控制与决策
Control and Decision

标签分布熵正则的模糊C均值平衡聚类方法

王哲昀, 胡文军, 徐剑豪, 胡天杰

引用本文:
王哲昀,胡文军,徐剑豪,胡天杰. 标签分布熵正则的模糊C均值平衡聚类方法[J]. 控制与决策, 2022, 37(9): 2274−2280.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0398

标签分布熵正则的模糊C均值平衡聚类方法

王哲昀, 胡文军†, 徐剑豪, 胡天杰

(1. 湖州师范学院 信息工程学院，浙江 湖州 313000；
2. 浙江省现代农业资源智慧管理与应用研究重点实验室，浙江 湖州 313000)

摘 要: 许多应用场景要求每个类别的数量相对平衡,而传统模糊C均值(FCM)聚类算法无法实现此功能. 为此,利用标签信息构造标签分布熵评价聚类的平衡度,然后将标签分布熵、模糊隶属度矩阵与标签矩阵之间的平方损失同时引入到传统FCM中,进而提出一种标签分布熵正则的模糊C均值平衡聚类方法(FCM_{LDE}). 同时,利用迭代方法和增广拉格朗日乘数法设计该模型的优化算法. 最后,利用6个真实数据集进行聚类实验,结果表明,所提方法在聚类性能和平衡性能上均具有很好的优势.

关键词: 平衡聚类；模糊C均值；标签分布熵；平方损失；迭代法；增广拉格朗日乘数法

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0398

引用格式: 王哲昀,胡文军,徐剑豪,等. 标签分布熵正则的模糊C均值平衡聚类方法[J]. 控制与决策, 2022, 37(9): 2274-2280.

Label distribution entropy regularized fuzzy C-means algorithm for balanced clustering

WANG Zhe-yun, HU Wen-jun†, XU Jian-hao, HU Tian-jie

(1. School of Information Engineering，Huzhou University，Huzhou 313000，China；2. Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources，Huzhou 313000，China)

Abstract: Many application scenarios require that the number of each category is relatively balanced, and the traditional fuzzy C-means (FCM) clustering algorithm cannot achieve this function. For this reason, we first design a label distribution entropy by using the label information , which can evaluate the balance degree of clustering. Then, the label distribution entropy and the square loss between the fuzzy membership matrix and the label matrix are simultaneously introduced into the traditional FCM, and then a fuzzy C-means balanced clustering method based on the regular label distribution entropy (FCM_{LDE}) is proposed. Besides, this paper designs an optimization algorithm to solve the proposed model through the iterative strategy and the augmented Lagrange multipliers method. Finally, clustering experiments are performed using six real data sets, and the results show that the proposed method has good advantages in clustering performance and balance performance.

Keywords: balanced clustering；fuzzy C-means (FCM)；label distribution entropy；square loss；iterative method；Augmented Lagrange Multipliers (ALMs) method

0 引 言

聚类是机器学习领域的一个热点问题,其目的是按照某种特定准则将相似样本分组到同一类别中[1],它是一种无监督学习方法[2]. 聚类问题广泛应用在经济学、生物学、人工智能等领域. 比如:在电子商务领域,聚类能帮助市场分析人员挖掘客户信息潜在知识,为企业管理者提供决策上的支持[3];在生物领域,对不同动植物的种类、特征、基因等进行分类,获得不同层次结构的认识[4-5];在人工智能领域,聚类能自动获取类别的关键词,实现智能搜索[6]. 另外,许多真实应用对聚类结果的平衡性提出了要求. 比如:在无线传感网络[7-8]中,每个类的样本数量应该大致相同,否则会造成网络不均衡的能源损耗,影响网络寿命;在课堂分组时要求每组人数应尽可能一样;自然情况下,世界男女性别比例大致相同. 因此,平衡聚类得到了广泛关注,其目的是在保证传统聚类性能时,每个

收稿日期: 2021-03-09；**录用日期:** 2021-06-03.
基金项目: 国家自然科学基金项目(61772198, U20A20228).
责任编委: 刘宝碇.
†通讯作者. E-mail: huwenjun@zjhu.edu.cn.

类的大小尽可能一致,避免数量过大或过小的样本分到同一个类中[9].

在过去几十年里,研究学者探索并提出许多聚类方法. K均值聚类[10-11](K-means)是一种基于划分的方法,其算法简单且聚类效率高,但该算法对初始聚类中心敏感,难以适用于数据划分不明确的情况. 此外,有一系列算法对K-means进行了不同角度的改进. 比如: 文献[12]通过改进初始聚类中心的选取方法,提出了K-means++算法; 文献[13]通过改进k参数选择,提出了一种改进的ISODATA(iterative self-organizing data)分析算法,提高了聚类的准确率. 文献[14]在1984年实现了模糊C均值算法(fuzzy C-means, FCM),其将隶属度定义为[0,1]而非0和1,通过隶属度以及样本到聚类中心之间的距离,采用迭代法进行聚类,该方法适合于具有近似超球体形状的数据集,对于不规则或不平衡的数据集则无能为力. 针对噪声污染的数据,文献[15-16]提出了基于密度的聚类算法(density-based spatial clustering, DBSC). 层次聚类[17]按照自顶向下和自底向上两种方式发现类的层次关系实现聚类,分别称为聚合型层次聚类和分裂型层次聚类.

