**Computer Science and Application** 计算机科学与应用, 2022, 12(4), 1031-1042
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124106

Hans汉斯

# 基于稀疏约束和对偶图正则化的受限概念分解算法及在数据表示中的应用

**翁宗慧，由从哲**

江苏理工学院，计算机工程学院，江苏 常州

收稿日期：2022年3月16日；录用日期：2022年4月15日；发布日期：2022年4月22日

## 摘　要

**概念分解算法(CF)是一种经典的数据表达方式，已经被广泛使用于机器视觉、模式识别等领域。基本的CF方法是一种无监督的学习算法，无法利用数据中存在的先验知识，没有考虑数据空间流形和特征空间流形的几何结构信息，同时分解结果也不具有稀疏性。为了解决以上缺陷，本文提出了一种基于稀疏约束和对偶图正则化的受限概念分解算法(DCCFS)。该算法通过保持样本数据空间和特征空间中内蕴的几何结构信息不变，使得算法可以更加有效提取数据的特征，增强了算法的数据表达能力；利用数据中天然存在的类别性息，增强算法的鉴别能力；添加$L_P$平滑范数提高了算法的稀疏性，使得分解结果更加准确、平滑。在COIL20图像数据集、PIE人脸数据集以及TDT2文本数据集上的聚类实验证明本文提出的DCCFS的聚类性能优于其他同类算法。**

## 关键词

**概念分解，标签信息，对偶图正则化，$L_P$平滑范数**

# Constrained Concept Factorization Based on Sparseness Constraints and Dual Graph Regularization for Data Representation

**Zonghui Weng, Congzhe You**

School of Computer Engineering, Jiangsu University of Technology, Changzhou Jiangsu

Received: Mar. 16th, 2022; accepted: Apr. 15th, 2022; published: Apr. 22nd, 2022

## Abstract

**Concept decomposition algorithm (CF) is a classical data representation that has been widely used**

文章引用: 翁宗慧, 由从哲. 基于稀疏约束和对偶图正则化的受限概念分解算法及在数据表示中的应用[J]. 计算机科学与应用, 2022, 12(4): 1031-1042. DOI: 10.12677/csa.2022.124106

**in machine vision, pattern recognition and other fields. In response to the fact that the basic CF method is an unsupervised learning algorithm that does not consider the geometric structure information and the class information of the samples present in the data space and feature space, and also does not take into account the sparsity of the decomposition results, this paper proposes a novel method named constrained concept factorization based on sparseness constraints and dual graph regularization for data representation (DCCFS) to overcome the above defects. This method constructs the geometric structure information in the sample data space and feature space unchanged, which extracts the features of the data more effectively and enhances the data expression ability of the algorithm; by using the natural label information in the data to enhance the identification ability of the algorithm; DCCFS adds the smooth sparse constraint to make the matrix factorization process more stable, smooth, which makes sure that the results are more accurate. The experimental results on COIL20 image dataset, PIE face dataset and TDT2 text dataset show that the DCCFS method can provide better representation for high-dimensional data and effectively improve the clustering performance.**

## Keywords

**Concept Decomposition Algorithm, Label Information, Dual Graph Regularized, $L_P$ Smoothness Constraint**



## 1. 引言

低秩矩阵分解是科学计算、数据挖掘和计算机视觉中最有用的工具之一。如何从高维数据中探索有效的数据低维表示方法至今仍是一个基础课题[1] [2] [3]。近些年来，低维数据表示技术受到了研究人员的关注，并被广泛应用于数据挖掘、计算机视觉和信息检索等领域。常用的低维数据表示技术是矩阵分解，其目的是获得两个或多个低维矩阵，并使其乘积尽可能接近原始数据矩阵。使用矩阵分解技术可以保留原始高维图像的特征信息、挖掘出图像蕴含的现在的语义结构，有利于机器学习、计算机视觉等领域的进一步发展。一些典型的矩阵分解算法有主成分分析(principal component analysis, PCA) [4]、矢量量化(vector quantization, VQ) [5]、奇异值分解(singular value decomposition, SVD) [6]、线性判别分析(Linear discriminant analysis, LDA) [7]、非负矩阵分解(Non-negative matrix factorization, NMF) [8]、概念分解(Concept Factorization, CF) [9]以及与深度学习技术相结合的一些列矩阵分解方法[10] [11]。在以上矩阵分解方法中，非负矩阵分解(NMF)和概念分解(CF)作为两个有效的数据降维工具，受到了研究人员的青睐，已被广泛应用于数据分析的任务中。

假设存在一个高维矩阵，NMF 的分解目标是将该高维矩阵转化为两个低维矩阵的乘积，使用两个低维矩阵(基矩阵和系数矩阵)的乘积来近似原始高维矩阵。同时，NMF 还有一个最基本的附加条件，即两个低维矩阵中的所有的元素都是非负，只进行加法运算而不进行减法运算，因此，NMF 是一个整体基于部分的图像表示方法，在心理和生理上具有较强的解释性。该附加条件使得 NMF 在处理图像、文档等聚类任务时优于 PCA、SVD、LDA 等传统数据降维算法。但是，NMF 无法处理数据中由噪声造成的含有负元素的数据矩阵，这使得 NMF 在处理实际问题时具有局限性。为了解决此类问题，Xu 等提出概念分解算法(CF)用于文本聚类以及图像表达。CF 不仅可以解决由噪声引起的矩阵中含负元素问题，无需考虑

正负，还可以处理高维数据，并可以很容易的对其进行核化，保留了 NMF 所有的优点。CF 算法的核心思想是建立一个聚类模型，其中每个聚类由数据点的线性组合表示，每个数据点由聚类中心的线性组合表示。

