控制与决策
Control and Decision

基于密度的模糊代表点聚类算法

周洁， 姜志彬， 张远鹏， 王士同

引用本文:
周洁, 姜志彬, 张远鹏, 等. 基于密度的模糊代表点聚类算法[J]. 控制与决策, 2020, 35(5): 1123−1133.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1179

文章编号: 1001-0920(2020)05-1123-11 DOI: 10.13195/j.kzyjc.2018.1179

基于密度的模糊代表点聚类算法

周 洁†, 姜志彬, 张远鹏, 王士同

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 江苏省媒体设计与软件技术重点实验室，江苏 无锡 214122)

摘 要: 结合密度聚类和模糊聚类的特点,提出一种基于密度的模糊代表点聚类算法. 首先利用密度对数据点成为候选聚类中心点的可能性进行处理,密度越高的点成为聚类中心点的可能性越大;然后利用模糊方法对聚类中心点进行确定;最后通过合并聚类中心点确定最终的聚类中心. 所提出算法具有很好的自适应性,能够处理不同形状的聚类问题,无需提前规定聚类个数,能够自动确定真实存在的聚类中心点,可解释性好. 通过结合不同聚类方法的优点,最终实现对数据的有效划分. 此外,所提出的算法对于聚类数和初始化、处理不同形状的聚类问题以及应对异常值等方面具有较好的鲁棒性. 通过在人工数据集和UCI真实数据集上进行实验,表明所提出算法具有较好的聚类性能和广泛的适用性.

关键词: 聚类；密度聚类；模糊聚类；代表点聚类；聚类中心；鲁棒性

中图分类号: TP181 **文献标志码:** A

A density-based fuzzy exemplar clustering algorithm

ZHOU Jie†, JIANG Zhi-bin, ZHANG Yuan-peng, WANG Shi-tong

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. Jiangsu Key Laboratory of Digital Design and Software Technology，Wuxi 214122，China)

Abstract: According to the characteristics of density-based clustering and fuzzy clustering, a density-based fuzzy exemplar clustering algorithm is proposed. Firstly, the possibility of data points becoming candidate clustering centers is processed by the density. The higher the density of the data point is, the greater the likelihood for the data point to become a clustering center is. The clustering centers are then selected using the fuzzy method. The final clustering centers are determined by merging the clustering centers. The proposed algorithm has great adaptability, which can deal with clustering problems of different shapes, it can not only automatically determine cluster centers, but also get better results with higher accuracy. It can automatically determine the real clustering centers with good interpretability and there is no need to preset the number of clusters in advance. By combining the advantages of different clustering methods, the effective division of data can be realized. In addition, it has better robustness to number of clusters and initialization, processing clustering problems of different shapes, and dealing with outliers. Experiments on synthetic datasets and UCI datasets show that the proposed algorithm has better clustering performance and wide applicability.

Keywords: clustering; density clustering; fuzzy clustering; exemplar clustering; clustering centers; robustness

0 引 言

聚类分析[1-2]是数据挖掘领域中一个重要的研究课题,其目标是将数据对象分组成多个类或簇,使得在同一簇内的对象之间具有较高的相似度,而不同簇中的对象差别较大. 聚类作为一种常用的数据分析方法,已广泛应用于文本分析、图像分割、人脸识别、模式识别等领域[3-8]. 目前已有的聚类算法大致可分为以下几类[9-15]:基于划分的聚类方法、基于层次的聚类方法、基于密度的聚类方法、基于网格的聚类方法、基于模糊方法的聚类方法等.

基于代表点的聚类算法是聚类中的一个研究热点,代表点是数据集中真实存在的样本. 近年来提出了许多基于代表点的聚类算法[16-18],如K-medoid[17]、AP(affinity propagation clustering)[18]等. K-medoid是一个常见的基于代表点的方法,通过最小化平方误差和的方法寻找局部最优解,适用于球状簇但聚类结果依赖于初始点的设定. AP算法通过假定每个样本点为候选聚类中心点,不断迭代更新吸引信

收稿日期: 2018-09-01；**修回日期:** 2018-12-11.

基金项目: 国家自然科学基金项目(61170122, 61272210, 81701793)；江苏省自然科学基金项目(BK20130155)；南通市科技计划项目(MS12017016-2).

责任编委: 陈家伟.

†通讯作者. E-mail: 799489588@qq.com.

息(responsibility)矩阵R和归属信息(availability)矩阵A的值确定聚类中心点,再对其他数据点的类别进行划分. AP算法具有较好的聚类效果,但其效果依赖于偏向参数的选取,且迭代过程中不够稳定,时间复杂度很高. 基于密度的聚类算法也是近年来的研究热点[19-21],这类方法可以发现任意形状的簇,且不受噪声数据的干扰. DBSCAN (density-based spatial clustering of applications with noise)[20]是一种广泛应用的基于密度的算法,利用对高密度连通区域的聚类划分方法实现对样本的聚类,可以在有噪声的空间数据中发现任意形状的聚类,但其方法效果取决于参数输入,需要较强的对数据的先验知识. 另外一种基于密度的聚类算法是密度峰值快速搜索聚类算法(CFSFDP)[21],通过对数据点的局部密度和距离两个维度确定聚类中心点,效果较好,缺点是需要人工挑选出适应的聚类中心点. 模糊聚类也是近年来聚类研究的重点[22-24]. 模糊C-均值聚类算法(FCM)[22]是一种基于模糊的聚类算法,它不同于早期对于样本或者属于某个类,或不属于某个类的硬性划分,增加了对样本之间隶属度相对性的思考,从概率和统计分析的角度实现对样本的软划分,利用模糊隶属度和聚类中心点的不断迭代更新使得目标函数达到最小值以划分聚类,聚类效果好且应用广泛,对实际问题的解决能力强,但其聚类方法无法自适应确定聚类个数,且聚类中心点为虚拟的样本点.

在实际应用中,对于样本量大且维度高的数据集需要进行预处理,如果直接对样本中的候选聚类中心点进行选取,并对其他样本点按照相似度进行划分,则可以得到整个聚类结果的方法,能够有效减少算法的时间复杂度,并将聚类中心点确定为实际存在的点,能够获得更好的聚类效果,可解释性强. 文献[25]证明了密度越高的点成为聚类中心点的可能性越大,受此启发,通过点密度寻找代表点也是一种有效的方法. 本文结合密度聚类和模糊聚类的优点,提出一种基于密度的模糊代表点聚类算法(a density-based fuzzy exemplar clustering algorithm, DFEC). 该方法具有强烈的自适应性,无需提前规定聚类个数,能够自动确定真实存在的聚类中心点,可解释性好,并能够按照样本点到各个聚类中心点的模糊隶属度确定各个样本点的类别,实现对样本的有效聚类.

1 相关工作

1.1 选取代表点的方法

本节介绍基于代表点的聚类方法,该方法表示为

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}+\varsigma_j)y_j. \tag{1}$$

其中:$\boldsymbol{x}_i$、$\boldsymbol{x}_j$为数据集$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]$中的数据点;$d_{ij}$为数据点$\boldsymbol{x}_i$到数据点$\boldsymbol{x}_j$的距离;$\varsigma_j$为该算法的损益值,一般取距离中值;$y_j$为数据点$\boldsymbol{x}_j$是否作为代表点的指示,取值为1表示该点可以作为代表点,取值为0表示该点不能作为代表点.

