Advances in Applied Mathematics 应用数学进展, 2021, 10(2), 433-443
Published Online February 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.102049

Hans汉斯

基于拉普拉斯约束的半监督模糊C均值算法

张　宁，马盈仓，朱恒东

西安工程大学理学院，陕西　西安
Email: 1123287175@qq.com, mayingcang@126.com, 2648906124@qq.com

收稿日期：2021年1月7日；录用日期：2021年2月11日；发布日期：2021年2月20日

摘　要

模糊聚类算法作为经典的无监督算法之一，在未提供先验信息的基础上容易陷入局部最优。为了能够将监督学习与无监督学习相结合，同时利用已标签数据和未标签数据共同进行训练学习，本文通过对目标函数进行拉普拉斯约束，通过验证隶属度的范围始终大于等于零，能够证明该算法是有效的。在其基础上加入先验信息来挖掘大量有用的信息，使之在未提供先验信息的基础上，算法能够合理、有效地利用部分已标识样本的类别信息对未标识样本产生影响，从而提高半聚类算法的聚类性能；最后，将文章中提出的两类改进算法与原始模糊c均值(FCM)进行聚类指标对比，能够显示其具有良好的聚类效果。

关键词

拉普拉斯约束，先验信息，隶属度，聚类

Semi-Supervised Fuzzy C-Means Algorithm Based on Laplace Constraint

Ning Zhang, Yingcang Ma, Hengdong Zhu

School of Science, Xi'an Polytechnic University, Xi'an Shaanxi
Email: 1123287175@qq.com, mayingcang@126.com, 2648906124@qq.com

Received: Jan. 7th, 2021; accepted: Feb. 11th, 2021; published: Feb. 20th, 2021

Abstract

As one of the classical unsupervised algorithms, fuzzy clustering algorithm is easy to fall into local optimum without providing prior information. In order to combine supervised learning with unsupervised learning and use both labeled and unlabeled data for training learning, this paper proved the effectiveness of the algorithm through Laplace constraint on the objective function and verification that the range of membership is always greater than or equal to zero. On this basis,

文章引用: 张宁, 马盈仓, 朱恒东. 基于拉普拉斯约束的半监督模糊 C 均值算法[J]. 应用数学进展, 2021, 10(2): 433-443. DOI: 10.12677/aam.2021.102049

prior information is added to mine a lot of useful information, so that the algorithm can reasonably and effectively use the category information of part of the identified samples to affect the unidentified samples, so as to improve the clustering performance of the semi-clustering algorithm. Finally, the two improved algorithms proposed in this paper are compared with the original Fuzzy C-means (FCM) for clustering index, and the results show that the proposed algorithm has good clustering effect.

Keywords

Laplacian Constraint Sparse, Prior Information, Membership, Clustering

1. 引言

近年来，聚类分析在模式识别、图像处理和数据挖掘中得到了广泛的应用。它试图将数据集划分为不同的组，使得同一集群中的数据点(模式、样本和主题)具有较高的相似性，而不同集群中的数据点具有较低的相似性。到目前为止，已经开发了许多聚类算法，包括层次聚类[1]、谱聚类[2]和模糊 c 均值聚类(FCM) [3]等。

作为半监督聚类，可以采用不同的方法来控制聚类过程。传统的模糊聚类算法对未知样本的使用率较低，针对于该问题，相关领域学者经过不断研究提出了半监督模糊聚类。2000 年，Wagstaff 引入了一个具有成对约束集群的修改版本，即“必须连接”和“不能连接”，以提高聚类性能[4]。由于模糊 c 均值(FCM)是最经典的算法之一，一些相关的工作已经被提出，来约束半监督模糊 c 均值，例如在隶属度中加入半监督项[6]。半监督模糊聚类算法通过将少量的数据类别标签作为监督信息[5]来加入到模糊聚类算法中，使其在整个聚类迭代优化过程中发挥一定的监督作用。SFCM 算法[7]是一种经典的半监督聚类算法，它以标签信息作为先验知识。该算法将已知的类别标签集成到隶属度矩阵中，指导隶属度矩阵的优化，约束项中所含的先验信息则会对隶属度矩阵的优化起监督作用，并创建最合理的模糊划分，以此提高聚类效果。Pedrycz 和 Waletzky [8]接受了改进的 FCM 算法，并将聚类问题的标记数据和未标记数据作为改进目标函数的途径。Luis 等人[9]提出了一种新的半监督模糊 c 均值算法，该算法利用基因本体注释作为先验知识来指导基因分组过程。同时，引入基于核的 FCM 方法(SSKFCM) [10] [11]将半监督学习技术与核方法相结合，提高了模糊划分的质量。该方法将半监督聚类扩展到核空间，使聚类在输入空间中划分成具有非线性边界的群。