Cai 等在流形学习的基础上，利用流形学习的思想对基本的 NMF 和 CF 的模型中进行改进，提出了 GNMF [12]和 LCCF [13]。主要工作是对基本的 NMF 和 CF 模型中施加图正则化以保证了数据间的几何流形结构不变。Liu 等人[14]为了数据间存在的标签信息，将基本的 CF 模型改进为一种半监督的学习算法，提出了一种基于局部受限的概念分解(LC-CF)方法，LC-CF 对基本的 CF 模型施加局部正则化约束，通过提取与部分已知标签信息一致的图像信息，并使具有相同标签的数据点在低维空间中被映射至同一点上。Tang [15]在考虑数据流形的基础上，还考虑了特征流形，提出了 ODGNMF 算法，扩展了 GNMF 的学习能力。但是，以上的均从数据的先验知识方面对 NMF 或 CF 加以改进，没有考虑到算法的稀疏性。研究表明[16] [17] [18]，对算法施加稀疏约束，使得算法可以学习数据的稀疏表示，有效提高算法的鲁棒性，进一步提高分解结果稳定性、平滑性，使得算法具有更高的学习效率。

因此，本文提出了一种新的方法，称为基于稀疏约束和对偶图正则化的受限概念分解算法(Constrained Concept Factorization Based on Sparseness Constraints and Dual graph regularization, DCCFS)用于数据表示。DCCFS 不仅利用流形学习的思想考虑了数据空间流形和特征空间流形上的内部几何结构，而且还利用了无参数的标记样本的标签信息，将基本的无监督 CF 算法扩展为半监督算法，增强了算法的鉴别能力，最后，加入了 $L_P$ 平滑约束以考虑算法的稀疏性。为了得到该算法的迭代式，我们提出了一种基于乘法更新算法的高效优化算法来解决所提出的模型。在几个公共数据集上的实验表明，我们提出的 DCCFS 方法优于相关的最先进方法。

## 2. 相关工作

### 2.1. 非负矩阵分解(NMF)

假设有一原始非负矩阵 $X=\left[x_{ij}\right]\in\mathbb{R}^{M\times N}$ 被分解成两个低维矩阵，分别为 $U=\left[u_{ik}\right]\in\mathbb{R}^{M\times K}$，$V=\left[v_{jk}\right]\in\mathbb{R}^{N\times K}$，两个低维矩阵得乘积无限逼近原始矩阵，则 NMF 的表达式如下：

$$O_{NMF}=\sum_{i=1}^{N}\sum_{j=1}^{M}\left(X_{ij}-\left(UV^{\mathrm{T}}\right)_{ij}\right)^2=\left\|X-UV^{\mathrm{T}}\right\|_F^2 \quad \text{s.t. } U\geq 0, V\geq 0 \tag{2.1}$$

其中 $\|\cdot\|_F$ 代表范德蒙范数，$K\ll\min(M,N)$。$M$ 表示维度，$N$ 表示样本总数，$U$ 表示基矩阵，$V$ 表示系数矩阵也称作权重矩阵。使用交替乘性更新算法可得到 NMF 的迭代规则如下：

$$u_{ik}\leftarrow u_{ij}\frac{(XV)_{ij}}{\left(UV^{\mathrm{T}}V\right)_{ij}} \tag{2.2}$$

$$v_{jk}\leftarrow v_{ij}\frac{\left(X^{\mathrm{T}}U\right)_{ij}}{\left(VU^{\mathrm{T}}U\right)_{ij}} \tag{2.3}$$

由于 $K\ll\min(M,N)$，此时基矩阵 $U$ 和系数矩阵 $V$ 的秩远小于原始矩阵 $X$ 的秩，此时便实现了由高维矩阵到低维矩阵的降维。

### 2.2. 概念分解(CF)

假设有一矩阵 $X=\left[x_{ij}\right]\in\mathbb{R}^{M\times N}$，CF 的任务是找到基矩阵 $V=\left[v_{ij}\right]\in\mathbb{R}^{N\times K}$，系数矩阵

$W=\left[w_{ij}\right]\in\mathbb{R}^{N\times K}$。$v_j$ 称为基向量，是数据样本的非线性组合，$v_j$ 可表示如下：

$$v_j=\sum_i W_{ij}X_i\left(W_{ij}\geq 0\right) \tag{2.4}$$

其中 $W=\left[w_{ij}\right]\in\mathbb{R}^{N\times K}$。因此 CF 的目的是找到低秩矩阵 $W$ 和 $V$，使得 $W$ 和 $V$ 的乘积满足以下关系：

$$X\approx XWV^{\mathrm{T}} \tag{2.5}$$

使用 Frobenius 范数，CF 的目标函数可表示为：

$$\begin{aligned} &O_{CF}=\left\|X-XWV^{\mathrm{T}}\right\|_F^2 \\ &\text{s.t. } W>0,V>0 \end{aligned} \tag{2.6}$$

使用交替乘性更新算法求解 CF 的目标函数，CF 的更新规则可以表示为：

$$W_{ij}^{t+1}\leftarrow W_{ij}^{t}\frac{\left(KV\right)_{jk}}{\left(KWV^{\mathrm{T}}V\right)_{jk}} \tag{2.7}$$

$$V_{ij}^{t+1}\leftarrow V_{ij}^{t}\frac{\left(KW\right)_{ij}}{\left(VW^{\mathrm{T}}KW\right)_{ij}} \tag{2.8}$$

其中，$K=X^{\mathrm{T}}X$。

## 3. 基于稀疏约束和对偶图正则化的受限概念分解算法(DCCFS)

### 3.1. 引言

流形学习[19]表明，数据中存在有潜在的低维子流形，称之为数据流形。数据点间的几何拓扑结构都会在数据流形上显示。研究表明[20]，数据间的特征同样也分布在一个低维子流形上，可将之称为特征流形[19]，因此在对数据降维时，需要同时考虑到原始数据空间中存在的数据流性几何结构信息以及特征空间上的特征流形几何结构信息。通过构造数据图和特征图模型，刻画数据流形和特征流形的几何结构，可以保证数据在低维空间中仍然保持了高维空间中几何结构信息，保证算法学习的质量，因此该方法也被称为构造对偶图。

但是，该方法没有考虑到数据中潜在的类别信息。现实中的数据含有潜在的类别信息，人为的对这些类别信息添加标签，可以使之转化为标记信息。有效的利用样本中存在的标记信息将会提高算法的精度与数据表达能力，因此，在流形学习的基础上结合标记信息，将基本的 CF 算法从无监督转化为半监督是很有必要的。最后，考虑到算法的稀疏性，使用 $L_P$ 平滑约束可以滤除无用特征信息、防止过拟合、得到更加稀疏的特征向量，算法得到一个准确、平滑的解。