1.2 模糊C均值聚类算法

模糊C均值聚类算法(FCM)[22]是一种典型的模糊聚类方法,其目标表达式为

$$J_m(\boldsymbol{U},\boldsymbol{V};\boldsymbol{X})=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^m d_{ij}^2;$$

$$\text{s.t. } u_{ij}\in[0,1],\ \sum_{j=1}^{c}u_{ij}=1. \tag{2}$$

其中:$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]\in\boldsymbol{R}^{r\times n}$为数据集;$m$为模糊指数且$m>1$,$m$值越大,表明模糊性越高;$d_{ij}$为第$i$个数据点$\boldsymbol{x}_i$到第$j$类聚类中心$\boldsymbol{v}_j$的距离,$d_{ij}=\|\boldsymbol{x}_i-\boldsymbol{v}_j\|$;$c$为类别数;$u_{ij}$为第$i$个数据点$\boldsymbol{x}_i$属于第$j$类的隶属度.

通过拉格朗日优化函数可得到如下迭代公式:

$$u_{ij}=\frac{1}{\displaystyle\sum_{k=1}^{c}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{v}_j\|}{\|\boldsymbol{x}_i-\boldsymbol{v}_k\|}\right)^{\frac{1}{m-1}}}, \tag{3}$$

$$\boldsymbol{v}_j=\frac{\displaystyle\sum_{i=1}^{n}\boldsymbol{x}_i u_{ij}^m}{\displaystyle\sum_{i=1}^{n}u_{ij}^m}. \tag{4}$$

其中$\boldsymbol{v}_j$为聚类中心点,但是为虚拟样本点,并不是真实的样本点.

1.3 密度对样本成为代表点的影响

文献[25]定理1指出,密度越高的点成为聚类中心点的可能性越大,其推导如下:对于给定的数据集$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]$,从中选择$M$个样本来估计$\boldsymbol{X}$中样本的概率密度,按照PW密度估计理论[26],$\boldsymbol{X}$中样本的概率密度可以表示为

$$p(\boldsymbol{x})=\frac{1}{M}\sum_{i=1}^{M}K_\sigma(\boldsymbol{x},\boldsymbol{x}_i).$$

若利用$\widehat{p}(\boldsymbol{x})$表示$\boldsymbol{X}$中样本真实概率密度,则$\widehat{p}(\boldsymbol{x})$与$p(\boldsymbol{x})$的积分累计误差为ISE$(\widehat{p}(\boldsymbol{x}),p(\boldsymbol{x}))$,且有

$$\text{ISE}(\widehat{p}(\boldsymbol{x}),p(\boldsymbol{x}))\approx\frac{1}{M}-\frac{2}{M}\sum_{i=1}^{M}p(\boldsymbol{x}_i)+C.$$

可见,ISE$(\widehat{p}(\boldsymbol{x}),p(\boldsymbol{x}))$仅与所选样本数量$M$和所选样本密度有关,换言之,若使ISE$(\widehat{p}(\boldsymbol{x}),p(\boldsymbol{x}))$越小,则

所选的M个样本的密度需要越高. 因此在聚类过程中,簇内代表点应该优先选择概率密度高的样本.

2 DFEC

本文从两个方面对聚类中心点的选取进行约束. 由文献[25]定理1可知,密度是衡量样本点是否具有成为聚类中心点资格的一个重要指标,但若仅通过密度一个维度来直接确定聚类中心点,则无法处理密度分布不均或者不平衡的复杂数据形式,不能得到理想的聚类效果. 因此,本文方法首先对样本进行密度计算,按照密度大小对各个数据点成为聚类中心点的可能性作相应的预测,并将这些具有成为聚类中心点资格的点作为候选聚类中心点,以便结合后续操作确定聚类中心. 其次,在对候选聚类中心点是否能够作为聚类中心点的挑选时采用模糊聚类的方法,以此寻找符合软划分的聚类中心点. 通过对密度和模糊同时约束聚类中心点的选取,能够合理有效地选出更具适应性的聚类中心点.

2.1 密度的计算

本文通过计算各个数据点的密度值实现对候选聚类中心点的初步判定. 数据集中所有样本点均可以通过计算其密度寻找具有较高密度的点作为候选聚类中心点.

将待聚类的数据集表示为$\boldsymbol{S} = \{\boldsymbol{x}_i\}_{i=1}^N \in \boldsymbol{R}^{r\times N}$, $\boldsymbol{I_S} = \{1,2,\dots,N\}$是数据集的指示集. 计算样本点的密度可以采用两种方法: 截断距离法和高斯核密度法.

通过截断距离计算样本点密度表示为

$$\rho_i = \sum_j \chi(d_{ij} - d_c). \tag{5}$$

其中: $\chi(\cdot)$为

$$\chi(x) = \begin{cases} 1, & x < 0; \\ 0, & x \geqslant 0; \end{cases} \tag{6}$$

d_{ij}为样本点$\boldsymbol{x}_i$到样本点$\boldsymbol{x}_j$的距离,本文取欧氏距离; d_c为截断距离,需要预先指定且其值大于0. 式(6)计算的是样本点$\boldsymbol{x}_i$在该点d_c半径范围内的样本个数,即样本点$\boldsymbol{x}_i$的局部密度值.

计算样本点的高斯核密度表示为

$$\rho_i = \sum_{j\in \boldsymbol{I_S}\setminus\{i\}} \mathrm{e}^{-\left(\frac{d_{ij}}{d_c}\right)^2}. \tag{7}$$

比较两种密度公式可知: 利用高斯核密度所计算的密度值是连续的,而通过截断距离计算的密度值是离散的; 利用高斯核密度对于样本点的密度计算更加精确,对样本点的密度表示发生冲突(多个样本点的密度相同情况)的概率更小. 因此本文采用高斯核密度法计算样本点的密度.

由文献[25]定理1,对样本点进行密度计算后,根据样本点的密度高低对样本点成为候选聚类中心点的可能性进行从高到低排序,并根据数据集情况舍弃排序靠后的点,以减少后续迭代次数.

2.2 利用模糊方法确定聚类中心点

前文通过密度对样本点成为候选聚类中心点的可能性进行了排序和初步处理,本节对标准FCM进行改进,提出具有模糊属性的聚类中心点选取方法. 对于DFEC,通过距离对算法进行限定,同时引入偏向参数f对算法进行调节. 此处f为该样本作为聚类中心点的惩罚,比采用固定惩罚量更合理,因此引入f能够使所提出算法更符合模糊思想,具有更好的适用性,更加符合实际需要.

在算法中加入样本点$\boldsymbol{x}_j$是否为聚类中心点的判别y_j,取值为$\{0,1\}$,有

$$y_j = \begin{cases} 1, & \boldsymbol{x}_j\text{为聚类中心}; \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

y_i取值为0代表该点不作为聚类中心点,即无需计算其隶属度函数的值;取值为1代表该点作为聚类中心点,需要对其进行隶属度函数值的计算.

