文章编号：1008-1542(2018)02-0159-07　　doi：10.7535/hbkd.2018yx02010

基于马氏距离的模糊聚类优化算法——KM-FCM

祖志文，李　秦

（兰州交通大学数理学院，甘肃兰州　730070）

摘　要：为了解决以欧氏距离作为相似性准则的传统模糊聚类算法对多维数据处理不利的问题，采用马氏距离代替欧氏距离，对基于马氏距离的模糊聚类算法进行优化研究，以增强基于马氏距离的模糊聚类算法的聚类效果和能力。通过构造启发式搜索与 k-means 算法结合的初始优化方法，利用可以自动调节最佳聚类数的有效性函数，提出了一种优化算法 KM-FCM，并将此新算法与 FCM，FCM-M，M-FCM 聚类算法在 3 个标准数据集上进行了实验。结果表明，KM-FCM 算法有效，聚类精度比 FCM，FCM-M，M-FCM 高，对高维数据聚类识别能力强，具有全局优化作用，并且聚类个数无需提前设定。新算法可为基于马氏距离的模糊聚类算法的优化提供参考。

关键词：算法理论；模糊聚类；马氏距离；初始优化；聚类个数

中图分类号：TP301.6　　**文献标志码**：A

KM-FCM：A fuzzy clustering optimization algorithm based on Mahalanobis distance

ZU Zhiwen，LI Qin

（College of Mathematics and Physics，Lanzhou Jiaotong University，Lanzhou，Gansu 730070，China）

Abstract：The traditional fuzzy clustering algorithm uses Euclidean distance as the similarity criterion，which is disadvantageous to the multidimensional data processing. In order to solve this situation，Mahalanobis distance is used instead of the traditional Euclidean distance，and the optimization of fuzzy clustering algorithm based on Mahalanobis distance is studied to enhance the clustering effect and ability. With making the initialization means by Heuristic search algorithm combined with k-means algorithm，and in terms of the validity function which could automatically adjust the optimal clustering number，an optimization algorithm KM-FCM is proposed. The new algorithm is compared with FCM algorithm，FCM-M algorithm and M-FCM algorithm in three standard data sets. The experimental results show that the KM-FCM algorithm is effective. It has higher clustering accuracy than FCM，FCM-M and M-FCM，recognizing high-dimensional data clustering well. It has global optimization effect，and the clustering number has no need for setting in advance. The new algorithm provides a reference for the optimization of fuzzy clustering algorithm based on Mahalanobis distance.

Keywords：algorithm theory；fuzzy clustering；Mahalanobis distance；initial optimization；clustering number

收稿日期：2017-12-24；修回日期：2018-02-20；责任编辑：张　军

基金项目：国家自然科学基金(11262009)

第一作者简介：祖志文(1993—)，女，河北保定人，硕士研究生，主要从事智能算法方面的研究。

通信作者：李　秦教授。E-mail：liq@mail.lzjtu.cn

祖志文，李秦.基于马氏距离的模糊聚类优化算法——KM-FCM[J].河北科技大学学报，2018，39(2)：159-165.

ZU Zhiwen，LI Qin. KM-FCM：A fuzzy clustering optimization algorithm based on Mahalanobis distance[J]. Journal of Hebei University of Science and Technology，2018，39(2)：159-165.

