基于超图正则化受限的概念分解算法

李 雪* 赵春霞 舒振球 郭剑辉
(南京理工大学计算机科学与工程学院 南京 210094)

摘 要：针对概念分解(Concept Factorization, CF)算法没有同时考虑样本中存在的类别信息及数据间多元几何结构信息的问题，该文提出一种基于超图正则化受限的概念分解(Hyper-graph regularized Constrained Concept Factorization, HCCF)算法。HCCF 算法通过构建一个无向加权的拉普拉斯超图正则项，提取数据间的多元几何结构信息，克服了传统图模型只能表达数据间成对关系的缺陷；同时采用硬约束的方式使样本的类别信息在低维空间中保持一致，充分利用了标记样本的类别信息。该文采用乘性迭代的方法求解 HCCF 算法的目标函数并证明了其收敛性。在 TDT2 库、Reuters 库和 PIE 库上的实验结果表明，HCCF 算法提高了聚类的准确率和归一化互信息，验证了算法的有效性。

关键词：信息处理；概念分解；聚类；硬约束；超图；流形学习
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2015)03-0509-07
DOI: 10.11999/JEIT140799

Hyper-graph Regularized Constrained Concept Factorization Algorithm

Li Xue Zhao Chun-xia Shu Zhen-qiu Guo Jian-hui
(*College of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094, *China*)

Abstract: The Concept Factorization (CF) algorithm can not take into account the label information and the multi-relationship of samples simultaneously. In this paper, a novel algorithm called Hyper-graph regularized Constrained Concept Factorization (HCCF) is proposed, which extracts the multi-geometry information of samples by constructing an undirected weighted hyper-graph Laplacian regularize term, hence overcomes the deficiency that traditional graph model expresses pair-wise relationship only. Meanwhile, HCCF takes full advantage of the label information of labeled samples as hard constraints, and it preserves label consistent in low-dimensional space. The objective function of HCCF is solved by the iterative multiplicative updating algorithm and its convergence is also proved. The experimental results on TDT2, Reuters, and PIE data sets show that the proposed approach achieves better clustering performance in terms of accuracy and normalized mutual information, and the effectiveness of the proposed approach is verified.

Key words: Information processing; Concept Factorization(CF); Cluster; Hard constraints; Hyper-graph; Manifold learning

1 引言

目前，矩阵分解方法在文本聚类、数据挖掘和信息检索等方面起着重要作用[1]。基于矩阵分解的算法在处理海量文本问题时，通常把文本数据描述为高维空间中的一个点。通过有效的数据表示得到的样本数据可以在低维空间中保持原始样本在高维空间时的几何流形结构，提高算法的鉴别能力[2−4]。常用的矩阵分解算法包括奇异值分解(Singular Value Decomposition, SVD)，非负矩阵分解(Non-negative Matrix Factorization, NMF)[1]和概念分解(Concept Factorization, CF)[5]等。

文献[1]提出的 NMF 算法用两个非负的低秩矩阵的乘积逼近原始高维数据。针对 NMF 算法无法进行核化的问题，文献[5]提出了 CF 算法，其思想是每个聚类中心可用数据的线性组合来表示，而每个数据又可以用聚类中心的线性组合来表示。CF 算法通过最小化数据间的重构误差，找到线性系数的非负解。近年来，文献[6]提出一种半监督的鉴别概念分解(Discriminative Concept Factorization, DCF)算法，DCF 算法进行分类器训练时考虑了样本中存

2014-06-17 收到，2014-10-15 改回
国家自然科学基金(61272220, 61101197, 90820306)，中国博士后科学基金(2014M551599)，江苏省社会安全图像与视频理解重点实验室基金(30920130122006)和江苏省普通高校研究生科研创新计划项目(KYLX_0383)资助课题
*通信作者：李雪 lixue_angel@163.com

在的类别信息，但没有考虑数据间几何结构信息；文献[7]提出双图正则化的概念分解(Dual-graph regularized Concept Factorization, GCF)算法，GCF 同时考虑基向量和特征向量的流形结构，但没有考虑样本类别信息；文献[8]提出一种局部一致性概念分解(Locally Consistent Concept Factorization, LCCF)算法，该算法通过构造一个传统图模型，使其在低维空间中保持了数据原有的流形结构信息，但 GCF, LCCF 算法均为无监督的，并且忽略了数据的高阶信息，破坏了数据内在关联性；文献[9]提出超图正则化的非负矩阵分解(Hyper-graph regularized Non-negative Matrix Factorization, HNMF)算法，实验证明 HNMF 聚类效果明显高于传统图模型的 NMF 算法。上述算法均没有同时考虑样本的类别信息和数据间的高阶关系，从而影响了最终的聚类效果。

为解决上述算法没有同时考虑类别信息和数据间多元关系的缺陷，本文提出一种基于超图正则化受限的概念分解算法，超图正则化受限的概念分解(Hyper-graph regularized Constrained Concept Factorization, HCCF)算法采用硬约束[10]方式把样本类别信息添加到目标函数中，同时，用 k 个具有相似属性的数据子集构建超边，建立拉普拉斯超图正则项模型，提取数据间多元几何结构信息[11]。本文采用乘性迭代方法求解 HCCF 的目标函数，并证明算法的收敛性，实验结果表明了算法的有效性和准确性。