基于密度的模糊代表点聚类算法(DFEC)目标表达式为

$$\begin{aligned} &\min_{u_{ij}} J = \sum_{i=1}^N\sum_{j=1}^N u_{ij}^m(d_{ij}+f)y_j + \sum_{j=1}^N y_j/\rho_j; \\ &\text{s.t.} \sum_{j=1}^N u_{ij} = 1, 0 \leqslant u_{ij} \leqslant y_j, m > 1, y_j \in \{0,1\}. \end{aligned} \tag{9}$$

其中: u_{ij}为样本点$\boldsymbol{x}_i$归属到聚类中心点$\boldsymbol{x}_j$的隶属度, d_{ij}为样本点$\boldsymbol{x}_i$到中心点$\boldsymbol{x}_j$的距离, f为偏向参数, y_j为是否选取样本点$\boldsymbol{x}_j$作为聚类中心点的标记值. 为了表述简单,此处将$d_{ij}+f$表示为$\overline{d}_{ij}$.

由DFEC的目标表达式可以看出,该方法无需对聚类个数进行提前设定,仅通过满足密度和模糊双重标准的目标表达式的最小化求解即可得到在密度和模糊约束下真实存在的聚类中心点. 式(9)第1部分$\sum\limits_{i=1}^N\sum\limits_{j=1}^N u_{ij}^m(d_{ij}+f)y_j$通过模糊思想确定$\boldsymbol{x}_j$点是否作为聚类中心点,以及在此情况下各个样本点$\boldsymbol{x}_i$到聚类中心点$\boldsymbol{x}_j$的模糊隶属度$u_{ij}$. 此时通过距离$d_{ij}$和偏向参数$f$检测样本$\boldsymbol{x}_j$成为整个数据集聚类中心点的可能性,相较仅通过距离$d_{ij}$限定该点的隶属度

u_{ij}和是否作为聚类中心点的标记y_j更具合理性. 偏向参数f代表了该样本作为聚类中心点的惩罚,调节偏向参数f能够对数据集的聚类中心点个数进行调节,更具实用价值. 为了计算简单,偏向参数f值为所有样本点相似度的中值. 式(9)第2部分$\sum\limits_{j=1}^{N} y_j/\rho_j$通过密度对聚类中心点的选择进行约束,该点的密度越高,$\sum\limits_{j=1}^{N} y_j/\rho_j$的值越小,目标表达式$J$的值也越小,点$\boldsymbol{x}_j$作为聚类中心点的可能性也越高. 通过密度和模糊两个标准组成的目标表达式能够充分反映聚类中心点选取的依据.

当$y_j = 1$时,该样本点被选为聚类中心点,需要计算其他点归属到该聚类中心点的隶属度u_{ij},利用拉格朗日乘子法对u_{ij}进行求解,得出u_{ij}的表达式为

$$u_{ij} = \frac{(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}}. \tag{10}$$

因此有

$$u_{ij} = \begin{cases} 0,\ y_j = 0; \\ \dfrac{(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}},\ y_j = 1. \end{cases} \tag{11}$$

求解过程如下.

当$y_j = 0$时,该样本点没有被选为聚类中心点,隶属度$u_{ij} = 0$;当$y_j = 1$时,该样本点被选为聚类中心点,需要计算其他点归属到该聚类中心点的隶属度u_{ij}. 对于DFEC目标函数,通过拉格朗日乘子法可得式(9)的拉格朗日函数为

$$L = \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ij}^m \bar{d}_{ij}y_j + \sum_{j=1}^{N} y_j/\rho_j + \sum_{i=1}^{N}\alpha_i\Big(\sum_{j=1}^{N} u_{ij} - 1\Big), \tag{12}$$

其中$\alpha_i \geqslant 0$为拉格朗日乘子.

令L对u_{ij}的偏导为零可得

$$\frac{\partial L}{\partial u_{ij}} = m u_{ij}^{m-1}\bar{d}_{ij}y_j + \alpha_i = 0. \tag{13}$$

求解可得

$$u_{ij} = \Big(\frac{-\alpha_i}{m\bar{d}_{ij}y_j}\Big)^{\frac{1}{m-1}} = \Big(\frac{-\alpha_i}{m}\Big)^{\frac{1}{m-1}}(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}. \tag{14}$$

由$\sum\limits_{j=1}^{N} u_{ij} = 1$可得

$$1 = \sum_{j=1}^{N}\Big(\frac{-\alpha_i}{m}\Big)^{\frac{1}{m-1}}(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}} = \Big(\frac{-\alpha_i}{m}\Big)^{\frac{1}{m-1}}\sum_{j=1}^{N}(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}. \tag{15}$$

注意到,当$y_j = 0$时,$u_{ij} = 0$. 为了表述方便,令$(\bar{d}_{ij}y_j)^{-1} = 0$. 于是有

$$\Big(\frac{-\alpha_i}{m}\Big)^{\frac{1}{m-1}} = \frac{1}{\sum\limits_{j=1}^{N}(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}} = \frac{1}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}}. \tag{16}$$

将式(19)代入(17),可得

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}}(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}} = \frac{(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}}. \tag{17}$$

因此有

$$u_{ij} = \begin{cases} 0,\ y_j = 0; \\ \dfrac{(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}}{\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}},\ y_j = 1. \end{cases} \tag{18}$$

将u_{ij}代回原目标表达式J,可得

$$J = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(\bar{d}_{ij}y_j)^{-\frac{1}{m-1}}}{\Big(\sum\limits_{k=1}^{N}(\bar{d}_{ik}y_k)^{-\frac{1}{m-1}}\Big)^m} + \sum_{j=1}^{N} y_j/\rho_j. \tag{19}$$

求解使得上述目标函数最小时所得到的聚类中心点即是在$y_j = 1$情况下的所有聚类中心点$\boldsymbol{x}_j$,样本点$\boldsymbol{x}_i$的类别由该点到聚类中心点$\boldsymbol{x}_j$的隶属度u_{ij}确定,从而实现对聚类中心点的确定和对数据集$\boldsymbol{S}$的模糊划分.

与FCM所得到的虚拟聚类中心点不同,所提出方法得出的聚类中心点是真实存在的样本点,在需要利用某个类的聚类中心点进行后续处理时,可以选取客观存在的聚类中心点,而对于虚拟的聚类中心点则无法进行利用. 此时,具有物理意义的真实聚类中心点更适应实际应用. 但在受噪声影响的样本中,真实聚类中心点受噪声影响会丧失其相应的意义,且聚类效果较虚拟聚类中心点的聚类算法所受的影响更大.

2.3 DFEC 算法描述

DFEC 首先通过对每个样本点的密度计算并降序排列以确定不同样本点作为候选聚类中心点的可能性; 然后通过迭代求解基于密度的模糊代表点聚类算法的目标函数的最小值进一步确定聚类的聚类中心点. 下面简要介绍算法步骤.

算法 1　求解聚类中心点算法.

输入: 训练数据集 $\boldsymbol{S}=\{\boldsymbol{x}_i\}_{i=1}^{N}$;

输出: 聚类中心点集 $\boldsymbol{D_S}=\{d_i\}_{i=1}^{N_d}$($N_d$ 为聚类中心点个数), 隶属度 $\boldsymbol{U}=\{u_{ij}\}_{i=1}^{N}{}_{j=1}^{N}$.

实验参数: 模糊系数 m, 偏向参数 f, 截断距离 d_c.

step 1: 对于数据集 $\boldsymbol{S}=\{\boldsymbol{x}_i\}_{i=1}^{N}$, 通过式(7)计算所有样本点 $\boldsymbol{x}_i$ 的密度 ρ_i, 并按照密度降序排列, 根据数据集情况选取一定量排列靠前的样本点加入聚类中心点序列 $\boldsymbol{D_S}$.

