基于稀疏约束的半监督非负矩阵分解算法

胡学考[1] 孙福明[1] 李豪杰[2]
(辽宁工业大学电子与信息工程学院 锦州 121001)[1] (大连理工大学软件学院 大连 116300)[2]

摘 要 矩阵分解因可以实现大规模数据处理而具有十分广泛的应用。非负矩阵分解(Nonnegative Matrix Factorization,NMF)是一种在约束矩阵元素为非负的条件下进行的分解方法。利用少量已知样本的标注信息和大量未标注样本,并施加稀疏性约束,构造了一种新的算法——基于稀疏约束的半监督非负矩阵分解算法。推导了其有效的更新算法,并证明了该算法的收敛性。在常见的人脸数据库上进行了验证,实验结果表明CNMFS算法相对于NMF和CNMF等算法具有较好的稀疏性和聚类精度。

关键词 非负矩阵分解,半监督,稀疏约束
中图法分类号 TP37 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.060

Constrained Nonnegative Matrix Factorization with Sparseness for Image Representation

HU Xue-kao[1] SUN Fu-ming[1] LI Hao-jie[2]
(School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China)[1]
(School of Software, Dalian University of Technology, Dalian 116300, China)[2]

Abstract Matrix decomposition is widely applied in many domains since it is exploited to process the large-scale data. To the best of our knowledge, nonnegative matrix factorization (NMF) is a non-negative decomposition method under the condition that constraint matrix elements are non-negative. By using the information provided by a few known labeled examples and large number of unlabeled examples as well as imposing a certain sparseness constraint on NMF, this paper put forward a method called constraint nonnegative matrix factorization with sparseness (CNMFS). In the algorithm, an effective update approach is constructed, whose convergence is proved subsequently. Extensive experiments were conducted on common face databases, and the comparisons with two state-of-the-art algorithms including CNMF and NMF demonstrate that CNMFS has superiority in both sparseness and clustering.

Keywords Nonnegative matrix factorization, Semi-supervised, Sparseness constraints

1 引言

在许多数据分析问题中,一个最基本的问题是:如何找到一种合理的数据表示方法。当数据是高维数据时,矩阵分解显得尤为重要。作为一种数据表示的基本工具,矩阵分解技术受到了越来越多的关注,它包含有多种方法,比如:PCA(主成分分析)、LDA(线性判别分析)、ICA(独立分量分析)、SVD(奇异值分解)等。这些方法通常是在一定的限制下对原始数据矩阵进行线性变换或分解,并且允许原始数据矩阵和分解结果中出现负值。然而,在大多实际情况下,负值是没有意义的。所以,一种新的矩阵分解方法——非负矩阵分解(Nonnegative Matrix Factorization,NMF)[1]应运而生。非负矩阵分解方法通常将原始矩阵分解成左、右两个非负矩阵。左面的矩阵被称为基矩阵,右面的矩阵被称为系数矩阵。因而,原始矩阵中的列向量可解释为基矩阵中所有列向量的加权之和,而权重系数即为系数矩阵。这样的向量组合具有非常直观的语义解释,符合人类思维中的整体由局部构成的概念[2]。

为了提高NMF算法的识别率和有效性,已经有不少人提出了改进算法。Cai等[3]考虑到原始数据中隐含的几何结构,将NMF和流形理论相结合,提出了基于图正则化的非负矩阵分解(Graph Regularized Nonnegative Matrix Factorization,GNMF)算法,其利用流形学习揭示数据中内在的几何结构性质。但这种算法忽略了已有标签的信息,并且稀疏度也不是很高。Hoyer[4]将NMF应用到稀疏编码理论中构造出了非负稀疏编码(Nonnegative Sparse Coding,NSC)算法,并进一步提出了一种可以较为精确控制的稀疏性NMF算法(Nonnegative Matrix Factorization with Sparseness Constraints,NMFSC),使得分解后的矩阵具有了较好的稀疏性。这不仅节省了存储空间,而且提高了运算效率;然而这种改进算法却并不能提高识别率。由于数据稀疏具有许多良好的特性,其已经广泛地应用在多个领域。比如,Sun等[5]将其应用在多标签图像分类中,同样取得了很好的效果。Li等[6]在标准NMF的基础上对基矩阵增加空间局部化限制后构造出了LNMF算法。

到稿日期:2014-08-28 返修日期:2014-11-17 本文受国家自然科学基金(61272214,61472059)资助。
胡学考(1989-),男,硕士生,主要研究领域为图像语义理解,E-mail:lghuxk@gmail.com;孙福明(1972-),男,博士,教授,CCF会员,主要研究领域为计算机视觉、图像语义理解,E-mail:sunwenfriend@hotmail.com;李豪杰(1973-),男,博士,副教授,主要研究领域为计算机视觉、图像语义理解,E-mail:hjli@dlut.edu.cn。