半监督聚类方法分为基于相似度的聚类方法和基于搜索的聚类方法，Zhang 等人[12]提出了一个框架，对由边缘信息构造的加权拉普拉斯矩阵进行优化更新。在文献[13]的基础上，提出了一种鲁棒稀疏模糊 k 均值聚类算法，该算法是对标准模糊 k 均值算法的一种改进，它采用了一个鲁棒函数而不是平方数据拟合项来处理离群点。该文章中提出了一种新的方法(FSCM)来实现谱聚类结构和数据的模糊相似矩阵的联合学习。更为重要的是，结合稀疏性的概念，进一步引入惩罚项，使每个样本的对象簇成员具有适当的稀疏性。该算法不仅保证了软聚类算法在实际应用中的鲁棒性，而且考虑到隶属度数量较少，避免了性能下降[14]。根据不同聚类评价算法的适用范围，提出了一种特征加权模糊半监督聚类算法(SFFD) [15] [16]。该算法基于完全自适应的距离函数、特征权重和两两约束构造一个统一的目标函数，用于在两两约

束下搜索最优原型参数和最优特征权重。同时，给出了四种不同的模糊聚类有效性评价算法，采用不同的算法来评估 SFFD 算法的有效性，得到不同输入数据集的最优聚类数，从而确定聚类形成过程中的聚类数。文章[17]中提出的半监督模糊聚类算法充分利用了已知的信息样本，以最小信息熵对应的聚类数作为整个样本的最优聚类数，以此得到的聚类中心是模糊聚类的原始聚类中心。

本文在研究模糊 c 均值聚类(FCM)算法的基础上，通过加入正则项来约束 FCM，提出了一种基于拉普拉斯约束的模糊 c 均值(FCML)算法，给出了 FCML 算法的迭代结果，并对其进行非负证明，即 u_{ij} 经过多次迭代后，其最终结果仍为非负数，以此来证明该算法的有效性。

文章第三节在研究基于拉普拉斯约束的模糊 c 均值(FCML)算法的基础上，提出了基于拉普拉斯约束的半监督模糊 c 均值(SFCML)算法，该算法通过引入一些监督信息来改进 FCML 算法，可以在不提供先验信息的情况下充分利用先验信息来对未标记样本进行部分标记，合理有效地利用部分已识别样本的类别信息，从而提高半聚类算法的聚类性能，其最终结果具有和 FCM 算法一样简洁的隶属度与聚类中心的迭代公式。

最后，将文章中提出的基于拉普拉斯约束的模糊 c 均值(FCML)算法及基于拉普拉斯约束的半监督模糊 c 均值(SFCML)算法与原始模糊 c 均值(FCM)的聚类性能进行了检验和评价。

2. 基于拉普拉斯约束的模糊 c 均值(FCML)

2.1. 目标函数设计

为了将数据点按一定的隶属度分组到每个聚类中，模糊 c 均值聚类的目标函数为：

$$\min \sum_{i=1}^{n}\sum_{j=1}^{c}\left\|x_i - v_j\right\|_2^2 u_{ij}^2$$

$$\text{st.} \sum_{i=1}^{c} u_{ij} = 1 \tag{1}$$

其中 X 为原始样本数据集，$V=\{v_1, v_2, \cdots, v_c\}$ 为 C 个聚类中心，$U=\left(u_{ij}\right)_{c\times n}$ 为隶属度矩阵，$\|\cdot\|$ 为欧式距离。

通过稀疏化模糊 c 均值去除大量的冗余变量，只保留与相应变量最相关的解释变量，简化了模型的同时却保留了数据集中最重要的信息，能够有效地解决高维数据集建模中的诸多问题，因此，本文在模糊 c 均值聚类的基础上引入了拉普拉斯约束：

$$\min \sum_{i=1}^{n}\sum_{j=1}^{c}\left\|x_i - v_j\right\|_2^2 u_{ij}^2 + 2\lambda Tr\left(U^{\mathrm{T}} L_S U\right) \tag{2}$$

式中 n 表示元素个数，c 表示聚类个数，d 为维数，S 为数据 X 的相似矩阵，$L_S = D - \dfrac{S^{\mathrm{T}}+S}{2}$ 为 S 的 Laplace 矩阵，$D=\sum_{j=1}^{n}\dfrac{\left(s_{ij}+s_{ji}\right)}{2}$ 为对角矩阵。