2 概念分解算法和超图简介

传统图模型在点与点之间建立连接关系的边，只考虑了数据间的成对关系，即二元关系。在实际应用中，数据分布是非常复杂的，因此，基于点对的传统图模型不能有效描述数据间的复杂关系。超图扩展了传统图模型中两个顶点组建边的构图方式，以具有某种相似属性的数据子集构建超边，从而可以有效刻画数据间的高阶关系。

给定一个非负矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{m\times n}$，$\boldsymbol{X}$ 的每一列表示一个样本，CF 算法要求每个基向量是样本向量 $\boldsymbol{x}_i$ 的非负线性组合，样本向量 $\boldsymbol{x}_i$ 是每个基向量的非负线性组合。CF 算法的目标函数表示为

$$\min_{W,V} O_{\mathrm{CF}}=\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}\right\|^2 \tag{1}$$

其中 $\|\cdot\|$ 表示 Frobenius 范数。令 $\boldsymbol{K}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$，更新规则如下：

$$w_{jk}^{t+1}\leftarrow w^t\frac{(\boldsymbol{K}\boldsymbol{V})_{jk}}{\left(\boldsymbol{K}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right)_{jk}},\ v_{jk}^{t+1}\leftarrow v_{jk}^t\frac{(\boldsymbol{K}\boldsymbol{W})_{jk}}{\left(\boldsymbol{V}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{W}\right)_{jk}} \tag{2}$$

3 超图正则化受限的概念分解(HCCF)算法

HCCF 算法结合流形学习和半监督学习的思想，采用 K 近邻[12](K-Nearest-Neighbor, KNN)方法选择 k 个顶点组成超边，构建超图[13,14]正则项保持数据的多元几何结构信息；同时把已标记样本的类别信息采用硬约束方式加入到 CF 算法的目标函数中，使得样本从高维空间映射到低维空间后类别信息仍保持一致。

3.1 构建超图正则项

超图 G 包含 N 个顶点，x_i 和 x_j 在高维空间中是近邻点，z_i 和 z_j 分别是低维空间中的近邻点，$\boldsymbol{V}$ 是 N 个顶点在低维空间的集合。文献[14]提出超边权重计算方法。定义 $\boldsymbol{D}_v$ 和 $\boldsymbol{D}_e$ 是对角矩阵，分别表示顶点的度和超边的度。数据映射到低维空间后，构建超图正则项 $\Re$：

$$\begin{aligned}\Re&=\frac{1}{2}\sum_{e\in E}\sum_{(i,j)\in e}\left\|z_i-z_j\right\|^2\frac{w(e)}{\delta(e)}\\&=\mathrm{Tr}\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D}_v\boldsymbol{V}\right)-\mathrm{Tr}\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{V}\right)=\mathrm{Tr}\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{h}}\boldsymbol{V}\right)\end{aligned} \tag{3}$$

$\mathrm{Tr}(\cdot)$ 表示矩阵的迹，$\boldsymbol{L}_{\mathrm{h}}$ 表示超图的拉普拉斯矩阵：$\boldsymbol{L}_{\mathrm{h}}=\left[l_{ij}\right]=\boldsymbol{D}_v-\boldsymbol{S}$，其中 $\boldsymbol{S}=\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\mathrm{T}}$。

为了尽可能使数据集在新的表示空间中保持光滑，需要最小化超图正则项 $\Re$。

3.2 构建 HCCF 算法的目标函数

数据集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$ 包含 c 类样本，前 d 个样本为标记样本，后 $n-d$ 个为未标记样本。分别定义标签矩阵 $\boldsymbol{C}^{d\times c}$ 和类别矩阵 $\boldsymbol{A}$ 为

$$c_{i,j}=\begin{cases}1, & \boldsymbol{x}_i\in c_j\\ 0, & \text{其它}\end{cases},\quad \boldsymbol{A}=\begin{pmatrix}\boldsymbol{C}^{d\times c} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}^{n-d}\end{pmatrix}。$$

其中矩阵 $\boldsymbol{I}^{n-d}$ 是大小为 $(n-d)\times(n-d)$ 维的单位矩阵。

在高维空间中，样本 $\boldsymbol{x}_i$ 的标签信息为 c_j，v_i 是 $\boldsymbol{x}_i$ 在低维空间中的表示，为确保 v_i 的标签信息仍为 c_j，添加辅助矩阵 $\boldsymbol{Z}$：

$$\boldsymbol{V}=\boldsymbol{A}\boldsymbol{Z} \tag{4}$$

为了同时考虑数据间多元几何结构信息和样本类别信息，HCCF 算法将超图正则项和样本类别信息同时添加到 CF 目标函数式(1)中，得到 HCCF 算法的目标函数为

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{Z}} O_{\mathrm{HCCF}}&=\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}(\boldsymbol{A}\boldsymbol{Z})^{\mathrm{T}}\right\|^2\\&\quad+\alpha\,\mathrm{Tr}\left((\boldsymbol{A}\boldsymbol{Z})^{\mathrm{T}}\boldsymbol{L}_{\mathrm{h}}(\boldsymbol{A}\boldsymbol{Z})\right)\end{aligned} \tag{5}$$