前述的标准 NMF 及其改进算法都属于无监督分解方法，没有考虑样本的标签信息。Liu 等[7]将部分标签约束强加到 NMF 分解中，提出了一种半监督有约束的非负矩阵分解(Constrained Nonnegative Matrix Factorization，CNMF)算法。但是，该算法由于没有施加稀疏约束，导致具有存储空间较大以及运算效率较低等特点。本文在 CNMF 的基础之上增加稀疏约束，提出了稀疏约束的半监督非负矩阵分解(Constrained Nonnegative Matrix Factorization with Sparseness，CNMFS)。该方法在增加标签信息约束的同时，对分解结果进一步稀疏，这样既节省了存储空间，提高了运算效率，又得到了较高的聚类精度。本文在两个标准的人脸数据库上对 CNMFS 进行了实验验证。

2 约束非负矩阵分解

2.1 非负矩阵分解

给定 n 个非负样本 $x_i, i=1,2,\cdots,n, x_i \in R^m$ 为列向量，构成矩阵 $X=[x_1,x_2,\cdots,x_n]\in R^{m\times n}$。NMF 算法的目的就是寻找两个非负矩阵 $U\in R^{m\times k}$ 和 $V\in R^{n\times k}$。其中，$k\leqslant mn/(m+n)$，并且矩阵中的元素都为非负值。

使 X 与 UV^{T} 之间的相似度最高，也就是最小化以下目标函数：

$$O_F=\| X-UV^{\mathrm{T}}\|^2 \quad \text{s.t.}\ U\geqslant 0, V\geqslant 0 \tag{1}$$

其中，$\|\cdot\|_F$ 是 Frobenius 范数。式(1)是欧氏空间的目标函数描述形式，也有另外一种散度描述方法：

$$O_{KL}=D(X\|UV^{\mathrm{T}})=\sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{(uv^{\mathrm{T}})_{ij}}-x_{ij}+(uv^{\mathrm{T}})_{ij}\right) \tag{2}$$

经过推导可证明，以上两个公式是收敛的；并且可得出目标函数(1)的乘性迭代更新公式如下：

$$u_{ik}\leftarrow u_{ik}\frac{(XV)_{ik}}{(UV^{\mathrm{T}}V)_{ik}} \tag{3}$$

$$v_{jk}\leftarrow v_{jk}\frac{(X^{\mathrm{T}}V)_{jk}}{(V^{\mathrm{T}}VU^{\mathrm{T}})_{jk}} \tag{4}$$

其中，$U=[w_{ik}]$，$V=[v_{jk}]$。在更新计算过程中，初始矩阵 U_0 和 V_0 可以随机生成。在给定迭代的终止条件后，按照式(3)和式(4)交替更新，当满足终止条件时可得到最终的矩阵 U 和 V。

2.2 半监督的非负矩阵分解

NMF 算法是无监督式的矩阵分解，没有用到数据样本的标签信息。如果用部分样本的标签信息对 NMF 算法进行约束，使得具有相同类别信息的样本在低维空间中投影为一个点，再进行识别和聚类，则会得到更好的效果[7]。

假设 n 个非负样本 $x_i(i=1,2,\cdots,n)$ 来自 c 类，前 l 个非负样本为已知标签信息，剩余样本的标签信息未知。分别定义索引矩阵 C 和标签约束矩阵 A 如式(5)和式(6)所示，其中 I_{n-l} 是单位矩阵。

$$c_{ij}=\begin{cases}1, & x_j \text{ 被标记为第 } i \text{ 类}\\ 0, & \text{其他}\end{cases} \tag{5}$$

$$A=\begin{pmatrix}C_{l\times c} & 0\\ 0 & I_{n-l}\end{pmatrix}_{n\times(c+n-l)} \tag{6}$$

在 NMF 目标函数中，引入标签约束矩阵 A，即 $X\approx UV^{\mathrm{T}}=U(AZ)^{\mathrm{T}}$。其中，$V=AZ$，$Z=[z_1,z_2,\cdots,z_k]\in R^{(c+n-l)\times k}$。很容易得知，如果 x_i 和 x_j 具有相同的标签，那么它们的加权系数矢量是相同的。CNMF 算法中的目标函数如式(7)所示。

$$O_F=\| X-UZ^{\mathrm{T}}A^{\mathrm{T}}\| \quad \text{s.t.}\ U\geqslant 0, Z\geqslant 0 \tag{7}$$

其迭代更新规则如下：

$$u_{ik}\leftarrow u_{ik}\frac{(XAZ)_{ik}}{(UZ^{\mathrm{T}}A^{\mathrm{T}}AZ)_{ik}} \tag{8}$$

$$z_{jk}\leftarrow z_{jk}\frac{(A^{\mathrm{T}}X^{\mathrm{T}}U)_{jk}}{(A^{\mathrm{T}}AZU^{\mathrm{T}}U)_{jk}} \tag{9}$$