DCCFS 通过 $L_P$ 平滑约束提高算法的稀疏性，构造偶图得到样本数据空间和特征空间的流形结构信息，提高了算法的数据的挖掘能力，最后将标签信息转化为硬约束，揭示隐藏语义，提高算法的学习能力。

### 3.2. 构造对偶图正则项

研究表明，样本的数据不仅存在于非线性的低维流形上(数据流形)，数据的特征也存在于非线性流形上，称为特征流形。因此，可以构造数据流形和特征流形的最邻近图的几何结构来有效建模数据流形和特征流形的几何结构。

假设存在一个 $k$ 最近邻数据图，其顶点对应于 $\left\{X_1,\cdots,X_N\right\}$，使用 0~1 加权方案来构造 $k$ 最近邻图，

并定义数据权重矩阵如下：

$$S_{ij}^{V}=\begin{cases}1, \text{ if } x_j\in N_p\left(x_i\right); \\ 0, \quad \text{otherwise}\end{cases} \quad i,j=1,\cdots,n \tag{3.1}$$

其中，$N_p\left(x_i\right)$是 $x_i$ 的 $p$ 邻近集，数据图的拉普拉斯矩阵定义为：

$$L^V=D^V-S^V \tag{3.2}$$

其中，$D^V$ 是对角线矩阵，$D_{ii}^V=\sum_j S_{ij}^V$ 。

假设存在矩阵 $V$ 为待求的低维数据表示，$V=\left[v_1^{\mathrm{T}},v_2^{\mathrm{T}},\cdots,v_n^{\mathrm{T}}\right]\in R^{n\times k}$，通过前文定义的 $S^V$，可以使用 $\Re_1$ 衡量数据在低维表示的空间中的光滑程度，定义如下：

$$\begin{aligned}\Re_1&=\frac{1}{2}\sum_{i,j=1}^{n}\left\|v_i-v_j\right\|^2 S_{ij}^V \\ &=\sum_{i=1}^{n}v_i v_i^{\mathrm{T}}D_{ii}^V-\sum_{i,j=1}^{n}v_i v_j^{\mathrm{T}}S_{ij}^V \\ &=Tr\left(V^{\mathrm{T}}D^V V\right)-Tr\left(V^{\mathrm{T}}S^V V\right) \\ &=Tr\left(V^{\mathrm{T}}L_V V\right)\end{aligned} \tag{3.3}$$

其中，$Tr(\cdot)$表示矩阵的迹。为了尽可能使得数据集在低维空间中保持光滑，需要最小化 $\Re_1$，即如果两个数据点 $x_i$ 和 $x_j$ 距离较近，则两点在低维空间中对应的点 $v_i$ 和 $v_j$ 距离也是彼此接近的。此时通过最小化 $\Re_1$，保证高维空间中映射到低维空间中的点保持光滑。因此，式(3.3)也被称为数据图的拉普拉斯正则项。

同样的方法，构造一个 $p$ 最近邻点特征图，其顶点对应于$\left\{x_1^{\mathrm{T}},\cdots,x_m^{\mathrm{T}}\right\}$，然后使用 0~1 加权方案来构造特征矩阵，并定义数据权重矩阵如下：

$$S_{ij}^{U}=\begin{cases}1, \text{ if } x_j^{\mathrm{T}}\in M_p\left(x_i^{\mathrm{T}}\right); \\ 0, \quad \text{otherwise}\end{cases} \quad i,j=1,\cdots,m \tag{3.4}$$

其中，$M_p\left(x_i^{\mathrm{T}}\right)$是 $x_i^{\mathrm{T}}$ 的 $p$ 邻近集，特征图的拉普拉斯矩阵定义为：

$$L_U=D^U-S^U \tag{3.5}$$

其中，$D^U$ 是对角矩阵，$D_{ii}^U=\sum_j S_{ij}^U$ 。

令 $U=\left[u_1^{\mathrm{T}},u_2^{\mathrm{T}},\cdots,u_m^{\mathrm{T}}\right]\in R^{m\times k}$ 为待求的低维数据表示，使用 $\Re_2$ 衡量数据在低维表示的空间中的光滑程度，定义如下：

$$\begin{aligned}\Re_2&=\frac{1}{2}\sum_{i,j=1}^{m}\left\|u_i-u_j\right\|^2 S_{ij}^U \\ &=\sum_{i=1}^{m}u_i u_i^{\mathrm{T}}D_{ii}^U-\sum_{i,j=1}^{m}u_i u_j^{\mathrm{T}}S_{ij}^U \\ &=Tr\left(U^{\mathrm{T}}D^U U\right)-Tr\left(U^{\mathrm{T}}S^U U\right) \\ &=Tr\left(U^{\mathrm{T}}L_U U\right)\end{aligned} \tag{3.6}$$

### 3.3. 构造受限矩阵

CF 是一种无监督学习算法，没有考虑样本的类别信息，不能直接应用于标签信息可用的情况。相关研究表明将数据中的标签信息转化为硬约束可以有效提高算法的鉴别能力以及算法的聚类性能。假设数据集$\left\{x_i\right\}_{i=1}^{n}$ 中前 $l$ 个样本已标记，剩余 $n-l$ 个为未标记样本，数据集 $X$ 可被划为 $C$ 个集簇。前 $l$ 个数据点

都被标记为这些集簇中的一个。由此建立个指示矩阵 $M_{l\times c}$，假设 $x$ 被标记为第 $j$ 个集簇，则 $m_{ij}=1$ 否则 $m_{ij}=0$，此时矩阵 $A$ 如下：

$$A=\begin{pmatrix} M_{l\times c} & 0 \\ 0 & I_{n-l} \end{pmatrix} \tag{3.7}$$

其中 $I$ 是单位矩阵。为了充分利用标签信息，通过引入辅助矩阵 $Z$ 施加标签约束。系数矩阵可以重写如下：

$$V=AZ \tag{3.8}$$

### 3.4. DCCFS 目标函数

为了充分利用数据的先验知识，如数据流形和特征流形的几何结构信息、标签信息以及算法的稀疏性，本文提出了一种新的图像表示方法 DCCFS。DCCFS 尽可能地考虑了更多的先验知识，因此与传统的 CF 方法相比具有更强的表示能力。拟议 DCCFS 方法的目标函数如下所示：

$$O_F=\left\|X-XWZ^{\mathrm{T}}A^{\mathrm{T}}\right\|_F^2+\lambda Tr\left(Z^{\mathrm{T}}A^{\mathrm{T}}L_ZAZ\right)+\alpha Tr\left(W^{\mathrm{T}}L_WW\right)+2\mu\left\|W\right\|^p \tag{3.9}$$