2.2. 理论分析

当 U 和 S 固定时，更新 V 可得到

$$V_j = \frac{\sum_{i=1}^{n} u_{ij}^2 x_i}{\sum_{i=1}^{n} u_{ij}^2} \tag{3}$$

当 V 和 S 固定时，更新 U，由目标函数可知与 U 有关的函数为：

$$\min \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 d_{ij}^2 + 2\lambda Tr\left(UL_S U^{\mathrm{T}}\right) \tag{4}$$

在聚类算法中，有一个经典的等式：

$$2Tr\left(UL_S U^{\mathrm{T}}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|u_i - u_j\right\|_F^2 s_{ij} \tag{5}$$

关于(5)式是否成立，给出了如下证明：

$$U^{\mathrm{T}} L_S U = U^{\mathrm{T}}\left(D - S\right)U = U^{\mathrm{T}} DU - U^{\mathrm{T}} SU$$

$$2Tr\left(U^{\mathrm{T}} L_S U\right) = \sum_{i=1}^{n} u_i^2 d_i - \sum_{i,j=1}^{n} u_i u_j s_{ij} = \frac{1}{2}\left(\sum_{i=1}^{n} u_i^2 d_i - 2\sum_{i,j=1}^{n} u_i u_j s_{ij} + \sum_{i=1}^{n} u_j^2 d_j\right)$$

$$= \sum_{i,j=1}^{n} \left\|u_i - u_j\right\|_F^2 s_{ij}$$

因此，目标函数(4)可以改写成：

$$J = \min \sum_{i=1}^{n}\sum_{j=1}^{c} \left\|x_i - v_j\right\|_2^2 u_{ij}^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{c} \left\|u_i - u_j\right\|_F^2 s_{ij} + \mu \sum_{i=1}^{n}\sum_{j=1}^{c} \left\|x_i - x_j\right\|_2^2 s_{ij}^2 \tag{6}$$

即(6)式为

$$J_u = \min \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 d_{ij}^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|u_i - u_j\right\|_F^2 s_{ij} \tag{7}$$

其中含 u_{ij} 的项为：

$$\begin{aligned} &u_{ij}^2 d_{ij}^2 + 2\lambda \sum_{l\neq j} \left\|u_l - u_j\right\|_F^2 s_{lj} \\ &= u_{ij}^2 d_{ij}^2 + 2\lambda\left(\left\|u_1 - u_j\right\|_F^2 s_{1j} + \cdots + \left\|u_{j-1} - u_j\right\|_F^2 s_{j-1,j} + \left\|u_{j+1} - u_j\right\|_F^2 s_{j+1,j} + \cdots + \left\|u_n - u_j\right\|_F^2 s_{nj}\right) \end{aligned} \tag{8}$$

同理，$\left\|u_1 - u_j\right\|_F^2 = \left(u_{11} - u_{1j}\right)^2 + \left(u_{21} - u_{2j}\right)^2 + \cdots + \left(u_{c1} - u_{cj}\right)^2$，其中只有 $\left(u_{i1} - u_{ij}\right)^2$ 中 u_{ij}，所以(8)式中含 u_{ij} 的值为：

$$u_{ij}^2 d_{ij}^2 + 2\lambda \sum_{l\neq j} \left(u_{ij} - u_{il}\right)^2 s_{lj} \tag{9}$$

对其进行拉格朗日求导可得

$$2u_{ij} d_{ij}^2 + 4\lambda \sum_{l\neq j} u_{ij} s_{lj} - 4\lambda \sum_{l\neq j} u_{il} s_{lj} - \xi_j = 0$$

其中 ξ_i 为 Lagrange 乘子，所以

$$u_{ij} = \frac{4\lambda \sum_{l\neq j} u_{il} s_{lj} + \xi_j}{2d_{ij}^2 + 4\lambda \sum_{l\neq j} s_{lj}} \tag{10}$$

由约束条件 $\sum_{k=1}^{c} u_{kj} = 1$ 可知(10)式为

$$\sum_{k=1}^{c} u_{kj} = \sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj} + \xi_j}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}} = 1 \tag{11}$$

即

$$\xi_j = \frac{1 - \sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}}{\sum_{k=1}^{c} \frac{1}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}} \tag{12}$$

