第50卷第7期 中国科学技术大学学报 Vol. 50, No. 7
2020年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2020
文章编号:0253-2778(2020)07-0993-10

图正则化的模糊局部坐标编码概念分解模型

张怿恺[1],彭 勇[1],孔万增[1],文益民[2]
(1.杭州电子科技大学计算机科学与技术学院,浙江杭州 310018;2.桂林电子科技大学计算机与信息安全学院,广西桂林 541000)

摘要:现有的基于矩阵分解聚类模型训练过程大多需要两个独立的步骤,一是通过自身的模型对数据集进行训练获得系数矩阵,二是对得到的系数矩阵进一步使用 K-means 方法来获得最终的聚类结果.这种两阶段模式一方面增加了计算消耗,也会因为 K-means 对初始聚类中心的敏感,会对聚类效果产生一定的影响.针对此问题,本文提出了一种图正则化的模糊局部坐标编码概念分解模型.该模型通过对系数矩阵添加约束使得系数矩阵行和为 1,从而避免了再次使用 K-means 方法进行二次训练,而直接由系数矩阵获得聚类结果.另外,由于此系数矩阵的约束.该模型实现了模糊聚类,增强了聚类结果的可解释性.本文通过对人工合成数据的测试,验证了该模型的模糊性与可解释性;同时在常用的标准数据集上,通过与现有的聚类方法相比较,同样获得了较好的聚类效果.

关键词:概念分解;局部坐标编码;模糊聚类;图正则化
中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.07.017

引用格式:张怿恺,彭勇,孔万增,等. 图正则化的模糊局部坐标编码概念分解模型[J]. 中国科学技术大学学报,2020,50(7):993-1002.
ZHANG Yikai, PENG Yong, KONG Wanzeng, et al. Fuzzy local coordinate concept factorization with graph regularization[J]. Journal of University of Science and Technology of China, 2020, 50(7): 993-1002.

Fuzzy local coordinate concept factorization with graph regularization

ZHANG Yikai[1], PENG Yong[1], KONG Wanzeng[1], WEN Yimin[2]
(1. *School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou* 310018, *China*;
2. *School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin* 541000, *China*)

Abstract: Matrix Factorization is an effective and efficient method to solve clustering problems in machine learning. However, for most traditional which factorization based models in clustering, there are two necessary steps to get the final assignments. First, original data can be decomposed to a basis matrix and a coefficient matrix through a certain model. Second, the learned coefficient matrix is fed into K-means to make discretization. This two-step paradigm causes extra computational burden and may have some side effect on the final results due to the sensitivity to initialization of K-means. To this end, a novel model termed fuzzy local coordinate concept factorization with graph regularizer (FLCCF-G) is proposed. Which avoids using K-means by enforcing the sum of each row of the non-negative coefficient matrix to equal to one. Then the final clustering results can obtained directly by checking the maximum value of each row of the coefficient matrix. In addition, through this constraint, our proposed model changes is a fuzzy clustering model rather than hard clustering, indicating that the model has better interpretability to data points in boundaries of different clusters. Extensive experimental results on synthetic and Benchmark data sets indicate the better performance of FLCCF-G on data clustering.

Key words: concept factorization; local coordinate coding; fuzzy clustering; graph regularizer

0 引言

矩阵分解是解决聚类问题的一个有效途径,其基本思想是通过构造多个低维的矩阵来近似表示原始的高维矩阵.非负矩阵分解(nonnegative matrix factorization, NMF)[1-3]是其中一个经典模型,它通过构造两个非负矩阵来近似原数据矩阵,其中一个非负矩阵表示数据的特征信息,称为基矩

收稿日期:2020-04-30;**修回日期**:2020-06-22
基金项目:国家自然科学基金(61971173, 61602140);浙江省科技计划(2017C33049);中国博士后科学基金(2017M620470);浙江省新苗人才计划(2019R407030)资助.
作者简介:张怿恺,男,1998年生,硕士生,研究方向:机器学习与模式识别.E-mail:yikaizhang@hdu.edu.cn
通讯作者:彭勇,博士/副教授.E-mail:yongpeng@hdu.edu.cn

阵,另一个非负矩阵表示样本与特征之间的系数,称为系数(权重)矩阵,在分析聚类结果时,通过对系数矩阵采用 K-means 算法来获得最终结果.这种计算模式通常需要两个独立的步骤来完成,一是通过模型特有算法的训练获得对原始数据分解后的系数矩阵,二是通过 K-means 模型对得出的系数矩阵进一步训练得出聚类结果.许多基于 NMF 的聚类模型同样采用了第二种计算模式,如概念分解模型(concept factorization,CF)[4];图正则化非负矩阵分解模型(graph regularized nonnegative matrix factorization, GNMF)[5];局部一致概念分解模型(locally consistent concept factorizatrion, LCCF)[6];非负局部坐标矩阵分解模型(nonnegative local coordinate factorization, NLCF)[7];局部坐标概念分解模型(local coordinate concept factorization,LCF)[8],在该种计算模式下,人们无法直接从系数矩阵中获得聚类信息,必须借由如 K-means 模型进行二次聚类,然而该两段式的计算模式通常具有以下缺点:①聚类效果受 K-means 初始聚类中心的影响较大,对聚类效果会产生一定影响;②增加了整个算法的计算消耗;③该模式为硬聚类,在某些应用方面缺乏一定的合理性.

针对上述问题,本文采用了模糊聚类[9-11]的思想,提出了一种直接通过系数矩阵得出聚类结果的模型:图正则化的模糊局部坐标编码概念分解模型(fuzzy local coordinate concept factorization with graph regularization, FLCCF-G),该模型除了在 LCF 模型的基础上结合了图正则化技术[5]之外,还对系数矩阵进行了约束,使得系数矩阵行和为 1[12].在系数矩阵中,每一行代表某个样本对各个聚类的隶属度,称为系数向量,当模型计算出系数矩阵后,只需通过寻找每个样本的系数向量中的最大值即可获得相应的聚类标签;若某个样本的系数向量中有多个非零值,说明该样本对多个样本都有一定的隶属度,为模糊聚类.

本文针对 FLCCF-G 给出了一种有效的求解方式,分析了其算法复杂度与收敛性,并依据算法的特性在人工拟合数据与真实数据上做了相关的实验与分析,获得了较好的聚类效果.

1 模型介绍

给定的数据集 $X=[x_1,x_2,\cdots,x_n]\in\mathbb{R}^{d\times n}$,NMF 模型[1]通过构造非负的基矩阵 U 和非负的系数矩阵 V 来近似表示原数据 X,有

$$X\approx UV^{\mathrm{T}} \tag{1}$$

式中,非负矩阵 $U=[u_{ik}]\in\mathbb{R}^{d\times c}$,非负矩阵 $V=[v_{jk}]\in\mathbb{R}^{n\times c}$,$c$ 为聚类个数.由此可以看出,U 中的每一列表示一个聚类中心,原数据 X 可由这 c 个聚类中心进行线性表示,其系数组成系数矩阵 V.

由于 NMF 只能针对非负数据进行训练,且无法对训练数据空间进行变换,这导致其在面对高度非线性数据时无法对其使用核技术,使其应用具有一定的局限性.CF 模型[4]针对以上问题对基矩阵 U 进行约束使得 $U\approx XW$,其中非负矩阵 $W\in\mathbb{R}^{n\times c}$,CF 模型可通过优化如下表达式获得

$$O=\|X-XWV^{\mathrm{T}}\|_F^2 \tag{2}$$

LCF 模型针对 CF 模型中解稀疏性不足的问题,提出了采用局部坐标编码的技术产生稀疏空间[8].局部坐标编码项的目标是使基矩阵 XW 更接近原始聚类中心且使得每个原始数据点只通过邻近的几个基础向量来线性表示,从而保证解的稀疏性,因此有[13]

$$\min_{W,V}\sum_{i=1}^{n}\sum_{k=1}^{c}v_{ik}\|x_i-\sum_{j=1}^{n}x_jw_{jk}\|_2^2 \tag{3}$$

综上可见,当数据点越接近基矩阵中的某个聚类中心,其对应的系数越大.