$\boldsymbol{W}$ 和 $\boldsymbol{Z}$ 均为非负矩阵，正则项参数 $\alpha\ge 0$。下面讨论 HCCF 算法目标函数的求解。

3.3 HCCF 目标函数求解

HCCF 的目标函数同时对于 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$ 来说是非凸函数，无法得到目标函数的全局最优解，但是对于单独的 $\boldsymbol{W}$ 或 $\boldsymbol{Z}$ 是凸函数，因此可以采用乘性迭代算法求解目标函数的局部最优解。根据矩阵性质：$\|\boldsymbol{A}\|^2=\mathrm{Tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$ ，$\mathrm{Tr}(\boldsymbol{AB})=\mathrm{Tr}(\boldsymbol{BA})$ ，$\mathrm{Tr}(\boldsymbol{A})=\mathrm{Tr}(\boldsymbol{A}^{\mathrm{T}})$，目标函数式(5)可化简为

$$\begin{aligned}O_{\mathrm{HCCF}}=&\mathrm{Tr}\left(\left(\boldsymbol{X}-\boldsymbol{XWZ}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\boldsymbol{X}-\boldsymbol{XWZ}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)\right)\\&+\alpha\,\mathrm{Tr}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{h}}\boldsymbol{AZ}\right)\\=&\mathrm{Tr}(\boldsymbol{K})-2\,\mathrm{Tr}\left(\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{K}\right)\\&+\mathrm{Tr}\left(\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{KWZ}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)\\&+\alpha\,\mathrm{Tr}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{h}}\boldsymbol{AZ}\right)\end{aligned}\tag{6}$$

分别对 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$ 求偏导，通过 Karush-Kuhn-Tucker 条件，得到 HCCF 算法的更新迭代规则：

$$w_{ij}^{t+1}\leftarrow w_{ij}^{t}\frac{(\boldsymbol{KAZ})_{ij}}{\left(\boldsymbol{KWZ}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZ}\right)_{ij}}\tag{7}$$

$$z_{ij}^{t+1}\leftarrow z_{ij}^{t}\frac{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{KW}+\alpha\boldsymbol{A}^{\mathrm{T}}\boldsymbol{SAZ}\right)_{ij}}{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{KW}+\alpha\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_{v}\boldsymbol{AZ}\right)_{ij}}\tag{8}$$

3.4 收敛性证明

上一小节对 HCCF 目标函数进行求解并求出更新规则，本节将证明目标函数式(5)在更新规则式(7)和式(8)下的迭代是收敛的。为证明收敛性，引入相关定义和引理。

定义 1　当函数 $G(x,x')$ 满足下列条件：$G(x,x')\geq F(x)$, $G(x,x)=F(x)$ 时，则称 $G(x,x')$ 是 $F(x)$ 的辅助函数。

引理 1　如果函数 G 是函数 F 的辅助函数，则 F 在下面条件下是非增的：

$$x^{(t+1)}=\arg\min_{x}G(x,x^{t})\tag{9}$$

对式(8)，定义 z_{ab} 是矩阵 $\boldsymbol{Z}$ 的元素，$F_{z_{ab}}$ 表示目标函数 O_{HCCF} 中与变量 z_{ab} 相关的函数，由于目标函数 O_{HCCF} 是逐个元素进行更新的，因此首先证明 $F_{z_{ab}}$ 在迭代式(8)下是非增的。

引理 2　函数

$$\begin{aligned}G\left(z,z_{ab}^{(t)}\right)=&F_{z_{ab}}\left(z_{ab}^{(t)}\right)+F'_{z_{ab}}\left(z_{ab}^{(t)}\right)\left(z-z_{ab}^{(t)}\right)\\&+\frac{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{KW}\right)_{ab}+\alpha\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_{v}\boldsymbol{AZ}\right)_{ab}}{z_{ab}^{(t)}}\\&\cdot\left(z-z_{ab}^{(t)}\right)^{2}\end{aligned}\tag{10}$$

是 $F_{z_{ab}}$ 的辅助函数。

证明　由定义 1 知：显然 $G(z,z)=F_{z_{ab}}(z)$；下面通过比较 $G(z,z_{ab}^{t})$ 和 $F_{z_{ab}}(z)$ 泰勒级数展开式的大小来证明 $G(z,z_{ab}^{(t)})\geq F_{z_{ab}}(z)$，即可证明引理 2。$F_{z_{ab}}(z)$ 的泰勒级数展开式：

$$\begin{aligned}F_{z_{ab}}(z)=&F_{z_{ab}}\left(z_{ab}^{(t)}\right)+F'_{z_{ab}}\left(z_{ab}^{(t)}\right)\left(z-z_{ab}^{(t)}\right)\\&+\frac{1}{2}F''_{z_{ab}}\left(z_{ab}^{(t)}\right)\left(z-z_{ab}^{(t)}\right)^{2}\end{aligned}\tag{11}$$