上述聚类方法已经成功应用到很多领域,但因为这些模型在聚类过程中没有考虑到各个类的大小,所以无法很好地解决平衡聚类问题. 平衡聚类可分为两类,即硬平衡聚类和软平衡聚类. 严格限制聚类结果中每个类的大小的方法称为硬平衡聚类. 与硬平衡聚类相比,软平衡聚类并非绝对的平衡状态,它是一种平衡趋势. 为了解决平衡聚类问题,文献[18]向K-means模型添加k个约束并提出了BKM (balanced K-means)算法. 文献[19]针对谱聚类[20](spectral clustering, SC)方法提出了平衡标签传播方法,旨在找到最佳的统一图分区. 然而,诸如K-means和FCM等简单有效的聚类模型,其聚类过程没有考虑到各类平衡而导致不能适用于平衡聚类问题. 为此,本文针对FCM算法,利用标签信息构造标签分布熵以评价聚类过程的平衡度,并通过软平衡聚类策略将标签分布熵引入到FCM,该方法称之为标签分布熵正则的FCM平衡聚类方法(label distribution entropy regularized fuzzy C-means algorithm for balanced clustering, $\mathrm{FCM_{LDE}}$).

1　模糊C均值

给定数据集$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]\in R^{d\times n}$. 其中: n为样本个数, d为样本的维度. 假定将样本数据集$\boldsymbol{X}$划分为c类,传统的FCM模型为

$$\begin{aligned}&\min\ J(\boldsymbol{W},\boldsymbol{C})=\sum_{i=1}^{n}\sum_{k=1}^{c}w_{ik}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2;\\&\text{s.t.}\ w_{ik}\geqslant 0,\ \sum_{k=1}^{c}w_{ik}=1,\\&\qquad i=1,2,\dots,n,\ k=1,2,\dots,c.\end{aligned}\tag{1}$$

其中: $\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\dots,\boldsymbol{c}_c]\in R^{d\times c}$为聚类中心矩阵; $\boldsymbol{W}=[w_{ik}]\in R^{n\times c}$为模糊隶属度矩阵, w_{ik}为第i个样本隶属于第k类的隶属度; $m(m\geqslant 1)$为模糊加权指数,当$m>1$时为FCM算法, $m=1$时FCM则退化为硬划分(hard C-means, HCM).

此优化问题可以通过拉格朗日方法进行求解,得到如下的隶属度和聚类中心迭代公式:

$$\boldsymbol{c}_k=\sum_{i=1}^{n}w_{ik}^{m}\boldsymbol{x}_i\Big/\sum_{i=1}^{n}w_{ik}^{m},\tag{2}$$

$$w_{ik}=\frac{1}{\displaystyle\sum_{j=1}^{c}\Big(\frac{\|\boldsymbol{x}_i-\boldsymbol{c}_k\|}{\|\boldsymbol{x}_i-\boldsymbol{c}_j\|}\Big)^{\frac{2}{m-1}}}.\tag{3}$$

为了更好地描述FCM和本文工作,这里给出FCM求解算法具体步骤.

算法1　FCM算法.

输入: 数据集$\boldsymbol{X}$、聚类数c、最大迭代次数$\max_t$以及阈值e;

输出: 模糊隶属度矩阵$\boldsymbol{W}$、聚类中心$\boldsymbol{C}$.

step 1: 随机初始化模糊隶属度矩阵,满足式(1)中的约束条件.

step 2: 利用式(2)计算聚类中心$\boldsymbol{C}^{(t)}$.

step 3: 利用式(3)更新模糊隶属度矩阵$\boldsymbol{W}^{(t)}$.

step 4: 根据当前的$\boldsymbol{W}$和$\boldsymbol{C}$计算目标函数式(1)的值. 若迭代次数大于$\max_t$或者相邻两次目标函数值差的绝对值小于阈值e,则算法停止; 否则,令$t=t+1$,转step 2.

2　标签分布熵正则的模糊C均值

FCM只考虑样本到聚类中心之间的距离,通过欧氏距离准则划分样本,其迭代过程依赖初始聚类中心的选取. 因此, FCM对离群点(即噪声)较为敏感,容易陷入局部最小值. 另外, FCM模型及其迭代优化过程未考虑到各类样本的分布情况,即每个类大小情况,故传统的FCM无法适用平衡聚类的应用场景. 为此,首先利用标签信息构造标签分布熵评价数据中类的平衡度; 然后使用一个平方损失项正则化模糊隶属度矩阵的划分过程,以评估模糊隶属度矩阵与标签矩阵之间的拟合误差; 最后将模糊隶属度矩阵与标签矩阵的平方损失、标签分布熵同时加入到FCM模型中,提出标签分布熵正则的模糊C均值平衡聚类,以满足集群平衡和高质量集群. 同

时,利用增广拉格朗日乘数法[21](augmented Lagrange multipliers, ALMs)提出求解模型的交替更新优化算法,并给出完整的算法流程和时间复杂度分析.