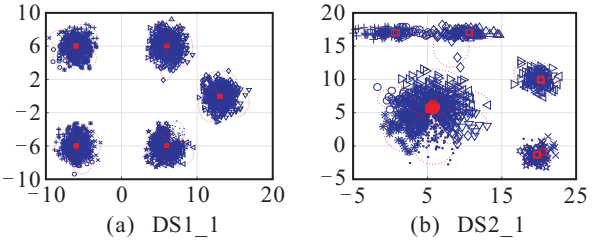
step 2: 选取具有最大密度值的点为聚类中心点, 通过式(19)计算初始目标函数 J 的值.

step 3: 将按照密度排序好的候选聚类中心点序列依次逐个加入聚类中心点集, 通过式(19)计算目标函数 J 的值, 若小于上次所计算的目标函数 J 的值, 则在聚类中心点集中保留该点, 否则将其从聚类中心点集中删除.

step 4: 当算法已经遍历过所有的可能聚类中心点时或者目标函数 J 的值不再变化时, 终止算法.

step 5: 依据算法中选出的聚类中心点集 $\boldsymbol{D_S}$ 和对于每个聚类中心点的隶属度矩阵 u_{ij} 确定每个样本点的类别.

下面通过两个数据集展示算法效果: 数据集 DS1 共有样本点 3 000 个, 距离 d_{ij} 采用欧氏距离进行计算, 聚类结果如图 1(a) 所示, 聚类后的聚类中心点共有 20 个, 每个类有 4 个代表点代表整个聚类. 数据集 DS2 中共有样本点 3 000 个, d_{ij}、d_c 取值与数据集 DS1 取值相同, 聚类结果如图 1(b) 所示, 聚类中心点有 33 个, 每个类中有多个聚类中心点来共同代表一个聚类.



(a) DS1_1　　(b) DS2_1

图 1　不同数据集选取聚类中心点结果

由实验可知, 该算法属于多中心点的聚类算法. 对于这种情况, 如果需要划分为每个类一个聚类中心点, 则可以利用距离连通的聚类中心点合并为一类并选取其中密度最大点的方式来确定最终的聚类中心点.

2.4 合并聚类中心点

由前文可以看出, 同一类中聚类中心点间的距离较小, 不同类之间聚类中心点距离较大, 因此可以设定一个距离阈值合并在一定距离范围内相互连通的聚类中心点, 具体步骤如下.

算法 2　聚类中心点的合并算法.

输入: 聚类中心点集 $\boldsymbol{D_S}=\{d_1,d_2,\ldots,d_{N_d}\}$, 距离临界值 ξ;

输出: 合并后聚类中心点集 $\boldsymbol{D_{S'}}$, 数据集类标 $T'=\{t_i\}_{i=1}^{N}$.

step 1: 对 $\boldsymbol{D_S}$ 中的聚类中心点按照密度从大到小排列, 选取第 1 个点加入搜索矩阵, 并标注为第 1 类, 该点作为第 1 类的聚类中心点.

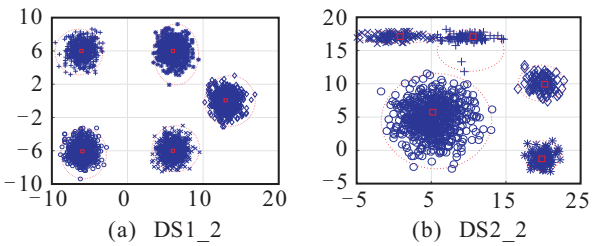
step 2: 依次计算搜索矩阵中的点与其他点的距离, 将距离小于 ξ 的点加入搜索矩阵并赋予搜索矩阵中点的类标, 同时在搜索矩阵中删除已扩展类标的点. 重复该做法直至搜索矩阵再次为空, 表明该类中所有聚类中心点寻找完毕, 并标记其数据集类标 T'.

step 3: 在剩余未标记点中选取密度最大的点作为第 2 类聚类中心点, 再次进行 step 2 的操作, 直至该类的所有聚类中心点确定完毕.

step 4: 重复 step 3, 直至聚类中心点集 $\boldsymbol{D_S}$ 中所有元素类别均已标注完毕.

step 5: 将聚类中心点的类标根据模糊聚类中的隶属度矩阵 u_{ij} 进行拓展, 实现整个数据集的标注.

由算法 2 可见, 本文算法无需提前确定聚类个数, 仅通过临界值 ξ 的选取即可做到对聚类个数的调节. 临界值 ξ 是对聚类中心点集 $\boldsymbol{D_S}$ 合并时选取的距离截断值, 即在进行类标传播的过程中将未标记数据集中的点 $\boldsymbol{x}_i$ 与已标记数据集中点的最短距离进行比对, 当其值小于临界值 ξ 时, 可将点 $\boldsymbol{x}_i$ 标记为已标记数据集中最近点的类标. 算法从具有最高密度值的点开始扩散, 直到该类所有点的半径 ξ 范围内不再有数据点时, 从未标记类标的剩余聚类中心点中选取密度最高值的点进行传播, 所有点类标确定时算法结束.



(a) DS1_2　　(b) DS2_2

图 2　不同数据集聚类中心合并后结果

图2为数据集DS1和DS2合并聚类中心点后的聚类结果. 由图2可见, DS1和DS2经过聚类中心点合并后聚类中心点个数均为5, 每个类中仅有1个聚类中心点, 其他数据点已按照其隶属度归属于相应的类别中.

2.5 算法复杂度及鲁棒性分析

根据算法1和算法2, 本文算法的时间复杂度可以分两部分进行分析. 根据算法1的描述, 其时间复杂度为$O(NN_d+N\log N+rN^2)$, N表示数据集样本个数, N_d表示聚类中心点集$\boldsymbol{D_S}$的规模, r为样本维数. 对于算法2, 其时间复杂度为$O(N_d^2N_c)$, N_c表示合并后聚类中心点集中的规模. 鉴于上述分析, 本文算法的时间复杂度可以表示为$O(NN_d+N\log N+rN^2+N_d^2N_c)$, 其中$N_d$和$N_c$均远小于$N$.

根据文献[27], 聚类方法的鲁棒性通常涉及3个方面: 1) 对初始化(聚类数和初始猜测)的鲁棒性; 2) 对簇体积的鲁棒性(检测不同数量的簇的能力); 3) 对噪声和异常值的鲁棒性(容忍噪声和异常值的能力). 本文从上述3方面分析DFEC的鲁棒性: 1) DFEC算法无需提前确定聚类个数, 且不需要初始化, 具有很好的鲁棒性; 2) DFEC算法继承了密度和代表点聚类算法的特点, 能够处理不同形状的聚类问题, 且对于数量不同的不平衡数据具有一定的处理能力和对簇体积的鲁棒性; 3) DFEC利用密度确定样本点成为候选聚类中心点的可能性, 能够将离群点等异常值排除在候选聚类中心点之外, 因此该算法对于离群点等异常值具有很好的鲁棒性.

因为DFEC的表达式与FCM类似, 具有与FCM相似的低抗噪性, 且DFEC中的聚类中心点是真实样本, 更易受到噪声影响. 所以DFEC对于异常值具有一定鲁棒性, 但对于噪声点的鲁棒性较差, 将在以后的工作中寻找解决方案以提出对噪声点也具有鲁棒性的算法.