3 稀疏约束的半监督非负矩阵分解

越来越多的数据呈现高维特性，因此数据的稀疏表示成为了当前的研究热点。稀疏可以理解为利用少量的元素有效地替代大量的数据或者向量。稀疏表示就是将一组数据分解为一系列基向量的线性组合。在半监督约束的非负矩阵分解算法的基础之上进行稀疏约束，就能够得到效率更高的分解，提高分解质量，并节省存储空间。

综合 CNMF 算法和稀疏理论，本文提出了 CNMFS。因为原始矩阵分解为基矩阵和系数矩阵，基矩阵反映了数据的特征，所以只对基矩阵稀疏化，从而获得新的最小化目标函数，见式(10)。

$$O_F=\| X-UZ^{\mathrm{T}}A^{\mathrm{T}}\|+\beta\|U\|_F^2 \quad \text{s.t.}\ U\geqslant 0, Z\geqslant 0 \tag{10}$$

式中，$\beta\in(0,1)$。利用最速下降法和迭代法，可以推导出最小化目标函数的乘性迭代规则：

$$\begin{aligned}O_F&=Tr((X-UZ^{\mathrm{T}}A^{\mathrm{T}})(X-UZ^{\mathrm{T}}A^{\mathrm{T}})^{\mathrm{T}})+\beta\|U\|_F^2\\&=Tr(XX^{\mathrm{T}})-2Tr(XAZU^{\mathrm{T}})+Tr(UZ^{\mathrm{T}}A^{\mathrm{T}}AZU^{\mathrm{T}})+\\&\quad\beta\|U\|_F^2\end{aligned}$$

定义拉格朗日乘子 φ 和 ϕ 分别约束 u 和 z，由拉格朗日定理得拉格朗日函数为：

$$L=O_F+Tr(\varphi U^{\mathrm{T}})+Tr(\phi Z^{\mathrm{T}})$$

对上式分别求 U 和 Z 的偏导数并令其等于 0，得：

$$\frac{\partial L}{\partial U}=-2XAZ+2UZ^{\mathrm{T}}A^{\mathrm{T}}AZ+\varphi+2\beta U=0$$

$$\frac{\partial L}{\partial Z}=-2A^{\mathrm{T}}X^{\mathrm{T}}U+2A^{\mathrm{T}}AZU^{\mathrm{T}}U+\phi=0$$

使用 KKT 条件 $\varphi_{ij}u_{ij}=0$ 和 $\phi_{ij}z_{ij}=0$，最终得到迭代的更新规则如下：

$$u_{ik}\leftarrow u_{ik}\frac{(XAZ)_{ik}}{(UZ^{\mathrm{T}}A^{\mathrm{T}}AZ)_{ik}+\beta U_{ik}} \tag{11}$$

$$z_{jk}\leftarrow z_{jk}\frac{(A^{\mathrm{T}}X^{\mathrm{T}}U)_{jk}}{(A^{\mathrm{T}}AZU^{\mathrm{T}}U)_{jk}} \tag{12}$$

式(10)中的目标函数 O_F 在迭代更新规则式(11)和式(12)下是收敛的。下面将证明 CNMFS 目标函数的收敛性。证明的过程中需要引入辅助函数和引理，具体如下：

定义 1 定义 $G(u,u')$ 是 $F(u)$ 的辅助函数，并且满足 $G(u,u')\geqslant F(u)$，$G(u,u)=F(u)$。

引理 1 如果 $G(u,u')$ 是 $F(u)$ 的辅助函数，那么在以下更新规则下 $F(u)$ 是非增的：

$$u^{t+1}=\arg\min_u G(u,u^t) \tag{13}$$

证明：很显然有 $u=u^{t+1}$ 时函数 $G(u,u^t)$ 取得最小值。由定义 1 知，$G(u^{t+1},u^t)$ 大于等于 $F(u^{t+1})$ 可以用不等式表示为：

$$F(u^{t+1})\overset{def}{\leqslant}G(u^{t+1},u^t)\overset{\min}{\leqslant}G(u^t,u^t)\overset{def}{=}F(u^t)$$

注意到 $F(u^{t+1})=F(u^t)$，只有当 u^t 为 $G(u,u^t)$ 的局部极小值时才成立。如果函数 F 存在导数且在 u^t 的一个微小领域内连续，则微分 $\nabla F(u^t)=0$。通过式(13)即可得到收敛到局部极小点 $u_{\min}=\arg\min_u F(u)$ 的序列：

$$F(u_{\min})\leqslant\cdots\leqslant F(u^{t+1})\leqslant F(u^t)\leqslant\cdots\leqslant F(u^1)\leqslant F(u^0)$$