2.1 标签分布熵

假定二元矩阵$\boldsymbol{Y}=[y_{ik}]\in R^{n\times c}$(0-1矩阵)为数据的标签矩阵,为了方便,二元特性用$\boldsymbol{Y}\in \text{Ind}$表示. 因此,标签矩阵$\boldsymbol{Y}$与模糊隶属度矩阵$\boldsymbol{W}$之间存在如下映射关系:

$$y_{ik}=\begin{cases}1,\ w_{ik}=\arg\max\{w_{ij}\}_{j=1}^{c};\\ 0,\ \text{otherwise}.\end{cases}\tag{4}$$

用$\boldsymbol{b}=[n_1,n_2,\dots,n_c]^{\mathrm{T}}\in R^c$表示类大小的向量,其中$n_k$为第$k$类的大小,由此可以得到$\boldsymbol{b}=\boldsymbol{Y}^{\mathrm{T}}\cdot\boldsymbol{1}(\boldsymbol{Y}\in\text{Ind},\boldsymbol{1}$为全1的列向量). 为了使每个类的大小接近$n/c$,即所有类大小差异达到最小化,将$n_k$视为随机变量,即$N\in\{n_1,n_2,\dots,n_c\}$,则其方差$D(N)$为

$$\begin{aligned}D(N)=&\frac{1}{c}\sum_{k=1}^{c}\Big(n_k-\frac{n}{c}\Big)^2=\\ &\frac{1}{c}\sum_{k=1}^{c}n_k^2-\frac{2n}{c^2}\sum_{k=1}^{c}n_k+\frac{n^2}{c^2}=\\ &\frac{1}{c}\Big(\|\boldsymbol{b}\|^2-\frac{n^2}{c}\Big).\end{aligned}\tag{5}$$

这里将式(5)中的$\|\boldsymbol{b}\|^2$定义为标签分布熵$H(\boldsymbol{Y})$,即

$$H(\boldsymbol{Y})=\|\boldsymbol{b}\|^2=\|\boldsymbol{Y}^{\mathrm{T}}\cdot\boldsymbol{1}\|^2=\mathrm{Tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{1}\cdot\boldsymbol{1}^{\mathrm{T}}\boldsymbol{Y}),\tag{6}$$

其中$\mathrm{Tr}(\cdot)$为矩阵的迹. 从式(5)和(6)可以观察到,当标签分布熵$H(\boldsymbol{Y})=n^2/c$时,方差$D(N)=0$,这意味着类是完全平衡的. $H(\boldsymbol{Y})$越大,对应的$D(N)$也越大,意味着不平衡类越多. 因此,$H(\boldsymbol{Y})$能够反映出类的不平衡程度.

2.2 FCM_{LDE}模型

FCM中的模糊隶属度矩阵$\boldsymbol{W}$可视为聚类的输出值,那么在其聚类时应该尽量减少输出值与标签矩阵$\boldsymbol{Y}$的损失. 另外,为了使得FCM能适用平衡聚类,要求其在聚类过程中充分考虑标签分布熵,即优化过程中要最小化标签分布熵. 为此,将传统FCM模型、平方损失项和标签分布熵有机结合起来,提出如下软平衡模型:

$$\begin{aligned}&\min_{\boldsymbol{Y}\in\text{Ind}}O(\boldsymbol{W},\boldsymbol{C},\boldsymbol{Y})=\\ &\sum_{i=1}^{n}\sum_{k=1}^{c}w_{ik}^m\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2+\lambda\|\boldsymbol{Y}-\boldsymbol{W}\|_F^2+\\ &\gamma\|\boldsymbol{Y}^{\mathrm{T}}\cdot\boldsymbol{1}\|^2;\\ &\text{s.t.}\ w_{ik}\geqslant 0,\ \sum_{k=1}^{c}w_{ik}=1,\\ &\qquad i=1,2,\dots,n,\ k=1,2,\dots,c.\end{aligned}\tag{7}$$

其中:λ为正则化参数,γ为平衡参数可控制模型的软平衡程度. 与传统的FCM算法一样,FCM_{LDE}算法遵循FCM最优化准则,距离越小则隶属度越大.

2.3 算法求解

在提出的模型中,有3个变量$\boldsymbol{W}$、$\boldsymbol{C}$和$\boldsymbol{Y}$需要求解,同时产生了3种不同类型的约束条件,包括不等式约束、等式约束和二元约束. 这是一个NP-hard问题,直接解决它是十分困难的. 为此使用交替更新策略求解上述问题. 具体如下:首先,固定$\boldsymbol{Y}$,求解$\boldsymbol{W}$和$\boldsymbol{C}$,将式(7)转化为最小平方损失项正则的FCM模型;然后,固定$\boldsymbol{W}$和$\boldsymbol{C}$,求$\boldsymbol{Y}$,将式(7)转化为一个等式约束问题,并且通过增广拉格朗日乘数法求解.