与式(10)相比，可得 $G(z,z_{ab}^{(t)})\geq F_{z_{ab}}(z)$。证毕

引理 3　函数

$$\begin{aligned}G\left(w,w_{ab}^{(t)}\right)=&F_{w_{ab}}\left(w_{ab}^{(t)}\right)+F'_{w_{ab}}\left(w_{ab}^{(t)}\right)\left(w-w_{ab}^{(t)}\right)\\&+\frac{\left(\boldsymbol{KWZ}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZ}\right)_{ab}}{w_{ab}^{(t)}}\left(w-w_{ab}^{(t)}\right)^{2}\end{aligned}\tag{12}$$

是 $F_{w_{ab}}$ 的辅助函数。

引理 3 的证明过程同引理 2 的证明，由于篇幅限制，此处具体证明参见引理 2。

定理 1　目标函数式(5)在更新迭代规则式(7)，式(8)下是非增的。当且仅当 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$ 是稳定点时，目标函数值是不变的。

证明　由引理 2 知：把式(10)代入式(9)得

$$\begin{aligned}z_{ab}^{(t+1)}=&z_{ab}^{(t)}-z_{ab}^{(t)}\frac{F'_{z_{ab}}\left(z_{ab}^{(t)}\right)}{2\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{KW}\right)_{ab}+2\alpha\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_{v}\boldsymbol{AZ}\right)_{ab}}\\=&z_{ab}^{(t)}\frac{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{KW}+\alpha\boldsymbol{A}^{\mathrm{T}}\boldsymbol{SAZ}\right)_{ab}}{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{AZW}^{\mathrm{T}}\boldsymbol{KW}+\alpha\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_{v}\boldsymbol{AZ}\right)_{ab}}\end{aligned}\tag{13}$$

由引理 1 知：因为 $G(z,z_{ab}^{t})$ 是 $F_{z_{ab}}(z)$ 的辅助函数，所以 $F_{z_{ab}}$ 在更新过程中是非增的。

同理，由引理 1 知：$G(w,w_{ab}^{t})$ 是 $F_{w_{ab}}$ 的辅助函数，所以 $F_{w_{ab}}$ 在更新过程中是非增的。定理 1 确保 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$ 分别在更新迭代规则式(7)和式(8)下目标函数式(5)是收敛的，并且可以找到局部最优解。

3.5 复杂度分析

算法的复杂度常用 O 表示，为了准确区分本文 HCCF 算法和其他对比算法的计算复杂度，本节使用算术运算的方法计算算法的复杂度。由更新迭代式(2)可得 CF 算法的复杂度 $O(n^2r)$，HCCF 算法需要计算核矩阵，复杂度为 $O(n^2m)$；HCCF 算法需要把具有相同属性的 k 个近邻点构建为一条超边，复杂度为 $O(n^2k)$。经过 t 次迭代更新后，HCCF 复杂度为 $O(tn^2r+n^2m+n^2k)$。表 1 总结了 HCCF 算法与 CF, LCCF, CCF 算法的复杂度计算，其中，n 为样本数目，m 是特征值数目，r 表示基向量个数，k 是构建边的近邻点数。

表 1 算法每次迭代的计算次数

算法	加法	乘法	除法	总计
CF	$4n^2r+4nr^2$	$4n^2r+4nr^2+2nr$	$2nr$	$O(n^2r)$
LCCF	$4n^2r+4nr^2+n(k+3)r$	$4n^2r+4nr^2+n(k+3)r$	$2nr$	$O(n^2r)$
CCF	$4n^2r+4nr^2+2nr$	$4n^2r+4nr^2+2nr$	$2nr$	$O(n^2r)$
HCCF	$4n^2r+4nr^2+n(k+3)r+2nr$	$4n^2r+4nr^2+n(k+3)r+2nr$	$2nr$	$O(n^2r)$

4 实验结果与分析

聚类实验中常用准确率(ACcuracy, AC)和归一化互信息(Normalized Mutual Information, NMI)[5]作为聚类算法的评价标准。本节重点评估本文HCCF 算法与 NMF[1], CF[5], CNMF[10], HNMF[9], LCCF[8], CCF[15]算法在 3 个数据集上的结果，进行比较分析，证明了算法的有效性。

4.1 在 TDT2 文本库上的实验

本文实验选取 TDT2 文本库中样本数目大于 10 的样本。表 2 描述的是在 TDT2 库上 7 种算法的平均 AC 和 NMI，其中，本文算法比 LCCF 算法平均 AC 和平均 NMI 分别提高了 5.04%和 6.46%，比传统 CF 算法的平均 AC 和平均 NMI 分别提高 12.67%和 15.53%。

4.2 在 Reuters 文本库上的实验

在 Reuters 文本库上的实验忽略属于多个类别的样本、选取样本数目大于 10 的类簇组成的实验数据集。表 3 描述的是在 Reuters 数据集上 7 种算法的实验结果。由表 3 可知，本文算法与 LCCF 相比，平均 AC 和 NMI 分别提高 15.14%和 9.07%。