2 图正则化的模糊局部坐标编码概念分解模型

2.1 模型的提出

为了实现直接聚类,本文在 LCF 的基础上对非负矩阵 V 引入约束条件:$v_i1=1$,其中 v_i 为矩阵 V 的第 i 行,1 为一个元素全为 1 的列向量,并结合图正则化来增强 V 的局部一致性,该模型表达式为

$$\begin{aligned}\min_{W,V}\|X-XWV^{\mathrm{T}}\|_F^2+\gamma\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|v_i^{\mathrm{T}}-v_j^{\mathrm{T}}\|_2^2+\\ \lambda\sum_{i=1}^{n}\sum_{k=1}^{c}v_{ik}\|x_i-\sum_{j=1}^{n}x_jw_{jk}\|_2^2,\\ s.t.\ W\geqslant 0,V\geqslant 0,v_i1=1.\end{aligned} \tag{4}$$

式中,s_{ij} 为相似矩阵 S 的第 i 行第 j 个元素,λ,γ 为正则化系数.$\sum_{j=1}^{n}x_jw_{jk}$ 为第 k 个聚类的聚类中心,v_{ik} 为第 i 个样本对第 k 个聚类的隶属度.显然,当样本越接近聚类中心,样本对该聚类的隶属度越大.由于模型确定了约束 $v_i1=1$ 即系数矩阵行和为 1,此时可以直接通过取样本对各个聚类的隶属度最大值确定每个样本的聚类结果,实现直接聚类.此外,对于处在多个聚类之间的边界样本,隶属度不再是非 0 即 1 状态,而是对多个聚类都有相应的处于(0,1)之间的隶属度值,从而实现了对样本的模糊聚类.

2.2 模型的求解

对于本文的 FLCCF-G 模型,我们通过交替优化的方法依此更新 W,V.

(Ⅰ)固定 V,求解 W

当 V 固定时,问题(4)可变换为

$$\begin{aligned}\min_{W}\|X-XWV^{\mathrm{T}}\|_F^2+\lambda\sum_{i=1}^{n}\|(x_i1^{\mathrm{T}}-XW)\Lambda_i^{\frac{1}{2}}\|_F^2,\\ s.t.\ W\geqslant 0\end{aligned} \tag{5}$$

式中,$\Lambda_i=diag(v_i)$.此时,式(5)的拉格朗日函数为

$$\begin{aligned}\mathscr{L}(W)=\lambda\sum_{i=1}^{n}\|(x_i1^{\mathrm{T}}-XW)\Lambda_i^{\frac{1}{2}}\|_F^2+\mathrm{Tr}(\Phi W^{\mathrm{T}})\\ +\|X-XWV^{\mathrm{T}}\|_F^2\end{aligned} \tag{6}$$

式中,Φ 为矩阵 W 的拉格朗日乘子.令 $\mathscr{L}(W)$ 关于 W 求导为 0 可得

$$\frac{\partial\mathscr{L}(W)}{\partial W}=2X^{\mathrm{T}}XWV^{\mathrm{T}}V-2X^{\mathrm{T}}XV+\Phi$$

$$-2\lambda\sum_{i=1}^{n}(X^{\mathrm{T}}x_i1^{\mathrm{T}}\Lambda_i-X^{\mathrm{T}}XW\Lambda_i)=0 \quad (7)$$

式(7)结合 KKT 条件：$w_{jk}\varphi_{jk}=0$，并令 $K=X^{\mathrm{T}}X$ 可得

$$(KVw_{jk}-(KWw_{jk}+\lambda(\sum_{i=1}^{n}X^{\mathrm{T}}w_{jk}=0 \quad (8)$$

通过式(8)，可以得到 W 的更新公式

$$w_{jk}\leftarrow w_{jk}\frac{(KV)_{jk}+\lambda(\sum_{i=1}^{n}X^{\mathrm{T}}x_i1^{\mathrm{T}}\Lambda_i)_{jk}}{(KWV^{\mathrm{T}}V)_{jk}+\lambda(\sum_{i=1}^{n}KW\Lambda_i)_{jk}} \quad (9)$$

由矩阵乘法法则可得 $XV=\sum_{i=1}^{n}x_i1^{\mathrm{T}}\Lambda_i$，定义对角矩阵 H，并令 $H_{ii}=\sum_{k=1}^{c}v_{ik}$，则式(9)变为

$$w_{jk}\leftarrow w_{jk}\frac{(\lambda+1)(KV)_{jk}}{(KWV^{\mathrm{T}}V)_{jk}+\lambda(KWH)_{jk}} \quad (10)$$

(Ⅱ)固定 W，更新 V

定义矩阵 E，并令 $e_{ik}=\frac{1}{2}\|x_i-\sum_{j=1}^{n}x_jw_{jk}\|_2^2$，则问题(4)变为

$$\min_V 2\mathrm{Tr}((\lambda E-KW)V^{\mathrm{T}})+\mathrm{Tr}(VW^{\mathrm{T}}KWV^{\mathrm{T}})+\gamma\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|v_i^{\mathrm{T}}-v_j^{\mathrm{T}}\|_2^2,\ \text{s.t. } V\geqslant 0,v_i1=1 \quad (11)$$

对于问题(11)，本文采用逐行求解的方法[14]计算 v_i，通过化简，可得其向量化形式为

$$\min_{v_i\geqslant 0,v_i1=1}v_iAv_i^{\mathrm{T}}-b^{\mathrm{T}}v_i^{\mathrm{T}} \quad (12)$$

式中，对称矩阵 $A=W^{\mathrm{T}}KW+\gamma\sum_{j=1}^{n}s_{ij}I_c\in\mathbb{R}^{c\times c}$，$I_c$ 为 $c\times c$ 的单位阵，向量 $b^{\mathrm{T}}=-2((\lambda E-KW)_i-\gamma\sum_{j=1}^{n}s_{ij}v_j)\in\mathbb{R}^{1\times c}$，$(\lambda E-KW)_i$ 为矩阵 $(\lambda E-KW)$ 的第 i 行. 再令 $x=v_i^{\mathrm{T}}$，并引入辅助变量 z，令 $z=x$，则问题(12)变为

$$\min_{x\geqslant 0,x^{\mathrm{T}}1=1,x=z}x^{\mathrm{T}}Az-b^{\mathrm{T}}x \quad (13)$$

基于增广的拉格朗日法(ALM)，式(13)可变换为

$$\min_{x\geqslant 0,x^{\mathrm{T}}1=1,x=z}x^{\mathrm{T}}Az-x^{\mathrm{T}}b+\frac{\mu}{2}\|x-z+\frac{\beta}{\mu}\|_2^2 \quad (14)$$

通过配方法，问题(14)可继续化简为

$$\min_{x\geqslant 0,x^{\mathrm{T}}1=1,x=z}\frac{\mu}{2}\|x-(z-\frac{\beta+Az-b}{\mu})\|_2^2 \quad (15)$$

式中，β 为拉格朗日乘数，μ 为增长系数. 同样，通过交替优化的方法来求解问题[15-17].