当$\boldsymbol{Y}$固定时,目标函数(7)转换为

$$\begin{aligned}&\min O(\boldsymbol{W},\boldsymbol{C})=\\ &\sum_{i=1}^{n}\sum_{k=1}^{c}w_{ik}^m\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2+\lambda\|\boldsymbol{Y}-\boldsymbol{W}\|_F^2;\\ &\text{s.t.}\ w_{ik}\geqslant 0,\ \sum_{k=1}^{c}w_{ik}=1,\\ &\qquad i=1,2,\dots,n,\ k=1,2,\dots,c.\end{aligned}\tag{8}$$

考虑到约束条件$\sum_{k=1}^{c}w_{ik}=1$,可以构造如下拉格朗日函数:

$$\begin{aligned}L(\boldsymbol{W},\boldsymbol{C},\boldsymbol{\alpha})=&\sum_{i=1}^{n}\sum_{k=1}^{c}w_{ik}^m\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2+\lambda\|\boldsymbol{Y}-\\ &\boldsymbol{W}\|_F^2+\sum_{i=1}^{n}\alpha_i\Big(1-\sum_{k=1}^{c}w_{ik}\Big),\end{aligned}\tag{9}$$

其中$\boldsymbol{\alpha}$为拉格朗日乘子. 对参数$\boldsymbol{c}_k$求偏导数可以得到如下方程:

$$\frac{\partial L}{\partial\boldsymbol{c}_k}=-2\sum_{i=1}^{n}w_{ik}^m(\boldsymbol{x}_i-\boldsymbol{c}_k)=0.\tag{10}$$

由式(10)可得聚类中心

$$\boldsymbol{c}_k=\sum_{i=1}^{n}w_{ik}^m\boldsymbol{x}_i\Big/\sum_{i=1}^{n}w_{ik}^m.\tag{11}$$

同样对w_{ik}求偏导数,得到如下方程:

$$\begin{aligned}&\frac{\partial L}{\partial w_{ik}}=\\ &mw_{ik}^{m-1}\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2-2\lambda(y_{ik}-w_{ik})-\alpha_i=0.\end{aligned}\tag{12}$$

由于式(12)涉及到模糊加权指数m,为了不失一般性,可以考虑两种特殊情况求解式(12),即$m=1$和$m=2$. 通过式(12)可以得到

$$w_{ik}=\begin{cases}(2\lambda y_{ik}+\alpha_i-\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2)/(2\lambda),\ m=1;\\ (2\lambda y_{ik}+\alpha_i)/(2\lambda+2\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2),\ m=2.\end{cases}\tag{13}$$

通过约束条件$\sum_{k=1}^{c} w_{ik}=1$和简单的数学推导可以获得拉格朗日乘子$\boldsymbol{\alpha}$的解,将其重新代回式(13)得到隶属度迭代式

$$w_{ik}=\begin{cases}\Big(2\lambda y_{ik}+\dfrac{1}{c}\sum\limits_{l=1}^{c}(d_{il}-d_{ik})\Big)\Big/(2\lambda),\ m=1;\\ \Big(1+\sum\limits_{l=1}^{c}\dfrac{\lambda(y_{ik}-y_{il})}{\lambda+d_{il}}\Big)\Big/\sum\limits_{l=1}^{c}\dfrac{\lambda+d_{ik}}{\lambda+d_{il}},\ m=2.\end{cases}\tag{14}$$

其中

$$d_{ik}=\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2.\tag{15}$$

上述求解过程得到的w_{ik}仅考虑了等式约束,故w_{ik}可能出现小于0的情况,这与$w_{ik}\geqslant 0$的约束条件相矛盾. 为此,可以做如下处理:

$$w_{ik}=\max(w_{ik},0),\tag{16}$$

$$\hat{w}_{ik}=w_{ik}\Big/\sum_{j=1}^{c} w_{ij}.\tag{17}$$

然后固定$\boldsymbol{W}$和$\boldsymbol{C}$,将目标函数(7)转换为

$$\min_{\boldsymbol{Y}\in\mathrm{Ind}} O(\boldsymbol{Y})=\lambda\|\boldsymbol{Y}-\boldsymbol{W}\|_F^2+\gamma\|\boldsymbol{Y}^{\mathrm{T}}\cdot\boldsymbol{1}\|^2.\tag{18}$$

由于标签矩阵$\boldsymbol{Y}$是一个二元矩阵,为此定义一个辅助变量$\boldsymbol{Z}\in R^{n\times c}$替换式(18)中的$\boldsymbol{Y}$,将式(18)转换为一个等式约束的优化问题,公式如下:

$$\begin{aligned}&\min_{\boldsymbol{Y}\in\mathrm{Ind}} O(\boldsymbol{Z})=\lambda\|\boldsymbol{Z}-\boldsymbol{W}\|_F^2+\gamma\|\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{1}\|^2;\\ &\text{s.t. }\boldsymbol{Y}-\boldsymbol{Z}=\boldsymbol{0}.\end{aligned}\tag{19}$$

利用增广拉格朗日乘数法将式(19)转化为无约束问题,具体如下:

$$\begin{aligned}\min L(\boldsymbol{Z},\boldsymbol{U},\mu)=&\lambda\|\boldsymbol{Z}-\boldsymbol{W}\|_F^2+\gamma\|\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{1}\|^2+\\ &\frac{\mu}{2}\Big\|\boldsymbol{Y}-\boldsymbol{Z}+\frac{1}{\mu}\boldsymbol{U}\Big\|_F^2.\end{aligned}\tag{20}$$

其中:$\boldsymbol{U}$为拉格朗日乘子,$\mu>0$为缩放因子,对$\boldsymbol{Z}$求偏导可得

$$\boldsymbol{Z}=((2\lambda+\mu)\boldsymbol{I}_n+2\gamma\boldsymbol{1}\cdot\boldsymbol{1}^{\mathrm{T}})^{-1}(2\lambda\boldsymbol{W}+\mu\boldsymbol{Y}+\boldsymbol{U}).\tag{21}$$