表 2 在 TDT2 库上的聚类实验(%)

p	准确率							归一化互信息						
	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**
2	86.32	87.44	87.89	88.74	94.28	94.78	**96.92**	66.42	66.79	67.53	67.71	84.81	85.32	**86.78**
3	80.68	81.32	82.06	82.87	88.81	89.65	**92.58**	65.38	67.01	66.78	69.26	78.64	79.78	**84.42**
4	76.29	78.29	78.34	78.61	87.51	88.07	**90.07**	67.16	67.82	67.62	68.31	78.15	78.48	**83.16**
5	69.91	76.46	77.10	77.11	83.72	84.03	**86.14**	62.41	63.78	63.96	64.16	72.32	76.27	**80.01**
6	70.11	72.62	73.06	73.38	80.46	81.95	**84.06**	66.86	67.06	67.23	67.51	74.57	75.18	**81.14**
7	68.76	69.01	69.81	70.96	79.01	79.79	**82.79**	64.31	65.13	66.34	66.46	73.30	74.06	**80.78**
8	65.96	66.78	68.71	68.71	74.06	75.83	**80.41**	63.75	64.09	64.42	64.77	68.67	71.99	**78.82**
9	67.41	66.09	66.79	67.42	74.81	75.34	**82.28**	61.94	63.77	63.85	67.15	70.14	71.13	**76.77**
10	65.32	65.12	66.32	66.59	69.16	71.76	**81.92**	61.09	62.14	62.84	64.42	68.62	70.32	**75.46**
平均	72.31	73.68	74.45	74.93	81.31	82.36	**86.35**	64.37	65.29	65.62	66.64	74.36	75.84	**80.82**

表 3 在 Reuters 库上的聚类实验(%)

p	准确率							归一化互信息						
	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**
2	83.11	83.76	83.19	83.26	86.75	88.04	**93.38**	43.62	44.85	44.08	44.25	50.24	51.78	**55.68**
3	72.26	73.01	74.23	75.22	76.55	80.35	**91.77**	41.14	42.37	42.09	42.47	46.64	48.23	**49.06**
4	68.76	69.06	69.83	70.68	75.71	77.89	**88.18**	48.57	49.36	51.63	51.98	54.83	56.24	**58.72**
5	59.11	60.81	62.97	64.11	71.15	73.46	**85.14**	42.56	43.14	43.27	43.62	49.12	50.43	**54.41**
6	58.74	59.58	60.68	61.59	67.87	69.59	**83.69**	48.63	49.62	49.21	49.75	51.92	53.32	**57.22**
7	53.95	54.25	56.62	58.57	61.32	64.74	**81.78**	45.01	46.47	45.99	46.68	49.28	49.96	**53.84**
8	46.32	46.89	47.21	47.42	59.79	61.38	**78.91**	36.82	37.48	38.43	38.54	45.34	47.13	**50.47**
9	45.71	46.36	46.23	46.11	57.41	59.46	**76.12**	39.12	39.98	39.97	40.20	45.62	46.42	**51.22**
10	48.54	49.01	49.13	49.26	58.77	59.53	**72.66**	45.98	46.62	47.32	47.48	50.67	52.06	**54.17**
平均	59.61	60.30	61.12	61.84	68.37	70.49	**83.51**	43.49	43.49	44.67	45.00	49.30	50.62	**53.87**

4.3 在 PIE 人脸库上的实验

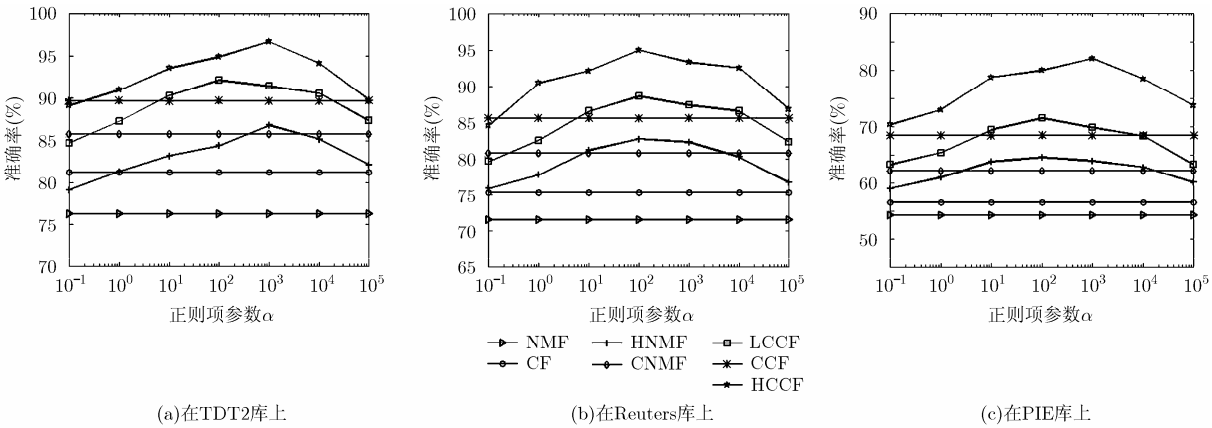
在 PIE 人脸库中，固定姿势和表情，在不同的照明条件下，选取 11554 张图像进行实验。实验结果由表 4 可知：本文算法与 CCF 算法相比，平均 AC 和 NMI 分别提高 6.40%和 5.41%。