(Ⅲ)固定 z 更新 x

此时，令 $m=z-\frac{\beta+Az-b}{\mu}\in\mathbb{R}^c$，则问题可变换为

$$\min_{x\geqslant 0,x^{\mathrm{T}}1=1}\frac{1}{2}\|x-m\|_2^2 \quad (16)$$

问题(16)的拉格朗日函数为

$$\mathscr{L}(x)=\frac{1}{2}\|x-m\|_2^2-\alpha(x^{\mathrm{T}}1-1)-\varphi^{\mathrm{T}}x \quad (17)$$

式中，α，φ 为拉格朗日乘子. 对 $\mathscr{L}(x)$ 关于 x 求导并令导数为 0 可得

$$x-m-\alpha 1-\varphi=0 \quad (18)$$

式(18)结合约束条件：$x^{\mathrm{T}}1=1$ 可得

$$\alpha=\frac{1-1^{\mathrm{T}}m-1^{\mathrm{T}}\varphi}{c} \quad (19)$$

把式(19)代回式(18)可得

$$x=m+\frac{1}{c}1-\frac{1^{\mathrm{T}}m}{c}1-\frac{1^{\mathrm{T}}\varphi}{c}1+\varphi \quad (20)$$

针对式(20)，定义如下变量

$$\begin{cases}u=m+\frac{1}{c}1-\frac{1^{\mathrm{T}}m}{c}1\\ \varphi^*=\frac{1^{\mathrm{T}}\varphi}{c}1\end{cases} \quad (21)$$

此时，式(20)可写为

$$x=u+(\varphi-\varphi^*) \quad (22)$$

根据 KKT 条件可得出如下结论

$$\begin{cases}x_k\geqslant 0\\ \varphi_k\geqslant 0\\ \varphi_kx_k=0\end{cases} \quad (23)$$

结合式(22)与式(23)可知，若 $u_k-\varphi_k^*<0$，则 $\varphi_k>0$，$x_k=0$；若 $u_k-\varphi_k^*\geqslant 0$，则 $\varphi_k=0$，$x_k=u_k-\varphi_k^*$. 因此式(22)可变换为

$$x_k=(u_k-\varphi_k^*)_+ \quad (24)$$

式中，$(a)_+=max(a,0)$.

式(24)结合约束条件 $x^{\mathrm{T}}1=1$ 可得

$$f(\hat{\varphi})=\sum_{k=1}^{c}(u_k-\hat{\varphi}_k)_+-1 \quad (25)$$

此时，对式(25)使用牛顿法即可求解 φ^* 的最优解，即

$$\hat{\varphi}_{k+1}=\hat{\varphi}_k-\frac{f(\hat{\varphi})}{f'(\hat{\varphi})} \quad (26)$$

(Ⅳ)固定 x 更新 z

此时，可直接对式(15)关于 z 求导并令导数为 0，可得到 z 的求解公式为

$$z=x+\frac{\beta-A^{\mathrm{T}}x}{\mu} \quad (27)$$

本文针对上述提出的问题提供了相应的算法流程，提供了解决子问题(16)的方法，见算法 2.1 所示；算法 2.2 提供了解决子问题(13)的方法；算法 2.3 提供了解决问题(4)的流程.

算法 2.1　解决子问题(16)的相关算法

输入：向量 m$\in\mathbb{R}^c$

输出：向量 x$\in\mathbb{R}^c$

步骤：

1 根据式(21)，计算向量 u；

2 根据式(26)，通过牛顿法计算 φ^*；

3 根据式(24),计算最优解 x;

算法 2.2 解决子问题(13)的相关算法

输入:矩阵 $A\in\mathbb{R}^{c\times c}$,向量 $b\in\mathbb{R}^{c}$

输出:向量 $x\in\mathbb{R}^{c}$

步骤:

1 初始化 x, μ, β, ρ

2 While 未达到完全收敛 do

3 根据式(27),更新辅助向量 z;

4 根据算法 2.1,更新向量 x;

5 更新向量 $\beta=\beta+\mu(x-z)$;

6 更新增长系数 $\mu=\mu\rho$;

7 End while

算法 2.3 解决问题(4)的相关算法

输入:数据集 $X\in\mathbb{R}^{d\times n}$,正则化参数 λ,γ,图的相似矩阵 $S\in\mathbb{R}^{n\times n}$

输出:基矩阵 $W\in\mathbb{R}^{n\times c}$,系数矩阵 $V\in\mathbb{R}^{n\times c}$

步骤:

1 初始化矩阵 V 并令 $V\geqslant 0$ 且 $V1=1$;

2 While 未达到完全收敛 do

3 根据式(10),更新矩阵 W;

4 For i=1:n do

5 计算矩阵 A 和向量 b;

6 根据算法 2.2,更新 v_i;

7 End for

8 End while

2.3 模型的时间复杂度与收敛性分析

对于复杂度分析,本文只计算该模型时间复杂度的渐进上界 O,同时假定如下关系成立:$n>d\gg c$. 此时,更新矩阵 W 的时间复杂度为 $O(n^2c)$. 对于矩阵 V 的更新,由于本文采用按行更新的策略,因此只需考虑子问题的时间复杂度,即更新 v_i 的时间复杂度 $O(c^2)$,因此更新矩阵 V 的时间复杂度为 $O(nc^2)$. 假设算法 2.3 中算法循环次数为 t_1,算法 2.2 中算法循环次数为 t_2,则整个模型的时间复杂度为 $O(t_1(n^2c+t_2nc^2))$.

对于收敛性的分析,在本文模型的优化框架中,需要对 W 和 V 两个变量进行迭代更新. 在更新 W 时,由于 V 的固定,式(5)与 LCF 模型[8]相同,此时模型随 W 更新的收敛性也是非增的,具体证明步骤可参见文献[8]. 在固定 V 更新 W 时,问题(11)形式的问题已在文献[18]被证明收敛于一个全局最优解.

3 实验

本节将介绍了 FLCCF-G 模型在合成数据和真实数据上的测试过程和测试结果,并评估该模型的模糊性、可解释性与聚类效果.

3.1 合成数据实验

设置一个人工合成的拟合数据,该数据由 5 个不同类标签的满足高斯分布的数据组成,不同类数据之间比例为 0.4∶0.15∶0.15∶0.15∶0.15,共有 500 个数据样本. 如图 1 所示,数据集中存在 200 个蓝色数据点(中间位置),其余 4 个颜色(4 角位置)分别存在 75 个不同的数据点,为了使模糊的现象更明显,本文使蓝色数据点(中间位置)被其余 4 色数据点(4 角位置)包围,产生更多的模糊点供测试.

图 2 所示为拟合数据经过 FLCCF-G 训练后的聚类结果分布,对于所有训练得到的模糊点进行了标注,其中模糊点的判断标准是对第 i 个样本,若它的真实类别是第 k 类,而 $V(i,k)\neq 1$,则该点为模糊点. 图 3 所示为系数矩阵 V 的可视化形式,图 3 中,纵轴表示 500 个数据样本,横轴表示数据样本在矩阵 V 中所属的隶属度值. 下面结合图 1、图 2 和图 3 对系数矩阵的模糊性以及本文提出的通过增加系数矩阵行和为 1 的约束使得模型可以直接通过寻找系数矩阵每行最大值来判断样本所属聚类方法的合理性并进行解释. 我们将模糊点分为两类进行分析:①第一类模糊点为如 64 号、248 号、183 号、202 号、171 号、31 号的数据点,以 31 号数据点为例,由图 1 可知,其原始的标签颜色为蓝色(中间位置),然而,在实际数据集中,其更接近绿色数据样本(左下位置)的区域,因此遵循“物以类聚,人以群分”的思想,该点应当被聚类为绿色数据点. 由图 3 可知,FLCCF-G 对 31 号数据点训练得到的隶属度为[0,0,0,1,0],成功地将其分至最符合解释的类中,对于其余提及的模糊点,仍有此种解释性. ②第二类数据点为如 277 号、233 号、242 号、384 号数据点,这类点有一共同特点,即它们是各类簇的边界点. 由于此类点特征的特殊性,其对其他聚类的隶属度往往不为 0. 以 242 号点为例,通过图 3 可知其隶属度为[0,0.736,0,0,0.264],显然,该点并不完全属于青色类(右上位置),仍有小部分属于红色类(右下位置),这种结果证明了该模型在聚类时具有良好的模糊性.