其中 $L_Z=L_V=D^Z-S^Z$，$L_W=X^{\mathrm{T}}L_UX=X^{\mathrm{T}}\left(D^U-S^U\right)X=D^W-S^W$，$A$ 代表标签约束矩阵，而 $Z$ 表示一个辅助矩阵，系数矩阵 $V=AZ$。$\lambda\geq 0$、$\alpha\geq 0$、$\mu\geq 0$，三者是正则化参数，均是用来平衡重建误差。$P$ 的取值在 1 和 2 间，$L_V$ 和 $L_U$ 分别是数据图和特征图的拉普拉斯矩阵。

### 3.5. DCCFS 目标函数求解

很明显，DCCFS 的目标函数是非凸的，因此无法找到全局最优解，使用乘法迭代算法可解决此问题，具体做法为固定一个变量的情况下，对另一个变量的目标进行优化。另外，根据矩阵迹的特点，$Tr\left(A^{\mathrm{T}}\right)=Tr\left(A\right)$，$Tr\left(AB\right)=Tr\left(BA\right)$，此时可以实现 DCCF 模型的局部最小值。那么，公式(3.9)可以进一步改写如下。

$$\begin{aligned} O_F&=\left\|X-XWZ^{\mathrm{T}}A^{\mathrm{T}}\right\|_F^2+\lambda Tr\left(Z^{\mathrm{T}}A^{\mathrm{T}}L_ZAZ\right)+\alpha Tr\left(W^{\mathrm{T}}L_WW\right)+2\mu\left\|W\right\|^p \\ &=Tr\left(\left(X-XWZ^{\mathrm{T}}A^{\mathrm{T}}\right)^{\mathrm{T}}\left(X-WZ^{\mathrm{T}}A^{\mathrm{T}}\right)\right) \\ &\quad+\lambda Tr\left(Z^{\mathrm{T}}A^{\mathrm{T}}L_ZAZ\right)+\alpha Tr\left(W^{\mathrm{T}}L_WW\right)+2\mu\left\|W\right\|^p \\ &=Tr\left(K\right)-2Tr\left(WT^{\mathrm{T}}KAZ\right)+Tr\left(W^{\mathrm{T}}KWZ^{\mathrm{T}}A^{\mathrm{T}}AZ\right) \\ &\quad+\lambda Tr\left(Z^{\mathrm{T}}A^{\mathrm{T}}L_ZAZ\right)+\alpha Tr\left(W^{\mathrm{T}}L_WW\right)+2\mu\left\|W\right\|^p \end{aligned} \tag{3.10}$$

利用乘性迭代算法，简化式(1.18)。由于 $U\geq 0,V\geq 0$，假设 $\Psi=\left[\psi_{ik}\right]$ 和 $\phi=\left[\varphi_{jk}\right]$ 分别式关于 $U$ 和 $Z$ 的拉格朗日乘数，因此，式(3.10)的拉格朗日函数 $L$ 可写成如下：

$$L=O_F+Tr\left(\psi W^{\mathrm{T}}\right)+Tr\left(\varphi Z^{\mathrm{T}}\right) \tag{3.11}$$

$L$ 对 $U$ 和 $Z$ 的偏导数如下所示：

$$\frac{\partial\ell}{\partial W}=-2KAZ+2KWZ^{\mathrm{T}}A^{\mathrm{T}}AZ+2\alpha L_WW+\psi+2\mu PW^{P-1} \tag{3.12}$$

$$\frac{\partial\ell}{\partial Z}=-2A^{\mathrm{T}}KW+2A^{\mathrm{T}}AZW^{\mathrm{T}}KW+2\lambda A^{\mathrm{T}}L_ZAZ+\Phi \tag{3.13}$$

根据 KKT 条件， $\varphi U=0$ ， $\phi V=0$ ，等式(3.12)和(3.13)可被改写为：

$$w_{ij}^{t+1} \leftarrow w_{ij}^{t} \frac{\left(KAZ+\alpha S^{W} W\right)_{ij}}{\left(KWZ^{\mathrm{T}} A^{\mathrm{T}} AZ+\alpha D^{W} W+\mu P W^{P-1}\right)_{ij}} \tag{3.14}$$

$$z_{ij}^{t+1} \leftarrow z_{ij}^{t} \frac{\left(A^{\mathrm{T}} KW+\lambda A^{\mathrm{T}} S^{Z} AZ\right)_{ij}}{\left(A^{\mathrm{T}} AZW^{\mathrm{T}} KW+\lambda A^{\mathrm{T}} DAZ\right)_{ij}} \tag{3.15}$$

### 3.6. DCCFS 算法具体步骤

本文提出的 DCCFS 算法迭代步骤如下表 1 所示：

**Table 1.** Iterative steps of DCCFS algorithm
**表 1.** DCCFS 算法迭代步骤

| DCCFS 迭代步骤 |
| --- |
| **输入**：数据矩阵 $X=\left[x_1,\cdots,x_n\right]\in R^{m\times n}$ ，其中标记样本 $X_L=\left[x_1,\cdots,x_l\right]$ 为标记样本， $X_U=\left[x_{l+1},\cdots,x_n\right]$ 为未标记样本，最大迭代次数 maxIter，正则项参数 $\lambda,\alpha,\mu,P$ 。<br>**输出**：$U$ 和 $Z$；<br>1) 通过等式(3.7)构建受限矩阵 $A$；<br>2) 随机初始化基矩阵 $U$ 和辅助矩阵 $Z$，执行以下步骤：<br>a) 通过迭代式(3.14)更新 $W$；<br>b) 通过迭代式(3.15)更新 $Z$；<br>c) 若达到最大迭代次数 maxIter，算法终止，否则返回执行步骤(a)。<br>3) 计算准确度、归一化互信息。<br>**End** |

