doi:10.11835/j.issn.1000-582X.2023.05.011

协方差测距算法在多维聚类分析中的优化研究

刘　云,张　轶,郑文凤
(昆明理工大学 信息工程与自动化学院,昆明 650500)

摘要:为了在多维聚类分析中运用有效距离度量方法表征数据对象的邻近度,提出一种协方差测距(covariance distance measure analysis ,CDM)算法,首先,采用模糊C均值(fuzzy c-means ,FCM)方法对数据对象赋予权值,得到每个样本点相对类别特征的隶属度,再依据隶属度计算每个样本的差异度;其次,为了使类别分离最大化,用样本点同关联类别的协方差距离度量代替模糊聚类中欧式距离度量作为优化问题的第一个标准,使相似数据对象更为接近;最后,用样本点间的协方差距离度量作为第二个优化标准,使相异数据相互隔开,交替固定变量迭代计算最优解,使聚类指标和距离度量学习参数同时得到优化,获得更好的聚类结果。在不同数据集上的实验结果表明,与FCM-Sig和UNCA算法相比,CDM算法在聚类准确性和算法收敛性方面均有更好表现。

关键词:聚类分析;协方差测距;模糊C均值;距离度量学习
中图分类号:TP312　　**文献标志码:**A　　**文章编号:**1000-582X(2023)05-102-09

Optimization of covariance distance measurement algorithm for multidimensional clustering analysis

LIU Yun, *ZHANG Yi*, *ZHENG Wenfeng*
(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, P. R. China)

Abstract: In order to use effective distance measurement methods to characterize the proximity of data objects in multi-dimensional clustering analysis, a covariance distance measurement (CDM) algorithm is proposed. First, fuzzy C-means (FCM) is used to assign weights to the data objects, so that the membership degree of each sample point relative to the category feature is obtained. Based on the membership degree, the difference degree of each sample is calculated. Then, as the first optimization criterion, the variance distance measure is used to replace the Euclidean distance measure in fuzzy clustering to make similar data objects closer. Finally, the covariance distance measure between the sample points is used as the second optimization criterion to make the different data objects separate from each other. The optimal solution is calculated iteratively with alternate fixed variables, so that the clustering index and distance measurement learning parameters are optimized at the same time, and better clustering results are obtained. Experimental results on different data sets show that compared with FCM-Sig and UNCA algorithms, CDM algorithm has better performance in clustering accuracy and algorithm convergence.

Keywords: cluster analysis; covariance distance; fuzzy C-means (FCM); distance metric learning

收稿日期: 2022-06-09
基金项目:国家自然科学基金资助项目(61761025);云南省重大科技专项计划资助项目(202002AD080002)。
Supported by National Natural Science Foundation of China(61761025) and Major Science and Technology Project of Yunnan Province(202002AD080002).

作者简介:刘云(1973—),男,副教授,主要从事数据挖掘分析、人工智能方向研究,(E-mail)liuyun@kmust.edu.cn。

距离度量学习作为一种有效表征数据结构的方法被广泛应用于聚类分析,通过学习的距离度量构建学习模型。基于带标签实例的可用性存在有监督和无监督的距离度量学习算法,有监督距离度量学习需要带标签的训练数据集。而在实际应用中,由于缺少类别标签信息,无监督距离度量学习对于先验信息有限的问题更为重要[1-2]。

针对无监督距离度量学习在多维聚类分析中的研究取得很多成果,传统算法仅在聚类前将距离度量用作单独的数据预处理步骤。文献[3]提出一种新距离度量模糊C均值(new distance metric for fuzzy c-means, FCM-Sig)算法,通过新的距离度量标准结合群集中的距离变化以规范数据点和群集中心的距离。将其应用于常规模糊C均值(fuzzy c-means ,FCM)聚类和高维特征空间的内核模糊C均值(kernel fuzzy c-means, KFCM)聚类,子空间选择和聚类之间的固有分隔可能会影响聚类的可分离性。另一种方法将距离度量学习和聚类结合到联合框架中,文献[4]提出无监督邻域成分分析(unsupervised neighborhood component analysis , UNCA)算法,通过最大化未标记数据的遗忘K近邻(k-nearest neighbor, KNN)随机变量同时运用距离度量学习和聚类,而未很好考虑数据间的固有关联信息。

为了选取更有效的距离度量方法提高聚类质量,提出一种协方差测距算法(covariance distance measure analysis, CDM) 。首先,采用模糊C均值[5]聚类对数据对象赋予权值,得到类别特征的隶属度计算出每个样本的差异度;其次,依据样本点与类别特征的协方差距离代替模糊聚类中的欧式距离,作为优化问题的第一标准使相似样本之间距离最小;最后,将样本点间的协方差距离[6]作为约束条件得到第二个标准隔离不相似样本,计算优化问题最优解。仿真结果表明,对比FCM-Sig和UNCA算法,CDM算法在聚类精度和算法收敛性方面均有提升。