此时,当$\boldsymbol{Z}$、$\boldsymbol{U}$和μ都固定时,得到$\boldsymbol{Y}$的优化解等价于

$$\boldsymbol{Y}=\min_{\boldsymbol{Y}\in\mathrm{Ind}}\Big\|\boldsymbol{Y}-\Big(\boldsymbol{Z}-\frac{1}{\mu}\boldsymbol{U}\Big)\Big\|_F^2=\min_{\boldsymbol{Y}\in\mathrm{Ind}}\|\boldsymbol{Y}-\boldsymbol{V}\|_F^2.\tag{22}$$

其中

$$\boldsymbol{V}=\boldsymbol{Z}-\frac{1}{\mu}\boldsymbol{U}.\tag{23}$$

因为$\boldsymbol{Y}$是二元矩阵,且每行只有一个元素为1,所以

$$\hat{y}_{ik}=\begin{cases}1,\ v_{ik}=\mathrm{argmax}\{v_{ij}\}_{j=1}^{c};\\ 0,\ \text{otherwise}.\end{cases}\tag{24}$$

综合上述过程,$\mathrm{FCM_{LDE}}$模型求解的具体算法流程如下.

算法2　$\mathrm{FCM_{LDE}}$算法.

输入:数据集$\boldsymbol{X}$,聚类数c,$\rho>1$,$\lambda>0$,$\gamma>0$,缩放因子$\mu>0$,初始$\boldsymbol{U}=0$,最大迭代次数$\max_t$,以及迭代停止时最小阈值e;

输出:标签矩阵$\boldsymbol{Y}$、聚类中心$\boldsymbol{C}$.

step 1: 随机初始化$\boldsymbol{W}\in R^{n\times c}$,同时满足约束条件$\sum_{k=1}^{c} w_{ik}=1$和$w_{ik}\geqslant 0$.

step 2: 利用式(4)求解$\boldsymbol{Y}^{(0)}$.

step 3: 利用式(11)求解聚类中心矩阵$\boldsymbol{C}^{(t)}$.

step 4: 利用式(15)求解d_{ik}.

step 5: 固定标签矩阵$\boldsymbol{Y}$,利用式(14)、(16)和(17)求解$\boldsymbol{W}^{(t)}$.

step 6: 固定标签矩阵$\boldsymbol{Y}$和模糊隶属度矩阵$\boldsymbol{W}$,利用式(21)求解$\boldsymbol{Z}^{(t)}$.

step 7: 固定$\boldsymbol{Z}$、$\boldsymbol{U}$和μ,利用式(23)和(24)求解$\boldsymbol{Y}^{(t)}$.

step 8: 更新拉格朗日乘子$\boldsymbol{U}$:$\boldsymbol{U}^{(t+1)}=\boldsymbol{U}^{(t)}+\mu^{(t)}(\boldsymbol{Y}^{(t)}-\boldsymbol{W}^{(t)})$,以及更新缩放因子$\mu$:$\mu^{(t+1)}=\rho\mu^{(t)}$.

step 9: 根据当前的$\boldsymbol{W}$、$\boldsymbol{C}$和$\boldsymbol{Y}$计算目标函数(7)的值. 若迭代次数大于$\max_t$或者相邻两次目标函数值差的绝对值小于阈值e,则算法停止;否则,令$t=t+1$,转step 3.

2.4　时间复杂度分析

传统的FCM算法的时间复杂度为$O(ndc^2)$. 本文提出的$\mathrm{FCM_{LDE}}$算法进行优化时,采用了一种交替更新策略,当拉格朗日乘子$\boldsymbol{U}$和缩放因子μ都固定时,可以在单次迭代后获得局部最优解. 单次迭代包括两个步骤,第1步是通过平方损失项将其正则化,从而优化FCM模型,第2步是通过增广拉格朗日乘数法优化平方损失项与标签分布熵的组合. 此外,在单次迭代中,主要的运算量在于对$\boldsymbol{C}$、$\boldsymbol{W}$和$\boldsymbol{Z}$的求解. 根据式(11)和(14)可以发现计算$\boldsymbol{C}$的时间复杂度为$O(ndc)$,计算$\boldsymbol{W}$的时间复杂度为$O(nd^2c)$. 通过式(21)可知求解涉及到矩阵的逆,容易得出它的逆为

$$\begin{aligned}&((2\lambda+\mu)\boldsymbol{I}_n+2\gamma\boldsymbol{1}\cdot\boldsymbol{1}^{\mathrm{T}})^{-1}=\\ &\frac{(2\lambda+\mu+2n\gamma)\boldsymbol{I}_n-2\gamma\boldsymbol{1}\cdot\boldsymbol{1}^{\mathrm{T}}}{(2\lambda+\mu)^2+2n\gamma(2\lambda+\mu)}.\end{aligned}\tag{25}$$

由此计算$\boldsymbol{Z}$的时间复杂度为$O(n^2c)$. 以上时间复杂度主要考虑乘法,因此单次迭代的总时间复杂度

为$O(ndc+nd^2c+n^2c)$. 虽然本文提出的FCM_{LDE}算法增大了时间开销,但是其聚类性能和平衡性能明显提高.