通过上述分析,与一般的硬聚类算法相比,本文提出的 FLCCF-G 模型对位于不同类簇边界的点具有更好的可解释性.

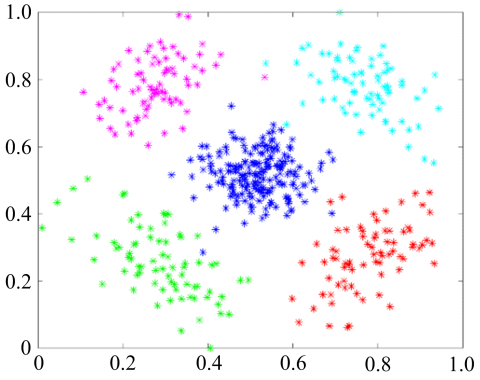


图 1 人工合成数据

Fig. 1 The visualization of Synthetic Gaussian distributed data

3.2 真实数据实验

与小节测试模型的模糊性与可解释性不同,实验着重测试了模型的聚类效果. 本小节设置了 4 个真实数据集和 3 个评估指标,通过与其他现有的聚类方法相互比较,验证 FLCCF-G 的聚类效果.

3.2.1 评估指标

本文通过设置 3 个评估指标来衡量模型的聚类效果,分别为:精度(accuracy, Acc),归一化互信息(Normalized Mutual Information, NMI),纯度(purity). 下面分别介绍这 3 种评估指标.

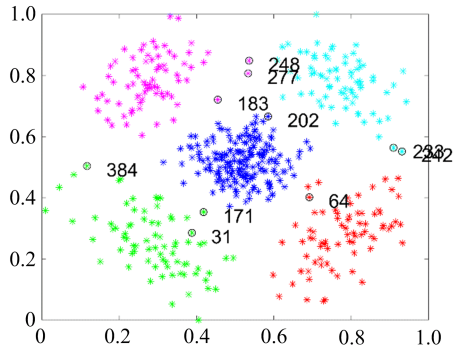


图 2　FLCCF-G 在人工合成数据集上的聚类结果

Fig. 2　The clustering results of FLCCF-G on synthetic Gaussian distributed data

精度指标表示为被正确聚类的数据样本占所有数据样本的比例. 精度表达式为

$$\mathrm{Acc}=\frac{\sum_{i=1}^{n}\delta(\mathrm{map}(r_i),l_i)}{n} \tag{28}$$

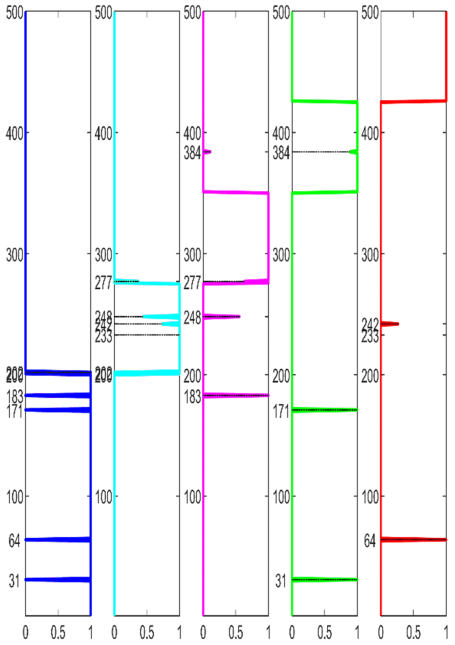


图 3　样本对不同类簇的隶属度

Fig. 3　The membership degree of each data point to different clusters

式中，n 表示为所有样本的个数，r_i 表示为第 i 个样本的聚类标签，l_i 表示为第 i 个样本的真实类簇标签，$\mathrm{map}(r_i)$ 表示为一个把 r_i 映射为与原始数据相对应的标签的函数，$\delta(a,b)$ 表示若 $a=b$ 则 $\delta(a,b)=1$，反之，$\delta(a,b)=0$.

归一化互信息指标用来衡量两个聚类之间的吻合度，表达式为

$$\mathrm{NMI}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{c}n_{ij}\log\frac{n_{ij}}{n_i\hat{n}_j}}{\sqrt{\left(\sum_{i=1}^{c}n_i\log\frac{n_i}{n}\right)\left(\sum_{j=1}^{c}\hat{n}_j\log\frac{\hat{n}_j}{n}\right)}} \tag{29}$$

式中，c 表示类的个数，n_i 为属于聚类 $C_i(1\leqslant i\leqslant c)$ 的样本数，n_j 为属于聚类 $L_j(1\leqslant j\leqslant c)$ 的样本数，n_{ij} 表示同时属于 C_i 与 L_j 的样本数.

纯度指标用来衡量聚类中样本与其他样本的关联度，表达式为

$$\mathrm{Purity}=\sum_{i=1}^{c}\frac{n_i}{n}P(S_i),P(S_i)=\frac{1}{n_i}\max_j(n_i^j) \tag{30}$$

式中，c 表示为类的个数，n 表示为所有样本的个数，n_i 表示为第 i 个簇中的样本个数，n_i^j 表示为在第 i 个簇中标签为 j 的样本个数.

3.2.2　数据集与实验设置

本实验设置了 4 个标准数据集以供测试，分别是 Yale，Caltech101，TDT2，dig1-10. Yale 数据集记录了 165 张 64 * 64 像素大小的人脸图片，共有 15 个不同的类簇；Caltech101 数据集记录了 101 张有不同标签的物体图片，每类有 40 至 800 张图片，每张图片的大小为 300 * 200 像素，本文选取了 10 类共 3044 张图片，并对每张图片的大小重构为 500 维；TDT2 数据集记录了 10212 个文档数据样本，本文选取了 1938 个类簇标签从 11 至 30 的数据样本，每个数据的大小为 36771 维；dig1-10 数据集记录了 1797 个从 0 至 9 的手写数字图片，每张图片的大小为 8 * 8 像素. 本文所使用的 4 个标准数据集的详细信息如表 1 所示.

表 1　实验使用的数据集

Tab. 1　The benchmark data sets for experiments

数据集	样本个数	样本特征数	类簇个数
Yale	165	1024	15
TDT2	1938	36771	20
dig1-10	1797	64	10
Caltech101	3044	500	10

为了更好地测试 FLCCF 和 FLCCF-G 模型的聚类效果，本文设置了 7 个不同的聚类方法用于比较，分别是 K 均值算法（K-means），模糊 C 均值算法（fuzzy C-means，FCM）[19]，非负矩阵分解模型（NMF），概念分解模型（CF），归一化切割（Normalized Cut，NCut）[20]，局部坐标概念分解模型（LCF），图正则化的局部坐标概念分解模型（LCF-G）.