1　距离度量模型

聚类分析的研究重点是采用有效的距离度量方法分析数据对象之间的离散性或相异性信息,用于数据分类。

1.1　欧式距离的模糊聚类模型

模糊C均值(FCM)是一种模糊聚类算法,其中每个数据点都具有多个类别属性。假设X是输入数据,$\boldsymbol{C}=\left\{c_l \middle| c_l \in R^p\right\}_{l=1}^{k}$和$\boldsymbol{Q}=\left\{\boldsymbol{q}_{il} \middle| \boldsymbol{q}_{il} \in R\right\}$分别为集群中心的位置和模糊隶属度矩阵,$k$为聚类数,其中每个$\boldsymbol{q}_{il}$是$\boldsymbol{x}_i$在集群$\boldsymbol{l}$中的隶属度 。FCM的目标函数定义为

$$\min_{Q,C} \operatorname{imize} \sum_{i=1}^{n}\sum_{l=1}^{k} q_{il}^{u} \left\| \boldsymbol{x}_i - c_l \right\|^2 \text{subject to:} \begin{cases} 0 < \sum_{i=1}^{n} \boldsymbol{q}_{il} < n \forall l \\ \sum_{l=1}^{k} \boldsymbol{q}_{il} = 1 \forall i \end{cases}, \tag{1}$$

u是模糊程度,$\left\| \boldsymbol{x}_i - c_j \right\|^2$为欧氏距离[7],表征数据点之间的离散程度。公式(1)是具有双凸目标函数的非凸优化问题,该问题可通过交替优化方案的方式解决。固定$\boldsymbol{C}$时,其相对于$\boldsymbol{Q}$凸,固定$\boldsymbol{Q}$时,其相对于$\boldsymbol{C}$凸。首先认为$\boldsymbol{C}$是固定的,在另一个参数上优化问题,然后对$\boldsymbol{Q}$重复此过程直到实现收敛,更新公式为

$$\begin{cases} \boldsymbol{q}_{il'} = 1 \Big/ \sum_{l=1}^{k} \left(\left\| \boldsymbol{x}_i - c_{l'} \right\|^2 \Big/ \left\| \boldsymbol{x}_i - c_l \right\|^2 \right)^{1/(u-1)} \\ c_l = \sum_{i=1}^{n} \boldsymbol{q}_{il}^{u} x_i \Big/ \sum_{i=1}^{n} \boldsymbol{q}_{il}^{u} \end{cases}, \tag{2}$$

FCM的时间复杂度为$O\left(\Gamma npk^2\right)$,$\Gamma$是迭代次数。

1.2 分类距离度量模型

聚类分析模型使用的所有数值和分类距离度量表示为

$$\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\begin{bmatrix}\boldsymbol{d}^{pp}\left(\boldsymbol{x}_{i.1},\boldsymbol{x}_{j.1}\right),...,\boldsymbol{d}^{pp}\left(\boldsymbol{x}_{i.r},\boldsymbol{x}_{j.r}\right)\\ \boldsymbol{d}^{pp}\left(\boldsymbol{x}_{i,r+1},\boldsymbol{x}_{j,r+1}\right),...,\boldsymbol{d}^{pp}\left(\boldsymbol{x}_{i,r+s},\boldsymbol{x}_{j,r+s}\right)\end{bmatrix}^{\mathrm{T}},$$

其中：$\boldsymbol{x}_i$和$\boldsymbol{x}_j$是2个数据点之间的点对点距离矢量；r和s分别是类别和数值属性的数量。

对于$f\in\{1,..,r\}$,$\boldsymbol{d}^{\mathrm{pp}}_{\mathrm{ORLP}}$和$\boldsymbol{d}^{\mathrm{pp}}_{\mathrm{FSK}}$分别称为重叠距离与ESK距离[8],定义如下

$$\boldsymbol{d}^{\mathrm{pp}}_{\mathrm{ORLP}}\left(\boldsymbol{x}_{i,f},\boldsymbol{x}_{j,f}\right)=\begin{cases}0,if\quad \boldsymbol{x}_{i,f}=\boldsymbol{x}_{j,f},\\ 1,if\quad \boldsymbol{x}_{i,f}\neq\boldsymbol{x}_{j,f}。\end{cases}\tag{3}$$

$$\boldsymbol{d}^{\mathrm{pp}}_{\mathrm{ESK}}\left(\boldsymbol{x}_{i,f},\boldsymbol{x}_{j,f}\right)=\begin{cases}0,if\quad \boldsymbol{x}_{i,f}=\boldsymbol{x}_{j,f},\\ 1-m_{f^2}\big/\left(m_{f^2}+2\right),\mathrm{if}\quad \boldsymbol{x}_{i,f}\neq\boldsymbol{x}_{j,f}。\end{cases}\tag{4}$$

m_f是第f个属性采用不同值的数值属性,任何数值属性$f\in\{r+1,..,r+s\}$的距离定义为

$$\boldsymbol{d}^{pp}\left(\boldsymbol{x}_{i,f},\boldsymbol{x}_{j,f}\right)=\boldsymbol{x}_{i,f}-\boldsymbol{x}_{j,f}.\ ,\tag{5}$$

每个关联点$\boldsymbol{x}_i\in X$到集群$c_l\in C$的距离表示为

$$\vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\begin{bmatrix}\boldsymbol{d}^{pc}\left(x_{i.1},c_{l.1}\right),...,\boldsymbol{d}^{pc}\left(\boldsymbol{x}_{i.r},c_{l.r}\right)\\ \boldsymbol{d}^{pc}\left(x_{i,r+1},c_{l\ ,r+1}\right),...,\boldsymbol{d}^{pc}\left(\boldsymbol{x}_{i,r+s},c_{l,r+s}\right)\end{bmatrix}^{\mathrm{T}},$$