3 实验与结论

为了更好地验证所提出算法的有效性, 采用人工数据集和UCI真实数据集对本文方法进行验证, 并与算法AP[18]、 K-mediods[17]、 KNNCLUST[28]、DBSCAN[20]对比.

3.1 数据集和评价标准

3.1.1 人工数据集

本文采用不同形状的2-D人工数据集对实验结果进行直观的判断, 生成4个人工数据集DS1 ~ DS4, 分布情况如图3所示. 数据集类别数分别为7、2、5、15, 大小分别为788、300、3 200、600.

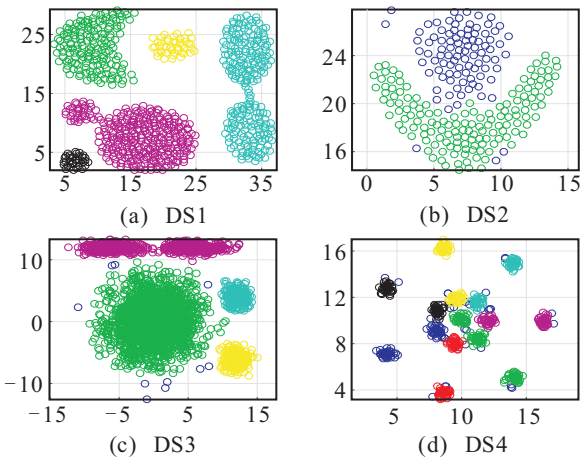


图3 人工数据集分布情况

3.1.2 UCI数据集

选取UCI中不同类型的真实数据集对本文聚类方法进行验证, 数据集样本个数、维度、类别数如表1所示.

表1 UCI数据集情况

数据集	样本个数	样本维度	类别数
lenses	24	4	3
heart	270	13	2
zoo	101	16	7
iris	150	4	3
vote	435	16	2
seeds	210	7	3

3.1.3 评价标准

利用NMI (normalized mutual information)[29]和ARI (adjusted rand index)[30]两种衡量标准对聚类结果进行评判. NMI和ARI分别定义为

$$\mathrm{NMI}=\frac{\sum\limits_i\sum\limits_j N_{ij}\log(N\times N_{ij}/N_i\times N_j)}{\sqrt{\Big(\sum\limits_i N_i\log(N_i\times N)\Big)\times\Big(\sum\limits_j N_j\log(N_j\times N)\Big)}},\tag{20}$$

$$\mathrm{ARI}=\Big(\sum_{ij}\begin{bmatrix}N_{ij}\\2\end{bmatrix}-\Big[\sum_i\begin{bmatrix}A_i\\2\end{bmatrix}\sum_i\begin{bmatrix}B_j\\2\end{bmatrix}\Big]I\begin{bmatrix}N\\2\end{bmatrix}\Big)\Big/\Big(\frac{1}{2}\Big[\sum_i\begin{bmatrix}A_i\\2\end{bmatrix}+\sum_i\begin{bmatrix}B_j\\2\end{bmatrix}\Big]-\Big[\sum_i\begin{bmatrix}A_i\\2\end{bmatrix}\sum_i\begin{bmatrix}B_j\\2\end{bmatrix}\Big]I\begin{bmatrix}N\\2\end{bmatrix}\Big).\tag{21}$$

其中

$$\begin{bmatrix} N \\ K \end{bmatrix} = \frac{N!}{K!(N-K)!},$$

$$A_i = \sum_j N_{ij}, B_j = \sum_i N_{ij}.$$

NMI 和 ARI 的取值区间分别为 $[0,1]$ 和 $[-1,1]$, 两个衡量标准的取值越高表示该聚类算法的准确率越高, 算法聚类结果越好.

3.1.4　参数设置及分析

对于所有对比方法, 通过网格搜索法对参数进行选取, 具体参数设置如表 2 所示. 实验所采用的硬件环境为: Intel I5-49503, 3 GHz×2, 8 GB RAM; 软件环境为: Windows 7 64 bit, Matlab R2016b.

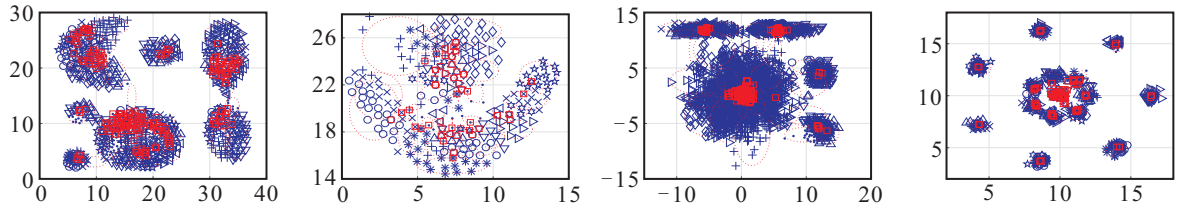
表 2　各方法的参数设置情况

方法	参数设置
AP	p 为相似度中值
K-mediod	K 为聚类个数
DBSCAN	$\min \text{Pts} \in [2:1:150], \text{esp} \in [0.001:0.001:1]$
KNNCLUST	$k \in [2:1:500]$
DFEC	$m=2, f$ 为距离中值 $d_c = \text{mean}[d_{q(0.01\times M)}, d_{q(0.02\times M)}]$, $D = \{d_1, d_2, \ldots, d_M\}$. $\xi \in \left[\min d_{ij} : \frac{\max d_{ij} - \min d_{ij}}{20} : \max d_{ij}\right]$, $i, j = 1, 2, \ldots, N_d$.

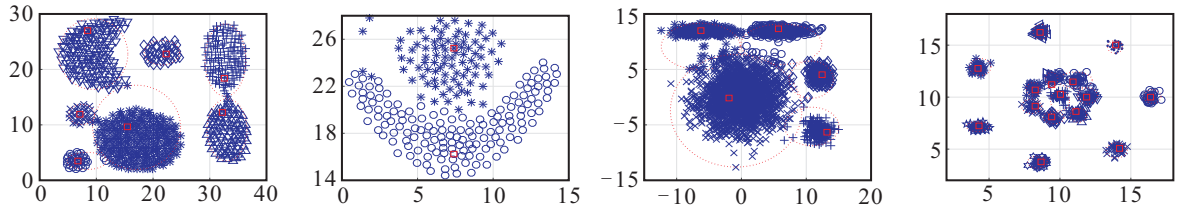
DFEC 中的截断距离 d_c 在密度选取过程中起重要作用, 选取了计算局部密度时的作用半径. 参照文献 [21] 的参数选取方式, 每个样本在 d_c 范围内的平均邻居个数约占样本总数的 1 %～2 %. 计算 $M = \frac{1}{2}N \times (N-1)$ 个 d_{ij} 的值, N 为样本个数, 按照升序排列为 $D = \{d_1, d_2, \ldots, d_M\}$. 则 d_c 的取值范围为 $d_c \in [d_{q(0.01\times M)}, d_{q(0.02\times M)}]$, 其中 $q(\cdot)$ 表示取整. 本文直接选取平均值进行计算, 即 $d_c = \text{mean}[d_{q(0.01\times M)}, d_{q(0.02\times M)}]$. DFEC 依据密度对样本作为候选聚类中心点的可能性排序, 参数 d_c 的值间接影响聚类的效果. 临界值 ξ 的取值范围为 $\xi \in [\min d_{ij}, \max d_{ij}]$. 当 $\xi < \min d_{ij}$ 时, 所有聚类中心点的 ξ 半径内均无其他数据点, 每个聚类中心点所代表的类均为单独的一类, ξ 失去意义. 当 $\xi \geqslant \max d_{ij}$ 时, 所有聚类中心点的 ξ 半径内均囊括所有样本点, 即所有样本点均成为同一类, 显然并不合理. 在 $\xi \in [\min d_{ij}, \max d_{ij})$ 时, 通过对 ξ 的调节能够对聚类的个数和精度进行调节, 本文等步长选取一定量在此区间内的数值对所提出方法进行实验, 由实验结果可以看出本文算法对该参数较为敏感.