针对不同的模型有如下的参数设置：对于 NCut，LCF-G，FLCCF-G 中的图结构，本文在生成图时设置近邻数为 5，并使用热核函数（Heatkernel）进行构造. 在 FCM 中，需要找到合适的阶才能得到良好的聚类效果，因此本文设置阶的范围为 $\{1.1,1.2,1.3,\cdots,2\}$，并使用取得最优聚类效果时的阶. 类似地，在 LCF，LCF-G，FLCCF，FLCCF-G 中，需要对正则化参数进行调参，基于网格搜索法，设置局部坐标编码项正则化参数 λ 的搜索范围为 $\{10-5,10-4,10-3,\cdots,10^5\}$，若模型涉及图结构，则图的正则化参数 λ 的范围也设置为 $\{10-5,10-4,10-3,\cdots,10^5\}$. 在本实验的 9 个模型中，NMF，CF，NCut，LCF，LCF-G 将执行 5 次

不同初始值的 K-means 模型进行后处理并记录最优结果，而 FCM，FLCCF，FLCCF-G 则直接通过系数矩阵 V 中的最大值获得最终聚类结果. 此外，为了更好地避免数据偶然性，实验中设置聚类数 c，对于不同的 c，从原始数据集中随机取出 20 个相应类标签种类数的样本集进行测试并取平均值.

3.2.3 实验结果分析

表 2～5 所示为各模型在 4 个真实数据集上的聚类效果评价表，每张表包含前述 9 个聚类方法在精度、归一化互信息和纯度 3 个指标下的评价结果，当聚类个数为数据集类簇个数时，表中对应数据为模型对整个数据集训练的最优结果；当聚类个数小于数据集类簇个数时，表中对应数据表示 20 个随机选择的测试集的平均值与标准差，最后取每个模型在所有不同取值下评价结果的平均值作为模型在该评价指标下的整体性能结果. 为了方便分析，本文对取不同值时 9 个聚类方法中最优结果进行了加粗处理.

从 4 张表格可以看出，FLCCF-G 模型的聚类效果在大多数情况下优于其他模型. 以 Caltech101 数据集为例，FLCCF-G 的整体性能相较 LCF-G，在精度、归一化互信息、纯度 3 个指标上分别提升了 4.49%、9.01%、3.95%，由此可见，模糊聚类通过系数矩阵 V 直接获得聚类结果的方法保留了聚类结果与原始数据之间关系的完整性，切实提高了聚类效果；比较 CF 模型与 FLCCF 模型可以看到，FLCCF 模型对 CF 模型有明显提升，分别为 10.38%、5.26%、3.26%，由此可见局部坐标编码技术对提升模型聚类效果有一定作用；同样地，通过比较 FLCCF 与 FLCCF-G 模型，发现图正则化对 FLCCF 模型可以提升 4.73%、1.78%、0.42%的聚类效果. 类似地，整体性能提升效果也可以在 TDT2、dig1-10、Yale 数据集上观察到. 在真实数据下，FLCCF-G 在 c 取个别值时聚类效果没有在所有方法中达到最优，结合表 2，表 3 和表 4 可以看出，在 Yale、Caltech101 和 TDT2 中，FLCCF-G 在聚类个数较少时性能完全优于其他模型，但随着聚类个数的增加，性能会略低于其他模型，但是其整体性能仍然最好，从表 5 可看出 FLCCF-G 在所有情况下都对其他方法的性能有所提升.

表 2 各模型在 Yale 数据集上的聚类效果

Tab. 2 Clustering results of different models on Yale data set

c	K-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					accuracy (%)				
2	80.23±16.37	66.14±14.93	67.50±13.93	82.05±11.84	87.59±12.08	90.83±5.35	71.36±13.36	82.73±10.70	**90.91±7.37**
4	58.64±9.17	58.64±7.89	56.59±9.06	66.36±10.57	69.98±10.20	**70.43±9.25**	54.77±9.47	62.27±7.89	69.55±8.58
6	46.52±6.95	48.64±7.92	46.29±6.15	51.97±8.37	55.79±6.34	**56.28±6.53**	46.36±6.49	50.23±6.03	54.62±6.60
8	45.17±4.99	47.10±4.99	43.64±6.96	49.43±4.97	53.15±4.12	53.14±3.15	45.23±4.26	49.94±4.70	**53.47±4.68**
10	45.59±5.75	44.14±4.38	43.64±4.93	49.18±4.51	50.24±4.16	49.52±5.07	46.27±5.79	46.77±5.26	**50.45±4.55**
12	42.77±3.62	40.23±4.06	36.70±4.03	44.36±3.69	**46.43±2.72**	44.26±3.67	41.10±3.63	42.46±3.20	46.36±3.01
15	38.79	37.58	40.00	39.39	45.35	41.50	43.03	46.06	**48.48**
Avg.	51.10	48.92	47.77	54.68	58.36	57.99	49.73	54.35	**59.12**
					normalized mutual information (%)				
2	42.05±30.70	16.49±24.74	17.30±24.41	41.50±27.51	55.98±28.54	59.49±18.24	21.74±24.98	40.85±25.52	**62.13±23.24**
4	40.42±11.93	40.78±11.57	36.46±13.37	46.29±10.55	50.24±13.05	50.26±13.08	35.84±14.91	42.95±14.33	**50.45±12.90**
					normalized mutual information (%)				
6	35.76±8.39	37.29±8.68	34.32±7.26	41.25±9.75	44.21±7.50	**44.28±8.49**	35.55±9.21	37.85±9.11	43.54±8.35
8	40.87±6.01	41.78±5.12	37.06±6.80	44.16±5.19	**47.53±4.46**	45.98±5.28	41.07±4.07	43.70±5.24	47.20±5.70
10	44.73±5.86	43.46±3.28	41.95±4.92	48.01±4.67	46.15±4.03	47.41±5.31	46.46±4.84	45.43±5.98	47.95±5.55
12	45.02±4.32	42.89±3.12	39.58±3.33	46.85±3.39	**48.92±3.25**	44.73±3.26	44.12±4.01	43.59±3.25	47.39±2.48
15	44.88	44.42	42.56	48.60	49.99	43.01	51.06	49.88	**51.23**
Avg.	41.96	38.16	35.60	45.24	49.00	47.88	39.41	43.46	**49.98**
					purity (%)				
2	80.23±16.37	66.14±14.93	67.50±13.93	82.05±11.84	87.59±12.08	90.83±5.35	73.86±12.15	82.73±10.70	**90.91±7.37**
4	59.77±9.15	59.66±8.24	57.27±9.26	67.05±9.71	69.98±10.20	**71.00±8.43**	68.52±10.79	62.73±8.26	69.55±8.58
6	48.11±6.56	49.55±7.39	47.20±5.88	53.11±8.14	55.25±6.07	57.25±7.09	**58.11±6.15**	50.68±6.11	54.92±6.88
8	46.99±4.85	48.35±4.77	44.20±6.46	50.68±5.49	53.63±4.51	53.48±3.51	**54.83±4.27**	49.94±4.70	53.58±4.68
10	46.77±5.39	45.73±4.02	44.55±4.69	49.86±4.51	49.53±4.58	50.08±4.28	**54.82±5.46**	47.14±5.16	50.55±4.53
12	44.05±3.66	41.89±3.38	38.48±3.51	45.53±3.63	47.78±2.73	45.29±3.19	**48.30±4.42**	43.03±3.28	46.82±3.15
15	41.21	38.79	40.00	40.00	46.76	41.17	**50.30**	48.48	49.09
Avg.	52.45	50.02	48.46	55.47	58.65	58.44	58.39	54.96	**59.35**

