第 43 卷 第 11 期 计 算 机 工 程 2017 年 11 月
Vol. 43 No. 11 Computer Engineering November 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)11-0134-06 文献标志码：A 中图分类号：TP391

基于 Hessian 正则化的多视图联合非负矩阵分解算法

王超锋，施 俊，吴金杰，朱 捷

（上海大学 通信与信息工程学院，上海 200444）

摘 要： 非负矩阵在表征多视图数据时没有考虑数据本身的流型结构，不能有效表达数据内部信息。为此，提出一种基于 Hessian 正则化的非负矩阵分解算法。利用 Hessian 泛函的 L2 模，保持样本局部拓扑结构，并扩展成基于 Hessian 正则化的联合非负矩阵分解算法，以对多视图数据进行变换。实验结果表明，基于 Hessian 正则化的非负矩阵分解算法和基于 Hessian 正则化的联合非负矩阵分解算法的聚类精度以及互信息值都有较大提高，2 种算法的数据变化性能都优于传统非负矩阵分解算法。

关键词： Hessian 正则化；回归模型；非负矩阵分解；多视图数据；聚类

中文引用格式： 王超锋，施 俊，吴金杰，等. 基于 Hessian 正则化的多视图联合非负矩阵分解算法[J]. 计算机工程，2017，43(11)：134-139.

英文引用格式： WANG Chaofeng，SHI Jun，WU Jinjie，et al. Hessian Regularization Based Factorization Algorithm Combining Multi-view and Non-negative Matrix[J]. Computer Engineering，2017，43(11)：134-139.

Hessian Regularization Based Factorization Algorithm Combining Multi-view and Non-negative Matrix

WANG Chaofeng，SHI Jun，WU Jinjie，ZHU Jie

（School of Communication and Information Engineering，Shanghai University，Shanghai 200444，China）

【Abstract】 Non-negative matrix does not consider the manifold of data when represents multi-view data，which results in the ineffective express of the data internal expression. In this paper，Hessian regularized Non-negative Matrix Factorization（NMF）is proposed. By using the L2 model of Hessian functional，the local topology of the sample is preserved and the algorithm is further extended into Hessian Regularized Joint Non-negative Matrix Factorization（HR-J-NMF）to work on multi-view data. Experimental results show that the Hessian regularized NMF and the HR-J-NMF have a great improvement in both clustering accuracy and mutual information value. The performance of the two algorithms is superior to that of the traditional NMF algorithm.

【Key words】 Hessian regularization；regression model；Non-negative Matrix Factorization（NMF）；multi-view data；clustering

DOI： 10.3969/j.issn.1000-3428.2017.11.022

0 概述

在图像和文本等数据的聚类处理中，矩阵分解技术已经成为数据表示的一种重要方法。由于非负矩阵分解（Non-negative Matrix Factorization，NMF）的分解结果中不出现负值，而且具有可解释性和明确的物理意义[1]，NMF 成为已经最有效的多维数据处理工具之一，成功应用于聚类等应用[2]。

原始的 NMF 回归模型中只采用 L2 正则项，没有考虑到数据本身的流形结构。由于图 Laplacian 矩阵能有效表达数据内在的几何信息，体现其低维流形[3]，因此文献[4]提出了基于 Laplacian 正则化的 NMF（Laplacian Regularized NMF，LR-NMF）算法，将图正则的流形学习与非负矩阵分解相结合，使得低维表示很好地保留了原始样本的近邻结构，其性能有了进一步的提高。

另一方面，现实世界数据往往是由多个视图或者表达组成，或者一个事物的多种描述构成了事物的多个视图，多视图的描述具备更强的数据特性表征能力。在数据聚类问题中，也存在着多视图聚类的问题。如携带疾病的基因数据在不同的片段中都会有所体现，文献检索数据通常也会存在不同类型

基金项目： 国家自然科学基金面上项目(61471231，81627804)。

作者简介： 王超锋(1992—)，男，硕士研究生，主研方向为机器学习；施 俊，教授；吴金杰、朱 捷，硕士研究生。

收稿日期： 2016-07-11 **修回日期：** 2016-11-04 **E-mail：** junshi@shu.edu.cn

的语言。在 NMF 算法成功应用聚类分析的基础上,针对多视图数据的 NMF 算法也已经被提出[5-7]。如文献[7]提出通过添加正则项的方式来整合不同视图的无标签信息,实现了联合 NMF 算法(J-NMF)。该算法用于多视图数据聚类,较原始 NNF 取得了更为优异的聚类效果。