将(12)式代入(10)式中可得

$$u_{ij} = \frac{4\lambda \sum_{l \neq j} u_{il} s_{lj} + \frac{1 - \sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}}{\sum_{k=1}^{c} \frac{1}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}}}{2d_{ij}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \tag{13}$$

本文中，我们对迭代后(13)式中 $u_{ij} \geq 0$ 是否成立进行了证明：

证明：
$$u_{ij} = \frac{4\lambda \sum_{l \neq j} u_{il} s_{lj} + \frac{1 - \sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}}{\sum_{k=1}^{c} \frac{1}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}}}}{2d_{ij}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \geq 0$$

由于 $2d_{ij}^2 + 4\lambda \sum_{l \neq j} s_{lj} \geq 0,\ 2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj} \geq 0,\ 4\lambda \sum_{l \neq j} u_{il} s_{lj} \geq 0$，因此在这里只需证：$1 - \sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \geq 0$

即

$$\sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \leq 1$$

因为

$$\frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \leq \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{4\lambda \sum_{l \neq j} s_{lj}} = \frac{\sum_{l \neq j} u_{kl} s_{lj}}{\sum_{l \neq j} s_{lj}}$$

可得

$$\sum_{k=1}^{c} \frac{4\lambda \sum_{l \neq j} u_{kl} s_{lj}}{2d_{kj}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \leq \sum_{k=1}^{c} \frac{\sum_{l \neq j} u_{kl} s_{lj}}{\sum_{l \neq j} s_{lj}}$$

因为

$$\sum_{k=1}^{c}\frac{\sum_{l\neq j}u_{kl}s_{lj}}{\sum_{l\neq j}s_{lj}}=\sum_{k=1}^{c}\frac{u_{k1}s_{1j}+u_{k2}s_{2j}+\cdots+u_{k,j-1}s_{j-1,j}+u_{k,j+1}s_{j+1,j}+\cdots+u_{kn}s_{nj}}{s_{1j}+s_{2j}+\cdots+s_{j-1,j}+s_{j+1,j}+\cdots+s_{nj}}$$
$$=\frac{s_{1j}\sum_{k=1}^{c}u_{k1}+s_{2j}\sum_{k=1}^{c}u_{k2}+\cdots+s_{j-1,j}\sum_{k=1}^{c}u_{k,j-1}+s_{j+1,j}\sum_{k=1}^{c}u_{k,j+1}+\cdots+s_{nj}\sum_{k=1}^{c}u_{kn}}{s_{1j}+s_{2j}+\cdots+s_{j-1,j}+s_{j+1,j}+\cdots+s_{nj}} \tag{14}$$

由约束条件可得 $\sum_{k=1}^{c}u_{k1}=1,\cdots,\sum_{k=1}^{c}u_{kn}=1$，从而 $\sum_{k=1}^{c}\frac{\sum_{l\neq j}u_{kl}s_{lj}}{\sum_{l\neq j}s_{lj}}=1$，因此 $u_{ij}\geq 0$ 得证。

固定 U 和 V，更新 S，(4)式可以看作：

$$\lambda Tr\left(U^{\mathrm{T}}L_S U\right)+\mu\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|x_i-x_j\right\|_2^2 s_{ij}^2=\lambda\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|u_i-u_j\right\|_2^2 s_{ij}+\mu\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|x_i-x_j\right\|_2^2 s_{ij}^2 \tag{15}$$

由于 i 独立，因此对(5)式求解，可以得到

$$s_i^{\mathrm{T}}Ds_i+\frac{\lambda}{\mu}s_i^{\mathrm{T}}V_i \tag{16}$$

2.3. 算法流程

基于拉普拉斯约束的模糊 c 均值(FCML)与其他聚类算法具有类似的框架，其流程可以总结为：

输入：原始样本数据集 X，聚类簇数 c 以及相似矩阵 S。

输出：聚类中心 V，迭代后的隶属度矩阵 U。

步骤1：按照约束条件初始化 U 并设置聚类簇数 c，迭代次数设置为100次。

步骤2：计算实际数据集的样本点与聚类中心的距离 d；

步骤3：利用公式(4)计算迭代后的隶属度矩阵 U；

步骤4：利用公式(3)计算 V；

步骤5：本次计算得出的聚类中心与上次相比，不发生变化或者满足迭代次数超出最大次数，则算法停止；否则，返回步骤2。

3. 基于拉普拉斯约束的半监督模糊 c 均值算法(SFCML)