所以，通过定义这样的辅助函数 $G(u,u^t)$，使得目标函数(10)相应的迭代规则满足 $u^{t+1}=\arg\min_u G(u,u^t)$。证毕。

引理 2 函数 F' 表示关于 Z 的一阶偏导，$F_{z_{ij}}$ 表示目标函数中仅与 Z 中的元素 z_{ij} 有关的部分。函数：

$$G(z,z_{ij}^t)=F_{z_{ij}}(z_{ij}^t)+F'_{z_{ij}}(z_{ij}^t)(z-z_{ij}^t)+\frac{(A^TAZU^TU)_{ij}}{z_{ij}^t}(z-z_{ij}^t)^2 \tag{14}$$

是 $F_{z_{ij}}$ 的辅助函数。

证明：显然 $G(z,z)=F_{z_{ij}}(z)$，根据所定义的辅助函数，只需要证明 $G(z,z_{ij}^t)\geqslant F_{z_{ij}}(z)$。

由 $F_{z_{ij}}(z)$ 的泰勒展开式为：

$$F_{z_{ij}}(z_{ij})=F_{z_{ij}}(z'_{ij})+F'_{z_{ij}}(z'_{ij})(z-z'_{ij})+\frac{1}{2}F''_{z_{ij}}(z'_{ij})(z-z'_{ij})^2 \tag{15}$$

其中，$F''_{z_{ij}}(z'_{ij})=2(A^TA)_{ii}(U^TU)_{jj}$。结合式(14)和式(15)可以将问题转化为证明式(16)：

$$\frac{(A^TAZU^TU)_{ij}}{z_{ij}^t}\geqslant\frac{1}{2}F''=(A^TA)(U^TU) \tag{16}$$

由于

$$(A^TAZU^TU)=\sum_l(A^TAZ)(U^TU)\geqslant(A^TAZ)(U^TU)\geqslant\sum_l(A^TA)z^t(U^TU)\geqslant z^t(A^TA)(U^TU)$$

因此，$G(z,z_{ij}^t)\geqslant F_{z_{ij}}(z)$。

综上所述，$G(z,z_{ij}^t)$ 是 $F_{z_{ij}}(z)$ 的辅助函数。

同理，也可以构造出目标函数中只关于 u_{ij} 部分的辅助函数。

引理 3 假设函数 F' 表示关于 U 的一阶偏导，$F_{u_{ij}}$ 表示目标函数中仅与 U 中的元素 u_{ij} 有关的部分。函数：

$$G(u,u_{ij}^t)=F_{u_{ij}}(u_{ij}^t)+F'_{u_{ij}}(u_{ij}^t)(u-u_{ij}^t)+\frac{(UZ^TA^TAZ-\beta U)_{ij}}{u_{ij}^t}(u-u_{ij}^t)^2 \tag{17}$$

是 $F_{u_{ij}}$ 的辅助函数。

其证明过程与引理 2 相似，由于篇幅限制在此就不给出具体证明过程。

定理 1 目标函数式(10)在迭代式(11)和式(12)更新条件下是非增的。

证明：将式(14)中 $G(z,z_{ij}^t)$ 应用到式(13)中，得到：

$$z_{ij}^{t+1}=\arg\min_z G(z,z_{ij}^t)=z_{ij}^t\frac{(A^TX^TU)_{ij}}{(A^TAZU^TU)_{ij}}$$

根据引理 2 可知，$G(z,z_{ij}^t)$ 是辅助函数，在此更新规则下 $F_{z_{ij}}(z)$ 则是非增的。

将式(17)代入式(13)中，得到

$$u_{ij}^{t+1}=\arg\min_u G(u,u_{ij}^t)=u_{ij}^t\frac{(XAZ)_{ij}}{(UZ^TA^TAZ)_{ij}}$$

根据引理 3 可知，$G(u,u_{ij}^t)$ 是辅助函数，在此更新规则下 $F_{u_{ij}}(u)$ 则是非增的。

因此，定理保证在式(11)和式(12)迭代更新后目标函数式(10)最后能收敛到一个局部最优值。

下面给出稀疏约束的半监督非负矩阵分解算法的具体步骤，见算法 1。

算法 1 稀疏约束的半监督非负矩阵分解算法

输入：数据集矩阵 X

输出：分解后的矩阵 U 和 Z

过程：

(1)初始化参数。设定参数 $\beta(0<\beta<1)$、分解维度 k 和最大迭代次数(nIterMax)，计算索引矩阵 $C_{n\times l}$ 并构造标签矩阵 $A_{(c+n-l)\times n}$。随机生成非负矩阵 $U_{m\times k}$ 和 $Z_{n\times k}$，并对原始矩阵 $X_{m\times n}$ 进行归一化。

(2)FOR nI=1:nIterMax(nI 是迭代次数)