值得一提的是,由于 Laplacian 正则化采用函数梯度的 L2 模,因此在做分类或者回归的时候,该项的极小化会导致函数趋向于常数函数,这不仅导致其不能较好地保持样本间局部拓扑结构,而且影响了其推测能力[8-9]。针对上述问题,一种新的正则算法——Hessian 正则化算法已经被提出[10]。Hessian 正则化利用了函数的 Hessian 泛函的 L2 模,其极小化会使得最优函数为流形上的线性函数,具有正则的零空间,而且可以使测地线函数随距离而变化,从而具有更好的保持样本局部拓扑结构的能力,能更准确地表征数据的内在局部几何特性[10-11]。Hessian 正则化不仅能恰当地表征由训练样本定义的区域内数据,而且能精确地预测区域边界外的数据点,特别是在小样本时具有很好的推断能力[10-11]。因此,基于 Hessian 正则化的半监督分类与回归、稀疏表达等算法表现出了较 Laplacian 正则化更为优异的性能[12-14]。而 Hessian 正则化也具有进一步应用于其他算法的可行性。

本文主要研究基于 Hessian 正则化的 NMF(Hessian Regularized NMF,HR-NMF)算法,提出一种针对多视图数据分解问题的基于 Hessian 正则化的联合 NMF 算法(HR-J-NMF)。

1　非负矩阵分解

1999 年文献[1]发表了 NMF,致力于分析非负的数据矩阵。NMF 的具体原理如下:给定矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]\in\mathbb{R}^{M\times N}$,其中,每一列是一个样本矩阵。NMF 目标是找到 2 个非负矩阵 $\boldsymbol{U}=[u_{ik}]\in\mathbb{R}^{M\times K}$ 和 $\boldsymbol{V}=[v_{jk}]\in\mathbb{R}^{N\times K}$,并且使得这两者的乘积可以很好地估计原始矩阵 $\boldsymbol{X}$:

$$\boldsymbol{X}\approx\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

其中,矩阵 $\boldsymbol{U}$ 称为系数矩阵;$\boldsymbol{V}$ 称为基矩阵。

为了减小计算误差,NMF 采用损失函数用来衡量估计的相似性,这里使用矩阵之间的欧式距离最小化(2 个矩阵的 F 范数的平方)[1]:

$$\boldsymbol{J}=\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|^2=\sum_{i,j}(x_{ij}-\sum_{k=1}^{K}u_{ik}v_{jk})^2 \tag{2}$$

文献[15]使用随机化方法初始化矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 值,这种初始化方法被后来的许多研究者应用。文献[16]通过研究非负矩阵的初始化问题,提出用 K-means 方法来初始化矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 的方法,能够加快算法收敛速度,并且能够提高算法结果的稳定性。

2　Hessian 正则化

Hessian 正则化项提供了简单建立映射和流型之间的关系,它由 Eells 能量得到。其具体过程为[10]:首先采用 k 近邻法或者其他方法构建局部连接图。设 $N_k(X_i)$ 为样本点 X_i 的 k 近领域数据集,在点集 $N_k(X_i)$ 上做 PCA 估计局部切空间 $\boldsymbol{T}_k(X_i)$。具体地,选取 X_i 的最近 k 个点,对这 $k+1$ 个点运用 PCA 估计出 m 维仿射子空间,其中,m 个特征向量作为该切空间的一组基。即 X_i 处的切空间矩阵 $\boldsymbol{T}_k(X_i)$ 就由该 m 个主特征向量所组成。在理想情况下,流形的采样足够稠密,主要特征值的数量应该等于 m。然而,真实数据通常是不够稠密的,导致数据的流形维数不能被自动地检测到。

在获得了提取过的切空间矩阵 $\boldsymbol{T}_k(X_i)$ 后,令 $\{u_r\}_{r=1}^m$ 为前 m 个归一化过的 PCA 特征向量,于是 $\{u_r\}_{r=1}^m$ 在 X_j 上的映射为:

$$x_r(X_j)=\langle u_r,X_j-X_i\rangle\frac{d_M(X_j,X_i)}{\sum_{r=1}^{m}\langle u_r,X_j-X_i\rangle^2} \tag{3}$$

其中,$\langle u_r,X_j-X_i\rangle$ 是不同的映射,X_j-X_i 是 $\boldsymbol{T}_k(X_i)$ 上的基,$\frac{d_M(X_j,X_i)}{\sum_{r=1}^{m}\langle u_r,X_j-X_i\rangle^2}$ 是归一化项,为了显示点 $X_j\in N_k(X_i)$ 到原始数据的距离等于流形矩阵 $\boldsymbol{M}$ 上 $X_j\sim X_i$ 的距离 $d_M(X_j,X_i)$,而:

$$\|x(X_j)\|^2=\sum_{r=1}^{m}x_r(X_i)^2=d_M(X_j,X_i)^2 \tag{4}$$

Eells 能量 $S_{\mathrm{Eells}},(f)$ 可以写成一个实函数 f:

$M\to R$

$$S_{\mathrm{Eells}},(f)=\int_M\|\nabla_a\nabla_b f\|^2_{T_x^*M\otimes T_x^*M}\mathrm{d}V(x) \tag{5}$$