3.2　人工数据集的结果

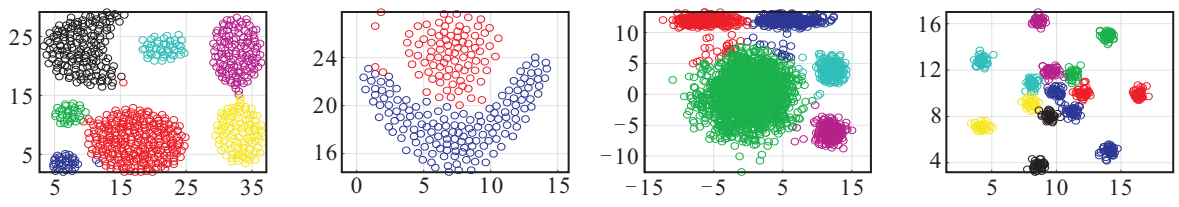
图 4 为本文聚类算法在人工数据集上的直接聚类结果和经过聚类中心点合并后的聚类中心及数据集聚类结果. 将 DFEC 与所有对比算法的效果进行对比, 结果如图 5 所示.



(a)　DFEC 对人工数据集的直接聚类结果

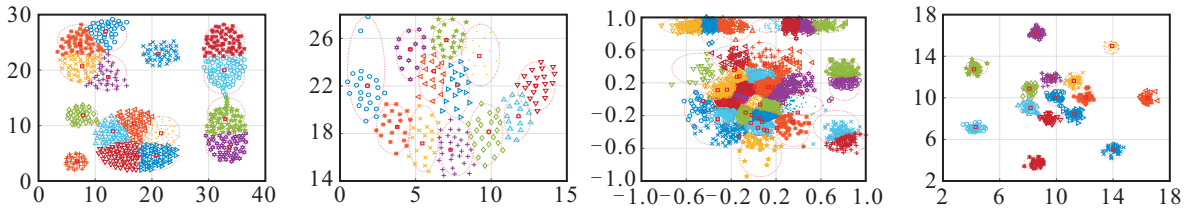


(b)　经过聚类中心点合并后的真实聚类中心

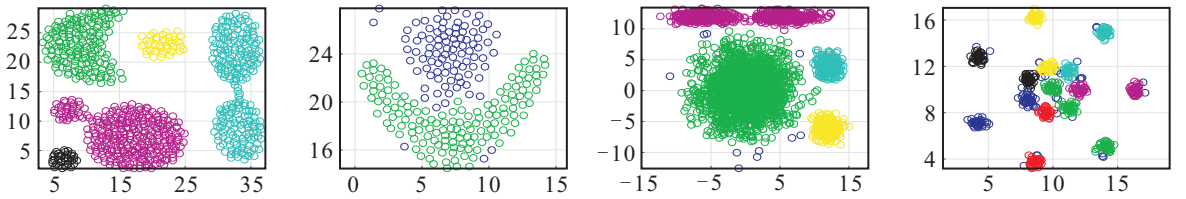


(c)　DFEC 经过聚类中心点合并后的聚类结果

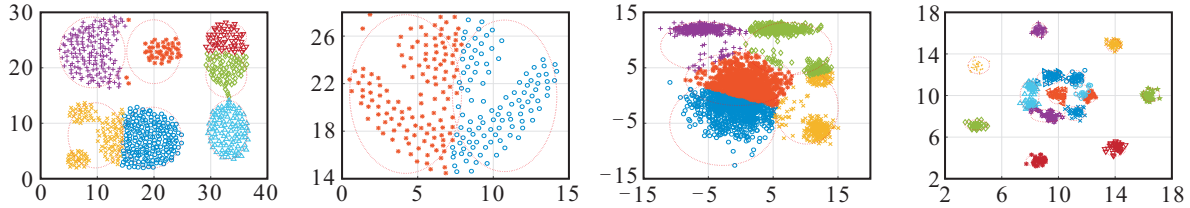
图 4　人工数据集聚类效果



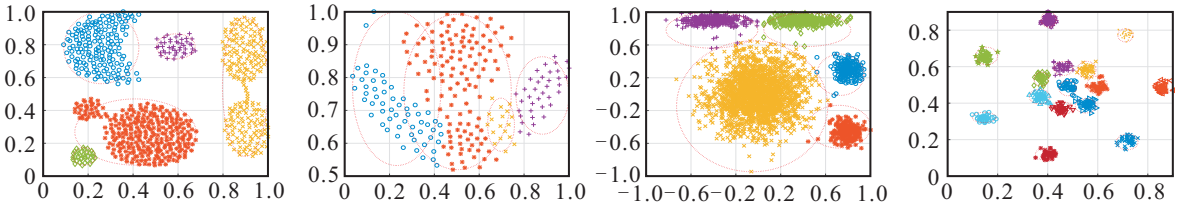
(a) AP算法



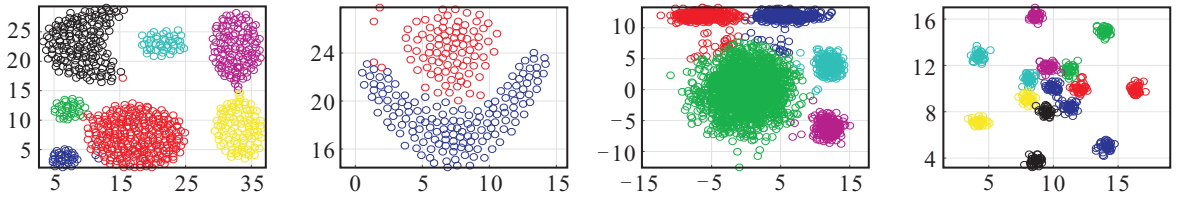
(b) DBCCAN算法



(c) *K*-mediod算法



(d) KNNCLUST算法



(e) DFEC算法

图 5 所有方法聚类效果

表 3 各种方法在DS1～DS4数据集上的NMI和ARI结果

方法	DS1			DS2			DS3			DS4		
	NMI	ARI	NC	NMI	ARI	NC	NMI	ARI	NC	NMI	ARI	NC
AP	0.775 0 (0)	0.391 1 (0)	16	0.4412 (0)	0.1293 (0)	12	0.568 6 (0)	0.084 5 (0)	35	0.994 2 (0)	0.992 8 (0)	15
K-mediod	0.789 9 (0.036 7)	0.650 1 (0.062 5)	7	0.014 1 (0.007 5)	0.007 1 (0.009 3)	2	0.707 6 (0.057 1)	0.541 9 (0.150 8)	5	0.903 5 (0.036 8)	0.765 3 (0.087 0)	15
DBSCAN	0.894 9 (0)	0.808 9 (0)	5	0.832 2 (0)	0.885 9 (0)	2	0.908 9 (0)	0.926 6 (0)	5	0.939 5 (0)	0.904 9 (0)	15
KNNCLUST	0.894 9 (0)	0.808 9 (0)	5	0.302 3 (0)	0.153 2 (0)	4	0.967 2 (0)	0.982 8 (0)	5	0.994 2 (0)	0.992 8 (0)	15
DFEC	0.988 7 (0)	0.991 0 (0)	7	0.888 3 (0)	0.933 7 (0)	2	0.926 1 (0)	0.947 4 (0)	5	0.970 9 (0)	0.944 1 (0)	15