相对$f\in\{1,..,r\}$,除重叠距离和ESK距离外,另一种距离度量计算为

$$\boldsymbol{d}^{\mathrm{pc}}_{\mathrm{Cheung}}\left(\boldsymbol{x}_{i,f},c_{l,f}\right)=1-\delta_{Af}=\boldsymbol{x}_{i,f}\left(c_{l,f}\right)\big/\delta_{Af\neq null}\left(c_{l,f}\right),\tag{6}$$

由$\delta_{Af=x_{i,j}}\left(c_{l,f}\right)$计算集群中$l$的数据点数,其中属性$A_f$的值为$\boldsymbol{x}_{i,j}$,与$\delta_{Af\neq null}\left(c_{l,f}\right)$的含义相同,但属性$A_f$不为null。此度量只能选取点到集群的距离,而不能为点到点的距离。因此,分别定义了2个不同的向量$\vec{\boldsymbol{d}}^{pc}$和$\vec{\boldsymbol{d}}^{pp}$,数值属性的距离度量记为

$$\vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_{i,f},c_{l,f}\right)=\boldsymbol{x}_{i,f}-c_{l,f}\forall f\in\{r+1,..,r+s\}。\tag{7}$$

1.3 马氏距离(协方差测距)模型

针对多维特征空间的聚类分析,数据属性间的关联关系需要采用有效的距离度量方法来表征。假设$X=\left\{x_i\middle|\ x_i\in R^P\right\}_{i=1}^{n}$是输入数据,其中：$x_i\in R^P$是第$i$个的数据点；$p$为属性数；$n$是数据点的数量。$S$和$D$分别代表相似和不相似的集合记为

$$S=\left\{\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\middle|\ \boldsymbol{x}_i,\boldsymbol{x}_j\ \text{belong to the same class}\right\},\tag{8}$$

$$D=\left\{\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\middle|\ \boldsymbol{x}_i,\boldsymbol{x}_j\ \text{belong to the different classes}\right\}。\tag{9}$$

距离度量的最大化可分离性定义假设$\boldsymbol{x}_i$和$\boldsymbol{x}_j$属于S,则它们应彼此靠近,属于D则它们应彼此分开。在线性方法中,通过学习线性变换并将数据投影到新空间中$\boldsymbol{L}:\boldsymbol{x}_i\leftarrow\boldsymbol{x}_j$。投影空间中的马氏距离(协方差测距)记为

$$\begin{aligned}\boldsymbol{d}_M\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)&=\left\|\boldsymbol{L}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\right\|_2^2=\\ &\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)=\\ &\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)^{\mathrm{T}}\boldsymbol{M}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right),\end{aligned}\tag{10}$$

$\boldsymbol{d}_M\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$为马氏距离,其本质是协方差测距[6],$\boldsymbol{M}$是一个半正定矩阵,去掉协方差矩阵,马氏距离就退化为欧式距离。对比欧式距离,马氏距离(协方差测距)是一种表征属性之间关联性且尺度无关的无监督度量学习方法,如图1所示。

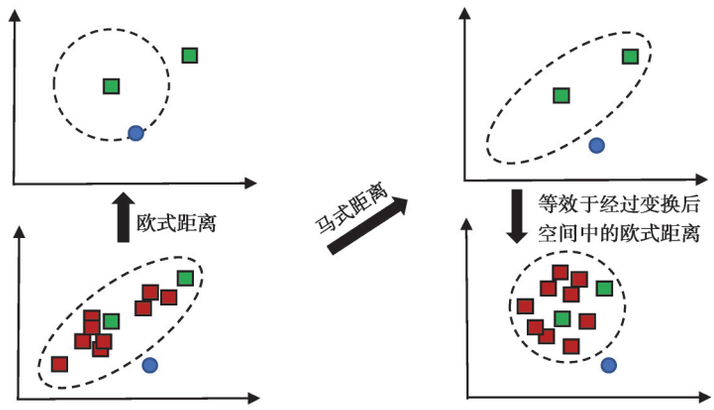

图 1　欧式距离与马氏距离度量

Fig. 1　Euclidean distance and Mahalanobis distance

观察图 1 中 2 个绿色点以及中间绿色点到蓝色点的距离。如果不考虑数据分布,则蓝色距离更近,这就是欧式距离度量。但实际需要考虑数据分布的影响,数据样本呈椭圆形分布,蓝色点在椭圆外,绿色在椭圆内,因此 2 个绿色点更为接近。马氏距离(协方差测距)度量可以有效的表征数据对象之间的邻近度[9],而欧式距离度量得到的是数据之间的离散度,不利于对多维数据之间的关联性进行分析。

2　CDM 算法

2.1　协方差测距的模糊聚类算法

运用协方差测距,模糊聚类算法获得相似集合 S 和不相似集合 D 的估计值。相同聚类的数据属于 S,不同聚类数据属于 D。$X=\left\{ \boldsymbol{x}_i \middle| \boldsymbol{x}_i \in R^P \right\}_{i=1}^{n}$ 是输入数据的集合,$(\boldsymbol{L}:\boldsymbol{x}_i \leftarrow \boldsymbol{L}\boldsymbol{x}_j)$ 为 $\boldsymbol{L}$ 的线性变换。