其中,$\mathrm{d}V(x)$ 是体积元素,$\nabla_a\nabla_b f$ 是 f 的二阶协方差导数,也就是映射 $f(x)$。

在中心点为 p 的正常坐标系 x_r 中:

$$\nabla_a\nabla_b f\big|_p=\sum_{r,s=1}^{m}\frac{\partial^2 f}{\partial x_r\partial x_s}\bigg|_p\mathrm{d}x_a^r\otimes\mathrm{d}x_a^s\Rightarrow$$

$$\|\nabla_a\nabla_b f\|^2_{T_P^*M\otimes T_P^*M}=\sum_{r,s=1}^{m}\left(\frac{\partial^2 f}{\partial x_r\partial x_s}\right)^2 \tag{6}$$

在 p 点,二阶协方差导数的范数就是 f 在标准坐标系中的 Hessian 的 Frobenius 范数。由此最终得到的泛函数就是 Hessian 正则化项 $S_{\mathrm{Hess}}(f)$。

Hessian 能量的估计也可以应用正则坐标的二阶导数的简化形式[11]。设 $N_k(X_i)$ 为样本点 X_i 的 k 个近领样本集合,运算 H 估计 f 在 X_i 处的 Hessian 可近似为:

$$\frac{\partial^2 f}{\partial x_r\partial x_s}\bigg|_{x_i}\approx\sum_{j=1}^{k}H_{rsj}^{(i)}f(X_j) \tag{7}$$

其中,$f=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ 在特征空间。

式(7)的求解可以通过对一个二阶多项式使用线性最小二乘法求解得到,标准线性最小二乘被用

于修正 f 的二阶泰勒展开式。于是通过式(7)得到理想的形式 $H_{rsj}^{(i)}$。Hessian 的 Frobenius 范数的估计由此给出:

$$\| \nabla_a \nabla_b f \|^2 \approx \sum_{r,s=1}^{m} \left(\sum_{\alpha=1}^{k} H_{rs\alpha}^{(i)} f_\alpha \right)^2 = \sum_{\alpha,\beta=1}^{k} f_\alpha f_\beta B_{\alpha\beta}^{(i)} \tag{8}$$

其中,$B_{\alpha\beta}^{(i)} = \sum_{r,s=1}^{m} H_{rs\alpha}^{(i)} H_{rs\beta}^{(i)}$ 完成所有的估计 Hessian 能量。

为了修正这个多项式,使用标准线性最小二乘:

$$\min_{\omega \in \mathbb{R}^P} \sum_{j=1}^{k} \left(\left(f(X_j) - f(X_i) \right) - (\varphi_\omega)_j \right)^2 \tag{9}$$

其中,$\boldsymbol{\varphi} \in \mathbb{R}^{k \times P}$ 是自定义矩阵,$P = m + m(m+1)/2$,$\omega \in \mathbb{R}^P$ 的解是 $\omega = \boldsymbol{\Phi}^{\dagger} f, f \in \mathbb{R}^k, f_j = f(X_j)$,且 $X_j \in N_k(X_i)$,$\boldsymbol{\Phi}^{\dagger}$ 表示 $\boldsymbol{\Phi}$ 的伪逆。

给出 Hessian 在点 X_i 处的欧式聚类表达式:

$$\hat{S}_{\text{Hess}}(f) = \sum_{i=1}^{n} \sum_{r,s=1}^{m} \left(\frac{\partial^2 f}{\partial x_r x_s} \bigg| X_i \right)^2 = \sum_{i=1}^{n} \sum_{\alpha \in N_k(X_i)} \sum_{\beta \in N_k(X_i)} f_\alpha f_\beta \boldsymbol{B}_{\alpha\beta}^{(i)} \tag{10}$$

其中,矩阵 $\boldsymbol{B}$ 是把矩阵 $\boldsymbol{B}^{(i)}$ 相加的累计矩阵,$\hat{S}_{\text{Hess}}(f)$ 是所有数据点的和,n 代表所有有标签和无标签数据的数量。因此,得出一个更简洁的目标函数:

$$\hat{S}_{\text{Hess}}(f) = \langle f, \boldsymbol{B} f \rangle \tag{11}$$

虽然 Hessian 正则化的推导过程复杂,但是最终的解的形式相对简洁,这使得 Hessian 正则化易于推广使用。有关 Hessian 正则化的详细内容请参考文献[10]。

3 基于 Hessian 正则化的多视图联合 NMF

3.1 基于 Hessian 正则化的 NMF

由于 Hessian 正则化的优异性能,本文提出基于 HR-NMF 算法。具体算法实现如下:

采用 Hessian 正则化 $\hat{S}_{\text{Hess}}(f) = \langle f, \boldsymbol{B} f \rangle$ 替代 Laplacian 正则化,那么 HR-NMF 的目标函数变为:

$$\boldsymbol{J} = \| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \|^2 + \lambda \mathrm{Tr}(\boldsymbol{V}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{V}) \tag{12}$$

为使得迭代过程中保证每次迭代过程的输出矩阵为负,需要把矩阵 $\boldsymbol{B}$ 拆分成正项和负项 $\boldsymbol{B} = \boldsymbol{B}_+ - \boldsymbol{B}_-$。使用拉格朗日方法求解目标函数得到以下新的迭代过程:

$$u_{i,k} \leftarrow u_{i,k} \frac{(\boldsymbol{X}\boldsymbol{V})_{i,k}}{(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})_{i,k}}$$

$$v_{i,k} \leftarrow v_{i,k} \frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + \lambda \boldsymbol{B}_- \boldsymbol{V})_{j,k}}{(\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \lambda \boldsymbol{B}_+ \boldsymbol{V})_{j,k}} \tag{13}$$

以上就是 HR-NMF 的求解过程。值得注意的是,乘性迭代规则的隐含条件是迭代式两边的值都需要为正值,这样才能保证最终迭代步骤的非负性质。所以,在 HR-NMF 的矩阵 $\boldsymbol{B}$ 需要拆分成 $\boldsymbol{B}_+$ 和 $\boldsymbol{B}_-$ 2 项。

3.2 多视图联合 NMF 介绍

为了提高 NMF 处理多视图数据的性能,在文献[7]提出的 J-NMF 的基础上,本文提出了针对多视图数据分解的基于 Hessian 正则化的联合 NMF(HR-J-NMF)算法。其动机在于:对于基于 J-NMF 的聚类而言,具有不同视图的同一个数据被聚为同一类的概率是很高的,因此,进行矩阵分解时,从不同视图学习获得的系数矩阵通过软正则到一个共同矩阵,而这个共同矩阵反映了不同视图所共享的潜在聚类结构[7];再考虑到同一数据的多个视图可能存在各自的流形结构,那么通过对每一个视图进行 Hessian 正则化可以充分挖掘其内在的流形结构;因此,将 Hessian 正则化与 J-NMF 相结合,在保持各视图样本的内在流形结构的同时,有效整合多个视图之间潜在的公共信息,则可以得到更合理有效的分解矩阵。

文献[7]提出的联合 NMF 的主要原理如下所示。

假设给定 n_v 个视图的数据:$\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(nv)}\}$,对于每个视图都做一次矩阵分解 $\boldsymbol{X}^{(v)} = \boldsymbol{U}^{(v)} (\boldsymbol{V}^{(v)})^{\mathrm{T}}$。对于标准的 NMF 来说,系数矩阵 $\boldsymbol{V}_j^{(v)}$ 是第 j 个数据点在新的基 $\boldsymbol{U}^{(v)}$ 上的低质逼近。对于不同的视图来说数据点数是相同的,但是数据的特征维数可以不同,对应的矩阵 $\boldsymbol{U}^{(v)}$ 行数允许不同,矩阵 $\boldsymbol{V}^{(v)}$ 的大小却是相同的。应用欧式距离到损失函数得:

$$D(\boldsymbol{V}^{(v)}, \boldsymbol{V}^*) = \| \boldsymbol{V}^{(v)} - \boldsymbol{V}^* \|_F^2 \tag{14}$$

基于 J-NMF,本文所提出的 HR-J-NMF 将优化以下联合最小化问题:

$$\begin{aligned} \boldsymbol{J} &= \sum_{v=1}^{n_v} \| \boldsymbol{X}^{(v)} - \boldsymbol{U}^{(v)} (\boldsymbol{V}^{(v)})^{\mathrm{T}} \|_F^2 + \\ &\sum_{v=1}^{n_v} \lambda_v \mathrm{Tr}((\boldsymbol{V}^{(v)})^{\mathrm{T}} \boldsymbol{B} \boldsymbol{V}^{(v)}) + \sum_{v=1}^{n_v} \beta_v \| \boldsymbol{V}^{(v)} - \boldsymbol{V}^* \|_F^2 \\ &\text{s.t. } \forall 1 \leqslant k \leqslant K, \| \boldsymbol{U}_{\cdot,k}^{(v)} \|_1 = 1 \\ &\boldsymbol{U}^{(v)}, \boldsymbol{V}^{(v)}, \boldsymbol{V}^* \geqslant 0 \end{aligned} \tag{15}$$