4.4 参数设置

HCCF 模型中需要确定创建超边时所选择的 k 个近邻点和正则项参数 α，k 取值从 2~10，正则项参数 α 取值 $\{10^{-1},10^{0},10^{1},10^{2},10^{3},10^{4},10^{5}\}$，通过搜索不同参数值对实验结果的影响进行评估。图 1，图 2 分别表明当正则项参数 α 变化时对聚类准确率和归一化互信息的影响。图 3，图 4 表明当 α 取实验效果最优的条件下，搜索不同 k 值对聚类准确率和归一化互信息的影响。

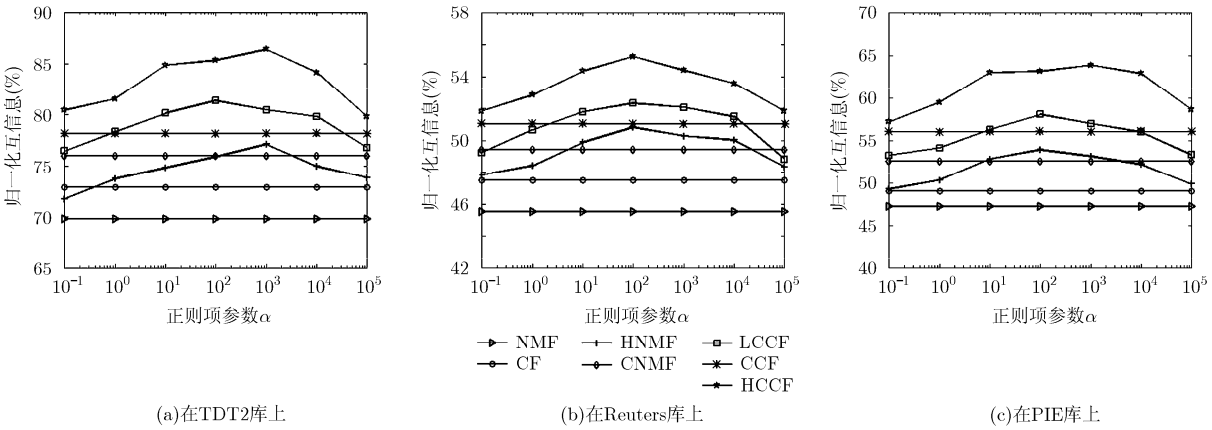
表 4 在 PIE 库上的聚类实验(%)

p	准确率							归一化互信息						
	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**	NMF	CF	HNMF	CNMF	LCCF	CCF	**HCCF**
2	56.14	57.31	57.40	57.46	63.76	66.59	**80.45**	46.62	48.96	53.13	55.42	60.32	62.31	**72.62**
3	53.31	56.62	57.96	58.32	60.78	64.49	**73.77**	44.32	47.31	52.67	53.78	58.88	60.04	**66.71**
4	54.12	56.04	55.11	55.38	61.32	63.52	**70.38**	45.56	46.76	50.66	51.02	54.45	56.76	**65.86**
5	52.59	54.61	56.36	58.62	61.78	62.36	**68.46**	44.16	44.14	46.56	49.69	50.42	52.67	**63.79**
6	50.88	53.79	54.78	57.33	59.57	61.86	**67.59**	42.25	43.77	44. 89	47.11	49.38	50.93	**52.78**
7	52.22	54.96	54.21	57.04	59.14	60.92	**65.26**	40.14	41.22	41.86	42.56	46.22	46.83	**47.96**
8	53.82	54.41	55.32	55.72	56.54	58.23	**63.78**	37.12	40.76	39.98	40.64	43.25	44.07	**45.64**
9	51.21	53.98	54.49	57.32	57.21	57.96	**62.14**	33.48	35.45	36.66	37.79	40.14	42.53	**46.71**
10	51.77	52.64	53.87	56.43	56.85	57.31	**58.96**	27.32	30.45	32.23	34.38	38.56	39.64	**42.42**
平均	52.90	54.93	55.50	57.07	59.66	61.47	**67.87**	40.10	42.09	42.30	45.82	49.07	50.64	**56.05**