聚类是数据处理的重要方法。模糊聚类建立了数据样本对于类别的不确定性的描述,表达了样本类属的模糊性,能够更客观地反映现实世界,具有较强的聚类效果与数据表达能力。

关于模糊聚类的研究应用最广的是基于目标函数的模糊聚类方法,此方法把聚类问题描述为一个带约束的优化问题,通过求解优化问题的解来确定数据集的模糊划分和聚类结果。FCM算法是最经典的基于目标函数的模糊聚类算法,但也存在一些问题,因此基于FCM算法的改进算法和应用被研究者所关注。蔡静颖[1]为改进FCM算法不能处理非球形簇且未考虑样本矢量各特征重要程度、处理高相关数据集时错误率增加的缺点,使用马氏距离得到FCM-M算法及MF-FCM算法;蔡威[2]提出自适应距离度量的GK算法。NATACHA等[3]研究了用马氏距离和闵可夫斯基距离来取代欧氏距离的模糊聚类的方法,以提高聚类检测能力,并对聚类结果进行了可视化分析。在应用方面,张敏等[4]将马氏距离和模糊 c-均值聚类结合,研究抠图算法;康韦晓[5]将基于马氏距离的PFCM算法应用于非线性系统故障诊断;赵泉华等[6]将马氏距离的模糊聚类算法用于遥感图像分割。另外,通过对FCM的初始化方法的改进研究逐渐形成了山峰函数法或势函数法[7]、减法聚类[8]等方法,进一步实现了快速减法聚类的模糊聚类算法,及基于密度函数的近似初始化方法[9]。上述文献中基于马氏距离的模糊聚类算法的研究多是对协方差估算方面的改进,并且已有的初始化方法对基于马氏距离的模糊聚类算法并不适用。为此,本研究通过对经典聚类算法和马氏距离特性的分析,提出具有新初始化方法的基于马氏距离模糊聚类的优化算法。

1 经典模糊聚类算法与马氏距离

1.1 经典模糊聚类算法(FCM)

在基于目标函数的模糊聚类算法中,模糊 c-均值算法(fuzzy c-means,FCM)的理论最为完善,应用最为广泛。FCM类型的算法最早是由“硬”聚类算法HCM导出的[10],DUNN[11]把它的目标函数 $J_1=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2,u_{ij}\in\{0,1\}$ 扩展到隶属度属于模糊情形的类内加权平均误差和函数 $J_2=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2,u_{ij}\in[0,1]$。后来BEZDEK[12]又引入了一个参数 m,把 J_2 推广到一个目标函数的无限簇,并给出了交替优化算法,即形成了经典的模糊 c- 均值算法。

FCM算法的核心思想如下:设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$ 为 n 元数据集合。FCM聚类方法就是把 $\boldsymbol{X}$ 划分为 c 个子集 $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_i,\cdots,\boldsymbol{S}_c$,若用 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_i,\cdots,\boldsymbol{v}_c\}$ 表示这 c 个子集的聚类中心,u_{ij} 表示元素 $\boldsymbol{x}_j$ 对 $\boldsymbol{S}_i$ 的隶属度,则FCM算法的优化目标函数为

$$J_{\mathrm{FCM}}^{m}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2}=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2,\tag{1}$$

u_{ij} 满足如下约束条件:

$$\begin{cases}\sum_{i=1}^{c}u_{ij}=1, & 1\leqslant j\leqslant n,\\ u_{ij}\geqslant 0, & 1\leqslant i\leqslant c,\quad i\leqslant j\leqslant n,\end{cases}\tag{2}$$

这里 $\boldsymbol{U}=\{u_{ij}\}$ 为 $c\times n$ 矩阵,$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_i,\cdots,\boldsymbol{v}_c\}$ 为 $s\times c$ 矩阵,d_{ij} 为 $\boldsymbol{x}_j$ 与 $\boldsymbol{v}_j$ 的距离,经典的FCM算法里使用的是欧氏距离,推荐使用 $m=2$。通过如上设定,最佳聚类结果可使得目标函数取得极小值。

FCM算法的具体步骤如下:

步骤1 设定聚类个数 $c(1<c<n)$ 和模糊指数 $m(1\leqslant m<+\infty)$;初始化各类中心 $\boldsymbol{V}^0$;设置收敛的精度 $\varepsilon>0$;令迭代次数 $L=0$,给出迭代数目的最大值 $L_{\max}$。

步骤2 根据 $u_{ij}=\dfrac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}}$ 来计算 $\boldsymbol{U}^{(L+1)}$。

步骤3 用 $\boldsymbol{V}_i=\dfrac{\sum_{j=1}^{n}(u_{ij})^m\boldsymbol{X}_j}{\sum_{j=1}^{n}(u_{ij})^m}$ 计算聚类中心矩阵 $\boldsymbol{V}^{(L+1)}$,令 $L=L+1$。