其中,参数 λ_v 来同时调节每个视图所占的比重和 2 个视图之间相关性的比重,参数 β_v 用来调节流形正则项的比重。$\sum_{v=1}^{n_v} \lambda_v \mathrm{Tr}((\boldsymbol{V}^{(v)})^{\mathrm{T}} \boldsymbol{B} \boldsymbol{V}^{(v)})$ 是 Hessian 正则化项,它约束了数据的空间分布模式;$\sum_{v=1}^{n_v} \beta_v \cdot \| \boldsymbol{V}^{(v)} - \boldsymbol{V}^* \|_F^2$ 是联合矩阵分解的正则性,它是不同视图下系数矩阵与共同矩阵的相似性,反映了属于同类的概率,概率越大越可能属于一类。因此,给出了不同视图下的潜在类结构。从某种意义上说,在对每一个视图的数据进行 Hessian 正则化挖掘其内在流形的同时,以加权的形式对多视图数据进行联合分解,从而实现多个视图数据的内在流形结构和信息的共享。

引入辅助变量 Q 代替矩阵 $\boldsymbol{U}^{(v)}$ 的约束,使得:

$$Q^{(v)}=\mathrm{Diag}\left(\sum_{i=1}^{M}\boldsymbol{U}_{i,1}^{(v)},\sum_{i=1}^{M}\boldsymbol{U}_{i,2}^{(v)},\cdots,\sum_{i=1}^{M}\boldsymbol{U}_{i,K}^{(v)}\right) \tag{16}$$

其中,Diag(·)表示括号中对角矩阵中非零元的个数。于是,目标函数 J 变为:

$$\begin{aligned}&\boldsymbol{J}=\sum_{v=1}^{n_v}\|\boldsymbol{X}^{(v)}-\boldsymbol{U}^{(v)}(\boldsymbol{V}^{(v)})^{\mathrm{T}}\|_F^2+\\&\sum_{v=1}^{n_v}\lambda_v\mathrm{Tr}((\boldsymbol{V}^{(v)})^{\mathrm{T}}\boldsymbol{B}\boldsymbol{V}^{(v)})+\sum_{v=1}^{n_v}\beta_v\|\boldsymbol{V}^{(v)}Q^{(v)}-\boldsymbol{V}^{*}\|_F^2\\&\text{s.t. }\forall 1\leqslant k\leqslant K,\|\boldsymbol{U}_{\cdot,k}^{(v)}\|_1=1\\&\boldsymbol{U}^{(v)},\boldsymbol{V}^{(v)},\boldsymbol{V}^{*}\geqslant 0\end{aligned} \tag{17}$$

为了解决这个最优化问题,本文使用迭代替换的方法:

1)修改矩阵 $\boldsymbol{V}^{*}$,通过求解矩阵 $\boldsymbol{U}^{(v)}$ 和矩阵 $\boldsymbol{V}^{(v)}$ 最小化目标函数 $\boldsymbol{J}$。

当矩阵 $\boldsymbol{V}^{*}$ 确定时,每个视图之间的矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 是相互独立的,因此,对于每个视图最小化以下函数:

$$\begin{aligned}&\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_F^2+\lambda\mathrm{Tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{V})+\lambda_v\|\boldsymbol{V}\boldsymbol{Q}-\boldsymbol{V}^{*}\|_F^2\\&\text{s.t. }\boldsymbol{U},\boldsymbol{V}\geqslant 0\end{aligned} \tag{18}$$

使用拉格朗日法求解以上问题,得到以下的迭代规则:

$$\begin{aligned}u_{ik}&\leftarrow u_{ik}\frac{(\boldsymbol{X}\boldsymbol{V})_{ik}+\lambda_v\sum_{j=1}^{N}\boldsymbol{V}_{j,k}\boldsymbol{V}_{j,k}^{*}}{(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})_{ik}+\lambda_v\sum_{l=1}^{M}\boldsymbol{U}_{l,k}\sum_{j=1}^{N}\boldsymbol{V}_{j,k}^{2}}\\v_{jk}&\leftarrow v_{jk}\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}+\lambda\boldsymbol{B}_{+}\boldsymbol{V})_{jk}+\lambda_v\boldsymbol{V}_{j,k}^{*}}{(\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}+\lambda\boldsymbol{B}_{-}\boldsymbol{V})_{jk}+\lambda_v\boldsymbol{V}_{j,k}}\end{aligned} \tag{19}$$

可以发现每次迭代以后矩阵 $\boldsymbol{U}_{i,k}$ 和矩阵 $\boldsymbol{V}_{i,k}$ 保持非负性。

2)修改矩阵 $\boldsymbol{U}^{(v)}$ 和矩阵 $\boldsymbol{V}^{(v)}$,通过求解矩阵 $\boldsymbol{V}^{*}$ 最小化目标函数 $\boldsymbol{J}$。

使用拉格朗日法对矩阵 $\boldsymbol{V}^{*}$ 得到以下解:

$$\boldsymbol{V}^{*}=\frac{\sum_{v=1}^{n_v}\lambda_v\boldsymbol{V}^{(v)}\boldsymbol{Q}^{(v)}}{\sum_{v=1}^{n_v}\lambda_v}\geqslant 0 \tag{20}$$