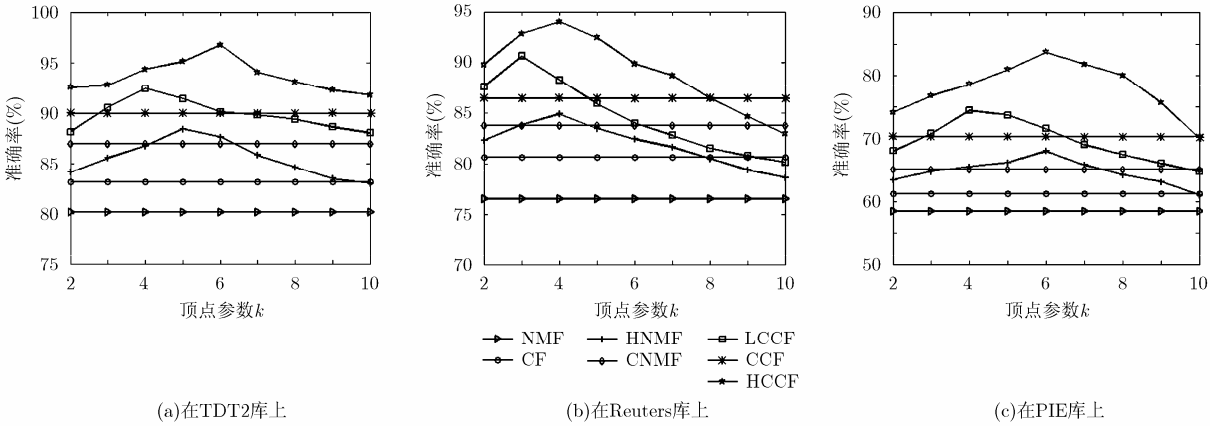
(a)在TDT2库上　(b)在Reuters库上　(c)在PIE库上

图 1 正则项参数 α 对 AC 的影响



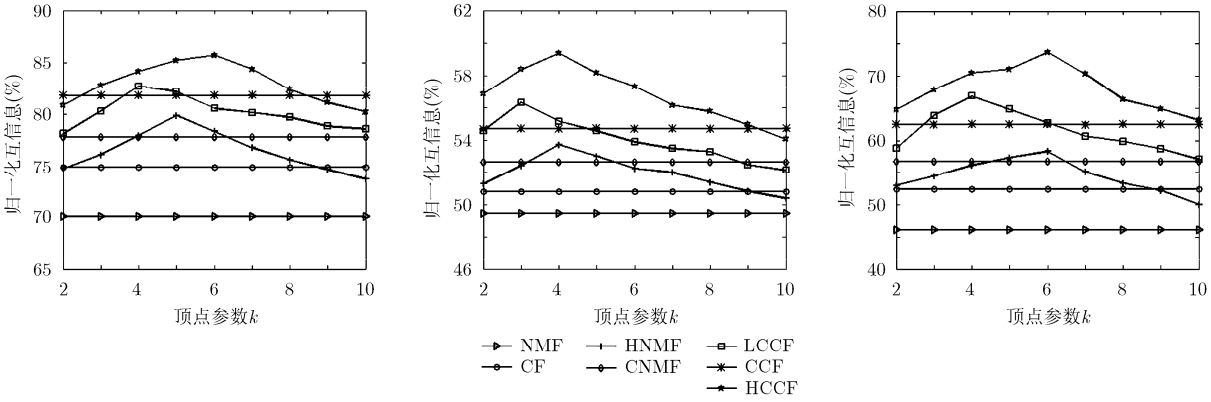
(a)在TDT2库上　(b)在Reuters库上　(c)在PIE库上

图 2 正则项参数 α 对 NMI 的影响



(a)在TDT2库上 (b)在Reuters库上 (c)在PIE库上

图 3 构建超边的顶点数 k 对 AC 的影响



(a)在TDT2库上 (b)在Reuters库上 (c)在PIE库上

图 4 构建超边的顶点数 k 对 NMI 的影响

4.5 结论分析

分析 4.3 节和 4.4 节实验结果可得如下结论：

(1)NMF, CF 算法没有考虑样本的类别信息，CNMF 和 CCF 算法分别对样本类别信息采用“硬约束”的方式，确保高维空间中属于同一类簇的样本在维数约简后仍属于同一类簇。与 NMF, CF 相比，添加了类别信息的 CNMF, CCF 算法的聚类 AC 和 NMI 在 3 个数据集上均优于 NMF, CF 算法，说明考虑样本的类别信息可以提高算法的鉴别能力，但是 CNMF, CCF 没有利用样本的几何结构信息；

(2)NMF, CF 算法没有考虑数据间的几何机构信息，HNMF 算法利用超图正则项获得数据间的多元几何结构信息，LCCF 算法在 CF 算法的目标函数中增加一个拉普拉斯图正则项，保持数据的几何流形结构信息，使得 HNMF 和 LCCF 算法的聚类 AC 和 NMI 在 3 个数据集上明显高于 NMF 和 CF 算法，说明考虑数据间潜在的流形结构可以提高算法的鉴别能力，但是 HNMF 和 LCCF 是无监督学习算法，忽略了样本中可能存在的类别信息；

(3)与 NMF, CF 算法相比，HNMF, LCCF 算法分别考虑了样本的几何结构信息，CNMF, CCF 算法分别考虑了样本的类别信息，从 3 个数据集实验结果知，CNMF 和 CCF 的平均 AC 和 NMI 分别优于 HNMF 和 LCCF 算法，说明聚类类别数小于 10 时，考虑样本的类别信息比考虑样本的几何结构信息更有利于提高算法的聚类准确率；

(4)本文 HCCF 算法同时考虑了样本的类别信息和样本的几何结构信息，从 3 个数据集的实验结果来看，HCCF 算法的平均 AC 和 NMI 优于其他对比算法，说明 HCCF 利用超图正则项保持了数据间高阶关系，因此 HCCF 具有更强的鉴别性；

(5)参数 k 大小与数据集样本分布有关，当样本分布相对分散时，较大的 k 值使得样本相似度降低，而当样本分布相对集中时，若参数 k 较小，使得具有相同结构信息的数据离散，故聚类准确率曲线先上升到最优值，如果 k 继续增大，会使聚类准确率下降；