步骤 4　判断 2 次聚类中心矩阵的欧氏距离与给定阈值的大小，如果满足终止条件：$\| \boldsymbol{V}^{(L)} - \boldsymbol{V}^{(L-1)} \| \leqslant \varepsilon, L \geqslant 1$，或者迭代次数不小于给定的最大迭代次数，则迭代停止，否则，重复步骤 2 和步骤 3。

在对比实验中，FCM 聚类的运行相对于 HCM 算法速度较慢，但聚错样本数明显减少，并且该算法的收敛性已经得以证明。但是由于经典模糊聚类算法就是反复修改聚类中心矩阵和隶属度矩阵的分类过程，并且 FCM 对聚类中心的初始化依赖，使得经典模糊聚类算法不能确保得到全局最优解。另一方面，与 HCM 算法一样，需要预先指定聚类数目 c，而实际中聚类数目通常都是未知的，并且使用欧氏距离只适于发现球状类型等非凸面形状的簇，不能处理椭球形结构簇，在处理高维数据时效果欠佳。

1.2　马氏距离

欧氏距离只适用于样本的属性在相互独立的条件下同等对待每个属性对聚类影响的情形。当属性之间相关时，对相关的属性计算欧氏距离时将产生重复数据，影响了聚类效果和最佳聚类数的确定。同时，欧氏距离受属性量纲的影响，对多维数据的处理是不利的。针对样本向量中各维特征对模式分类的不同影响，李洁等[13] 提出了基于特征加权的模糊聚类新算法，但收敛速度有所下降。用马氏距离来取代欧式距离，可以有效解决以上困扰。马氏距离是一种有效计算 2 个未知样本集相似度的方法，与欧氏距离不同的是它考虑到各种特性之间的联系，并且是与尺度无关的。在文献[14] 中说明了当聚类算法在用于入侵检测[15] 时，马氏距离比欧式距离具有明显的优势，并且马氏距离的模糊聚类在图像分割上也比欧氏距离的效果更优[16]。

设样本集合为 $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m\}$，共有 m 个样本，$\boldsymbol{x}_i$ 是 n 维特征矢量，$i \in \{1,2,\cdots,m\}$，令 $\boldsymbol{X}$ 代表 $m \times n$ 的输入矩阵，每行为一个样本，则样本的均值、自相关矩阵和协方差矩阵可用矩阵表示为

$$\boldsymbol{\mu} = E\{\boldsymbol{X}\} = \boldsymbol{X}^{\mathrm{T}}\left(\frac{1}{m}\right)_{m\times 1}; \quad \boldsymbol{S} = \left(\frac{1}{m}\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}; \quad \boldsymbol{\Sigma} = E\{(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}\} = \frac{1}{m}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}},$$

其中 $\left(\frac{1}{m}\right)_{m\times 1}$ 代表元素均为 $\frac{1}{m}$ 的 m 维列矢量。

样本 $\boldsymbol{X}_i$ 到样本总体 $\boldsymbol{X}$ 的马氏距离定义为 $d^2(\boldsymbol{X}_i - \boldsymbol{X}) = (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$。

由于对于聚类样本的总体分布通常是未知的，而样本协方差是总体协方差的无偏估计，常用样本协方差矩阵代替总体协方差：$\boldsymbol{\Sigma}_i = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{x})(\boldsymbol{x}_i - \bar{x})'$，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$。若协方差矩阵是奇异的，将导致无法直接求马氏距离，故大量文献对基于马氏距离的研究多是对计算其数据协方差的逆不存在的情况提出解决方案，王振丽[17] 为改进马氏距离使用加权 MP 马氏距离进行研究；吴香华等[18] 对马氏距离聚类分析中协方差矩阵的估算进行改进；赵小强等[19] 通过改进协方差矩阵的估计来提高马氏距离聚类分析效率。