传统的 FCM 没有清晰地定义数据的邻近度,而是提供了模糊形式。$\boldsymbol{Q}=\left\{ \boldsymbol{q}_{il} \middle| \boldsymbol{q}_{il} \in R \right\}$ 是模糊隶属度矩阵,$\boldsymbol{C}=\left\{ c_l \middle| c_l \in R^P \right\}_{l=1}^{k}$ 是集群中心。$\tilde{\boldsymbol{d}}_{ij} \in R$ 是数据点 i 和 j 的差异度,通过模糊隶属度值计算为

$$\tilde{\boldsymbol{d}}_{ij}=1-\boldsymbol{q}_i\boldsymbol{q}_j^{\mathrm{T}} \Big/ \left(\left\| \boldsymbol{q}_i \right\|_2 \times \left\| \boldsymbol{q}_j \right\|_2 \right), \tag{11}$$

其中:$\boldsymbol{q}_i$ 和 $\boldsymbol{q}_j$ 是矩阵 $\boldsymbol{Q}$ 的第 i 行和第 j 行,基于第 i 个数据点和第 j 个数据点之间的相似度 A,由隶属度是否相似来定量估计。如果数据属于不同的集群,则所有群集($\boldsymbol{q}_i$ 和 $\boldsymbol{q}_j$)的值不同。$\boldsymbol{q}_i$ 和 $\boldsymbol{q}_j$ 的关联条目对($\boldsymbol{q}_{il}$ 和 $\boldsymbol{q}_{jl}$)不同,至少一个接近于零。因此,它们的差异度 $\tilde{\boldsymbol{d}}_{ij}$ 接近于 1。

为了使类分离最大化,运用协方差测距代替欧式距离改进模糊聚类的优化问题[10]。该期望度量的一个标准是所有相似数据点之间距离最小。在同一群集 S 中的所有数据(即群集 $\boldsymbol{l}$ 中的 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$)都被视为配对,以群集中心 c_l 的方向进行迁移。由于 S 是一个模糊的相似集,该约束应与赋予群集 $\boldsymbol{l}$ 中 $\boldsymbol{x}_i$ 的隶属度 $\boldsymbol{q}_{il}$ 成比例满足。为了实现这一目标,用协方差测距(10)代替 FCM 目标函数(1)中的欧氏距离。第一个标准被表示为损失函数,表示为

$$\underset{\boldsymbol{Q},\boldsymbol{C},\boldsymbol{M}}{\min\operatorname{imize}} \sum_{i=1}^{n}\sum_{l=1}^{k} \boldsymbol{q}_{il}^{u} \vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{c}_l \right)^{\mathrm{T}} \boldsymbol{M} \vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{c}_l \right) 。 \tag{12}$$

第二个标准通过添加约束条件使 D 中不相似的数据点相互隔开

$$\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j \right)^{\mathrm{T}} \boldsymbol{M} \vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,c_l \right) \geqslant \varepsilon \forall \boldsymbol{x}_i,\boldsymbol{x}_j \in D 。 \tag{13}$$

ε 是一个大于零的常数,因 D 是模糊的不相似关系,所以这个新添加的约束必须与 $\tilde{\boldsymbol{d}}_{ij}$ 成比例地满足。引入与此约束相关的松弛变量 $\zeta_{ij} \geqslant 0$ 来衡量其违反量。第二个标准由所有 ζ_{ij} 的总和定义,其中每个 ζ_{ij} 乘以 $\tilde{\boldsymbol{d}}_{ij}$,问题的解决方法如式(14)。

$$\min\limits_{Q,C,M,\zeta}\imize(1-\alpha)\sum_{i=1}^{n}\sum_{l=1}^{k}\boldsymbol{q}_{il}^{u}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)+$$

$$\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\boldsymbol{d}}_{ij}\zeta_{ij}\ \text{subject}\quad\text{to:}$$

$$\begin{cases}\vec{\boldsymbol{d}}^{pp}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pp}(x_i,x_j)\geqslant\varepsilon-\zeta_{ij},\forall i,j\in\{1,..,n\},\\ \zeta_{ij}\geqslant 0,\\ \boldsymbol{M}\geqslant 0,\\ 0<\sum\limits_{i=1}^{n}\boldsymbol{q}_{il}<n,\forall l\in\{1,...,k\},\\ \sum\limits_{l=1}^{k}\boldsymbol{q}_{il}=1,\qquad\forall i\in\{1,...,n\},\end{cases}\tag{14}$$

式(14)中:α是2个项之间的折衷参数,是大于零的常数;$\boldsymbol{M}$是半正定矩阵且$\boldsymbol{M}\geqslant 0$。优化问题中,如果$\tilde{d}_{ij}=0$,则对应项为0。另一方面,如果$\tilde{\boldsymbol{d}}_{ij}=1$,$\zeta_{ij}$的值是对第二项的贡献。

FCM的运用与K均值(k-means)不同,K均值方法提供了清晰的标签信息,通过学习聚类指标来调整转换矩阵以适应聚类指标。在第二次迭代中,新的聚类指标将保持与前次相同,因此该方法存在局部优化,无法在更新迭代中学习新的转换矩阵,造成快速收敛到局部最优的问题[11]。FCM中数据点并不完全相似或相异,根据模糊隶属度值$\boldsymbol{q}_{il}$和差异度$\tilde{\boldsymbol{d}}_{ij}$得到目标函数(14)中2个准则的满意程度,避免了收敛到局部最优[12]。