表 3　各模型在 Caltech101 数据集上的聚类效果

Tab. 3 Clustering results of different models on Caltech101 data set

c	*K*-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					accuracy (%)				
2	67.86±23.11	67.95±10.17	70.32±14.23	74.83±16.28	75.68±11.25	84.88±8.43	67.44±9.82	75.29±12.00	**87.07±7.97**
3	60.49±19.86	59.40±14.26	61.24±13.95	66.23±16.22	67.29±15.20	72.96±12.76	60.85±14.40	67.63±14.08	**77.08±12.26**
4	56.90±17.50	54.61±10.83	59.83±12.89	64.20±11.79	61.32±11.56	**69.54±8.46**	54.64±9.86	66.35±11.70	64.76±14.89
5	57.87±14.25	54.84±9.67	51.66±10.59	57.64±10.04	58.10±10.73	65.22±6.21	57.21±11.19	62.75±10.00	**73.85±8.50**
6	51.14±10.26	50.00±7.13	46.73±8.26	49.86±6.96	52.68±7.22	56.27±3.08	51.83±6.45	53.83±6.66	**62.74±5.73**
7	49.66±9.67	45.15±8.45	43.42±7.58	52.75±8.09	46.42±7.39	53.01±5.19	50.81±7.31	**54.38±7.67**	52.08±6.76
8	48.35±6.86	46.56±3.35	42.61±6.33	51.52±4.85	47.39±3.89	50.39±4.37	50.70±4.83	53.98±5.95	**57.52±2.94**
9	45.14±2.68	43.89±2.43	39.01±4.25	46.25±3.43	45.53±3.10	45.13±2.04	49.33±5.25	50.96±3.48	**51.54±3.15**
10	46.91	42.81	31.54	43.53	42.01	44.57	50.07	54.63	**55.72**
Avg.	53.81	51.69	49.60	56.31	55.16	60.22	54.76	59.98	**64.71**
					normalized mutual information (%)				
2	12.18±23.11	11.47±21.94	19.83±24.09	28.49±31.65	15.78±22.59	23.61±31.17	12.14±22.58	23.45±25.19	**35.01±31.59**
3	23.01±19.86	21.38±19.26	27.37±19.08	33.64±22.87	31.89±24.10	30.79±20.01	22.56±19.96	33.15±21.21	**39.13±21.78**
4	27.31±17.50	23.91±15.31	35.05±13.55	38.35±12.41	33.59±16.25	34.24±17.88	23.05±14.69	**39.23±13.21**	36.91±21.96
5	35.42±14.25	31.35±13.57	33.97±12.43	36.72±12.20	37.10±14.59	26.08±15.67	34.63±15.03	41.11±10.64	**41.23±12.67**
6	30.00±10.26	28.73±9.66	31.11±7.84	33.57±7.75	29.21±13.56	24.10±14.35	30.50±8.72	**34.45±8.38**	33.78±9.76
					normalized mutual information (%)				
7	35.10±9.67	30.04±9.95	33.25±8.16	**37.79±9.09**	33.98±9.59	24.51±14.22	35.91±8.85	37.00±8.30	33.86±9.10
8	36.39±6.86	35.18±3.52	34.40±4.63	38.77±4.30	36.02±6.45	26.37±13.95	37.63±4.04	38.90±4.67	**41.38±2.34**
9	35.80±2.68	33.69±3.90	32.32±3.07	36.51±2.98	36.54±2.26	31.53±11.04	37.72±3.56	37.43±2.36	**38.60±2.41**
10	36.66	32.38	28.84	35.83	34.99	37.16	**39.74**	38.75	39.61
Avg.	30.21	27.57	30.68	35.52	32.12	28.71	30.43	35.94	**37.72**
					purity (%)				
2	79.20±9.41	79.10±9.44	81.95±7.84	84.67±9.88	81.88±7.59	84.88±8.43	80.22±10.50	82.70±8.21	**87.60±7.76**
3	67.87±16.07	67.21±16.27	72.03±12.63	74.70±13.98	72.35±14.79	73.12±12.73	77.00±10.45	74.41±13.23	**77.93±11.86**
4	65.77±12.18	63.39±9.25	70.75±8.18	72.17±9.12	69.99±11.26	71.45±9.51	70.50±7.36	**72.82±9.49**	66.30±15.42
5	69.64±9.12	66.80±8.08	70.91±8.08	71.70±9.08	70.50±7.56	67.09±8.09	69.39±8.13	74.60±7.71	**74.86±7.57**
6	64.00±5.83	63.44±5.56	66.89±4.77	67.77±5.05	64.00±6.32	63.51±7.18	65.29±5.88	**69.00±4.70**	68.49±4.87
7	63.58±7.07	59.16±7.40	64.20±7.17	67.61±7.93	62.10±6.69	59.93±8.25	61.75±6.54	**67.03±6.55**	65.14±5.95
8	61.62±5.62	60.13±3.17	63.32±4.00	65.93±3.73	61.09±4.65	58.57±9.34	59.11±3.62	66.14±3.72	**67.21±3.65**
9	60.02±2.55	57.79±1.49	60.05±2.15	62.98±2.34	60.45±2.15	59.18±8.26	56.89±3.98	63.59±1.94	**64.11±1.24**
10	59.95	55.68	54.53	59.86	57.55	64.50	56.8	63.67	**66.10**
Avg.	65.74	63.63	67.18	69.71	66.66	66.91	66.33	70.44	**70.86**

表 4　各模型在 TDT2 数据集上的聚类效果

Tab. 4　Clustering results of different models on TDT2 data set

c	*K*-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					accuracy (%)				
6	94.68±7.69	94.32±7.57	86.88±9.62	93.19±7.61	95.46±7.05	90.34±9.14	93.39±10.96	99.52±0.30	**99.61±0.41**
8	92.51±7.00	85.47±9.32	82.46±12.61	85.27±5.31	91.59±6.59	89.15±6.43	90.66±9.52	99.49±0.40	**99.52±1.15**
10	91.36±6.30	85.74±7.95	85.69±8.30	82.68±6.55	90.00±6.23	85.73±6.27	89.35±7.01	99.44±0.36	**99.48±0.62**
12	87.74±4.72	80.19±5.93	75.55±7.39	79.79±6.34	86.12±5.54	80.92±5.43	88.61±8.19	97.18±3.61	**97.25±2.15**
14	86.79±4.63	83.01±5.61	77.82±6.54	78.66±4.59	83.59±5.26	78.69±5.13	88.12±8.31	92.20±3.50	**95.05±3.02**
16	87.98±3.65	79.64±6.80	74.98±6.18	75.71±4.18	82.54±4.57	77.84±4.72	86.49±4.06	92.19±2.95	**93.62±2.58**
18	85.51±4.36	73.61±5.80	73.10±3.89	77.72±5.55	82.10±4.32	77.23±4.15	84.19±5.08	91.29±2.79	**92.31±2.26**
20	85.76	80.19	80.13	70.95	79.59	74.21	85.50	**94.38**	92.16
Avg.	89.04	82.77	79.58	80.50	86.37	81.76	88.29	95.71	**96.13**