在对数据集DS1的聚类结果中,DFEC能够对不平衡数据体现出一定优势,对数量较小的类别仍然有所区分. 在对数据集DS2的聚类结果中可以看出,DFEC与DBSCAN相似,体现出了DFEC采用密度聚类分割线性不可分聚类的优势. 在对数据集DS3的聚类结果中,DFEC能够兼顾分布不平衡的聚类,但效果稍逊于KNNCLUST方法. 此外,由图5可以看出DS4数据集是常规的团状数据,不同类别分布比较均匀,大多数聚类方法对于这种数据都能得到较好结果,因此相对较为容易聚类,提升空间不大,5种方法得到的结果几乎一致.

为了更直观地显示出DFEC与各个对比算法的聚类结果情况,将各种方法对于不同人工数据集的结果NMI指标和ARI指标列于表3(其中NC表示类别数).

本文方法DFEC与对比算法对各个人造数据集的NMI和ARI结果如图6所示.

由表3和图6可见,AP算法在固定偏向参数后聚类个数与预期结果相差较多,K-mediod不适用于不平衡聚类和线性不可分聚类,DBSCAN的综合聚类性能较好,KNNCLUST在团状数据集的聚类效果好,但在复杂数据集中表现很差. 虽然本文所提出方法在团状数据集表现不是最佳的,但准确度仍然很高,

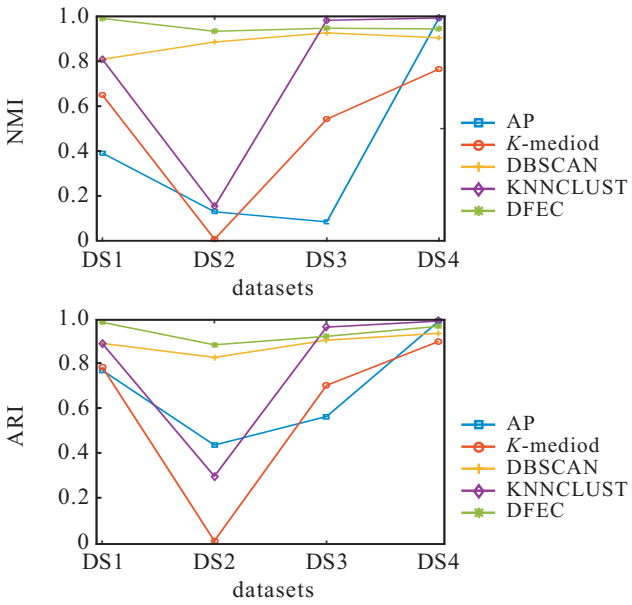


图6　DS1～DS4数据集上的NMI和ARI折线图

且具有更广泛的聚类适应性. 本文方法DFEC聚类得到的类别数与期望类别数一致,对非线性可分数据集体现出了一定的优势,具有自适应的能力,且聚类准确度高.

3.3　UCI数据集的结果

选取UCI数据集对本文方法进行验证,表4为各种方法在不同UCI数据集上的NMI和ARI的值.

表4　各种方法在UCI数据集上的NMI和ARI结果

方法	lenses			heart			zoo		
	NMI	ARI	NC	NMI	ARI	NC	NMI	ARI	NC
AP	0.508 6 (0)	−0.007 2 (0)	23	0.347 8 (0)	0 (0)	269	0.797 9 (0)	0.573 6 (0)	9
K-mediod	0.243 0 (0.111 9)	0.099 7 (0.101 6)	3	0.147 7 (0.098 0)	0.185 7 (0.130 4)	2	0.704 9 (0.037 3)	0.546 2 (0.164 0)	7
DBSCAN	0.549 8 (0)	0.152 9 (0)	3	0.190 7 (0)	0.058 4 (0)	17	0.447 4 (0)	0.043 7 (0)	13
KNNCLUST	0.549 8 (0)	0.152 9 (0)	8	0.229 7 (0)	0.011 9 (0)	74	0.695 2 (0)	0.622 1 (0)	3
DFEC	0.798 1 (0)	0.745 0 (0)	3	0.800 9 (0)	0.808 9 (0)	2	0.789 7 (0)	0.769 7 (0)	7

方法	iris			vote			seeds		
	NMI	ARI	NC	NMI	ARI	NC	NMI	ARI	NC
AP	0.464 7 (0)	0 (0)	149	0.331 4 (0)	0 (0)	434	0.540 9 (0)	0.279 3 (0)	12
K-mediod	0.815 0 (0.086 1)	0.789 3 (0.164 0)	3	0.352 5 (0.102 8)	0.418 9 (0.131 6)	2	0.630 3 (0.076 0)	0.657 0 (0.103 6)	3
DBSCAN	0.772 4 (0)	0.792 8 (0)	17	0.398 9 (0)	0.438 7 (0)	2	0.469 7 (0)	0.293 5 (0)	13
KNNCLUST	0.761 2 (0)	0.568 1 (0)	2	0.386 8 (0)	0.471 1 (0)	2	0.701 0 (0)	0.749 8 (0)	3
DFEC	0.818 0 (0)	0.795 5 (0)	3	0.403 8 (0)	0.471 1 (0)	2	0.702 2 (0)	0.739 6 (0)	3

图7为各方法对不同UCI数据集的NMI和ARI结果折线图.

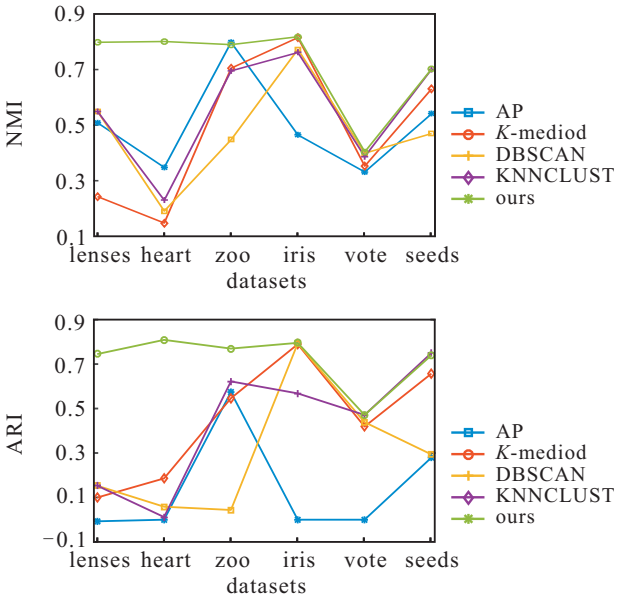


图7 UCI数据集上的NMI和ARI折线图

综合表4和图7可见:DFEC在UCI真实数据集中的聚类效果较其他对比方法更好,NMI和ARI值更高. 在不需要提前确定聚类个数的情况下,DFEC得到的聚类数与数据集真实类别数相同,具有良好的自适应性;AP、DBSCAN和KNNCLUST方法聚类结果的聚类数大多多于真实类别数,NMI和ARI性能也较差;K-mediod算法尽管能够得到正确的聚类个数,但性能很差. 综合以上各方法的UCI数据集实验结果可以验证本文方法的有效性和适用性.