3.1. 目标函数设计

半监督聚类是一种监督聚类和无监督聚类相结合的聚类方法，既可以使用有标记的数据，也可以使用无标记的数据。在上一节中，我们通过对 FCM 算法进行拉普拉斯约束，得到了 FCML 算法，该过程是简单的，它通常可以达到预期的性能。然而，FCML 算法没有嵌入可在某些应用中收集和有用的先验知识，因此算法容易陷入局部最优。本节中，我们在数据集 X 上运行 FCML 算法，给定原始数据集 X，第 l 个样本与它们的集群标签 y 构成一个标记子集 Y，其余的 $n-l$ 个样本构成一个未标记子集，通过将先验知识引入到 FCML 中，此时 SFCML 算法的目标函数为：

$$J=\min\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|x_i-v_j\right\|_2^2 u_{ij}^2+\lambda\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|u_i-u_j\right\|_F^2 s_{ij}+\alpha\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}-f_{ij}b_i\right)^2 d_{ij}^2$$
$$\text{st.}\sum_{i=1}^{n}u_{ij}=1,u_{ij}\geq 0 \tag{17}$$

式中，$\alpha(\alpha \geq 0)$是反映保真度项重要性的参数，α 的具体值和$\frac{\text{总样本数}L}{\text{标记样本数}L_1}$成正比例关系；$f_{ij}$ 为标签样本的隶属度矩阵，其值具体表示为 x_i 归于 c_j 的程度大小；b_i 为一个布尔型的二值向量，根据其实际值可以判断 x_i 是否是已经标记的数据。b_i 需要满足的条件如下：

$$\begin{cases} b_i = 1, x_i \text{被标记} \\ b_i = 0, \text{其他} \end{cases} \tag{18}$$

3.2. 理论分析

当 U 和 S 固定时，更新 V 可得到

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^2 x_i}{\sum_{i=1}^{n} u_{ij}^2} \tag{19}$$

当 V 和 S 固定时，更新 U，由目标函数可知与 U 有关的函数为：
其中含 u_{ij} 的项为：

$$J = \min \sum_{i=1}^{n}\sum_{j=1}^{c} \left\| x_i - v_j \right\|_2^2 u_{ij}^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{c} \left\| u_i - u_j \right\|_F^2 s_{ij} + \alpha \sum_{i=1}^{n}\sum_{j=1}^{c} \left(u_{ij} - f_{ij} b_i \right)^2 d_{ij}^2 \tag{20}$$

$$\begin{aligned} & u_{ij}^2 d_{ij}^2 + 2\lambda \sum_{l \neq j} \left\| u_l - u_j \right\|_F^2 s_{lj} + \alpha \sum_{i=1}^{n}\sum_{j=1}^{c} \left(u_{ij} - f_{ij} b_i \right)^2 d_{ij}^2 \\ & = u_{ij}^2 d_{ij}^2 + 2\lambda \left(\left\| u_1 - u_j \right\|_F^2 s_{1j} + \cdots + \left\| u_{j-1} - u_j \right\|_F^2 s_{j-1,j} + \left\| u_{j+1} - u_j \right\|_F^2 s_{j+1,j} + \cdots + \left\| u_n - u_j \right\|_F^2 s_{nj} \right) \\ & \quad + \alpha \left(\left(u_{11} - f_{11} b_1 \right)^2 d_{11}^2 + \cdots + \left(u_{1j} - f_{1j} b_1 \right)^2 d_{1j}^2 + \cdots + \left(u_{ij} - f_{ij} b_i \right)^2 d_{ij}^2 + \cdots + \left(u_{nc} - f_{nc} b_n \right)^2 d_{nc}^2 \right) \end{aligned} \tag{21}$$

由于只有 $\left(u_{i1} - u_{ij} \right)^2$ 及 $\alpha \left(u_{ij} - f_{ij} b_i \right) d_{ij}^2$ 中有 u_{ij}，所以(21)式中含 u_{ij} 的值为：

$$u_{ij}^2 d_{ij}^2 + 2\lambda \sum_{l \neq j} \left(u_{ij} - u_{il} \right)^2 s_{lj} + \alpha \sum_{i=1}^{n}\sum_{j=1}^{c} \left(u_{ij} - f_{ij} b_i \right)^2 d_{ij}^2 \tag{22}$$

