文章编号:1001 - 9081(2010)05 - 1280 - 04

基于加权的不完备非负矩阵分解算法

杨志君, 叶东毅
(福州大学 数学与计算机科学学院,福州 350108)
(flistorm@ gmail. com)

摘　要:非负矩阵分解(NMF)作为一种特征提取与数据降维的新方法,相较于一些传统算法,具有实现上的简便性,分解形式和分解结果上的可解释性等优点。但当样本矩阵不完备时,NMF 无法对其进行直接分解。提出一种基于加权的不完备非负矩阵分解(NMFI)算法,该算法在处理不完备样本矩阵时,先采用随机修复的方法降低误差,再利用加权来控制各样本的权重,尽量削弱缺损数据对分解结果产生的干扰。此外,NMFI 算法使用区域权重来进一步减少关键区域数据缺损对分解产生的影响。实验结果表明,NMFI 算法能有效提取样本中残余数据的信息,减少缺损数据对分解结果的影响。

关键词:非负矩阵分解;不完备数据集;随机修复;加权;区域权重
中图分类号: TP391　　**文献标志码**:A

Weighted non-negative matrix factorization for incomplete dataset

YANG Zhi-jun, YE Dong-yi
(*College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian* 350108, *China*)

Abstract: Nonnegative Matrix Factorization (NMF) is a new method for feature extraction and data dimension reduction. It has an advantage over traditional algorithms in the simple implementation and the interpretability of factorization form and factorization result. But NMF could not decompose the samples matrix when it is incomplete. However, when dealing with incomplete dataset, NMFI (Weighted Non-negative Matrix Factorization for Incomplete Dataset) made use of random repair to decrease the error and weighted method to control weights of the samples, which could weaken the disturbance of missing data as much as possible. In addition, NMFI used regional weight for further reducing the impact of missing data in critical region. The experimental results demonstrate that NMFI can effectively extract information from retained data and reduce the influence of missing data.

Key words: Nonnegative Matrix Factorization (NMF); incomplete dataset; random repair; weighting; regional weight

《计算机应用》Journal of Computer Applications

0　引言

非负矩阵分解(Nonnegative Matrix Factorization, NMF)[1]是机器学习领域中一种高效的数据降维方法,也是一种特征提取的新方法,与传统的矩阵分解方法不同,它的分解过程和结果都是非负的,这在心理学上符合人对整体的感知是由对组成整体的部分的感知构成的(纯加性的)。近年来非负矩阵分解方法在聚类分析、人脸识别、语音识别、语音分离[2-4]等领域取得了很好的应用。但 NMF 也存在一些缺陷,当样本矩阵由于各种原因而缺损时,NMF 无法对其进行直接处理。另外,若其部分数据由于某种原因而产生较大偏差,则对分解结果的准确性会有一定程度的干扰。

针对以上问题,本文提出了一种基于加权的不完备非负矩阵分解算法 NMFI。算法首先对不完备数据进行随机修复,使用其他样本中对应位置的未缺失数据来修复缺失数据,使修复之后的样本与原始样本误差尽量小。此外,考虑到样本的重要性各不相同,即无损或缺损程度较低的样本重要性应当比缺损程度高的样本高,算法采用加权的方法来控制各样本的权重,使缺失越大的样本权重越低,以避免缺损数据对分解结果产生较大干扰。最后,考虑了区域权重。所谓区域权重即在单个样本中某些数据的重要性明显低于其他数据,这些数据缺失时对分解结果的影响比其他数据缺失时的影响要小,即对于该样本的权重而言,这些数据的贡献较小。

1　非负矩阵分解

给定一个 n 行 m 列的非负矩阵 $\boldsymbol{V}$,将其近似地分解成两个非负矩阵之积,分别记为 $\boldsymbol{W}$ 与 $\boldsymbol{H}$,使得 $\boldsymbol{V} \approx \boldsymbol{WH}$,即:

$$V_{ij} \approx (\boldsymbol{WH})_{ij} = \sum_{a=1}^{r} W_{ia} H_{aj} \tag{1}$$

分解之后的矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{H}$ 的维数分别为 $n \times r$ 和 $r \times m$,r 的选取与 $\boldsymbol{V}$ 的秩有关,$r \ll \mathrm{rank}(\boldsymbol{V})$。