(3)运用迭代规则(11)和(12)进行迭代运算。

(4)根据式(10)计算最小化目标函数。

(5)END FOR

(6)返回最终迭代后的矩阵 U 和 Z。

4 实验与结果分析

4.1 数据集

PIE 数据集(PIE_pose27)和 ORL 数据集的相关信息见表 1。PIE_pose27 数据集包含 68 名志愿者的面部表情图像，共有 42 种不同的光照条件，2856 幅图片。图 1(a)显示了该数据集中在不同光照条件下前 10 名志愿者的不同姿态表情的图像。ORL 人脸数据集包含 40 名志愿者的图像，其中每人有 10 种不同的表情，共计 400 幅图像。部分志愿者的图像含有姿态表情和面部饰物的变化。该数据集中的部分图片如图 1(b)所示。

表 1 数据集信息

数据集	样本个数	特征维数	类别数(k)
PIE_pose27	2856	1024	68
ORL	400	1024	40

(a)PIE 数据库

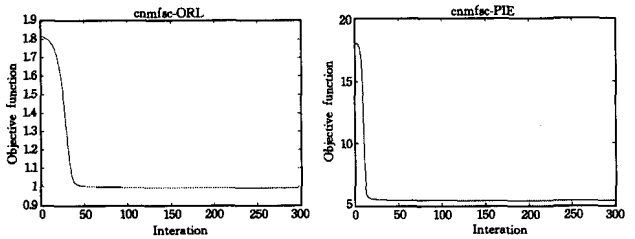
(b)ORL 数据

图 1 数据库中部分图像示例

4.2 类别选择及收敛性

首先从数据集中选取 $k(k=2,3,\cdots,10)$ 类数据(由表 1 可知，数据集 PIE_pose27 和 ORL 的维数远远大于 k 值，满足 $k\leqslant mn/(m+n)$)。然后，在每个类中随机选取 10% 的样本用于已知标签的约束条件。不过，由于 ORL 中每个类只有 10

个样本，如果选取10%的样本，已知样本的个数为1，这样样本数量太少，很难提取出共同的特征。为此，从每个类中任选两个样本。两个数据集目标函数的收敛曲线分别如图2所示（$k=10$）。



（a）ORL数据集目标函数的收敛曲线　（b）PIE_pose27数据集目标函数的收敛曲线

图2　该算法目标函数的收敛曲线

由图2可以看出，两个数据集的目标函数在迭代初始时下降非常快；随着迭代次数的增加，目标函数下降越来越缓慢，基本上在迭代50次以内，就已经收敛得非常好了。考虑到算法的复杂度（程序运行时间），可设置最大迭代次数为300。

4.3　评价指标及参数选择

将数据分解成基矩阵和系数矩阵后，即可对系数矩阵进行聚类[8]。本文利用两个评估指标：准确率（AC）和互信息（MI）来验证CNMFS算法的有效性。

准确率（AC）用来计算聚类后的标签与原数据自带标签之间的相似度，定义见式（18）。

$$AC=\frac{\sum_{i=1}^{n}\delta(r_i,map(l_i))}{n} \tag{18}$$

其中，$\delta(x,y)$是关于δ的函数，即如果$x=y$，则$\delta(x,y)=1$；否则，$\delta(x,y)=0$。$map(l_i)$是利用Kuhn-Munkres算法实现聚类标签与原标签映射的函数[9]。利用该函数可以将聚类标签转化为图像自带的标签。

互信息（MI）是用来度量两个聚类结果的相似性。给定两种聚类结果C和C'，MI的定义如式（19）所示。

$$MI(C,C')=\sum_{c_i\in C,c_j'\in C'}p(c_i,c_j')\cdot\log\frac{p(c_i,c_j')}{p(c_i)\cdot p(c_j')} \tag{19}$$

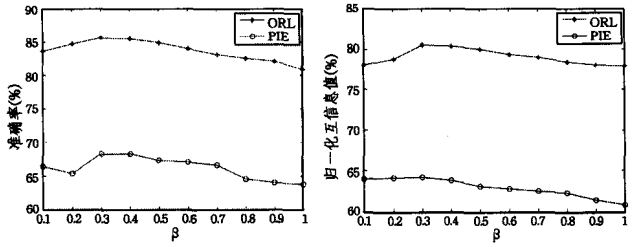
其中，$p(c_i)$，$p(c_j')$分别表示样本属于第c_i类和第c_j'类的概率，$p(c_i,c_j')$表示样本同时属于第c_i类和第c_j'类的联合概率。即

$$MI(C,C')\in[0,\max(H(C),H(C'))]$$

其中，$H(C)$，$H(C')$分别表示聚类C和C'的熵。当$MI(C,C')=\max(H(C),H(C'))$时，表示聚类$C$和$C'$结果一致。当$MI(C,C')=0$时，表示聚类$C$和$C'$完全独立。其中，$MI(C,C')$有一个重要的性质，即对于一种聚类值的各种排列求互信息值时，结果总是固定的。也就是说，互信息值与聚类值的排列顺序无关。本文进一步对其值按式（20）进行了归一化处理：