将(22)式改写为

$$L\left(u_{ij}, \eta, \beta_{ij} \right) = u_{ij}^2 d_{ij}^2 + 2\lambda \sum_{l \neq j} \left(u_{ij} - u_{il} \right)^2 s_{lj} + \alpha \sum_{i=1}^{n}\sum_{j=1}^{c} \left(u_{ij} - f_{ij} b_i \right)^2 d_{ij}^2 - \eta \left(\sum_{i=1}^{n} u_{ij} - 1 \right) - \beta_{ij} u_{ij} \tag{23}$$

对(23)式中 u_{ij} 求偏导数可得：

$$\frac{\partial J_u}{\partial u} = 2u_{ij} d_{ij}^2 + 4\lambda \sum_{l \neq j} \left(u_{ij} - u_{il} \right) s_{lj} + 2\alpha \left(u_{ij} - f_{ij} b_i \right) d_{ij}^2 - \eta - \beta_{ij}$$

对其进行拉格朗日求导可得

$$2u_{ij} d_{ij}^2 + 4\lambda \sum_{l \neq j} u_{ij} s_{lj} - 4\lambda \sum_{l \neq j} u_{il} s_{lj} - 2\alpha \left(u_{ij} - f_{ij} b_i \right) d_{ij}^2 - \eta - \beta_{ij} = 0$$

其中 η 和 $\beta_{ij} \geq 0$ 为 Lagrange 乘子，所以

$$u_{ij} = \left(\frac{4\lambda \sum_{l \neq j} u_{il} s_{lj} - 2\alpha f_{ij} b_i d_{ij}^2 + \eta + \beta_{ij}}{2d_{ij}^2 - 2\alpha d_{ij}^2 + 4\lambda \sum_{l \neq j} s_{lj}} \right)_+ \tag{24}$$

由约束条件$\sum_{k=1}^{c}u_{kj}=1$可知(24)式为

$$\sum_{k=1}^{c}u_{kj}=\sum_{k=1}^{c}\frac{4\lambda\sum_{l\neq j}u_{kl}s_{lj}-2\alpha f_{kj}b_k d_{kj}^2+\xi_j}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}=1 \tag{25}$$

因此

$$\xi_j=\frac{1-\sum_{k=1}^{c}\frac{4\lambda\sum_{l\neq j}u_{kl}s_{lj}}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}+\sum_{k=1}^{c}\frac{2\alpha f_{kj}b_k d_{kj}^2}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}}{\sum_{k=1}^{c}\frac{1}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}} \tag{26}$$

将(25)式代入(24)式中可得

$$u_{ij}=\frac{4\lambda\sum_{l\neq j}u_{il}s_{lj}-2\alpha f_{ij}b_i d_{ij}^2+\frac{1-\sum_{k=1}^{c}\frac{4\lambda\sum_{l\neq j}u_{kl}s_{lj}}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}+\sum_{k=1}^{c}\frac{2\alpha f_{kj}b_k d_{kj}^2}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}}{\sum_{k=1}^{c}\frac{1}{2d_{kj}^2-2\alpha d_{kj}^2+4\lambda\sum_{l\neq j}s_{lj}}}}{2d_{ij}^2-2\alpha d_{ij}^2+4\lambda\sum_{l\neq j}s_{lj}} \tag{27}$$

3.3. 算法 2

基于拉普拉斯约束的半监督模糊c均值(SFCML)与其他半监督聚类算法具有类似的框架，其流程为：

输入：需要聚类的数据对象集合 X，聚类的类别数 c 以及带有标签信息的约束集 Y。

输出：更新后的聚类中心 V 及隶属度矩阵 U。

步骤1：通过计算每个集群中标记样本的平均值来初始化集群中心 V_0，按照约束条件随机初始化 U，设置聚类个数 c，由标记信息计算标记信息的初始隶属度矩阵 F_0；

步骤2：计算实际数据集 X 的样本点与聚类中心的距离 d，标签样本集 Y 与聚类中心的距离 d_y；

步骤3：利用公式(27)计算迭代后的隶属度矩阵 U；

步骤4：利用公式(19)计算聚类中心 V；

步骤5：本次计算得出的聚类中心与上次相比，不发生变化或者满足迭代次数超出最大次数，则算法停止；否则，返回步骤2。

4. 实验结果分析

4.1. 实验设计

为评估聚类结果，采用聚类准确率(*ACC*)、*NMI* 指标及兰德指数(简称 *RI*)这三种被广泛使用的聚类性能指标。*ACC* 指标可以发现聚类结果和真实类标签之间的一对一关系，且测量每个聚类所包含的来自对应类别的数据点的多少，其计算式为

$$ACC=\frac{\sum_{i=1}^{n}\delta\left(map\left(r_i\right),q_i\right)}{n}$$

NMI 指标用于确定聚类的质量，给定一个聚类结果，则

$$NMI = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij} \lg \frac{n_{ij}}{n_i \hat{n}_j}}{\sqrt{\left(\sum_{i=1}^{n} n_i \lg \frac{n_i}{n}\right)\left(\sum_{j=1}^{n} \hat{n}_j \lg \frac{\hat{n}_j}{n}\right)}}$$