## 4. 数值实验

在本节中，为了评估所提出的 DCCFS 模型的有效性，我们在 PIE、COIL20、TDT2 三个公共数据集上进行了数值的实验。我们将我们提出的方法与其他方法包括 KM、NMF、GNMF、CF、LCCF、GCF 和 DCCFS 进行比较。在本实验中，将使用准确度(AC)和归一化互信息(NMI)作为所有方法的聚类评估指标。

### 4.1. 数据集介绍

本实验将在人脸数据集 PIE、物体数据集 COIL20 以及文本数据集 TDT2 上进行。在本实验中，我们从 PIE 数据集中选取不同对象的 11,554 幅图像，图像分辨率为 32 × 32，每张图片都是在不同的姿势和光照变化条件下收集的；COIL20 数据集这包含了 20 种不同物体的图像(如具鸭、杯子等)，每个物体 360 度，每 5 度拍摄一张图片，因此每个物体拥有 72 张图片，20 个不同的物体则拥有 1440 张图片，图像大小为 32 × 32；TDT2 文本数据集共包含了 56 类，10,021 个文档，在该数据集上本实验选取了每类数目大于 10 的样本用于聚类实验。

### 4.2. 实验介绍

在本实验中，我们从 PIE 数据集、COIL20 数据集、TDT2 数据集中随机选择 $K$ 类样本进行实验，$K$ 的取值均为 2 到 10。实验将所有挑选出来的实验图像混合放入集合 $X$ 中，对于每个 $K$ 值，所有方法运行 20 次取平均值，从而获取每种方法的准确度(AC)、归一化互信息(NMI)。实验结果如下表 2~4 所示：

**Table 2.** Clustering experimental results on PIE database
**表 2.** PIE 数据库上的聚类实验结果

| *K* | Accuracy/% | | | | | Normalized Mutual Information/% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | NMF | CF | LCCF | GCF | DCCFS | NMF | CF | LCCF | GCF | DCCFS |
| 2 | 69.27 | 73.23 | 74.85 | 78.23 | 80.45 | 70.13 | 72.23 | 73.34 | 74.48 | 75.46 |
| 3 | 65.45 | 72.62 | 73.75 | 78.65 | 81.75 | 63.30 | 65.46 | 67.73 | 70.94 | 73.73 |
| 4 | 63.51 | 71.72 | 73.24 | 77.37 | 78.27 | 66.17 | 65.73 | 67.43 | 71.66 | 73.47 |
| 5 | 62.34 | 73.16 | 75.68 | 76.82 | 79.34 | 65.89 | 67.66 | 67.56 | 72.83 | 75.84 |
| 6 | 61.29 | 74.35 | 76.76 | 79.46 | 81.83 | 60.54 | 63.46 | 65.25 | 69.63 | 74.57 |
| 7 | 59.47 | 71.47 | 72.36 | 76.84 | 78.23 | 61.29 | 63.56 | 66.72 | 68.47 | 77.37 |
| 8 | 59.81 | 72.72 | 74.59 | 75.52 | 78.73 | 58.00 | 62.41 | 65.34 | 66.84 | 75.43 |
| 9 | 59.62 | 74.48 | 76.97 | 75.36 | 78.81 | 60.02 | 63.00 | 65.63 | 66.47 | 73.75 |
| 10 | 58.68 | 71.46 | 72.39 | 74.46 | 77.49 | 58.46 | 62.26 | 64.76 | 65.27 | 74.33 |
| Avg | 62.16 | 72.80 | 74.51 | 76.97 | 79.43 | 62.64 | 65.09 | 67.08 | 69.62 | 74.88 |

**Table 3.** Clustering experimental results on COIL20 database
**表 3.** COIL20 数据库上的聚类实验结果

| *K* | Accuracy/% | | | | | Normalized Mutual Information/% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | NMF | CF | LCCF | GCF | DCCFS | NMF | CF | LCCF | GCF | DCCFS |
| 2 | 88.20 | 88.73 | 90.37 | 91.84 | 94.35 | 68.50 | 71.01 | 74.06 | 79.66 | 82.45 |
| 3 | 76.68 | 78.35 | 83.27 | 84.72 | 92.74 | 60.67 | 63.09 | 68.24 | 75.61 | 78.27 |
| 4 | 71.60 | 72.05 | 77.95 | 82.05 | 89.23 | 65.12 | 66.26 | 70.18 | 76.69 | 79.66 |
| 5 | 69.92 | 70.34 | 73.09 | 78.59 | 86.72 | 63.32 | 67.55 | 71.77 | 77.82 | 82.85 |
| 6 | 63.05 | 74.22 | 79.63 | 82.26 | 89.33 | 65.59 | 65.21 | 68.36 | 74.15 | 79.46 |
| 7 | 63.54 | 62.86 | 69.28 | 72.98 | 82.59 | 67.39 | 66.55 | 70.12 | 74.57 | 79.86 |
| 8 | 63.34 | 63.65 | 74.97 | 65.43 | 79.47 | 67.65 | 67.16 | 70.22 | 75.71 | 80.72 |
| 9 | 63.69 | 61.87 | 69.48 | 67.80 | 83.36 | 68.90 | 66.28 | 69.41 | 71.97 | 82.34 |
| 10 | 64.26 | 61.16 | 61.79 | 64.18 | 80.73 | 69.14 | 66.15 | 68.24 | 69.89 | 80.56 |
| Avg | 69.36 | 70.36 | 75.54 | 76.65 | 86.50 | 66.25 | 66.58 | 70.07 | 75.12 | 80.69 |