续表 4

c	*K*-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					normalized mutual information (%)				
6	94.15±5.47	93.64±6.18	90.00±5.70	93.38±5.27	94.36±6.23	89.88±7.54	93.63±7.38	98.51±0.85	**98.62±1.01**
8	93.94±4.71	89.07±6.22	87.92±7.28	88.64±3.04	91.78±6.49	88.99±5.06	92.71±6.17	**98.65±1.02**	98.41±2.39
10	92.84±4.03	88.41±6.43	88.24±5.29	87.39±4.56	91.00±5.26	88.43±4.89	92.30±4.26	**98.68±0.88**	98.32±1.12
12	90.31±3.38	86.45±3.59	83.22±4.72	85.07±4.35	89.52±3.31	85.06±4.21	91.82±4.52	96.79±2.59	**97.23±1.73**
14	90.24±3.28	88.02±2.76	84.52±3.54	84.65±3.45	87.31±3.76	84.79±4.30	91.00±4.58	92.13±2.95	**94.99±2.00**
16	91.46±2.28	86.10±3.76	83.51±3.66	83.94±3.60	86.77±2.45	84.25±3.57	90.28±2.82	92.74±2.82	**93.84±2.06**
18	89.68±2.44	82.66±3.46	82.67±2.37	84.27±3.82	87.21±3.29	84.10±3.26	88.96±2.08	92.23±1.69	**93.21±1.11**
20	89.17	85.27	85.83	80.75	85.22	85.81	89.00	**94.65**	92.78
Avg.	91.47	87.45	85.74	86.01	89.15	86.41	91.21	95.55	**95.93**
					purity (%)				
6	96.03±5.17	95.60±5.54	91.33±5.94	94.87±5.43	95.99±5.16	92.48±6.10	96.50±5.32	99.52±0.30	**99.61±0.41**
8	94.83±4.66	89.97±6.15	88.99±7.23	88.74±3.70	93.78±5.89	91.26±4.28	95.30±4.33	99.49±0.40	**99.52±1.15**
10	93.56±4.40	89.47±5.78	89.60±5.42	86.72±5.12	92.45±4.67	89.57±4.22	94.49±3.50	99.44±0.36	**99.48±0.62**
12	90.62±3.39	86.27±3.48	82.68±4.95	84.77±4.42	90.11±3.92	86.24±4.81	94.03±3.97	97.61±2.78	**97.73±1.97**
14	90.13±3.60	87.31±3.70	83.81±3.78	83.58±3.60	88.23±3.98	83.97±3.93	93.36±3.78	93.26±2.73	**95.62±2.33**
16	91.15±2.38	85.17±4.32	81.82±4.03	81.85±3.32	86.54±2.90	83.48±3.05	92.03±2.26	93.31±2.60	**94.61±2.05**
18	89.06±2.82	80.60±4.00	80.44±2.44	82.46±4.52	86.32±3.59	83.30±3.00	91.11±2.27	92.59±1.78	**93.64±1.59**
20	87.56	85.04	83.13	77.66	84.12	82.13	91.12	**95.41**	93.40
Avg.	91.62	87.43	85.23	85.08	89.69	86.55	93.49	96.33	**96.70**

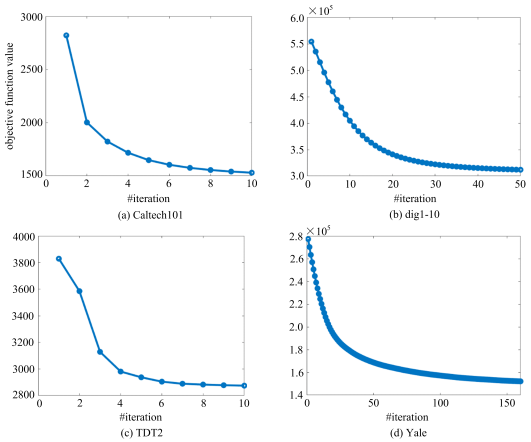


图 4 FLCCF-G 在 4 个标准数据集上的收敛曲线

Fig. 4 Convergence curves of FLCCF-G on four benchmark data sets

图 4 所示为 FLCCF-G 模型在 4 个标准数据集上的收敛曲线，由图 4 可以看出，模型在 Caltech101、TDT2 上最快收敛速度在 10 次迭代以内，在 dig1-10 上收敛速度在 50 次以内，在 Yale 上收敛速度在 150 次以内.

图 5 所示为 FLCCF-G 模型中正则化参数 λ 和 γ 的变化对聚类效果的影响，其中，λ 和 γ 的取值范围都为 $\{10^{-5},10^{-4},10^{-3},\cdots,10^{5}\}$，由图 5 可以看出，$\lambda$ 和 γ 的变化对聚类效果有一定影响. 在本实验中，当 λ 取较大的数并且 γ 取较小的数时聚类效果达到最优. 对于 λ 趋向于较大值而 γ 趋向于较小值的现象，可以看出式(4)中的第 3 项，即本文构造的模糊的局部坐标编码项对模型的性能影响强于图结构，对模型的聚类效果产生主要影响. 本文建议 λ 取值范围为 $[10^{1},10^{5}]$，γ 取值范围为 $[10^{-5},10^{-1}]$.

表 5 各模型在 dig1-10 数据集上的聚类效果

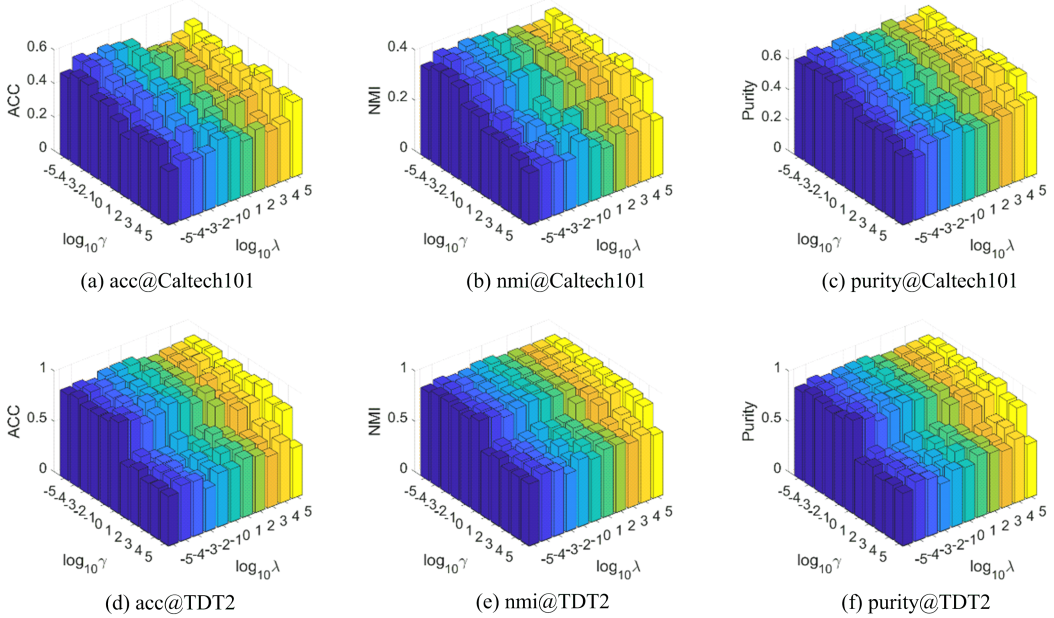
Tab. 5 Clustering results of different models on dig1-10 data set

c	*K*-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					accuracy (%)				
2	99.05±1.04	98.18±2.49	98.26±2.49	95.61±13.34	98.96±1.56	99.41±0.98	96.77±10.50	98.89±1.25	**99.70±0.33**
3	94.30±3.91	92.59±4.19	85.53±15.23	89.87±14.12	94.37±3.69	95.26±3.24	92.93±9.00	93.97±4.21	**96.02±2.41**
4	90.61±8.18	86.49±9.62	82.02±12.41	89.63±11.69	90.57±7.48	92.87±4.95	88.15±10.35	91.42±5.80	**94.58±3.58**
5	87.89±9.47	81.18±10.26	83.17±9.39	80.76±10.65	87.77±8.52	90.66±6.05	84.39±10.87	88.90±7.49	**92.67±4.16**
6	85.67±9.39	80.18±9.28	75.07±9.25	83.25±7.20	86.59±6.76	86.25±8.73	81.46±10.89	86.68±8.44	**91.24±5.29**
7	79.65±6.56	75.54±6.79	73.34±5.57	80.54±9.27	84.26±4.29	82.43±4.19	79.68±6.69	80.84±6.30	**88.39±2.05**
8	80.61±8.07	74.83±5.90	71.70±7.45	79.29±8.31	80.26±5.93	81.05±6.35	79.51±5.83	81.38±5.66	**87.09±4.79**
9	77.77±4.12	72.73±4.47	72.05±5.65	76.33±5.26	79.21±5.06	80.29±4.27	77.41±5.64	78.67±4.69	**83.21±3.43**
10	70.01	76.91	74.40	62.60	76.98	74.53	78.41	75.90	**82.43**
Avg.	85.06	82.07	79.50	81.99	86.55	86.97	84.30	86.29	**90.59**