兰德指数(简称 *RI*)算法的性能根据该算法获得的决策数量的正确率来评估。它需提供定实际类别信息 *C*，假设 *K* 是聚类结果，*a* 表示在 *C* 和 *k* 中都是同类别的元素对数，*d* 表示在 *C* 与 *k* 中都是不同类别的元素对数，则 *RI* 参数为：$RI = \frac{a+d}{C_2^n \text{samples}}$，其中 C_2^nsamples 数据集中可以组成的总元素对数。

4.2. 真实数据集实验结果

我们在真实数据集中选取 iris 数据集进行实验。实验采用 FCM、FCML 和 SFCML 进行测试和比较。为了验证测试结果的可靠性，在 FCM、FCML 和 SFCML 实验中均使用欧氏距离。此外，算法的标签样本点个数为10个，标签样本点的初始隶属度相同。采用三种算法对真实数据集进行依次测试，每种算法测试10次。最终实验结果见表1。

Table 1. Accuracy of FCM, FCML and SFCML algorithms on Iris data set
表 1. Iris 数据集上 FCM、FCML、SFCML 算法准确率

迭代次数	FCM	FCML	SFCML
1	0.8933	0.9067	0.9133
2	0.8797	0.8933	0.9067
3	0.9067	0.9184	0.9600
4	0.8900	0.9034	0.9307
5	0.7400	0.9700	0.9900
6	0.9700	0.9800	0.9900
7	0.9015	0.9184	0.9600
8	0.8400	0.9700	1
9	0.9700	0.9900	1
10	0.9700	0.9800	1

由表 1 可以清楚看出，FCM 算法、FCML 算法和 SFCML 算法的准确率大小依次为：SFCML 算法最大，FCML 算法其次，最后是 FCM 算法，与理论分析结果一致。和 FCM 算法相比，FCML 算法能够有效提取有用信息进行聚类，而 SFCML 算法在 FCML 算法的基础上加入监督信息，使得算法准确率进一步提升。

表 2 分别对 FCM、FCML 和 SFCML 算法进行聚类质量 NMI 及聚类效果 RI 对比。

Table 2. Clustering effect comparison of FCM, FCML and SFCML on Iris data set
表 2. Iris 数据集上 FCM、FCML、SFCML 聚类效果对比

评估标准	FCM	FCML	SFCML
NMI	0.9067	0.9184	0.9390
RI	0.9600	0.9703	0.9800

从聚类准确度方面分析，SFCML 算法达到最大值，为 0.9800；而 FCM 算法因为没有监督信息集拉普拉斯约束项做指导，聚类的准确率是 3 种算法中最小的，为 0.9600；FCML 算法居中，为 0.9703。综合来看，无论用 3 个方面的哪一个评价，SFCML 算法的聚类性能都要优于其他 2 种聚类算法。

5. 总结与展望

本文在经典 FCM 算法的基础上引入了拉普拉斯算法进行约束，提高聚类的抗噪性能以及提取重要的属性特征，并将最终迭代结果进行非负验证。其次，利用少量标记信息进行数据预处理，构造半监督聚类算法 SFCML 来对 FCML 算法进行改进。此外，由于 SFCML 的目标函数是基于 FCM 的，它继承了聚类算法 FCM 的大部分优点。本文在真实数据集上进行算法对比实验，实验结果进一步验证了本文提出的 SFCML 算法的有效性。之后对半监督的研究将从对相似矩阵 S 进行半监督学习，并验证是否能达到改进已有算法的效果。与多个领域进行融合，在不同领域内运用半监督聚类算法的思想，加入不同领域的知识，可以得到更加优化的效果。