4 结 论

本文提出了一种基于密度的模糊代表点聚类算法,该方法通过密度优化候选聚类中心点排列,并结合模糊的方法以迭代方式对数据集的聚类中心点进行确定,实现对整个数据集的划分. 通过在人工数据集和UCI真实数据集进行实验,表明所提出方法较其他对比方法有更好的自适应性和聚类准确性. 与其他聚类算法相比,所提出算法具有以下几点优势:

1) 所提出算法有机结合了密度聚类与模糊聚类的优点,提出了一种基于密度的模糊代表点聚类算法. 该方法应用文献[25]的定理1提出了与模糊聚类相结合的选取聚类中心点的方法,较仅通过密度聚类的方法有更强的理论支持,且通过模糊求解最优值的方法可以确保结果能够更加客观地符合实际需要.

2) 所提出算法与直接选取候选聚类中心点的聚类方法不同,通过密度将对成为候选聚类中心点可能性的预处理与基于模糊思想确定聚类中心点相结合,实现了对数据的软划分,且通过迭代方式直接求解聚类中心点,使得算法简单有效,更具有实用性.

3) 所提出算法是一种新的基于代表点的聚类方法,不仅能够确定真实存在的聚类中心点,而且具有很好的自适应性,无需提前规定聚类个数,且能够按照样本点到各个聚类中心点的相似度确定各个样本点的类别,实现对样本的有效聚类.

本文方法对于一些常规的团状数据的聚类结果较个别方法有略微差距,需要在后续研究中进行完善. 此外,本文方法对于噪声点的鲁棒性也较差,将在之后的工作中寻找解决方案以提出对噪声点也具有鲁棒性的更好的方法.

参考文献(References)

[1] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[2] Zhang Y, Lv D, Guo R, et al. Data mining: Practical machine learning tools and techniques[J]. Journal of Software Engineering, 1997, 11(1): 97-136.

[3] Xu R, Wunsch D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[4] Zhang W, Yoshida T, Tang X, et al. Text clustering using frequent itemsets[J]. Knowledge-Based Systems, 2010, 23(5): 379-388.

[5] Cades I, Smyth P, Mannila H. Probabilistic modeling of transactional data with applications to profiling, visualization and prediction, sigmod [C]. Proceedings of the 7th ACM SIGKDD. San Francisco: ACM Press, 2001: 37-46.

[6] 王杰, 梁吉业, 郑文萍. 一种面向蛋白质复合体检测的图聚类方法[J]. 计算机研究与发展, 2015, 52(8): 1784-1793.
(Wang J, Liang J Y, Zheng W P. A graph clustering method for detecting protein complexes[J]. Journal of Computer Research and Development, 2015, 52(8): 1784-1793.)

[7] Chuang K S, Tzeng H L, Chen S, et al. Fuzzy c-means clustering with spatial information for image segmentation[J]. Computerized Medical Imaging and Graphics, 2006, 30(1): 9-15.

[8] 王骏, 王士同, 邓赵红. 聚类分析研究中的若干问题[J]. 控制与决策, 2012, 27(3): 321-328.
(Wang J, Wang S T, Deng Z H. Survey on challenges in clustering analysis research[J]. Control and Decision, 2012, 27(3): 321-328.)

[9] Qian W N, Zhou A Y. Analyzing popular clustering algorithms from different viewpoints[J]. Journal of

Software, 2002, 13(8): 1382-1394.

[10] Hong Y, Kwong S. Learning assignment order of instances for the constrained k-means clustering algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2009, 39(2): 568-574.

[11] Fred A L N, Leitao J M N. Partitional vs hierarchical clustering using a minimum grammar complexity approach[C]. Proceedings of the Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition. Berlin: Springer, 2000: 193-202.

[12] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure[C]. Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. Philadelphia: ACM Press, 1999: 49-60.

[13] Zhao Y C, Song J. GDILC: A grid-based density isoline clustering algorithm[C]. Proceedings of the Internet Conference on Info-Net. Beijing: IEEE Press, 2001: 140-145.

[14] 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.

[15] 陈爱国, 王士同. 基于多代表点的大规模数据模糊聚类算法[J]. 控制与决策, 2016, 31(12): 2122-2130.
(Chen A G, Wang S T. Fuzzy clustering algorithm based on multiple medoids for large-scale data[J]. Control and Decision, 2016, 31(12): 2122-2130.)

[16] Krishnapuram R, Joshi A, Nasraoui O, et al. Low-complexity fuzzy relational clustering algorithms for web mining[J]. IEEE Transactions on Fuzzy Systems, 2001, 9(4): 595-607.

[17] Zhou J, Pan Y, Chen C L P, et al. K-medoids method based on divergence for uncertain data clustering[C]. Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics. Budapest: IEEE, 2016: 002671-002674.

[18] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[19] Ma W M, Chow E, Tommy W S. A new shifting grid clustering algorithm[J]. Pattern Recognition, 2004, 37(3): 503-514.

[20] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of the International Conference on Knowledge Discovery in Databases and Data Mining. Portland: AAAI Press, 1996: 226-231.

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

[22] Ruspini E H. New experimental results in fuzzy clustering[J]. Information Science, 1973, 18(2): 273-287.

[23] Lyer N S, Kandel A, Schneider M. Feature-based fuzzy classification for interpretation of mammograms[J]. Fuzzy Sets and System, 2000, 114(2): 271-280.

[24] Yang M S, Hu Y J, Lin K C R, et al. Segmenttation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithm[J]. Magnetic Resonance Imaging, 2002, 20(2): 173-179.

[25] Zhang Y P, Chung F L, Wang S T. Fast exemplar-based clustering by gravity enrichment between data objects[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 55(1): 163-178.

[26] Parzen E. On estimation of a probability density function and mode[J]. Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.

[27] Yang M S, Wu K L. A similarity-based robust clustering method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(4): 434-448.

[28] Tran T N, Wehrens R, Buydens L M C. KNN-kernel density-based clustering for high-dimensional multivariate data[J]. Computational Statistics & Data Analysis, 2006, 51(2): 513-525.

[29] Jiang Y Z, Chung F L, Wang S T, et al. Collaborative fuzzy clustering from multiple weighted views[J]. IEEE Transactions on Cybernetics, 2015, 45(4): 688-701.

[30] Qian P J, Chung F L, Wang S T, et al. Fast graph-based relaxed clustering for large data sets using minimal enclosing ball[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2012, 42(3): 672-687.

作者简介

周洁 (1992–) 女, 博士生, 从事人工智能与模式识别、机器学习的研究, E-mail: 799489588@qq.com;

姜志彬 (1991–), 男, 博士生, 从事人工智能与模式识别的研究, E-mail: jnuszmtjzb@163.com;

张远鹏 (1984–), 男, 讲师, 博士生, 从事人工智能与模式识别的研究, E-mail: 155297131@qq.com;

王士同 (1964–), 男, 教授, 博士生导师, 从事人工智能与模式识别、机器学习等研究, E-mail: wxwangst@jiangnan.eud.cn.

(责任编辑: 郑晓蕾)