基于马氏距离的聚类算法有如下特点。

1）协方差矩阵本身的意义是在多维向量之间找出一个自适应的权重，即马氏距离的计算是建立在总体样本基础上的，有利于加强聚类的准确性。

2）在计算马氏距离过程中，要求总体样本个数大于样本的维数，否则得到的总体样本协方差逆矩阵不存在。

3）在实际应用中，“总体样本个数大于样本的维数”这个条件很容易满足，即在绝大多数的情况下，马氏距离是可以顺利计算的，从而有关基于马氏距离的模糊聚类的优化可以先忽略协方差问题，针对高维大样本的数据做出优化的改进算法。但是马氏距离的计算是不稳定的，不稳定来源于协方差矩阵，这也是马氏距离与欧氏距离的最大差异之处，故在研究中应注意结合用加强稳定性的优化方案来弥补马氏距离的不足。

针对以上问题，在基于马氏距离的模糊聚类算法的研究基础上，本研究并不侧重对马氏距离协方差问题的优化，而是另辟蹊径，提出了新的初始化优化方法，形成一种具有全局优化性能、利于处理高维数据的、新的基于马氏距离的模糊聚类算法。

2　模糊聚类优化算法 ——KM-FCM

2.1　新的聚类初始化方法

由于基于山峰函数和减法聚类的初始化方法的实现性并不令人满意，近几年对初始化方法的研究出现了结合进化及生物仿生算法进行的初始化方法，李静[20] 结合粒子群算法初始聚类中心，使得聚类结果以较

快收敛速度接近最优解，并且无需给定聚类数目，但其缺点是引入了其他参数；NAIK 等[21] 使用基于教学学习(TLBO) 的优化算法，聚类结果以较快收敛速度接近最优解，但是局限于欧氏距离且需要给定聚类数目。

还有一种聚类初始化方法是将已有的复杂度低的聚类算法结果作为模糊聚类的初始聚类中心，如直接运用 k-means 算法选出初始聚类中心[9]；结合概率抽样方法的 k-means ++ 算法[22-23] 是比前者初始效果更好的聚类初始化方法，但引入了一个缺乏理论支撑的参数 p。

考虑到基于马氏距离的模糊聚类算法比经典的模糊聚类算法运行速度慢，为减少算法速度负担，在此提出的初始化方法也是将 k-means 聚类算法结果作为模糊聚类的初始聚类中心。由于 k-means 对聚类中心初始化敏感，易陷入局部最优且需给定聚类数，故给出基于启发式搜索算法与 k-means 算法结合的聚类初始化方法，对基于马氏距离的模糊算法进行优化。

新的初始化方法主要思想如下。为实现全局优化并且避免重复初始聚类中心，首先考虑在一定的聚类个数范围内用启发式搜索聚类中心，然后利用 k-means 聚类算法得出初始聚类中心。最佳聚类数通常满足 $c_{\max} \leqslant \sqrt{n}$，$n$ 为样本数据个数，本研究也作这样的设定。

新的初始化方法的具体步骤如下。

第 1 步：求出样本集合 O 中两两样本之间的距离，找出距离最小的一对较小样本 x_1，x_2；

第 2 步：把 x_1，x_2 放到新的集合 A_1 中，并将它们从样本集合 O 中删除；

第 3 步：计算 A_1 中样本的均值 a_1，求出 a_1 与 O 中每个样本的距离，找出 O 中与 a_1 最近的较小样本 x_3，将 x_3 并入集合 A_1，且从 O 中删除，如此重复，直到 A_1 中样本个数达到 $\dfrac{n}{\sqrt{3n}}$ 的向下取整值；

第 4 步：再把 O 中现有样本中距离最小的一对较小样本 x'_1，x'_2 放入新的集合 A_1 中，并将它们从 O 中删除，重复第 3 步的过程，直到形成 m 个集合 $A_1, A_2, \cdots, A_m$，m 预设成比 $\sqrt{n}$ 大的值，n 为数据集样本的个数，这里设定成 $m = \sqrt{3n}$；