以上即为 HR-J-NMF 算法的推导过程。为了对比 HR-J-NMF 算法的性能,本文还提出了基于 Laplacian 正则化的 J-NMF(LR-J-NMF)算法。由于推导过程与 HR-J-NMF 相似,在此不再详述。

4　实验结果与分析

4.1　单视图聚类实验与结果

本文首先进行单视图数据的聚类实验来验证所提出的 HR-NMF 算法的性能。采用了文献[1]所采用的 COIL20 数据库[17]、ALLAML 数据库[18]和 Digit 数据库。表 1 所示为 3 个数据库的有关数据信息,详细信息请参见参考文献[1]。

表 1　3 个数据库的样本与聚类个数

数据库	样本个数	聚类个数
COIL20	1 440	20
Digit	1 000	10
ALLAML	38	3

HR-NMF 和原始 NMF、LR-NMF 算法进行对比。分别使用 3 种 NMF 算法对原始数据进行矩阵分解,得到的系数矩阵作为降维后的数据,维数选择能够使聚类结果达到最好的。最后的聚类算法采用文献[1]提出的 LR-NMF 算法时进行验证时所用的 K-means 聚类。评价指标采用聚类评估中采用的聚类准确率和互信息。

表 2 和表 3 所示为 3 种不同 NMF 算法对不同数据集进行聚类后得到的准确率和互信息结果。由于 3 种方法对应的聚类结果收敛性很好且非常稳定,因此只统计一次结果。由表中结果可以发现,HR-NMF 对于 COIL20、Digit 和 ALLAML 这 3 个数据集的聚类精度分别为 82.08%,72.90% 和 97.37%,互信息分别为 0.904,0.683 和 0.899。上述结果中,除了 Digit 数据集的互信息值低于 LR-NMF 算法,其他结果都远远高于 LR-NMF 和原始 NMF 算法的结果,表明了 HR-NMF 的优异性能。

表 2　不同数据集 NMF 算法的聚类精度结果　%

算法	COIL20	Digit	ALLAML
NMF 算法	66.74	67.20	89.47
LR-NMF 算法	72.22	69.20	92.11
HR-NMF 算法	82.08	72.90	97.37

表 3　不同数据集 NMF 算法的互信息值结果

算法	COIL20	Digit	ALLAML
NMF 算法	0.784	0.613	0.715
LR-NMF 算法	0.876	0.729	0.777
HR-NMF 算法	0.904	0.683	0.899

4.2　多视图聚类实验与结果

在验证了 HR-NMF 算法有效性的基础上,进行了多视图数据聚类实验来评估 HR-J-NMF 算法的性能。本文采用了在文献[7]中的多视图聚类中使用的 UCI 手写数字数据库(Digit 数据库)和路透社多语言数据集(Reuters 数据集)[4]。对于 Digit 数据库,视图 1 的特征是 76 点的傅里叶系数,视图 2 的特征是 240 像素点。表 4 所示为多视图聚类实验中要用到的数据集信息。而对于 Reuters 数据集,把原始英文文件作为第 1 个视图,法语文件以及德语文件作为第 2 个和第 3 个视图。详细信息见参考文献[6]。

表 4　数据集信息

数据集	样本个数	视图个数	聚类个数
Reuters	600	3	6
Digit	1 000	2	10

HR-J-NMF 算法与 LR-J-NMF 以及原始的 J-NMF算法进行对比。同样地,分别使用 3 种 NMF 算法对原始数据进行矩阵分解,得到的系数矩阵作为降维后的数据,维数选择能够使聚类结果达到最好的。最后的聚类算法仍然采用 K-means 聚类。所有结果运行 3 次,其平均值为最终结果。为了便于做各种多视图方法之间的比较,本文首先都设置参数 β 为 0.01。随后再针对参数 β 进行讨论,把 β 从 10^{-3} ~ 10^{-1} 之间变动,并在图中显示对应的结果。

表 5 和表 6 所示为 3 种不同多视图 NMF 算法对不同数据集进行聚类后得到的准确率和互信息结果。从表 5 结果可以发现,HR-J-NMF 对于 Digit 和 Reuters 这 2 个数据集的聚类精度分别为 90.73% 和 48.25% ,互信息分别为 0.835 和 0.268。上述结果中,HR-J-NMF 的结果都高于 LR-NMF 和原始 NMF 算法的结果,表明了 HR-NMF 的有效性。

表 5 多视图 NMF 算法对不同数据集的聚类精度结果 %

算法	Digit	Reuters
NMF 算法	87.03 ± 2.55	46.11 ± 0.04
LR-NMF 算法	89.70 ± 1.13	46.17 ± 0.04
HR-NMF 算法	90.73 ± 1.07	48.25 ± 0.00