3 实验与分析

为了验证FCM_{LDE}算法的有效性,所有实验均使用一台CPU为Intel(R) Core(TM)5-8500 3.00 GHz、内存为8 GB、操作系统为Windows 10的64位电脑,实验平台为Matlab R2017b. 因为本文针对的实际应用场景是平衡聚类问题,即要求每个类别的数量相对平衡,所以使用6个平衡数据集,包括COIL-20、Isolet1、ORL_32×32、UMIST、PIE和AR,用于比较K-means、FCM、K-means++、ISODATA、BKM和FCM_{LDE}算法的聚类性能和平衡性能.

3.1 数据集及预处理

6个数据集包括: COIL-20、Isolet1、ORL_32×32、UMIST、PIE和AR,实验前对每个数据集特征进行归一化处理. 上述数据集维度高,含有较多噪声,为此利用局部保留投影(locality preserving projection, LPP)[22]方法对原始数据做降噪处理,涉及到的近邻参数$k=5$,热核带宽参数t通过计算样本点之间的距离均值得到,为防止广义特征方程求解出现奇异值,采用tikhonov正则化方法,设正则化系数$\sigma=0.1$.

COIL-20包含20个对象,且每个对象从不同角度拍摄了72张照片,共1 440张,每个图像调整为32×32像素,能从http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php下载得到. Isolet 1为Isolet语音字母识别数据的第1部分,总共选取150位受试者,分为5组,每组30人,让受试者从字母表中说2次,这样每个对象可得52个样本,每个样本的特征为617,能从http://archive.ics.uci.edu/ml/datasets/ISOLET下载得到. ORL_32×32包含每个对象不同时间、不同光照、不同面部表情和面部细节拍摄的10张照片,共400张,每个图像调整为32×32像素,能从http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html下载得到. UMIST包含20个对象从侧面到正面的各种样貌、种族和姿势的564张人脸照片,每个对象随机选取19张照片,每个图像为92×112像素,能从http://images.ee.umist.ac.uk/danny/database.html下载得到. PIE包含53个对象在不同姿态、光照条件和表情下的1 166张人脸照片,每个图像为32×32像素,能从http://www.ri.cmu.edu/projects/project_418.html下载得到. AR人脸数据集包含126位对象在不同光照下不同表情的照片,选取20个对象总共280张照片,每个图像为32×32像素,能从http://rvl1.ecn.purdue.edu/aleix/aleix_face_DB.html下载得到. 实验数据集的具体信息如表1所示.

表1 实验数据集

数据集	类型	样本数	特征数	类别数	降维后的特征数
COIL-20	物体	1 440	1 024	20	220
Isolet 1	语音	1 560	617	26	470
ORL_32×32	人脸	400	1 024	40	40
UMIST	人脸	380	10 304	20	400
PIE	人脸	1 166	1 024	53	200
AR	人脸	280	1 024	20	110

3.2 评价指标

本文选择以下3个常见的评价指标评价所有聚类算法的聚类性能和平衡性能. 使用准确率(accuracy, ACC)和标准互信息(normalized mutual information, NMI)评价聚类的有效性,使用标准熵(normalized entropy, NE)评估集群的平衡状态.

ACC[23]定义如下:

$$\text{ACC}=\sum_{i=1}^{n}\delta(s_i,\text{map}(q_i))/n. \tag{26}$$

其中: n为数据集中的所有样本数目; $\delta(x,y)$为一个δ函数,若$x=y$,则输出为1,否则为0; s_i为先验知识中数据原始标签; q_i为实验得到的聚类标签; $\text{map}(q_i)$是将聚类标签与原始标签匹配的映射函数. ACC值越大,表明聚类结果越准确.

MI[24]用来评价聚类结果与数据集原分布的近似程度,定义如下:

$$\text{MI}(X,Y)=\sum_{x_i\in X,y_j\in Y}p(x_i,y_j)\cdot\log_2\frac{p(x_i,y_j)}{p(x_i)\cdot p(y_j)}. \tag{27}$$

其中: X为先验知识中的类集, Y为实验中得到的类集, $p(x_i)$为数据集中的样本点属于类x_i的概率, $p(y_j)$为数据集中的样本点属于类y_j的概率, $p(x_i,y_j)$为任意选择的样本点同时属于x_i和y_j的联合概率. 在实验中,一般使用标准互信息NMI如下:

$$\text{NMI}(X,Y)=\frac{\text{MI}(X,Y)}{\max(E(X),E(Y))}, \tag{28}$$

其中$E(X)$和$E(Y)$分别为类集X和Y的熵. NMI越大,表明算法性能越好.

NE[25]衡量类间的平衡程度,定义如下:

$$\text{NE}=-\frac{1}{\log c}\sum_{k=1}^{c}\frac{n_k}{n}\log\frac{n_k}{n}. \tag{29}$$

其中: c为聚类数目, n_k为第k类样本数目, n为数据集所有样本的数目. $\text{NE}=1$表明完全平衡的集群,越接近0表示类越不平衡.