2.2 CDM算法实现

CDM算法优化公式(14)不是凸面的,这使得寻找最优解变得困难。通过固定某一变量,则它在每个变量中都是凸的,可以迭代计算最优解[13]。

1)固定$\boldsymbol{M},\zeta,\boldsymbol{Q}$并更新$\boldsymbol{C}$

当固定除$\boldsymbol{C}$外的所有参数时,目标函数(14)的第二项将变为常数,对参数$\boldsymbol{C}$没有任何约束。对于数值属性,可以通过式(2)计算更新每个聚类中心与原始聚类中心。为了更新分类属性的聚类中心$\boldsymbol{c}_l$,采用

定理1 模糊k模式更新方法[14]:由分类属性$A_1,A_2,...,A_r$和$\mathrm{Domain}(A_f)=\{a_f^{(1)},a_f^{(2)},...,a_f^{(n_f)}\}$定义分类对象$X^c=\{x_i^c\}_{i=1}^{n}$,$n_f$是属性$A_f$由$1\leqslant f\leqslant r$的类别数。聚类中心$\boldsymbol{c}_l$由$1\leqslant l\leqslant k$的$[c_{l,1},c_{l,2},...,c_{l,r}]$表示,当$c_{l,f}=a_f^{(t')}\in\mathrm{Domain}(A_f)$时,最小值为$\sum_{i=1}^{n}\sum_{l=1}^{k}\boldsymbol{q}_{il}^{u}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)^{\mathrm{T}}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)$,记为

$$\sum_{i=1,x_{i,f}=a_f^{(t')}}^{n}\boldsymbol{q}_{il}^{u}\geqslant\sum_{i=1,x_{i,f}=a_f^{(t)}}^{n},\qquad\boldsymbol{q}_{il}^{u},1\leqslant t\leqslant n_f,[1\leqslant f\leqslant r]。\tag{15}$$

根据该定理,分类属性的聚类中心c_l中每个属性类别均由所求总和最大的类别给出,从而对所有类别进行聚类。

2)固定M,ζ,C并更新$\boldsymbol{Q}$

当除$\boldsymbol{Q}$以外的参数固定,可以获得问题(14)的最佳$\boldsymbol{Q}$。在这种情况下,目标函数(14)的第二项变为常数,$\boldsymbol{Q}$的最值优通过对优化问题(14)中第一项求导得出。每个$\boldsymbol{q}_{i,l'}$的最优值定义为

$$\boldsymbol{q}_{il'}=1\Bigg/\sum_{l=1}^{k}\left(\frac{\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_{l'})^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_{l'})}{\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)}\right)^{\frac{1}{u-1}}。\tag{16}$$

目标函数(14)的第二项中$\vec{\boldsymbol{d}}_{i,j}$的定义取决于$q_{il'}$,参数在上一次迭代设置,在此步骤中值视为常数。

3)固定$\boldsymbol{Q},\boldsymbol{C}$并更新$\boldsymbol{M},\zeta$

当固定集群成员矩阵$\boldsymbol{Q}$和集群中心矩阵$\boldsymbol{C}$,通过解决以下优化问题来计算最佳$\boldsymbol{M}$和ζ

$$\min\limits_{C,M}\imize(1-\alpha)\sum_{i=1}^{n}\sum_{l=1}^{k}\boldsymbol{q}_{il}^{u}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pc}(\boldsymbol{x}_i,\boldsymbol{c}_l)$$

$$+\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\boldsymbol{d}}_{ij}\zeta_{ij}\ \text{subject}\quad \text{to:} \tag{17}$$

$$\begin{cases}\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)^{\mathrm{T}}\boldsymbol{M}\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\geqslant\varepsilon-\zeta_{ij},\forall i,j\in\left\{1,..,n\right\},\\ \zeta_{ij}\geqslant 0,\\ \boldsymbol{M}\geqslant 0。\end{cases}$$

优化公式定义新的松弛变量$\zeta_{i,j}$将线性不等式约束转换为线性等式约束,该优化公式为半正定规划问题(SDP)[15],可通过现有的在线程序包求解。

2.3　CDM 算法分析

基于以上分析,迭代算法解决优化问题(14)的伪代码如下所示。在此方法的每个步骤中,所有变量均根据其相应公式进行更新。此过程迭代进行,直到变量收敛为止。

CDM 算法代码如下

输入:混合数据 X,α,ε。

输出:$\boldsymbol{Q},\boldsymbol{C},\boldsymbol{M},\zeta$。

Step1:每个 $\boldsymbol{q}_{il}$ 和 ζ_{ij} 初始化 $\boldsymbol{Q},\boldsymbol{C},\zeta$,满足(14)中的约束 2,4 和 5,

Step2:设置 $\boldsymbol{M}\leftarrow I.$ 由式(13)计算 $\tilde{\boldsymbol{d}}_{ij},\forall i,j\cdot$,

Step3:迭代:

1)固定 $\boldsymbol{M},\zeta,\boldsymbol{Q},\tilde{\boldsymbol{d}}_{ij}$ 由式(6)更新 $\boldsymbol{C}$,

2)固定 $\boldsymbol{M},\zeta,\boldsymbol{C},\tilde{\boldsymbol{d}}_{ij}$ 由式(16)更新 $\boldsymbol{Q}$,然后由(11)设置 $\tilde{\boldsymbol{d}}_{ij},\forall i,j$,

3)固定 $\boldsymbol{Q},\boldsymbol{C},\tilde{\boldsymbol{d}}_{ij}$ 根据优化目标函数(17)来更新 $\boldsymbol{M},\zeta$,