第 5 步：分别求出 $A_1, A_2, \cdots, A_n$ 的均值 $\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_n$，对数据集 $\{\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_n\}$ 运用 k-means 算法，得到的 $\sqrt{n}$ 个聚类中心作为用于模糊聚类算法的初始聚类中心。

针对每个聚类数目进行模糊聚类时，都要重新初始化所造成的聚类数目不稳定的情况，经过算法迭代后，采用合并聚类中心的方式。这使得聚类个数自适应，无需给定聚类数目 c。

2.2 KM-FCM 算法

首先，在聚类相似性准则方面，在经典 FCM 算法的目标函数中用马氏距离替代欧氏距离，并且在目标函数上引进一个协方差调节因子：$-\ln|\boldsymbol{\Sigma}_i^{-1}|$，得到的优化目标函数为

$$J^m(\boldsymbol{U},\boldsymbol{C},\boldsymbol{X}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m [(\boldsymbol{x}_j - \boldsymbol{c}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{c}_i) - \ln|\boldsymbol{\Sigma}_i^{-1}|],$$

$$\text{s.t.} \quad u_{ij} \in [0,1]; \ \sum_{i=1}^{c} u_{ij} = 1; \quad 0 < \sum_{j=1}^{n} u_{ij} < n; \quad i = 1,2,\cdots,c; \quad j = 1,2,\cdots,n,$$

该优化问题的拉格朗日乘子式为

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m [(\boldsymbol{x}_j - \boldsymbol{c}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{c}_i) - \ln|\boldsymbol{\Sigma}_i^{-1}|] + \sum_{j=1}^{n} \lambda_j \left(1 - \sum_{i=1}^{c} u_{ij}\right),$$

最小化 J，对 $\boldsymbol{c}_i$，u_{ij}，$\boldsymbol{\Sigma}_i$ 求偏导，并令其结果等于零，得：

$$u_{ij} = \left[\sum_{s=1}^{c} \left[\frac{(\boldsymbol{x}_j - \boldsymbol{c}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{c}_i)}{(\boldsymbol{x}_j - \boldsymbol{c}_s)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{c}_s)}\right]^{\frac{1}{m-1}}\right]^{-1}, \quad i = 1,2,\cdots,c; \quad j = 1,2,\cdots,n, \tag{3}$$

$$\boldsymbol{c}_i = \left[\sum_{j=1}^{n} u_{ij}^m\right]^{-1} \left[\sum_{j=1}^{n} u_{ij}^m \boldsymbol{x}_j\right], \quad i = 1,2,\cdots,c, \tag{4}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{n} u_{ij}^m (\boldsymbol{x}_j - \boldsymbol{c}_i)(\boldsymbol{x}_j - \boldsymbol{c}_i)'}{\sum_{j=1}^{n} u_{ij}^m}, \quad i = 1,2,\cdots,c \ 。$$

其次，在交替优化方面，结合新的聚类初始化方法，并引用文献[9] 中的利用粒度分析原理的 GD 有效性准则函数在算法循环中合并聚类方法，得到基于马氏距离的模糊聚类优化算法 ——KM-FCM。

1）取定最大初始聚类个数 $c_{\max}=\sqrt{\text{data\_n}}$，data_n 为样本数据个数，模糊加权指数 $m=2$，迭代停止阈值 $L_{\min}=1\times10^{-5}$，权重因子 $\alpha=0.6$，最大迭代次数 $L_{\max}=100$，并令迭代计数器 $L=0$；

2）由设定的初始聚类个数 $c_{\max}$，运用新的初始化方法，得到初始聚类中心结果矩阵 $\boldsymbol{C}$ 并输出；

3）用式(3) 计算隶属矩阵 $\boldsymbol{U}$；

4）用式(4) 更新聚类中心矩阵 $\boldsymbol{C}$；

5）若聚类中心达到迭代停止阈值，则输出模糊分类隶属矩阵 $\boldsymbol{U}$ 和聚类中心 $\boldsymbol{C}$，若未达到迭代停止阈值，则令 $L=L+1$，转向步骤 3)；