如果用 $\boldsymbol{v}$ 和 $\boldsymbol{h}$ 分别表示矩阵 $\boldsymbol{V}$ 与 $\boldsymbol{H}$ 中对应的列向量,由式(1)可得 $\boldsymbol{V} \approx \boldsymbol{Wh}$。由此可见,原始矩阵 $\boldsymbol{V}$ 中的列向量 $\boldsymbol{v}$ 可以通过矩阵 $\boldsymbol{W}$ 以一定的权重 $\boldsymbol{h}$ 表示出来。因此,矩阵 $\boldsymbol{W}$ 可以看作是基矩阵,其中的每一列代表基向量,而矩阵 $\boldsymbol{H}$ 中的每一列代表与矩阵 $\boldsymbol{W}$ 中基向量相对应的编码系数。为了找到一个矩阵分解 $\boldsymbol{W}$ 与 $\boldsymbol{H}$,使得 $\| \boldsymbol{V} - \boldsymbol{WH} \|^2$ 最小,且满足 $\boldsymbol{W} \geqslant 0, \boldsymbol{H} \geqslant 0$,可以使用如下的规则进行迭代计算[1]:

$$H_{aj} \leftarrow H_{aj} \frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})_{aj}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH})_{aj}} \tag{2}$$

$$W_{ia} \leftarrow W_{ia} \frac{(\boldsymbol{VH}^{\mathrm{T}})_{ia}}{(\boldsymbol{WHH}^{\mathrm{T}})_{ia}} \tag{3}$$

收稿日期:2009 - 10 - 09;**修回日期**:2010 - 01 - 19。　**基金项目**:国家自然科学基金资助项目(60805042)。
作者简介:杨志君(1985 -),男,福建莆田人,硕士研究生,主要研究方向:非负矩阵分解;　叶东毅(1964 -),男,福建南安人,教授,博士生导师,主要研究方向:计算智能、数据挖掘、神经网络。

在计算过程中,随机选取非负的初始矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{H}$,定义适当的迭代中止条件之后,按照式(2)和(3)进行迭代计算即可得到 $\boldsymbol{W}$ 和 $\boldsymbol{H}$。

2　本文算法的基本思想

在实际应用中,经常由于某些主观或客观因素导致数据的缺失,如采集到的图像由于镜头存在污点而产生的缺损,道路监控中由于物体遮挡而产生的数据缺失等,对于这些数据 NMF 算法无法对其进行直接处理。鉴于此,本文提出一个基于加权的不完备非负矩阵分解算法 NMFI。NMFI 算法的目标是当样本不完备时,尽量提取样本中残余数据的信息,减少缺损数据对分解结果的影响。

2.1　算法的基本思想

在样本数据的采集和存储过程中,时常会出现数据缺失的现象,而 NMF 算法无法对这些不完备数据进行直接分解并分析,它要求数据矩阵必须是完备的。因此,首先要对不完备样本进行修复。一个简单的方法是采用单个数据值来修复缺失数据,再进行 NMF 分解。但由于样本的取值是多样性的,使用单一值来代替缺失数据会使修复后的样本与原始样本误差较大,进而影响分解结果。鉴于此,本文采用随机修复的方法对缺失数据进行修复,即使用其他样本中对应位置的未缺失数据来修复缺失数据,从而减少修复后的样本和原始样本之间的误差,避免缺损数据对分解结果产生较大干扰。

将不完备样本修复为完备样本后,若直接对其进行 NMF 分解,效果显然是不理想的。由于样本的缺失程度不尽相同,缺损较大的样本重要性应比无损或缺损较小的样本低,如此才能尽量减少缺损数据对分解结果的影响;反之,即加强较完备样本的影响。本文采用对各样本加权的方式,即控制各样本的权重来控制样本对分解结果的影响。

2.2　单一值修复与随机修复

为比较单一值修复和随机修复之间的差异,首先对使用单一值修复后的不完备样本矩阵进行 NMF 分解。文中所有实验均在 Windows Vista 环境下进行,CPU 为酷睿双核 2.20 GHz,内存 2 GB,Matlab 为 7.0 版。且 NMF 算法的迭代终止条件皆按如下定义。