3.3 实验结果

对于FCM_{LDE}算法,提供两种不同的模糊加权指数,即$m=1$和$m=2$,用$\text{FCM}_{\text{LDE}}{}^{m=1}$和$\text{FCM}_{\text{LDE}}{}^{m=2}$

分别表示. 在进行ALMs优化时,更新率ρ应略大于1,因此设置$\rho = 1.005, \mu = \{0.1, 0.01\}$,并且通过调整$\lambda$和$\gamma$找到适合于数据集的最佳参数,以及最佳的聚类结果;迭代停止条件参数设置最小阈值$e = 1\times100^{-10}$,最大迭代次数设为1 000次. 正则化参数λ可从$\{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}$中选择最优值;平衡参数$\gamma$从$\{10^{-5},10^{-4},10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$中选择最优值. 所有比较算法的聚类性能都与初始聚类中心有关,因此每个算法随机初始化10次,记录每次最佳参数下的运行结果,最终以最佳结果作为聚类结果. 实验结果如表2～表4所示,最佳的结果以粗体显示(除表4中的BKM).

表2　准确率ACC　　%

数据集	K-means	FCM	K-means++	ISODATA	BKM	$\text{FCM}_{\text{LDE}}{}^{m=1}$	$\text{FCM}_{\text{LDE}}{}^{m=2}$
COIL-20	59.58	62.99	61.74	45.42	74.38	65.42	**76.32**
Isolet 1	65.40	62.31	63.78	48.68	60.26	**66.47**	**68.01**
ORL_32×32	62.75	59.75	59.50	53.00	57.75	**66.25**	**67.00**
UMIST	48.68	52.11	50.79	46.58	**56.32**	53.16	56.05
PIE	66.04	58.32	63.64	55.06	69.81	59.09	**90.14**
AR	61.95	60.71	62.50	51.52	66.07	**75.36**	**76.43**

表3　标准互信息NMI　　%

数据集	K-means	FCM	K-means++	ISODATA	BKM	$\text{FCM}_{\text{LDE}}{}^{m=1}$	$\text{FCM}_{\text{LDE}}{}^{m=2}$
COIL-20	71.33	72.53	73.97	61.25	**82.13**	72.63	80.01
Isolet1	79.95	75.79	77.50	64.94	71.12	77.03	**80.01**
ORL_32×32	78.31	75.17	71.18	69.91	75.89	**78.35**	**79.97**
UMIST	63.56	64.40	61.85	60.78	**68.20**	65.17	68.09
PIE	86.42	81.85	81.66	73.46	88.12	79.87	**93.96**
AR	75.75	74.04	70.65	65.01	78.31	**81.98**	**83.96**

表4　标准熵NE　　%

数据集	K-means	FCM	K-means++	ISODATA	BKM	$\text{FCM}_{\text{LDE}}{}^{m=1}$	$\text{FCM}_{\text{LDE}}{}^{m=2}$
COIL-20	92.81	97.55	93.38	86.68	100.00	**99.54**	**99.63**
Isolet1	94.75	96.35	93.09	90.96	100.00	**97.11**	95.86
ORL_32×32	96.02	96.21	95.72	95.38	100.00	**99.49**	**99.39**
UMIST	94.79	98.35	93.95	97.20	100.00	**99.79**	98.31
PIE	96.70	96.88	96.20	94.58	100.00	**98.02**	**99.86**
AR	94.96	97.72	93.61	93.19	100.00	**98.33**	**99.55**

1) 从表2和表3中可以看出,对于准确率,BKM只在UMIST数据集达到最佳,然而$\text{FCM}_{\text{LDE}}{}^{m=1}$在3个数据集(Iolet 1、ORL_32×32和AR)和$\text{FCM}_{\text{LDE}}{}^{m=2}$在5个数据集(COIL-20、Isolet 1、ORL_32×32、PIE和AR)中都表现得更好. 对于标准互信息,BKM在COIL-20和UMIST数据集都达到最佳,$\text{FCM}_{\text{LDE}}{}^{m=1}$在两个数据集(ORL_32×32和AR)以及$\text{FCM}_{\text{LDE}}{}^{m=2}$在4个数据集(Isolet1、ORL_32×32、PIE和AR)中表现出最佳的效果. 由此可知,所提出的模型在聚类任务中是有效的.

2) 由表4可知,BKM算法是硬平衡聚类,因此所有数据集的平衡性能都达到了100%. 相较于其他传统聚类算法,$\text{FCM}_{\text{LDE}}{}^{m=1}$在所有数据集中平衡性能都是最佳的,而$\text{FCM}_{\text{LDE}}{}^{m=2}$也在4个数据集(COIL-20、ORL_32×32、PIE和AR)中表现出了最佳结果. 该算法在FCM模型基础上进行改进,也可以分析出引入标签分布熵有助于FCM模型更能均匀划分.

4　结　论

本文将FCM模型、平方损失项和标签分布熵相结合,提出了一种软平衡聚类算法FCM_{LDE}. 其特点是: 在应用于COIL-20、Isolet1、ORL_32×32、UMIST、PIE和AR这6个数据集时,与其他算法相比,该算法在聚类性能和平衡性能上均展现出良好优势. 但是,由于引入平方损失项和标签分布熵,并使用增广拉格朗日乘数优化算法,导致产生了新的参数选择. 实验过程中参数选择会直接影响到算法性能,因此如何合理选取算法中的参数,是进一步要解决的问题.