6）计算有效性函数 $GD(\boldsymbol{C})=\alpha CD(\boldsymbol{C})+(1-\alpha)\dfrac{1}{SD(\boldsymbol{C})}$，

其中：

$$CD(\boldsymbol{C})=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2},\quad i=1,2,\cdots,c;\quad j=1,2,\cdots,n,$$

$$SD(\boldsymbol{C})=\frac{\sum_{i,s=1,i\neq s}^{c}d_{is}^{2}}{[c(c-1)]/2},\quad i,s=1,2,\cdots,c,$$

将 GD 值保存起来；

7)将类间两两之间的距离最小的两类合并为一类，得到 $c-1$ 个聚类中心；

8)令 $c=c-1$，若 $c<2$，则转向步骤 9)，否则转向步骤 3)，直到达到最大迭代次数 $L_{\max}=100$；

9)选择最小值 GD 对应的聚类结果，即为最佳聚类结果，算法结束。

3　结果与分析

为了证明应用本研究提出的初始化优化后的 KM-FCM 算法比 FCM 算法及未优化的基于马氏距离的模糊聚类算法 FCM-M[1]，M-FCM[19] 在多属性的样本聚类实验中具有优势，笔者将 FCM，FCM-M，M-FCM，KM-FCM 算法应用于来自 UCI 数据库的 Iris，Wine 和 Pima 等 3 个标准数据集。Iris 数据集由 4 维空间的 150 个样本组成，分为 3 个类别，第 1 类与其他 2 类完全分离，而第 2 类与第 3 类之间有交叉，有很多学者认为 Iris 数据也可以分为两类[24-25]；Wine 数据由 13 维空间的 178 个样本组成，分为 3 个完全分离的类别；Pima 数据集由 8 维空间的 768 个样本组成，分为 2 个相互交叠的类别。

分别使用 FCM，FCM-M，M-FCM 聚类算法与 KM-FCM 聚类算法对 Iris 数据各进行 20 次实验，取 20 次实验的平均值。其中，FCM，FCM-M，M-FCM 聚类算法设定聚类类别数，4 种算法实验中相同参数设置如下：隶属度矩阵的指数 $m=2$，最大迭代次数 $L_{\max}=100$，迭代停止准则为 $L_{\min}=1\times10^{-5}$。对 3 种数据集的 4 种聚类算法比较分别如表 1，表 2 和表 3 所示。

由表 1 可知，KM-FCM 算法对低维属性的 Iris 数据集进行聚类是有效的；由表 2 可知，在高维属性的模糊聚类中，FCM 的聚类效果明显下降，而其他基于马氏距离的模糊聚类算法聚类精度较高，KM-FCM 算法聚错个数明显减少；由表 3 可知，KM-FCM 算法在有交叠的、类边界并不清晰的较大样本 Pima 数据集上聚类效果依然良好。对于 3 个不同的数据集，从时间上来看，由于文献[1]中的 FCM-M 算法并未使用优化技巧，运行速度大于文献[19]中用到数据标准化处理、总体协方差估计优化方法的 M-FCM 算法以及本研究提出的 KM-FCM 算法。但是，经过初始化优化的 KM-FCM 算法聚类精度比其他 3 种聚类算法的聚类精度都要好，误分个数少，具有全局优化效果，并且无需设定聚类个数。在 20 次实验中，由于马氏距离的不稳定性，并且未进行协方差奇异问题处理，KM-FCM 算法在有交叠样本数据中的聚类稳定性有待进一步研究。

表 1 对于 Iris 数据集的 20 次实验 4 种聚类算法比较

Tab.1 Comparison of the four clustering algorithms for 20 experiments on Iris datasets

算法	误分个数	平均耗时/s	聚类精度/%
FCM	16	0.163 7	89.33
FCM-M	8	0.210 5	94.67
M-FCM	14	0.231 4	91.23
KM-FCM	7	0.220 8	95.33