表 6 多视图 NMF 算法对不同数据集的互信息值结果

算法	Digit	Reuters
NMF 算法	0.792 ± 0.024	0.255 ± 0.023
LR-NMF 算法	0.832 ± 0.090	0.254 ± 0.023
HR-NMF 算法	0.835 ± 0.015	0.268 ± 0.000

图 1 ~ 图 4 所示为 2 个数据库中,不同 NMF 算法中参数 β 对聚类精度和互信息结果的影响。结果表明,MultiHR-NMF 的结果在 10^{-2} 附近稳定,而且在大部分情况下它的效果要优于其他 2 种方法。实验结果再次表明了 HR-J-NMF 的优异性能。

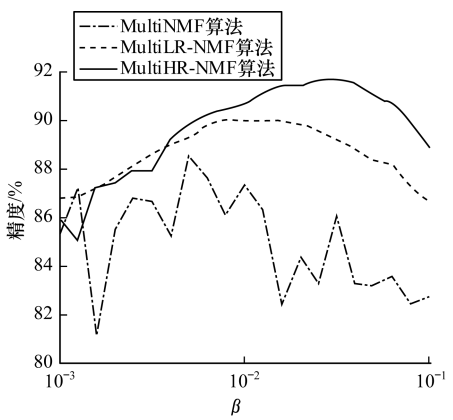

图 1 Digit 中 β 对不同 NMF 算法的聚类精度影响

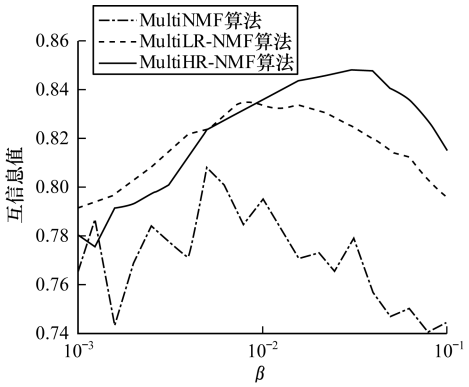

图 2 Digit 中 β 对不同 NMF 算法互信息值的影响

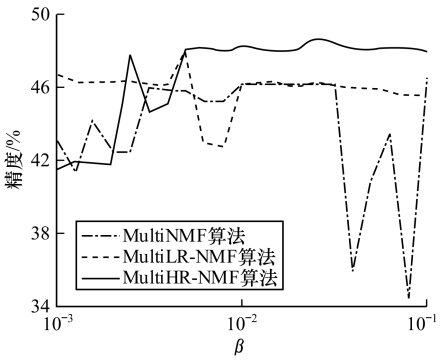

图 3 Reuters 中 β 对不同 NMF 算法的聚类精度影响

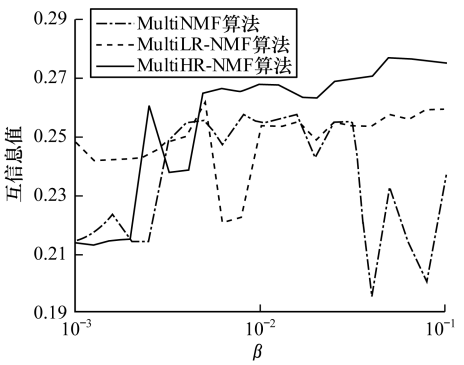

图 4 Reuters 中 β 对不同 NMF 算法互信息值的影响

5 结束语

基于 Hessian 正则化与传统的 Laplacian 正则化相比,能更好地表征数据流形的内在局部几何信息,因此,本文提出基于 Hessian 正则化的 NMF 算法,并针对多视图数据给出基于 Hessian 正则化的联合 NMF 算法。实验结果表明,本文所提出的 HR-NMF 和 HR-J-NMF 算法能获得较好的聚类精度和互信息结果。下一步将研究大样本量下该算法的聚类效果。

参考文献

[1] DANIEL L, SEBASTIAN S. Learning the Parts of Objects by Nonnegative Matrix Factorization [J]. Nature, 1999, 401(6755): 788-791.

[2] 马慧芳,赵卫中. 基于非负矩阵分解的双重约束文本聚类算法[J]. 计算机工程,2011,37(24):161-163.