设样本数为 m 时,算法第 i 次迭代各样本所产生的平均误差为:

$$T_i = \| \boldsymbol{V} - \boldsymbol{W}_i\boldsymbol{H}_i \|^2/m$$

其中 $\boldsymbol{W}_i$,$\boldsymbol{H}_i$ 分别为第 i 次迭代分解产生的基矩阵和编码矩阵,则迭代中止条件为 $1-(T_{i+1}/T_i) \leqslant 0.01\%$。

实验采用 Swimmer 数据集[5],该数据集包含一系列黑白图像,每个图像有 4 个可移动的部分(四肢),每个部分可以旋转到 4 个不同的位置(关节)。且每个黑白图像的中心有一个包含 12 个像素的躯干和四个包含 6 个像素的可移动部分,每个可移动部分可位于 4 个不同位置,因此总共有 256 张图像包括所有的肢体位置组合。每张图像包含 32×32 个像素。图 1 给出了该数据集的一个子集。

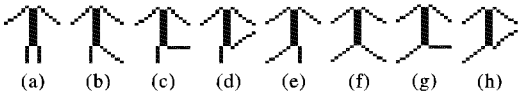


图 1　Swimmer 数据集

该数据集在文献[5]中是用来展示 NMF 算法发现局部特征(四肢)的能力。它遵守一些预定义的规则,如可分性和完全因子取样,即它是通过一种具有 NMF 样式的模型产生的。

实验时,取数据集中的前 20 张图像,保留前 2 张图像完整,对后 18 张图像采用随机去除元素的方法产生缺损图像。则使用单一值“1”修复缺损图像,并对其进行 NMF 分解后,分解产生的基图像如图 2。实验中基图像个数取为 12。

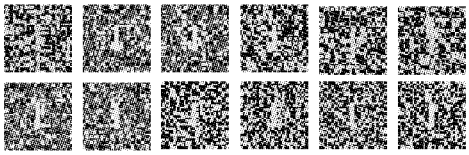


图 2　单一值修复时 NMF 分解的基图像

从图 2 可以看出,分解产生的基图像中较难分辨出四肢和躯干。这是由于使用单一值修复时,修复后的图像和原始图像之间存在较大误差,缺损数据会对分解结果产生较大影响。

NMFI 算法采用随机修复的方法对缺损数据进行修复,即用其他样本中无缺损的数据来修复样本中同一位置的缺损数据。在实际应用中,进行 NMF 分解的各样本间通常具有一定程度的相似性,因此该方法可以有效减少修复后的图像和原始图像之间的误差,减少缺损数据对分解结果的影响。图 3 是对缺损图像进行随机修复后,使用 NMF 分解产生的基图像。

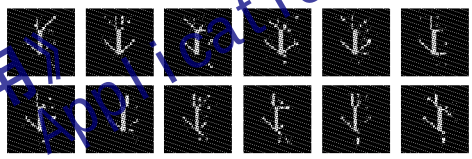


图 3　随机修复时 NMF 分解的基图像

从图 3 中可以看出,随机修复后分解产生的基图像四肢和躯干都比较明显。由于修复后的图像和原始图像之间误差较小,因此受到缺损数据的干扰较小。

2.3　样本加权

在不完备样本矩阵中,各样本的缺损程度不尽相同,缺损越大的样本其误差也越大,对分解结果的影响则应更小,即重要性更低。本文采用加权的方式来区分不同样本的重要性,重要的样本权重较高,反之则较低。

2.3.1　相关公式及定义

假设 $\boldsymbol{V}_n$ 为需分解的样本矩阵,n 为样本数,$k \leqslant n$ 为完备样本数,完备样本构成的样本矩阵为 $\boldsymbol{V}_k$,$(n-k)$ 个不完备样本的权重分别为 $p_{k+1},\cdots,p_n$,且 $0 \leqslant p_i \leqslant 1, i = k+1, k+2, \cdots, n$。$\boldsymbol{W}$ 为分解产生的基矩阵,$\boldsymbol{H}_k$ 为 k 个完备样本的编码矩阵,$\boldsymbol{h}_{k+1},\boldsymbol{h}_{k+2},\cdots,\boldsymbol{h}_n$ 为不完备样本基于基矩阵的编码向量,则算法的目标函数为:

$$F = \| \boldsymbol{V}_k - \boldsymbol{W}\boldsymbol{H}_k \|^2 + \sum_{i=k+1}^{n} p_i \| \boldsymbol{v}_i - \boldsymbol{W}\boldsymbol{h}_i \|^2 \tag{4}$$