致　谢

首先要感谢我的论文指导老师、西安工程大学理学院研究生院的马盈仓老师。马老师对我论文的研究方向做出了指导性的意见和建议，在写这篇论文的过程中，他对我遇到的困难和疑惑及时给予了认真的指导，提出了许多有益的改进建议，投入了大量的心血和精力。衷心感谢马老师对我的帮助和关心！同时，我也要感谢西安工程大学理学院研究生院计算数学专业的老师们和全体同学们，我们互相学习，互相帮助，度过了一段美好而难忘的时光。此外，我还要感谢我的朋友和同学们在论文的准备过程中给予我的大力支持和帮助，给我带来了很大的启发。也要感谢参考文献中的作者，他们的研究文章对我的研究课题有了一个很好的起点。最后，谢谢论文评阅老师们的辛苦工作。衷心感谢我的家人、朋友和同学对我的鼓励和支持，让我顺利完成了这篇论文。

基金项目

国家自然科学基金项目(61976130)；陕西省重点研发计划项目(2018KW-021)；陕西省自然科学基金项目(2020JQ-923)。

参考文献

[1] Johnson, S.C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, **32**, 241-254. https://doi.org/10.1007/BF02289588

[2] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002) On Spectral Clustering: Analysis and an Algorithm. *The Conference and Workshop on Neural Information Processing Systems*, Vol. 14, 849-856.

[3] Bezdek, J.C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York. https://doi.org/10.1007/978-1-4757-0450-1

[4] Wu, L., Hoi, S., Jin, R., *et al*. (2010) Learning Bregman Distance Functions for Semi-Supervised Clustering. *IEEE Transactions on Knowledge & Data Engineering*, **24**, 478-491. https://doi.org/10.1109/TKDE.2010.215

[5] 白福均, 高建瓴, 宋文慧, 等. 半监督模糊聚类算法的研究与改进[J]. 通信技术, 2018, 317(5): 71-75.

[6] Brodinova, S., Filzmoser, P., Ortner, T., *et al*. (2019) Robust and Sparse K-Means Clustering for High-Dimensional Data. *Advances in Data Analysis & Classification*, **13**, 905-932. https://doi.org/10.1007/s11634-019-00356-9

[7] 朱乐为. 模糊 C-means 聚类算法的拓展研究[J]. 云南民族大学学报(自然科学版), 2019, 28(3): 64-70.

[8] Pedrycz, W. and Waletzky, J. (1997) Fuzzy Clustering with Partial Supervision. *IEEE Transactions on Systems Man and Cybernetics Part B—Cybernetics*, **27**, 787-795. https://doi.org/10.1109/3477.623232

[9] Tari, L., Baral, C. and Kim, S. (2009) Fuzzy c-Means Clustering with Prior Biological Knowledge. *Journal of Biomedical Informatics*, **42**, 74-81. https://doi.org/10.1016/j.jbi.2008.05.009

[10] Zhang, H.X. and Lu, J. (2009) Semi-Supervised Fuzzy Clustering: A Kernel-Based Approach. *Knowledge-Based Systems*, **22**, 477-481. https://doi.org/10.1016/j.knosys.2009.06.009

[11] Zhang, D.Q., Zhou, Z.H. and Chen, S.C. (2007) Semi-Supervised Dimensionality Reduction. *Proceedings of the Seventh Siam International Conference on Data Mining*, Minneapolis, 26-28 April 2007, 629-634. https://doi.org/10.1137/1.9781611972771.73

[12] Zhang, R., Nie, F. and Li, X. (2017) Self-Weighted Spectral Clustering with Parameter-Free Constraint. *Neurocomputing*, **241**, 164-170. https://doi.org/10.1016/j.neucom.2017.01.085

[13] Zhang, R., Nie, F., Guo, M., *et al*. (2019) Joint Learning of Fuzzy k-Means and Nonnegative Spectral Clustering with Side Information. *IEEE Transactions on Image Processing*, **28**, 2152-2162. https://doi.org/10.1109/TIP.2018.2882925

[14] Wang, D., Nie, F. and Huang, H. (2015) Feature Selection via Global Redundancy Minimization. *IEEE Transactions on Knowledge & Data Engineering*, **27**, 2743-2755. https://doi.org/10.1109/TKDE.2015.2426703

[15] 李龙龙, 何东健, 王美丽. 模糊半监督加权聚类算法的有效性评价研究[J]. 计算机技术与发展, 2016, 26(6): 65-68.

[16] Li, L.L., Jonathan, G., He, D.J., *et al*. (2015) Semi-Supervised Fuzzy Clustering with Feature Discrimination. *PLoS ONE*, **10**, 131-160. https://doi.org/10.1371/journal.pone.0131160

[17] 郭新辰, 郗仙田, 樊秀玲, 等. 基于半监督的模糊 C-均值聚类算法[J]. 吉林大学学报: 理学版, 2015, 53(4): 705-709.