4)首先,使用 $\boldsymbol{M}$ 的 Cholesky 分解,当 $\boldsymbol{M}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ 时设置 $\boldsymbol{L}$,接着,通过 $\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\leftarrow\boldsymbol{L}\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$ 来更新 $\vec{\boldsymbol{d}}^{pp}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$。$\vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{c}_l\right)$ 由 $\vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{c}_l\right)\leftarrow\boldsymbol{L}\vec{\boldsymbol{d}}^{pc}\left(\boldsymbol{x}_i,\boldsymbol{c}_l\right)$ 更新,

5) $\boldsymbol{M}\leftarrow I$。

Step4:直到收敛。

在步骤 4)中,定义了额外变量 $\boldsymbol{L}$ 来计算新的点对点和点对集群的距离矢量。在每次迭代中学习到新的 $\boldsymbol{M}$ 之后,将更新这些距离向量,并在这些新向量的基础上继续执行该过程。

3　仿真分析

3.1　仿真环境和方法

为了验证所提出的算法,多维聚类分析常选用 UCI machine learning repositor 中的 Iris,Wine 和 Breast Tissue3 个数据集作为基准[16],另选用真实数据集 Mechanical Analysis 评估 CDM 算法解决实际问题的能力。Mechanical Analysis 是一个多变量工业数据集,任务是基于机械组件的属性信息,预测机电设备的故障诊断情况[17-19]。表 1 为仿真数据集信息,实验使用 SDP 在线程序包进行,SDP 为求解半正定规划问题的 MATLAB 程序包。运行环境为 Windows10 ,2.6GHzCPU,8 G 内存。

表 1　数据集信息

Tab. 1　Data set information

数据集	实例数	属性数	类别数
Breast Tissue	106	9	6
Wine	178	13	3
Iris	150	4	3
Mechanical Analysis	209	8	19

仿真实验将每个数据点的预测标记与其真实标记进行比较来评估聚类结果的准确性。“聚类精度”和“标准化互信息(NMI)”被用作比较不同算法的 2 个仿真指标[4]。

“聚类精度”通过对正确分配的数据点总数进行计数,并除以所有数据的数量来计算分配的准确性

$$\text{Accuracy}=\left(\sum_{i=1}^{n}\delta\left(y_i,\text{map}\left(x_i\right)\right)\right)\Big/n\times 100\,, \tag{18}$$

其中:n是数据点的数量;y_i是正确的标签;map是一个函数;当$s=t$将标签分配给x_i其群集和增量功能的多数标签$\delta(s,t)=1$,否则为0。

NMI的计算如下

$$\text{NMI}(Y,I)=200\times\frac{\sum_{y_i\in Y,\text{clust}_j\in I}p\left(y_i,\text{clust}_j\right)\cdot\log\dfrac{p\left(y_i,\text{clust}_j\right)}{p\left(y_i\right)p\left(\text{clust}_j\right)}}{H(Y)+H(I)}, \tag{19}$$

其中:Y和I是真实标签和聚类指标的集合;$P\left(y_i\right)$和$p\left(\text{clust}_j\right)$,$p\left(y_i,\text{clust}_j\right)$分别是概率随机分布在$y_i$,$\text{clust}_j$类和$y_i$与$\text{clust}_j$交集之间的概率;函数$H(Y)$和$H(I)$分别是$Y$和$I$的熵。

3.2 聚类数量对聚类精度的影响分析

为了研究聚类数对聚类结果的影响,在Breast Tissue(BT), Wine和Mechanical Analysis数据集中评估CDM算法与FCM-Sig和UNCA算法准确性结果。聚类数设置为与类数相等,并累加到类数的四倍,仿真结果如图2所示。