**Table 4.** Clustering experimental results on TDT2 database
**表 4.** TDT2 数据库上的聚类实验结果

| *K* | Accuracy/% | | | | | Normalized Mutual Information/% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | NMF | CF | LCCF | GCF | DCCFS | NMF | CF | LCCF | GCF | DCCFS |
| 2 | 82.34 | 82.18 | 91.23 | 93.19 | 95.34 | 60.54 | 62.93 | 77.86 | 79.87 | 82.74 |
| 3 | 75.54 | 76.35 | 85.46 | 88.98 | 92.83 | 58.32 | 64.98 | 71.45 | 74.64 | 81.57 |
| 4 | 72.64 | 74.79 | 83.84 | 86.37 | 91.47 | 63.12 | 64.24 | 72.69 | 73.25 | 81.24 |
| 5 | 65.78 | 70.13 | 80.12 | 85.25 | 90.93 | 57.47 | 59.56 | 65.70 | 69.67 | 80.26 |
| 6 | 67.56 | 71.21 | 78.09 | 82.64 | 88.47 | 62.31 | 62.73 | 68.26 | 70.48 | 78.19 |
| 7 | 65.78 | 67.50 | 76.17 | 77.78 | 84.28 | 61.34 | 66.26 | 67.14 | 68.89 | 77.34 |
| 8 | 63.37 | 63.03 | 71.73 | 72.09 | 82.53 | 60.28 | 61.53 | 63.56 | 65.67 | 75.63 |
| 9 | 63.85 | 64.27 | 72.34 | 70.00 | 80.01 | 62.61 | 63.35 | 64.49 | 65.15 | 74.47 |
| 10 | 62.12 | 65.44 | 66.04 | 69.28 | 78.66 | 62.56 | 64.12 | 63.04 | 65.37 | 73.80 |
| Avg | 68.78 | 70.54 | 78.34 | 80.62 | 87.17 | 60.95 | 63.30 | 68.24 | 70.33 | 78.36 |

### 4.3. 实验结果及实验结果分析

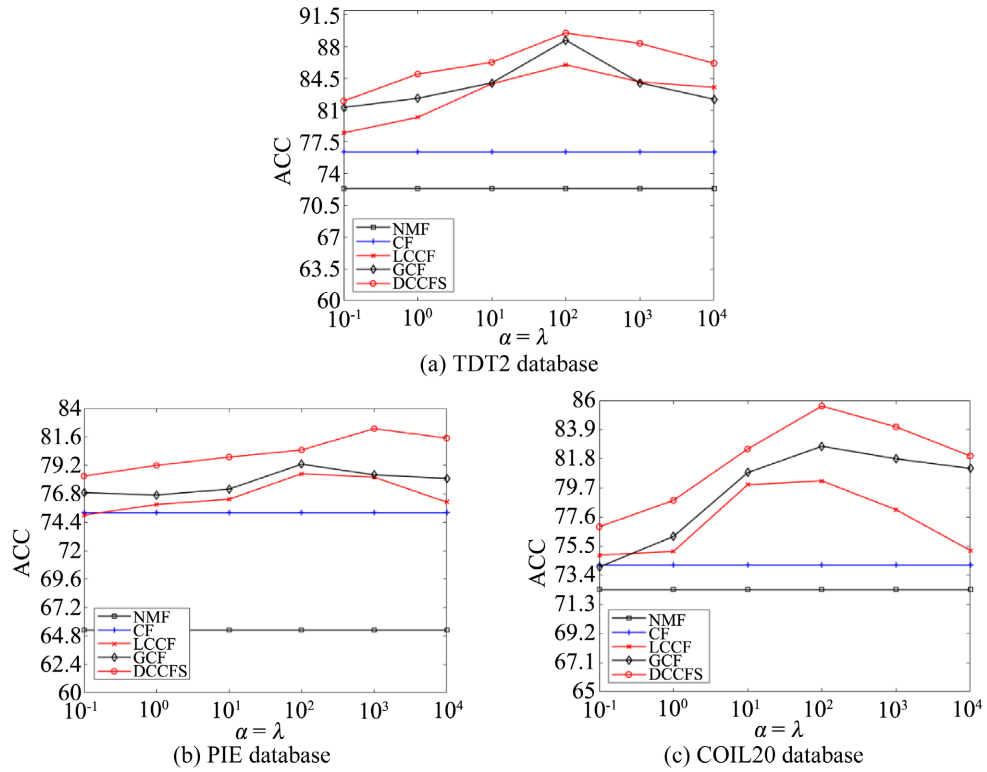
1) 表 2~4 分别为 PIE、COIL20、TDT2 三个数据集上的聚类实验。从表 2 中可以看出，DCCFS 的平均 AC 和人平均 NMI 比基本的 CF 分别高 6.63%和 8.98%；比 LCCF 分别高出 4.92%和 7.8%；比 GCF 分别高 2.46%和 5.26%。从表 3 和表 4 中也能得到类似的数据，充分说明了 DCCFS 的优越性。

2) 从表 2~4 可以看出，LCCF 的聚类性能比基本的 CF 算法优秀，这是由于 LCCF 采用了流形学习的思想，保证了数据内部的几何结构不变，因此实验所展现出的 LCCF 聚类性能较 CF 优越；GCF 在 LCCF 的基础上，不仅考虑了单边流形，同时考虑了双边流形，这说明了不仅保持原有数据空间的几何结构信息不变可以提高算法的学习质量，保持特征空间的几何结构信息同样也能增强效果，使得算法具有了更加优秀的数据挖掘能力，因此性能比 LCCF 更好。

3) 本章提出的 DCCFS 比 GCF 的聚类性能更好，主要原因是 DCCFS 考虑到了数据的标签信息以及算法分解的稀疏性。前者通过构造受限矩阵对标签信息施加硬约束，使得高维空间中属于同一类的样本可以投影到低维空间中的同一点上，将原本无监督的算法改造成了半监督的算法，提高了算法的鉴别能力和学习能力，后者通过人为添加 $L_p$ 平滑约束， $L_p$ 利用 $L_1$ 范数可以有效增加稀疏， $L_2$ 范数可以有效增加平滑的特点提高了算法的稀疏性，使得算法分解平稳、平滑、精确。

### 4.4. 参数选择

本文提出的 DCCFS 模型包含了四个参数 $\alpha,\lambda,\mu,P$ ，因此需要研究模型对四个参数的敏感性。在 CDNMFS 模型中 $\alpha,\lambda$ 是对偶图参数，为方便起见，这里只需二者的值相等，即 $\alpha=\lambda$ 。设置 $P$ 的取值为 $1<P<2$ ，设置 $\mu=(0,1,1,10,100,1000,2000)$ 。实验结果如图 1~3 所示，可以得到如下结论：