[3] 杨　剑,王　珏,钟　宁. 流形上的 Laplacian 半监督回归[J]. 计算机研究与发展,2007,44(7):1121-1127.
[4] DENG C, HE Xiaofei, HAN Jiawei, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33 (8): 1548-1560.
[5] KIM J, RENATO D C M, HAESUN P. Group Sparsity in Nonnegative Matrix Factorization [C]//Proceedings of SIAM International Conference on Data Mining. Anaheim, USA: Society for Industrial and Applied Mathematics, 2012: 851-862.
[6] ZEYNEP A, CHRISTIAN T, CHRISTIAN B. Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction[C]//Proceedings of the 16th Computer Vision Winter Workshop. Mitterberg, Austria: 2011: 27-34.
[7] LIU Jialu, WANG Chi, GAO Jing, et al. Multi-view Clustering via Joint Nonnegative Matrix Factorization[C]// Proceedings of SIAM International Conference on Data Mining. Austin, USA: Society for Industrial and Applied Mathematics, 2013: 252-260.
[8] 卢桂馥,万鸣华. Hessian 正则化的低秩矩阵分解算法[J]. 小型微型计算机系统, 2016, 37 (10): 2296-2299.
[9] 刘红丽,刘伟锋,王延江,等. Hessian 正则化 Logistic 回归模型[J]. 计算机工程与应用, 2016, 52 (5): 236-240.
[10] DAVID L D, CARRIE G. Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-dimensional Data[J]. Proceedings of National Academy of Sciences of the United States of America, 2003, 100(10): 5591-5596.
[11] KWANG I K, FLORIAN S, MATTHIAS H. Semi-supervised Regression Using Hessian Energy with an Application to Semi-supervised Dimensionality Reduction [C]//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: DBLP, 2009: 979-987.
[12] TAO Dapeng, JIN Lianwen, LIU Weifeng, et al. Hessian Regularized Support Vector Machines for Mobile Image Annotation on the Cloud [J]. IEEE Transactions on Image Processing, 2013, 15(4): 833-844.
[13] LIU Weifeng, TAO Dacheng. Multiview Hessian Regularization for Image Annotation [J]. IEEE Transactions on Image Processing, 2013, 22 (7): 2676-2687.
[14] LIU Weifeng, MA Tengzhou, TAO Dapeng, et al. HSAE: A Hessian Regularized Sparse Auto-encoders [J]. Neurocomputing, 2016, 187: 59-65.
[15] PENTTI P, UNTO T. Positive Matrix Factorization: A Nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values [J]. Environmetrics, 1994, 5(2): 111-126.
[16] STEFAN W, WRITTEN S W, JAMES C, et al. Seeding Non-negative Matrix Factorizations with the Spherical K-means [EB/OL]. (2010-11-21). https://www.researchgate.net/publication/247469258_Seeding_NonNegative_Matrix_Factorizations_with_the_Spherical_K-Means_Clustering.
[17] PRITHVIRAJ S, GALILEO N, GALILEO N, et al. Collective Classification in Network Data [J]. AI Magazine, 2008, 29: 93-106.
[18] LI Yifeng, ALIOUNE N. The Non-negative Matrix Factorization Toolbox for Biological Data Mining[J]. Source Code for Biology and Medicine, 2013, 8(10): 1-26.

编辑　刘　冰

(上接第 127 页)

[7] ZHANG Fangfang, HUANG Heqing, ZHU Sencun, et al. ViewDroid: Towards Obfuscation-resilient Mobile Application Repackaging Detection [C]//Proceedings of 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks. New York, USA: ACM Press, 2014: 25-36.
[8] CRUSSELL J, GIBLER C, CHEN Hao. Attack of the Clones: Detecting Cloned Applications on Android Markets [M]//FORESTI S, YUNG M, MARTINELLI F. Computer Security-ESORICS 2012. Berlin, Germany: Springer, 2012: 37-54.
[9] MILO R, SHENORR S, ITZKOVITZ S, et al. Network Motifs: Simple Building Blocks of Complex Networks[J]. Science, 2002, 298(5594): 824-827.
[10] VALVERDE S, SOLÉ R V. Network Motifs in Computational Graphs: A Case Study in Software Architecture[J]. Physical Review E, 2005, 72(2).
[11] 杨广亮,龚晓锐,姚　刚,等. 一个面向 Android 的隐私泄露检测系统[J]. 计算机工程, 2012, 38(23): 1-6.
[12] 陈　凯,王　鹏,LEE Y,等. 面向海量软件的未知恶意代码检测方法[J]. 信息安全学报, 2016, 1(1): 24-38.
[13] ZHOU Wu, ZHOU Yajin, JIANG Xuxian, et al. Detecting Repackaged Smartphone Applications in Third-party Android Marketplaces[C]//Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy. New York, USA: ACM Press, 2012: 317-326.
[14] AppBrain[EB/OL]. [2016-06-01]. http://www.appbrain.com/stats.
[15] NG Y Y, ZHOU Hucheng, JI Zhiyuan, et al. Which Android App Store Can Be Trusted in China? [C]// Proceedings of the 38th Annual Computer Software and Applications Conference. Washington D. C., USA: IEEE Press, 2014: 509-518.
[16] 吴雪平,张大方,苏　欣,等. 基于流量相似度的 Android 二次打包应用的检测技术研究[J]. 小型微型计算机系统, 2015, 36(5): 954-958.

编辑　金胡考