$$MI_n(C,C')=\frac{MI(C,C')}{\max(H(C),H(C'))}\in[0,1] \tag{20}$$

为了确定稀疏系数β，分别在两个数据集上进行了实验，研究了β对准确率和归一化互信息值的影响，实验结果如图3所示。依据图3，选择$\beta=0.32$。



（a）β与准确率的关系　（b）β与归一化互信息的关系

图3　β与准确率及归一化互信息的关系

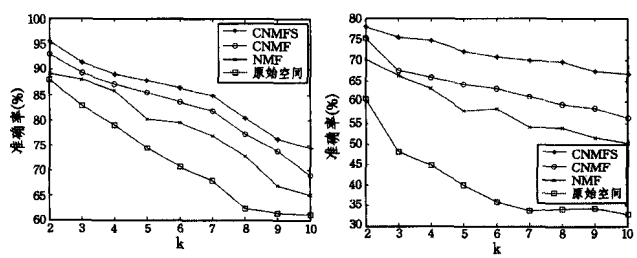
本文的对比实验包括：原始数据集上进行的聚类实验；分别利用NMF算法和CNMF算法分解后的聚类实验。对每个k值都运行10次后取平均值作为最终的结果，这样处理所得实验结果更加合理，各算法在两个数据集上聚类的准确率见表2和表3，相对应的聚类准确度的曲线如图4所示；聚类的归一化互信息见表4和表5，相对应的曲线如图5所示。实验中的对比算法CNMF算法与CNMFS算法同样都是随机取k（$k=2,3,\cdots,10$）类进行实验。

表2　不同算法在ORL数据集上聚类的准确率AC

k	聚类的准确率(AC)/(%)			
	原始空间	NMF空间	CNMF空间	CNMFS空间
2	87.96	89.23	93.00	95.61
3	82.91	87.99	89.33	91.42
4	79.00	85.79	87.07	89.12
5	74.40	80.20	85.44	87.81
6	70.76	79.50	83.52	86.37
7	67.89	76.86	81.86	84.86
8	62.26	72.75	77.25	80.38
9	61.33	67.01	73.89	76.33
10	61.02	65.00	69.00	74.40
平均值	71.95	78.26	82.26	85.14

表3　不同算法在PIE_pose27数据集上聚类的准确率AC

k	聚类的准确率(AC)/(%)			
	原始空间	NMF空间	CNMF空间	CNMFS空间
2	60.72	70.31	75.31	78.00
3	48.02	66.32	67.57	75.59
4	44.89	63.32	65.88	74.88
5	40.00	57.95	64.29	72.24
6	35.87	58.40	63.25	70.85
7	33.77	54.03	61.32	70.00
8	34.18	53.89	59.51	69.76
9	34.36	51.55	58.65	67.29
10	32.92	50.29	56.27	66.68
平均值	40.53	58.45	63.56	71.70



（a）ORL数据集　（b）PIE_pose27数据集

图4　聚类准确率

由表2和表3可知，NMF算法、CNMF算法和CNMFS算法在两个数据集上的聚类准确率平均值相对于原始数据聚类有较大的改善。在ORL数据集中，CNMFS算法比原始空间上的聚类准确率平均高13.19%，比NMF算法平均高

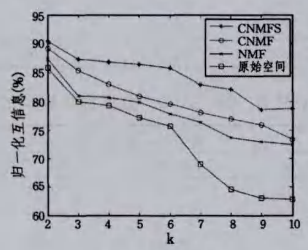
6.8%，也比 CNMF 算法的聚类效果好。同样地，在 PIE_pose27 数据集中，CNMFS 算法比原始空间上聚类准确率平均高 31.17%，比 NMF 算法平均高 13.25%，比 CNMF 算法平均高 8.14%。结合图 4 可以看出，CNMFS 算法的聚类效果最好，CNMF 算法的其次，在原始数据基础上的聚类效果是最差的。

表 4　不同算法在 ORL 数据集上的互信息

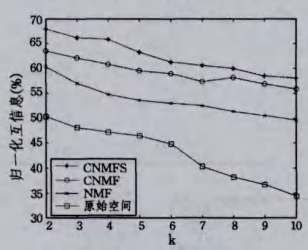
k	聚类的互信息值/(%)			
	原始空间	NMF 空间	CNMF 空间	CNMFS 空间
2	85.89	87.34	89.23	90.45
3	80.02	81.01	85.39	87.42
4	79.33	80.77	83.09	86.82
5	77.14	79.89	80.92	86.46
6	75.71	77.90	79.58	85.83
7	68.99	76.34	78.06	83.00
8	64.58	73.71	76.90	82.17
9	63.02	72.98	75.94	78.60
10	62.85	72.42	73.33	78.81
平均值	73.06	78.04	80.27	84.40