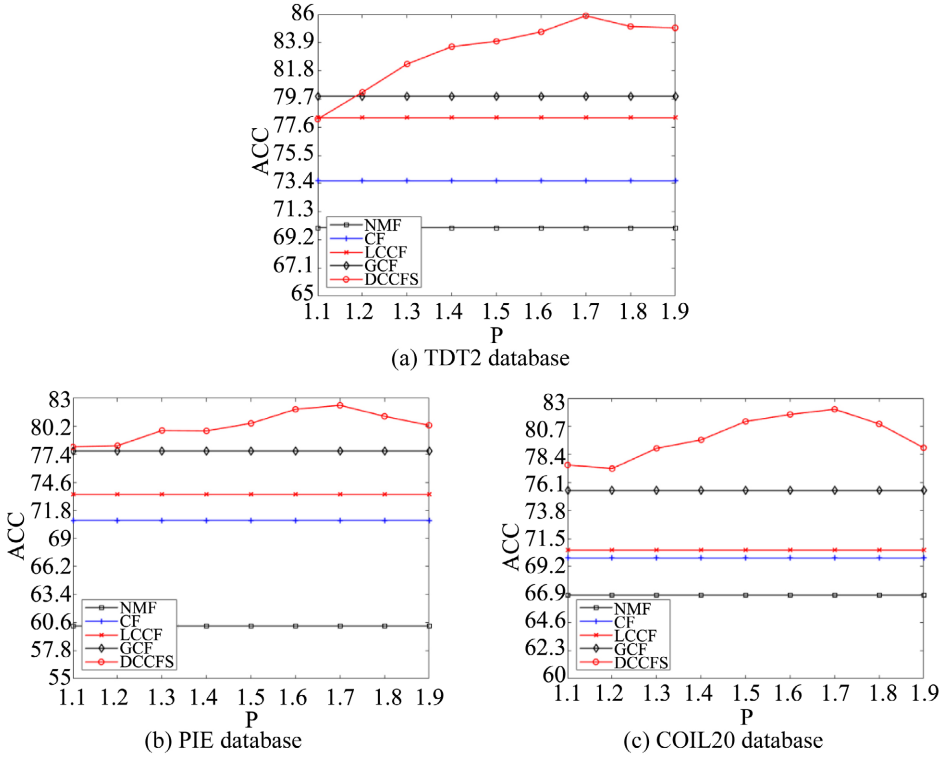

(a) TDT2 database

(b) PIE database

(c) COIL20 database

**Figure 1.** Evaluation experiment of regular term parameter $\alpha=\lambda$

**图 1.** 正则项参数 $\alpha=\lambda$ 的评估实验



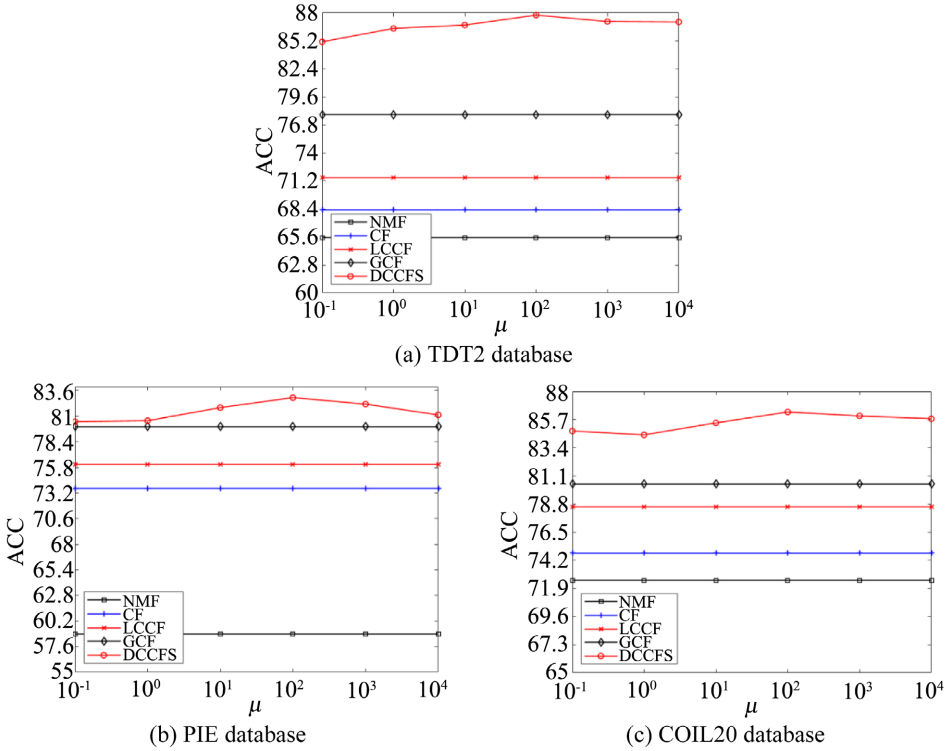

(a) TDT2 database

(b) PIE database

(c) COIL20 database

**Figure 2.** Evaluation experiment of regular term parameter $P$

**图 2.** 正则项参数 $P$ 的评估实验




(a) TDT2 database

(b) PIE database

(c) COIL20 database

**Figure 3.** Evaluation experiment of regular term parameter $\mu$

**图 3.** 正则项参数 $\mu$ 的评估实验

1) 本文提出的 DCCFS 对于正则项参数 $\alpha$ 和 $\lambda$ 具有较好的不敏感性。从图 1 中可以看出，当两正则项参数取 1~1000 时，DCCFS 可以取到较好的聚类性能，当超过 1000 时算法性能开始下降，但比其他同类算法优秀。

2) NMF、CF、LCCF、GCF 的算法不含平滑约束，因此模型中不含 $P$ 参数，所以这四种算法的聚类性能不随 $P$ 值的变化而变化。从图 2 中可以看出，DCCFS 的性能在大部分情况都很稳定且具有很高的性能，说明了 DCCFS 在 $P$ 取值为 1 至 2 之间时实验效果理想。

3) 同理，NMF、CF、LCCF、GCF 算法性能也不随 $\mu$ 的变化而变化。从图 3 中可以看出，无论 $\mu$ 取何值，DCCFS 的聚类性能均优于其他算法，充分证明了对算法施加稀疏约束的设想是合理的。