参考文献(References)

[1]　章永来, 周耀鉴. 聚类算法综述[J]. 计算机应用, 2019, 39(7): 1869-1882.
(Zhang Y L, Zhou Y J. Review of clustering algorithms[J]. Journal of Computer Applications, 2019,

39(7): 1869-1882.)

[2] Aggarwal C C, Reddy C K. Data clustering: Algorithms and applications[M]. London: Taylor and Francis Group, 2014: 4-7.

[3] Li H J, Bu Z, Wang Z, et al. Dynamical clustering in electronic commerce systems via optimization and leadership expansion[J]. IEEE Transactions on Industrial Informatics, 2020, 16(8): 5327-5334.

[4] Vijayarajeswari R, Nagabhushan M, Parthasarathy P. An enhanced symptom clustering with profile based prescription suggestion in biomedical application[J]. Journal of Medical Systems, 2019, 43(6): 1-6.

[5] 孙佳敏, 朱嘉富, 杨伏长, 等. 大规模生物网络马尔可夫聚类的并行化算法[J]. 计算机应用, 2019, 39(1): 66-71.
(Sun J M, Zhu J F, Yang F C, et al. Parallel algorithm of Markov clustering for large-scale biological networks[J]. Journal of Computer Applications, 2019, 39(1): 66-71.)

[6] 韦美峰, 王亚民. 基于后缀树聚类的主题搜索引擎研究[J]. 情报理论与实践, 2017, 40(12): 123-127.
(Wei M F, Wang Y M. Research on the focused search engine based on suffix tree clustering[J]. Information Studies: Theory & Application, 2017, 40(12): 123-127.)

[7] Rajpoot P, Dwivedi P. Optimized and load balanced clustering for wireless sensor networks to increase the lifetime of WSN using MADM approaches[J]. Wireless Networks, 2020, 26(1): 215-251.

[8] Patooghy A, Kamarei M, Farajzadeh A, et al. Load-balancing enhancement by a mobile data collector in wireless sensor networks[J]. International Journal on Smart Sensing and Intelligent Systems, 2020, 7(5): 1-5.

[9] Han J W, Liu H Y, Nie F P. A local and global discriminative framework and optimization for balanced clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(10): 3059-3071.

[10] 潘成胜, 张斌, 吕亚娜, 等. 改进灰狼优化算法的K-Means文本聚类[J]. 计算机工程与应用, 2021, 57(1): 188-193.
(Pan C S, Zhang B, Lyu Y N, et al. k-means text clustering based on improved gray wolf optimization algorithm[J]. Computer Engineering and Applications, 2021, 57(1): 188-193.)

[11] 杨华晖, 孟晨, 王成, 等. 基于目标特征选择和去除的改进K-means聚类算法[J]. 控制与决策, 2019, 34(6): 1219-1226.
(Yang H H, Meng C, Wang C, et al. Improved K-means clustering algorithm based on feature selection and removal on target point[J]. Control and Decision, 2019, 34(6): 1219-1226.)

[12] Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding[C]. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. New York: ACM, 2017: 1027-1035.

[13] Li M C, Han S, Shi J. An enhanced ISODATA algorithm for recognizing multiple electric appliances from the aggregated power consumption dataset[J]. Energy and Buildings, 2017, 140: 305-316.

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

[15] Wang L M, Li M Y, Han X M, et al. An improved density-based spatial clustering of application with noise[J]. International Journal of Computers and Applications, 2018, 40(3): 1-7.

[16] 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.
(Zhou J, Jiang Z B, Zhang Y P, et al. A density-based fuzzy exemplar clustering algorithm[J]. Control and Decision, 2020, 35(5): 1123-1133.)

[17] Johnson S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241-254.

[18] Malinen M I, Fränti P. Balanced K-means for clustering[M]. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2014: 32-41.

[19] Ugander J, Backstrom L. Balanced label propagation for partitioning massive graphs[C]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2013: 507-516.

[20] Janani R, Vijayarani S. Text document clustering using spectral clustering algorithm with particle swarm optimization[J]. Expert Systems With Applications, 2019, 134: 192-200.

[21] Bertsekas D P. Constrained optimization and Lagrange multiplier methods[M]. New York: Academic, 1982: 104-125.

[22] He X F. Locality preserving projections[J]. Advances in Neural Information Processing Systems, 2003, 16(1): 186-197.

[23] Cai D, He X, Han J. Document clustering using locality preserving indexing[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624-1637.

[24] He X F, Ji M, Zhang C Y, et al. A variance minimization criterion to feature selection using Laplacian regularization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 2013-2025.

[25] Zhong S, Ghosh J. Model-based clustering with soft balancing[C]. Proc of the 3rd IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2003: 459-466.

作者简介

王哲昀(1996−), 男, 硕士生, 从事机器学习与模式识别的研究, E-mail: wzy17888225402@163.com;

胡文军(1977−), 男, 教授, 博士, 从事数据挖掘、机器学习与模式识别等研究, E-mail: huwenjun@zjhu.edu.cn;

徐剑豪(1995−), 男, 硕士生, 从事机器学习与模式识别的研究, E-mail: xjhao603111366@163.com;

胡天杰(1997−), 男, 硕士生, 从事机器学习与模式识别的研究, E-mail: 15957598801@163.com.

(责任编辑：闫 妍)