表 2 对于 Wine 数据集的 20 次实验 4 种聚类算法比较

Tab.2 Comparison of the four clustering algorithms for 20 experiments on Wine datasets

算法	误分个数	平均耗时/s	聚类精度/%
FCM	92	0.825 8	48.31
FCM-M	48	0.913 1	73.03
M-FCM	49	0.924 3	72.03
KM-FCM	41	0.938 2	76.97

表 3 对于 Pima 数据集的 20 次实验 4 种聚类算法比较

Tab.3 Comparison of the four clustering algorithms for 20 experiments on Pima datasets

算法	误分个数	平均耗时/s	聚类精度/%
FCM	316	0.902 1	58.89
FCM-M	277	0.975 6	63.93
M-FCM	238	0.984 2	68.90
KM-FCM	233	0.996 3	69.66

4 结 语

通过对经典聚类算法和马氏距离特性的研究，以及观察关于初始化方法改进的最新动态，本研究给出了适用于马氏距离模糊聚类算法的初始化方法，提出基于马氏距离模糊聚类的优化算法 KM-FCM。用马氏距离替换经典的模糊聚类算法中的欧氏距离，采用启发式搜索与 k-means 算法结合的初始化方法，与 FCM，FCM-M，M-FCM 等 3 种算法在 3 个标准数据集上进行仿真对比实验，验证了新的算法的有效性和全局优化作用。研究结果可为基于马氏距离的模糊聚类算法的优化提供参考。

参考文献/References：

[1] 蔡静颖.模糊聚类算法及应用[M].北京：冶金工业出版社，2015.

[2] 蔡威. 模糊聚类算法在数据挖掘中的应用研究[D]. 兰州：兰州交通大学，2012.
CAI Wei. Research on the Application of Fuzzy Clustering Algorithm in Data Mining[D]. Lanzhou：Lanzhou Jiaotong University，2012.

[3] NATACHA G，IREN V，GEORGE G，et al. Fuzzy C-means clustering with mahalanobis and minkowski distance metrics[J]. Procedia Computer Science，2017，114：224-233.

[4] 张敏，闵乐泉，张群，等. 基于马氏距离和模糊 C 均值聚类的抠图算法与应用[J].北京科技大学学报，2014，36(5)：688-694.
ZHANG Min，MIN Lequan，ZHANG Qun，et al. Matting algorithm and application based on Mahalanobis distance and the fuzzy C-means clustering algorithm[J]. Journal of University of Science and Technology Beijing，2014，36(5)：688-694.

[5] 康韦晓. 基于马氏距离的 PFCM 算法的非线性系统故障诊断方法[D]. 哈尔滨：哈尔滨工业大学，2016.
KANG Weixiao.Fault Diagnosis Method for Nonlinear System based on PECM Algorithm with Mahalanobis Distance[D]. Harbin：Harbin Institute of Technology，2016.

[6] 赵泉华，李晓丽，赵雪梅，等. 结合马氏距离的区域化模糊聚类遥感图像分割[J]. 中国矿业大学学报，2017，46(1)：222-228.
ZHAO Quanhua，LI Xiaoli，ZHAO Xuemei，et al. Remote sensing image segmentation algorithm with regional fuzzy cluster and Mahalanobis distance[J]. Journal of China University of Mining & Technology，2017，46(1)：222-228.

[7] YAGER R R，FILEV D P.Approximate clustering via the mountain method[J]. IEEE Transaction on Systems Man & Cybemetics，

2002，24(8)：1279-1284.

[8] CHIU S L. Fuzzy model identification based on cluster estimation[J]. Journal of Intelligent and Fuzzy System，1994，2(3)：267-278.

[9] 陈东辉. 基于目标函数的模糊聚类算法关键技术研究[D]. 西安：西安电子科技大学，2012.
CHEN Donghui. Research of Key Techniques in Fuzzy Clustering Based on Objective Function[D]. Xi'an：Xidian University，2012.

[10] 陈新泉.聚类算法中的优化方法应用[M].成都：电子科技大学出版社，2014.