## 5. 结论

本文主要做了以下工作：针对传统的 CF 算法无法在低维空间中保持原始空间中的几何结构信息、无法有效利用样本中实际存在的类别信息以及没有足够的稀疏性，本文提出了一种基于稀疏约束和对偶图正则化的受限概念分解算法(DCCFS)。本方法结合流形学习的思想在算法进行图像表达时添加对偶图正则项，进一步利用样本中的先验知识，考虑数据空间与特征空间的几何结构信息，通过构造了对偶图，使得算法的数据表示能力进一步得到了提升，可以挖掘出数据中蕴含的本质特征，提高了算法的聚类性能；为了利用数据中存在的标签性息，通过构造受限矩阵对标签信息施加硬约束，使得高维空间中属于同一类的样本可以投影到低维空间中的同一点上，将原本无监督的算法改造成半监督的算法。最后考虑到 CF 算法对稀疏性考虑的不够，DCCFS 通过添加 $L_p$ 平滑约束，有效增加稀疏、有效增加平滑的特点使得算法可以学习数据的稀疏表示，使分解后的样本具有较高的稀疏性，算法分解平稳，分解结果平滑、精确。在三个公共数据集上的实验证明了本文提出 CDDFS 的有效性。

## 基金项目

国家自然科学青年基金，项目批准号：61902160，项目名称：基于潜在表示的不完整多视图子空间学习方法研究。

## 参考文献

[1] Wang, Q., He, X., Jiang, X., *et al*. (2020) Robust Bi-Stochastic Graph Regularized Matrix Factorization for Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 390-403. https://doi.org/10.1109/TPAMI.2020.3007673

[2] Peng, C., Zhang, Z., Kang, Z., *et al*. (2020) Two-Dimensional Semi-Nonnegative Matrix Factorization for Clustering. *Information Sciences*, **590**, 106-141. https://doi.org/10.1016/j.patcog.2020.107683

[3] Jiang, B., Ding, C. and Luo, B. (2018) Robust Data Representation Using Locally Linear Embedding Guided PCA. *Neurocomputing*, **275**, 523-532. https://doi.org/10.1016/j.neucom.2017.08.053

[4] Deutsch, H.P. (2004) Principle Component Analysis. Palgrave Macmillan, London. https://doi.org/10.1057/9781403946089_35

[5] Gray, R.M. (1990) Vector Quantization. In: *Readings in Speech Recognition*, Morgan Kaufmann Publishers, Burlington, 75-100. https://doi.org/10.1016/B978-0-08-051584-7.50011-5

[6] Strang, G. (2003) Introduction to Linear Algebra. Wellesley-Cambridge Press, Wellesley.

[7] Abramson, N., Braverman, D.J. and Sebestyen, G.S. (2006) Pattern Recognition and Machine Learning. *Publications of the American Statistical Association*, **103**, 886-887. https://doi.org/10.1198/jasa.2008.s236

[8] Lee, D. (2000) Algorithms for Non-Negative Matrix Factorization. *Proceedings of the* 13*th International Conference on Neural Information Processing Systems*, Denver, January 2000, 535-541.

[9] Xu, W. and Gong, Y. (2004) Document Clustering by Concept Factorization. *Proceedings* 27*th ACM*/*SIGIR*, Sheffield, 25-29 July 2004, 202-209. https://doi.org/10.1145/1008992.1009029

[10] Trigeorgis, G., Bousmalis, K., Zafeiriou, S. and Schuller, B.W. (2017) A Deep Matrix Factorization Method for

Learning Attribute Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 417-429. https://doi.org/10.1109/TPAMI.2016.2554555

[11] Su, X., Hu, L., You, Z., *et al*. (2021) A Deep Learning Method for Repurposing Antiviral Drugs against New Viruses via Multi-View Nonnegative Matrix Factorization and Its Application to SARS-CoV-2. *Briefings in Bioinformatics*, **23**, 1-15. https://doi.org/10.1093/bib/bbab526

[12] Cai, D., He, X.F., Han, J.W., *et al*. (2011) Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans on Pattern Analysis and Machine Intelligence*, **33**, 1548-1560. https://doi.org/10.1109/TPAMI.2010.231

[13] Cai, D., He, X.F., Han, J.W., *et al*. (2011) Locally Consistent Concept Factorization for Document Clustering. *IEEE Trans on Knowledge and Data Engineering*, **23**, 902-913. https://doi.org/10.1109/TKDE.2010.165

[14] Liu, H., Yang, G., Wu, Z., *et al*. (2014) Locality-Constrained Concept Factorization for Image Representation. *IEEE Transactions on Cybernetics*, **44**, 1214-1224. https://doi.org/10.1109/TCYB.2013.2287103

[15] Tang, J. and Wan, Z. (2021) Orthogonal Dual Graph-Regularized Nonnegative Matrix Factorization for Co-Clustering. *Journal of Scientific Computing*, **87**, Article No. 66. https://doi.org/10.1007/s10915-021-01489-w

[16] Ke, Q. and Kanade, T. (2005) Robust $L_1$ Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming. *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, Vol. 1, 739-746.

[17] Shen, B., Liu, B.D., Wang, Q.F., *et al*. (2014) Robust Nonnegative Matrix Factorization via L1 Norm Regularization by Multiplicative Updating Rules. In: *Proceedings of the International Conference on Image Processing*, IEEE Computer Society Press, Los Alamitos, 5282-5286. https://doi.org/10.1109/ICIP.2014.7026069

[18] Leng, C., Zhang, H., Cai, G., *et al*. (2019) Graph Regularized Lp Smooth Non-Negative Matrix Factorization for Data Representation. *IEEE/CAA Journal of Automatica Sinica*, **6**, 584-595. https://doi.org/10.1109/JAS.2019.1911417

[19] Seung, H. and Lee, D. (2000) The Manifold Ways of Perception. *Science*, **290**, 2268-2269. https://doi.org/10.1126/science.290.5500.2268

[20] Gu, Q. and Zhou, J. (2009) Co-Clustering on Manifolds. *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), Paris, 28 June-1 July 2009, 359-368. https://doi.org/10.1145/1557019.1557063