-),女,河北保定人,博士生,CCF 会员(74244G),研究方向为人工智能、机器学习和聚类分析。**E-mail**:ly1992bigdata@163.com

LI Ye, born in 1992, PhD candidate, CCF member(74244G), her research interests include artificial intelligence, machine learning, and clustering analysis.

李存金(1962-),男,内蒙古土默特左旗人,博士,教授,研究方向为现代组织管理理论与管理创新。**E-mail**:lixuan@bit.edu.cn

LI Cun-jin, born in 1962, PhD, professor, his research interests include modern organization management theory, and management innovation.