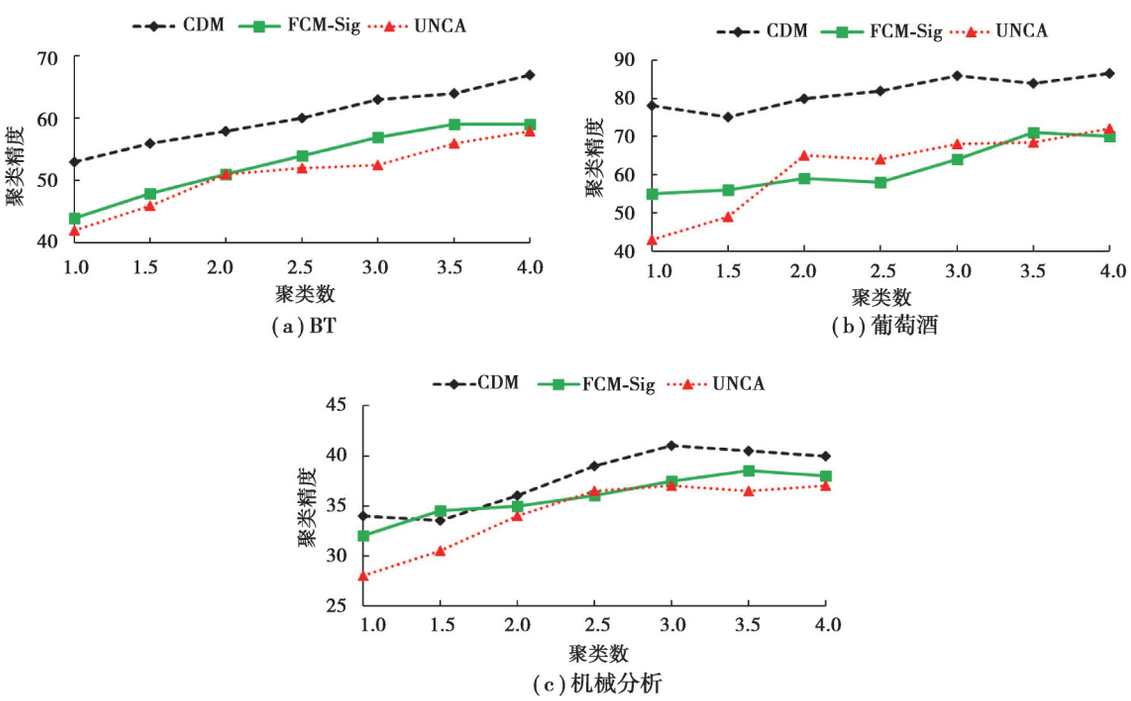

图2 聚类数对聚类精度影响分析

Fig. 2 Variation of clustering accuracy with the number of clusters

图2显示了算法在3个数据集中的性能,3种算法的聚类精度都随聚类数的增加逐渐提高。BT数据集中CDM算法的聚类精度明显优于对比算法。葡萄酒数据集中CDM算法的聚类精度保持在80%上下波动对比FCM-Sig和UNCA算法,能够保持平稳高精度的聚类性能。在机械分析数据集下,因多维数据的复杂分布使聚类精度有所降低,但CDM仍能保持更高的聚类精度,FCM-Sig次之,UNCA最低。

3.3 迭代次数对算法准确性和收敛性分析

为了进一步评估CDM算法的聚类性能,将目标函数,NMI和聚类精度作为迭代函数,在3个数据集中进行仿真分析如图3所示。

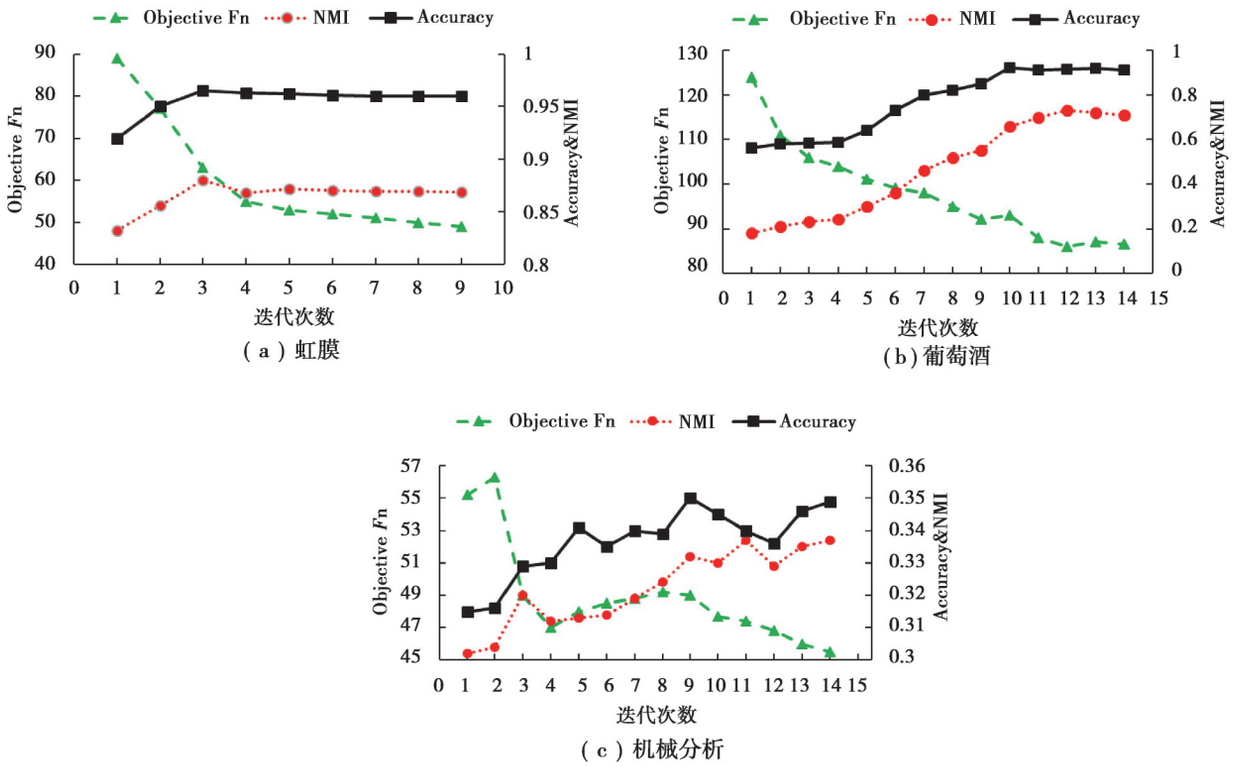

图 3　聚类精度,NMI和目标函数的迭代分析

Fig.3　Clustering accuracy, NMI, and the value of objective function in terms of iterations