表 5　不同算法在 PIE_pose27 数据集上的互信息

k	聚类的互信息值/(%)			
	原始空间	NMF 空间	CNMF 空间	CNMFS 空间
2	50.26	60.34	63.48	68.00
3	48.03	57.00	62.07	66.17
4	47.16	54.69	60.77	65.89
5	46.44	53.61	59.52	63.30
6	44.84	53.00	58.98	61.29
7	40.36	52.50	57.32	60.54
8	38.29	51.33	58.16	60.02
9	36.76	50.58	56.90	58.46
10	34.38	49.61	55.84	58.00
平均值	42.95	53.63	59.23	62.41



(a)ORL 数据集



(b)PIE_pose27 数据集

图 5　聚类归一化互信息

由表 4 和表 5 可以看出，NMF 算法、CNMF 算法和 CNMFS 算法的归一化互信息平均值比原始空间中要高得多。在 ORL 数据集中，CNMFS 算法比原始空间上的聚类互信息平均高 11.34%，比 NMF 算法平均高 6.36%，比 CNMF 算法平均高 4.13%。在 PIE_pose27 数据集中，CNMFS 算法比原始空间上聚类准确率平均高 19.46%，比 NMF 算法平均高 8.78%，比 CNMF 算法平均高 3.18%。由图 5 可更直观地看出 CNMFS 算法的效果是最好的。

总体来看，CNMFS、CNMF 及 NMF 在维数约简后数据上的聚类效果远比在原始数据集上的聚类效果好；改进的 NMF 算法比原 NMF 算法的聚类效果好。实验表明，所提出的 CNMFS 取得了最好的聚类效果。

4.4　基图像的稀疏度

一种稀疏度的度量方法是利用了向量 1-范数和 2-范数的关系[10]，定义如式(21)。

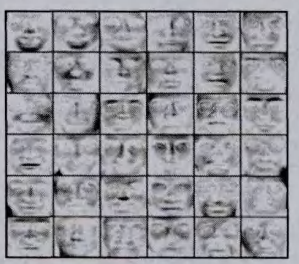
$$sparseness(x)=\frac{1}{n-1}[n-(\|x\|_1/\|x\|_2)^2] \tag{21}$$

其中，$\|\cdot\|_1$ 是向量的 1 范数，$\|\cdot\|_2$ 是向量的 2 范数，n 是向量 x 的维度。$0\leqslant sparseness(x)\leqslant 1$，当且仅当 x 仅有一个非零元时，$sparseness(x)=1$；当所有元素都相等且不为零时，$sparseness(x)=0$。

针对经 NMF、CNMF 和 CNMFS 得到的基向量 U，利用式(21)计算其稀疏度。

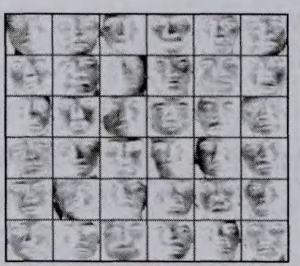
(a)NMF 稀疏度 0.4882

(b)CNMF 稀疏度 0.5125

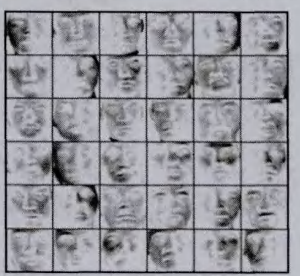


(c)CNMFS 稀疏度 0.5642

图 6　ORL 数据集的基图像



(a)NMF 稀疏度 0.5635



(b)CNMF 稀疏度 0.6742

(c)CNMFS 稀疏度 0.7363

图 7　PIE_pose27 数据集的基图像

(下转第 304 页)

试集上达到了89.3%的识别准确率，置信度和几何模型的使用对识别性能都有一定程度的贡献。另外，所提出的压缩字符分类器方法效果尤为明显，压缩后字符分类器存储在外部存储设备上时所占用的空间和运行时在内存中所占用的存储空间都大幅度减小，且压缩后对识别准确率和识别速度的影响都不大。本文所研究的叠写识别系统在Android平台上予以部署，实现了一个支持叠写中文输入功能的输入法应用，软件运行流畅，效果良好。

为了使叠写识别系统有更高的识别准确率，进一步的研究工作可以从获取真实的叠写样本和在实现几何模型时使用更多有效的几何信息这两个方面来考虑。

参 考 文 献

[1] Shimodaira H, Sudo T, Nakai M, et al. On-line Overlaid-Handwriting Recognition Based on Substroke HMMs[C]// Seventh International Conference on Document Analysis and Recognition. Washington DC: IEEE, 2003: 1043-1047