使用梯度下降法来取得 F 的局部最小值。

$$W_{ia} \leftarrow W_{ia} - \mu_W \frac{\partial F}{\partial W_{ia}} \tag{5}$$

$$H_{ia} \leftarrow H_{ia} - \mu_H \frac{\partial F}{\partial H_{ia}} \tag{6}$$

当步长 μ_H 取式(7)的值时,式(5)转换为式(8)所示的乘法更新规则。其中 $\boldsymbol{w}_i$ 为 $\boldsymbol{W}$ 的第 i 列,$\boldsymbol{h}_t$ 为 $\boldsymbol{H}$ 的第 t 列。

$$\mu_H = \frac{W_{ia}}{\sum_{t=1}^{k} \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{h}_t H_{at} + \sum_{t=k+1}^{n} p_t \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{h}_t H_{at}} \tag{7}$$

Journal of Computer Applications 计算机应用

$$W_{ia} \leftarrow W_{ia} \frac{\sum_{t=1}^{k} V_{it} H_{at} + \sum_{t=k+1}^{n} p_t V_{it} H_{at}}{\sum_{t=1}^{k} \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{h}_t H_{at} + \sum_{t=k+1}^{n} p_t \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{h}_t H_{at}} \tag{8}$$

即：

$$W_{ia} \leftarrow W_{ia} \frac{(\boldsymbol{V}\boldsymbol{H}'^{\mathrm{T}})_{ia}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}'^{\mathrm{T}})_{ia}} \tag{9}$$

其中 $\boldsymbol{H}' = [\boldsymbol{H}_k \quad p_{k+1}\boldsymbol{h}_{k+1} \quad p_{k+2}\boldsymbol{h}_{k+2} \quad \cdots \quad p_n \boldsymbol{h}_n]$。

当步长 μ_H 取式(10)的值时，式(6)转换为式(11)所示的乘法更新规则，其中 m 为样本的维数。

$$\mu_H = \frac{H_{ia}}{\sum_{t=1}^{m} p_a \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{h}_a W_{ti}} \tag{10}$$

$$H_{ia} \leftarrow H_{ia} \frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{v}_a)_{ia}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}_a)_{ia}} \tag{11}$$

即：

$$H_{ia} \leftarrow H_{ia} \frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})_{ia}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H})_{ia}} \tag{12}$$

其中 $\boldsymbol{v}_a$ 为 $\boldsymbol{V}$ 的第 a 个列向量。

以上更新规则的收敛性证明请参阅文献[6]中所采用的设计辅助函数的方法，在此不作赘述。

为计算各样本的权重，作如下定义。

定义 1 缺损度。样本中缺损的数据个数和样本总数据个数的比值称为样本的缺损度。

定义 2 高完整数。若样本 x 的缺损度为 θ，样本矩阵中缺损度小于 θ 的样本个数为 β，则称 β 为样本 x 的高完整数。

2.3.2 加权对分解结果的影响

从图 3 所示的基图像中可以看出，除四肢和躯干外，仍有由缺损数据引起的分解误差。这是由于在这个实验中不完备样本和完备样本的权重取值相同，即各样本的重要性没有差异，导致分解误差仍然较大。对样本的重要性产生影响的因素除了样本的缺损程度外，还与高完整数有关。若在一个样本矩阵中，完备样本的个数远大于不完备样本，显然，直接根据完备样本分解得到基图像和整个样本矩阵分解得到的基图像几乎是相同的。此时，不完备样本的权重可以为 0。若绝大部分的样本缺损度均小于某个值 θ，仅有少数样本高于 θ，则高于 θ 的少数样本权重可以为 0。因此样本的权重和高完整数是密切相关的。

下面采用式(9)和(12)的乘法更新规则对图 3 中已修复的样本矩阵进行分解，各样本的权重取值为：

$$p = ((1-\theta)/\beta)^2 \tag{13}$$

其中：θ 为样本缺损度；β 为样本的高完整数。加权后的迭代中止条件均按如下定义。

设样本数为 m 时，算法第 i 次迭代各样本所产生的平均误差为：

$$T_i = \left(\| \boldsymbol{V}_k - \boldsymbol{W}\boldsymbol{H}_k \|^2 + \sum_{i=k+1}^{n} p_i \| \boldsymbol{v}_i - \boldsymbol{W}\boldsymbol{h}_i \|^2 \right)/m$$