图 3 显示了 CDM 算法在不同数据集中的性能,仿真每次迭代都对应着参数 $\boldsymbol{Q}$,$\boldsymbol{C}$,$\boldsymbol{M}$ 的更新。在 3 个数据集实验中,随着迭代次数的增加,总的趋势是 CDM 算法的目标函数值逐渐降低,而聚类精度和 NMI 值逐步提高。在机械分析数据集下目标函数值在迭代过程中有提升,这是由于为了更新 $\boldsymbol{Q}$,公式(14)中第二项约束被忽略,随着 $\boldsymbol{Q}$ 的迭代更新目标函数在收敛过程中有所起伏,但最后仍能达到收敛。结合 3 次仿真,针对不同的数据集,CDM 算法均能在有效提升 NMI 与聚类精度的同时保证算法的收敛性。

4　结　论

为了在聚类分析中选取有效的距离度量表征数据间的关联信息,提出了一种协方差测距算法(CDM)。首先,由模糊 C 均值聚类得到数据类别特征的隶属度,并计算出每个样本的差异度;其次,采用协方差测距代替模糊聚类中的欧式距离作为第一个优化标准使相似样本之间距离最小;最后,将样本点间的协方差测距作为第二个优化标准使不相似样本距离最大,交替固定变量迭代计算最优解。仿真结果表明,对比 FCM-Sig 和 UNCA 算法,CDM 算法在聚类精确性和算法收敛性方面均有提升。下一步,面对更为复杂的数据结构和分析需求,将研究更有效的距离度量方法。

参考文献

[1] Ahmed I, Dagnino A, Ding Y. Unsupervised anomaly detection based on minimum spanning tree approximated distance measures and its application to hydropower turbines[J]. IEEE Transactions on Automation Science and Engineering, 2019, 16 (2): 654-667.

[2] Zhu X B, Pedrycz W, Li Z W. Fuzzy clustering with nonlinearly transformed data[J]. Applied Soft Computing, 2017, 61: 364-376.

[3] Wei L Y. A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting[J]. Applied Soft Computing, 2016, 42: 368-376.

[4] Qin C, Song S J, Huang G, et al. Unsupervised neighborhood component analysis for clustering[J]. Neurocomputing, 2015,

168: 609-617.

[5] 李鹏华, 刘晶晶, 冯辉宗, 等 . 改进测度下的模糊C均值三元催化器故障诊断方法[J]. 重庆大学学报, 2018, 41(1): 88-98.
Li P H, Liu J J, Feng H Z, et al. Fault diagnosis of three-way catalytic converter using improved fuzzy C-means clustering[J]. Journal of Chongqing University, 2018, 41(1): 88-98.(in Chinese)

[6] Li P H, Liu J, Feng H, et al. Fault diagnosis of three-way catalytic converter using improved fuzzy C-means clustering[J]. Chongqing Daxue Xuebao/Journal of Chongqing University, 2018, 41(1): 88-98.

[7] Bai Z X, Zhang X L, Chen J D. Speaker verification by partial AUC optimization with mahalanobis distance metric learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1533-1548.

[8] Cardarilli G C, Di Nunzio L, Fazzolari R, et al. N-dimensional approximation of euclidean distance[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020, 67(3): 565-569.

[9] Xue B, Zhang L H, Yu Y, et al. Locating the nodes from incomplete euclidean distance matrix using Bayesian learning[J]. IEEE Access, 2019, 7: 37406-37413.

[10] Chang Z P, Chen W H, Gu Y P, et al. Mahalanobis-taguchi system for symbolic interval data based on kernel mahalanobis distance[J]. IEEE Access, 2020, 8: 20428-20438.

[11] dos Santos T R L, Zárate L E. Categorical data clustering: what similarity measure to recommend?[J]. Expert Systems With Applications, 2015, 42(3): 1247-1260.

[12] Hou C P, Nie F P, Yi D Y, et al. Discriminative embedded clustering: a framework for grouping high-dimensional data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(6): 1287-1299.

[13] 刘洲洲, 李彬 . 基于动态多子族群自适应群居蜘蛛优化算法[J]. 四川大学学报(自然科学版), 2017, 54(4): 721-727.
Liu Z Z, Li B. An adaptation social spider optimization algorithm based on dynamic multi-swarm strategy[J]. Journal of Sichuan University (Natural Science Edition), 2017, 54(4): 721-727.(in Chinese)

[14] Sun Y N, Yen G G, 0001 Z Y. IGD indicator-based evolutionary algorithm for many-objective optimization problems[J]. IEEE Trans Evolutionary Computation, 2019, 23(2): 173-187.

[15] Zhao X W, Liang J Y, Dang C Y. Clustering ensemble selection for categorical data based on internal validity indices[J]. Pattern Recognition, 2017, 69: 150-168.

[16] Pakazad S K, Hansson A, Andersen M S, et al. Distributed semidefinite programming with application to large-scale system analysis[J]. IEEE Transactions on Automatic Control, 2018, 63(4): 1045-1058.

[17] BacheK, LichmanM. UCI machine learning repository,2013,[EB/OL],Available:http://archive.ics.uci.edu/ml

[18] Li P H, Zhang Z J, Xiong Q Y, et al. State-of-health estimation and remaining useful life prediction for the lithium-ion battery based on a variant long short term memory neural network[J]. Journal of Power Sources, 2020, 459(C): 228069.

[19] 余萍, 曹洁 . 深度学习在故障诊断与预测中的应用[J]. 计算机工程与应用, 2020, 56(3): 1-18.
Yu P, Cao J. Deep learning approach and its application in fault diagnosis and prognosis[J]. Computer Engineering and Applications, 2020, 56(3): 1-18.(in Chinese)

（编辑 侯 湘）