[2] Wan Xiang, Liu Chang-song, Zou Yan-ming. On-line Chinese Character Recognition System for Overlapping Samples[C]// 2011 International Conference on Document Analysis and Recognition. Washington DC: IEEE, 2011: 799-803

[3] Zou Yan-ming, Liu Ying-fei, Liu Ying, et al. Overlapped handwriting input on mobile phones[C]// 2011 International Conference on Document Analysis and Recognition. Washington DC: IEEE, 2011: 369-373

[4] Wang Qiu-feng, Yin Fei, Liu Cheng-lin. Handwritten Chinese Text Recognition by Integrating Multiple Contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1469-1481

[5] Yin Fei, Wang Da-han, Wang Qiu-feng. CASIA Online and Offline Chinese Handwriting Databases[C]// 2011 International Conference on Document Analysis and Recognition. Washington DC: IEEE, 2011: 37-41

[6] Wang Da-han, Liu Cheng-lin, Zhou Xiang-dong. An approach for real-time recognition of online Chinese handwritten sentences [J]. Pattern Recognition, 2012, 45(10): 3661-3675

[7] Liu Cheng-lin. Classifier Combination Based on Confidence Transformation[J]. Pattern Recognition, 2005, 38(1): 11-28

[8] Teng Long, Jin Lian-wen. Building compact MQDF classifier for large character set recognition by subspace distribution sharing [J]. Pattern Recognition, 2008, 41(9): 2916-2925

[9] Wang Yong-qiang, Huo Qiang. Building compact recognizers of handwritten Chinese characters using precision constrained Gaussian model, minimum classification error training and parameter compression[J]. International Journal on Document Analysis and Recognition, 2011, 14(3): 255-262

[10] 杨军. 聚类分析及其在大类别汉字识别中的应用[D]. 广州：华南理工大学，2007
Yang Jun. Application of the Clustering Analysis in the Large vocabulary Chinese Character Recognition [D]. Guangzhou: South China University of Technology, 2007

(上接第284页)

实验中ORL和PIE_pose27的类数皆取36。稀疏后的基图像如图6和图7所示。由图6和图7可知，NMF算法的稀疏度较差，CNMF的稀疏度也不高，CNMFS算法的稀疏度最高。由此表明，CNMFS算法可以更好地得到图像的局部表示。

结束语 本文提出了稀疏约束的半监督非负矩阵分解，并给出了相应的迭代公式以及收敛性的证明。以流行的人脸数据库为实验对象，应用本文提出的新方法进行了实验，利用两个评价指标(聚类准确率和归一化互信息)来衡量所提算法识别精度的好坏。实验表明，算法得出了更好的识别率和聚类性能。进一步考察算法的稀疏度，结果显示所提出的算法的稀疏度最高。该算法能够得到图像的最佳局部表示，使得基图像具有更好的判别能力。

参 考 文 献

[1] Lee D D, seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791

[2] 杜世强，石玉清，王维兰，等. 基于图正则化的半监督非负矩阵分解[J]. 计算机工程与应用，2012，48(36)：194-200
Du Shi-qiang, Shi Yu-qing, Wang Wei-lan, et al. Graph regularized-based semi-supervised non-negative matrix factorization [J]. Computer Engineering and Applications, 2012, 48(36): 194-200

[3] Cai Deng, He Xiao-fei, Han Jia-wei, et al. Graph regularized nonnegative matrix factorization for data representation[J]. IEEE Trans on Pattern Anal Mach Intell, 2011, 33(8): 1548-1560

[4] Hoyer P O. Non-negative matrix factorization with sparseness constrains[J]. Journal of Machine Learning Research, 2004, 5(9): 1457-1469

[5] Sun Fu-ming, Tang Jin-hui, Li Hao-jie, et al. Multi-label image categorization with sparse factor representation [J]. IEEE Transaction on Image Processing, 2014, 23(3): 1028-1037

[6] Li S Z, Hou Xin-wen, Zhang Hong-jiang, et al. Learning spatially localized, parts-based representation [C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos, California, USA, 2001 (1): 207-212

[7] Liu Hai-feng, Wu Zhao-hui, Li Xue-long, et al. Constrained nonnegative matrix factorization for image representation [J]. IEEE Trans on Pattern Anal Mach Intell, 2012, 34(7): 1299-1311

[8] Shahnaza F, Berrya M W, Paucab V, et al. Plemmonsb. Document clustering using nonnegative matrix factorization[J]. Information Processing Management, 2006, 42(2): 373-386

[9] Lovasz L, Plummer M. Matching Theory [M]. North Holland, 1986

[10] Michael W, Shakhina A, Stewart G W. Computing sparse reduced-rank approximations to sparse matrices [J]. ACM Transactions on mathematical software, 2004, 19(3): 231-235