[11] DUNN J C.Well-separated clusters and optimal fuzzy partitions[J].Journal of Cybernetics，1974，4(1)：95-104.

[12] BEZDEK J C.Pattern Recognition with Fuzzy Objective Function Algorithms[M].New York：Plenum Press，1981.

[13] 李洁，高新波，焦李成.基于特征加权的模糊聚类新算法[J].电子学报，2006，34(1)：89-92.
LI Jie，GAO Xinbo，JIAO Licheng. A new feature weighted fuzzy clustering algorithm[J]. Acta Electronica Sinica，2006，34(1)：89-92.

[14] 易倩，滕少华，张巍. 基于马氏距离的 K 均值聚类算法的入侵检测[J]. 江西师范大学学报(自然科学版)，2012，36(3)：284-287.
YI Qian，TENG Shaohua，ZHANG Wei. Mahalanobis distance-based on K-means clustering algorithm for intrusion detection [J]. Journal of Jiangxi Normal University (Natural Science)，2012，36(3)：284-287.

[15] LIU J，CHEN H，ZHONG Z，et al. Intrusion detection algorithm for the wormhole attack in Ad Hoc network[C]// Proceedings of International Conference on Computer Science and Information Technology Advances in Intelligent Systems and Computing[S.l.]：[s.n.]. 2014：147-157.

[16] LIU H C，JENG B C，YIH J M，et al. Fuzzy C-means algorithm based on standard Mahalanobis distances[C]. Proc of the 2009 international Symposium on Information Processing. Huangshan：[s.n.]，2009：422-427.

[17] 王振丽. 基于加权 MP 马氏距离的 GS 方法研究[D]. 南京：南京理工大学，2016.
WANG Zhenli. Research about GS Method Based on the Weighted MP Mahalanobis Distance[D]. Nanjing：Nanjing University of Science & Technology，2016.

[18] 吴香华，牛生杰，吴诚鸥，等. 马氏距离聚类分析中协方差矩阵估算的改进[J]. 数理统计与管理，2011，30(2)：240-245.
WU Xianghua，NIU Shengjie，WU Chengou，et al. An improvement on estimating covariance matrix during clusteranalysis using mahalanobis distance[J]. Journal of Applied Statistics and Management，2011，30(2)：240-245.

[19] 赵小强，李雄伟. 基于改进马氏距离的模糊 C 聚类研究[J]. 中南大学学报(自然科学版)，2013，44(sup2)：195-198.
ZHAO Xiaoqiang，LI Xiongwei. A fuzzy C-means clustering algorithm based on improved Mahalanobis distance[J]. Journal of Central South University (Science and Technology)，2013，44(sup2)：195-198.

[20] 李静. 模糊聚类算法的研究及应用[D]. 无锡：江南大学，2014.
LI Jing. Research and Application of Fuzzy Clustering Algorithm[D].Wuxi：Jiangnan University，2014.

[21] NAIK A，SATAPATHY S C，PARVATH K. Improvement of initial cluster center of C-means using teaching learning based optimization [J]. Procedia Technology，2012，6(4)：428-435.

[22] CELEBI M E，KINGRAVI H A，VELA P A. A comparative study of efficient initialization methods for the k-means clustering algorithm [J]. Expert Systems with Applications，2013，40(1)：200-210.

[23] STETCO A，ZENG Xiaojun，KEANE J. Fuzzy C-means++：Fuzzy C-means with effective seeding initialization[J]. Expert Systems with Applications，2015，42(21)：7541-7548.

[24] 汪西莉，焦李成.一种基于马氏距离的支持向量快速提取算法[J].西安电子科技大学学报(自然科学版)，2004，31(4)：639-643.
WANG Xili，JIAO Licheng. A fast algorithm for extracting the support vector on the Mahalanobis distance[J]. Journal of Xidian University，2004，31(4)：639-643.

[25] RUIZ A，LÓPEZDETERUEL P E. Nonlinear kernel-based statistical pattern analysis[J]. IEEE Trans Neural Netw，2001，12 (1)：16-32.