(6)当参数 α 过大(大于 10000)或过小(小于 10)时，过分强调或忽略了样本的几何结构信息和类别

信息，使得聚类 AC 下降。当α在 10～10000 范围内变化时在 3 个数据库上均可取的较好结果，说明 HCCF 算法具有一定的鲁棒性。

5　结束语

根据流形学习和半监督学习的思想，本文提出了基于超图正则化受限的概念分解算法。HCCF 算法选择k个近邻点构建超边，计算每条超边上的权重，通过构建一个无向加权的拉普拉斯超图正则项，获得数据间固有的多元几何结构信息，解决传统图模型只能表达数据间成对关系的缺陷；同时，HCCF 算法采用硬约束的方式，使得已标记样本的类别信息在低维空间中保持一致，与软约束[16]方法相比，硬约束的半监督学习没有增加参数，降低了重构误差。HCCF 算法同时考虑了数据的高阶几何结构信息和样本的类别信息，增强了算法的鉴别能力。本文还给出了 HCCF 目标函数的求解方法、收敛性证明、算法复杂度分析以及参数选择分析，并在 TDT2, Reuters 和 PIE 数据集上进行实验，证明了 HCCF 算法的有效性。但是，HCCF 模型中参数k和超图正则项参数α需要通过区间搜索得到最优值，因此如何自适应地选择 k 个节点构建超边以及有效选择α是今后研究的重点方向之一。

参 考 文 献

[1] Xu Wei, Liu Xin, and Gong Yi-hong. Document clustering based on non-negative matrix factorization[C]. Annual ACM SIGIR Conference, Toronto, Canada, 2003: 267–273.

[2] Li Ze-chao, Liu Jing, and Lu Han-qing. Structure preserving non-negative matrix factorization for dimensionality reduction[J]. *Computer Vision and Image Understanding*, 2013, 117(9): 1175–1189.

[3] Yu Jun, Liu Dong-quan, Tao Da-cheng, *et al.*. Complex object correspondence construction in two-dimensional animation[J]. *IEEE Transactions on Image Processing*, 2011, 20(11): 3257–3269.

[4] Yu Jun, Tao Da-peng, Li Jonathan, *et al.*. Semantic preserving distance metric learning and applications[J]. *Information Sciences*, 2014, 281(10): 674–686.

[5] Xu Wei and Gong Yi-hong. Document clustering by concept factorization[C]. ACM SIGIR, Sheffield, UK, 2004: 202–209.

[6] Hua Wei and He Xiao-fei. Discriminative concept factorization for data representation[J]. *Neurocomputing*, 2011, 74(10): 3800–3807.

[7] Ye Jun and Jin Zhong. Dual-graph regularized concept factorization for clustering[J]. *Neurocomputing*, 2014, 138(3): 120–130.

[8] Cai. Deng, He Xiao-fei, and Han Jia-wei. Locally consistent concept factorization for document clustering[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2011, 23(6): 902–913.

[9] Zeng Kun, Yu Jun, Li Cui-hua, *et al.*. Image clustering by hyper-graph regularized non-negative matrix factorization[J]. *IEEE Transactions on Neurocomputing*, 2014, 138(22): 209–217

[10] Liu Hai-feng, Wu Zhao-hui, Li Xue-long, *et al.*. Constrained non-negative matrix factorization for image representation[J]. *IEEE Transctions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1299–1311.

[11] Yu Jun, Rui Yong, and Chen Bo. Exploiting Click Constraints and multiview features for image re-ranking[J]. *IEEE Transactions on Multimedia*, 2014, 16(1): 159–168.

[12] Yu Jun, Tao Da-cheng, and Wang Meng. Adaptive hypergraph learning and its application in image classification[J]. *IEEE Transactions on Image Processing*, 2012, 21(7): 3262–3272.

[13] Hong Chao-qun, Yu Jun, Li Jonathan, *et al.*. Multi-view hypergraph learning by patch alignment framework[J]. *IEEE Transctions on Neurocomputing*, 2013, 118(2013): 79–86.

[14] Huang Yu-chi, Liu Qing-shan, Zhang Shao-ting, *et al.*. Image retrieval via probabilistic hypergraph ranking[C]. Proceedings of the International Conference on Computer Vision and Pattern Recognition, San Francisco, 2010: 3376–3383.

[15] Liu Hai-feng, Yang Gen-mao, Wu Zhao-hui, *et al.*.Constrained concept factorization for image representation[J]. *IEEE Transactions on Cybernetics*, 2014, 44(7): 1214–1224.

[16] He Yang-cheng, Lu Hong-tao, Huang Lei, *et al.*. Non-negative matrix factorization with pair-wise constraints and graph Laplacian[J]. *Neural Processing Letters*, 2014, 12(7): 82–91.

李　雪：　女，1989 年生，博士生，研究方向为模式识别、图像处理等.

赵春霞：　女，1964 年生，教授，研究方向为模式识别、机器人控制、人工智能、图像处理等.

舒振球：　男，1985 年生，博士生，研究方向为机器学习、模式识别.

郭剑辉：　男，1983 年生，副教授，研究方向为机器学习、智能机器人、目标跟踪及数据融合.