表 5　各模型在 dig1-10 数据集上的聚类效果

Tab. 5 Clustering results of different models on dig1-10 data set

c	K-means	NMF	CF	NCut	LCF	LCF-G	FCM	FLCCF	FLCCF-G
					normalized mutual information (%)				
2	93.48±6.55	89.60±12.03	90.06±12.26	90.18±28.53	92.82±7.43	95.41±5.82	89.11±21.86	92.54±7.63	**97.55±3.06**
3	81.13±10.76	76.16±11.96	70.86±18.35	83.92±18.74	80.69±10.87	84.87±8.72	79.62±14.46	80.13±11.61	**85.15±8.11**
4	79.54±11.63	71.29±13.71	68.84±14.19	87.64±11.94	77.64±12.43	80.93±11.59	77.35±12.69	78.58±11.17	**84.03±8.09**
5	78.08±11.04	70.26±10.56	70.91±10.09	81.48±9.44	76.98±10.83	80.25±11.09	75.72±10.89	77.40±10.54	**82.63±9.92**
6	77.63±9.96	69.98±9.36	67.67±8.46	84.05±6.43	76.24±8.47	77.00±10.24	74.88±9.81	76.59±9.95	**82.08±8.16**
7	74.19±5.17	67.70±5.16	66.51±4.68	82.66±5.94	73.29±3.51	72.06±6.32	73.81±5.14	72.47±5.19	**78.24±4.29**
8	74.90±6.16	67.34±4.70	66.21±5.28	82.49±5.85	72.51±5.28	73.15±6.71	73.73±5.16	72.95±5.52	**77.39±6.55**
9	73.10±2.56	66.49±3.49	66.47±3.46	81.74±2.59	71.57±4.01	72.39±5.20	72.64±3.49	71.68±3.54	**76.04±3.11**
10	70.91	70.07	67.45	74.24	70.89	70.21	72.11	71.97	**75.03**
Avg.	78.11	72.10	70.55	83.16	76.96	78.47	76.55	77.15	**82.02**
					purity (%)				
2	99.05±1.04	98.18±2.49	98.26±2.49	95.61±13.34	98.96±1.56	99.41±0.98	97.86±5.68	98.89±1.25	**99.70±0.33**
3	94.30±3.91	92.59±4.19	86.68±13.02	90.12±13.64	94.37±3.69	95.26±3.24	93.51±6.61	93.97±4.21	**96.02±2.41**
4	90.61±8.18	86.53±9.55	83.21±10.97	90.00±11.11	90.57±7.48	92.87±4.95	90.11±7.73	91.42±5.80	**94.58±3.58**
5	88.16±8.95	82.10±9.05	83.99±7.97	83.47±8.43	87.99±8.77	90.66±6.05	88.32±6.54	89.03±7.16	**92.67±4.16**
6	86.12±8.94	80.72±8.86	77.73±7.59	84.73±6.07	86.79±7.54	86.35±8.16	86.00±7.32	86.93±8.01	**91.24±5.29**
7	81.10±5.60	77.25±5.32	75.87±4.16	83.68±7.15	84.24±3.81	86.14±4.22	84.69±4.09	81.58±5.56	**88.39±2.05**
8	81.76±6.81	76.28±4.86	74.64±5.45	81.98±6.47	80.37±5.34	82.08±5.82	83.40±4.55	81.90±5.22	**87.09±4.79**
9	78.65±2.92	74.03±3.63	74.15±4.60	80.33±3.73	79.15±4.21	80.45±4.31	82.03±2.71	79.49±4.07	**83.11±3.08**
10	72.95	77.69	75.74	70.28	78.00	75.39	81.02	77.69	**82.06**
Avg.	85.86	82.82	81.14	84.47	86.72	87.62	87.44	**86.77**	90.54



(a) acc@Caltech101　(b) nmi@Caltech101　(c) purity@Caltech101

(d) acc@TDT2　(e) nmi@TDT2　(f) purity@TDT2

图 5　FLCCF-G 在不同 λ 和 γ 下的聚类效果

Fig. 5　Clustering performance of FLCCF-G with different λ and γ

4　结论

本文针对传统的两步式的基于矩阵分解的聚类模型由于 K-means 方法的缺陷而影响聚类精度的问题，提出了一种图正则化的模糊局部坐标编码概念分解模型，构造了一个新的系数矩阵，基于该系数矩阵，不仅可以直接获得聚类结果，也可以获得各样本对不同聚类的隶属度，从而对聚类结果进行有效的解释. 本文对所提模型进行了验证测试，在人工拟合数据上验证了模型的模糊性与可解释性，在真实数据上，与其他 7 种模型相比较，获得了良好的聚类效果.

参考文献(References)

[1] XU W, LIU X, GONG Y. Document clustering based on non-negative matrix factorization[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2003: 267-273.

[2] WANG Y X, ZHANG Y J. Nonnegative matrix factorization: A comprehensive review [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(6): 1336-1353.

[3] HE Y C, LU H T, HUANG L, et al. Non-negative matrix factorization with pairwise constraints and graph Laplacian[J]. Neural Processing Letters, 2015, 42(1): 167-185.

[4] XU W, GONG Y. Document clustering by concept factorization [C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004: 202-209.

[5] CAI D, HE X, HAN J, et al. Graph regularized nonnegative matrix factorization for data representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(8): 1548-1560.

[6] CAI D, HE X, HAN J. Locally consistent concept factorization for document clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 23(6): 902-913.

[7] CHEN Y, ZHANG J, CAI D, et al. Nonnegative local coordinate factorization for image representation[J]. IEEE Transactions on Image Processing, 2012, 22(3): 969-979.

[8] LIU H, YANG Z, YANG J, et al. Local coordinate concept factorization for image representation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 25(6): 1071-1082.

[9] 祁宏宇,吴小俊,王士同,杨静宇. 一种协同的 FCPM 模糊聚类算法[J]. 模式识别与人工智能,2010,23(01): 120-126.

[10] 马文萍,黄媛媛,李豪,等. 基于粗糙集与差分免疫模糊聚类算法的图像分割[J]. 软件学报,2014,25(11): 2675-2689.

[11] 苏冬雪,吴小俊. 基于多特征模糊聚类的图像融合方法[J]. 计算机辅助设计与图形学学报,2006,18(6): 838-843.

[12] YANG B, FU X, SIDIROPOULOS N D. Learning from hidden traits: Joint factor analysis and latent clustering [J]. IEEE Transactions on Signal Processing, 2016, 65(1): 256-269.

[13] YU K, ZHANG T, GONG Y. Nonlinear learning using local coordinate coding [C]//Advances in Neural Information Processing Systems. 2009: 2223-2231.

[14] NIE F, SHI S J, LI X. Semi-supervised learning with auto-weighting feature and adaptive graph [J]. IEEE Transactions on Knowledge and Data Engineering, 2019.

[15] KYRILLIDIS A, BECKER S, CEVHER V, et al. Sparse projections onto the simplex[C]//International Conference on Machine Learning. 2013: 235-243.

[16] NIE F, YANG S, ZHANG R, et al. A general framework for auto-weighted feature selection via global redundancyminimization[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2428-2438.

[17] CHEN X, YUAN G, NIE F, et al. Semi-supervised feature selection via sparse rescaled linear square regression [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 165-176.

[18] NIE F, HUANG H, CAI X, et al. Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization[C]// Advances in neural information processing systems. 2010: 1813-1821.

[19] 沈浩,王士同. 按风格划分数据的模糊聚类算法[J]. 模式识别与人工智能,2019,32(3):204-213.

[20] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.