则迭代中止条件为 $1 - (T_{i+1}/T_i) \leqslant 0.01\%$。

分解产生的基图像如图 4。从图 4 中可以看出因数据缺损而产生的分解误差明显减少，这是由于降低了缺损度较高的样本的权重，使得缺损数据对分解结果的影响进一步降低。

在某些情况下使用随机修复和使用单一值修复产生的样本矩阵，它们与原始矩阵的误差相差不甚明显。这时权重的影响尤为明显，合适的权重在降低缺损数据影响的同时，可以尽量提取其余数据中的有用信息。利用下面的实验来解释这点。

在文献[7]中，temporal image 数据集被用来验证 ONMF 算法在满秩分解条件下处理动态数据集的有效性。图 5 是该数据集的部分生成因子，每个生成因子对应于一张 10×10 像素的图片，图片中包含一条横线或竖线。该数据集中的每张图片均是生成因子的某种线性组合，且产生的样本间没有明显的相似性，这也是实验采用该数据集的原因。

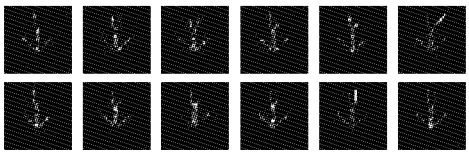


图 4 加权分解的基图像

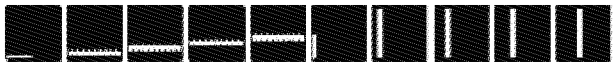


图 5 temporal image 数据集的生成因子

从该数据集中取出 200 张图像，保留前 100 张图像无缺损，剩余的 100 张图像随机去除元素生成不完备样本。基图像个数取为 20，随机修复后对其进行 NMF 分解得到的基图像如图 6。

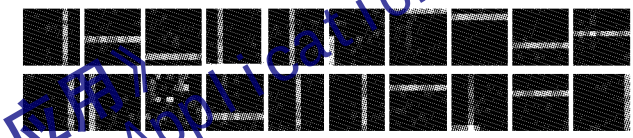


图 6 NMF 分解产生的基图像

从图 6 中可以看出，缺损数据对分解产生的基图像有较大的影响，所有基图像均有除生成因子之外的分解误差。

若采用式(9)和(12)的乘法更新规则并对其采用加权的方式进行分解，各样本的权重取值为：

$$p = ((1-\theta)/\beta)^2$$

则产生的基图像如图 7。

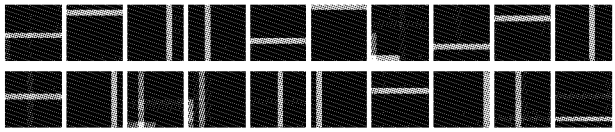


图 7 采用加权方式后分解产生的基图像

从图 7 中可以看出，由于降低了不完备样本的权重，基图像中受缺损数据干扰而产生的分解误差已经明显降低。

2.4 区域性加权对分解结果的影响

所谓区域权重是指在一个样本中，某些区域对于分解产生的基图像具有较大的影响，而其余区域如背景则影响较小。如 Swimmer 数据集包含的样本中，除躯干和四肢围绕关节活动的区域外，其余均可视为背景。若有两个缺损度相同的样本，一个主要是关键区域数据缺损，而另一个则主要是非关键区域如背景数据缺损，则显然第二个样本的重要性要高于第一个样本。在类似的应用中，需要考虑区域性权重。

实验时，从 Swimmer 数据集中取 20 个样本，通过随机去除元素使所有样本的缺损度在 0.4 左右，其中前 10 个样本主要是背景区域缺损，后 10 个样本主要是关键区域缺损。另外，在单个样本中，背景区域元素个数为 688，关键区域元素个数为 366。背景区域与关键区域的重要程度之比为 1∶9。

此时，若直接通过缺损元素个数来计算样本的缺损度，并按照式(13)得出各样本的权重，则随机修复后样本矩阵分解得到的基图像如图 8。

Journal of 计算机应用 Computer Applications

从图 8 中可以看出,基图像中关键区域受到缺损数据的干扰较大,四肢也出现了不同程度的误差。这是由于实验时背景区域缺损的数据和关键区域缺损的数据对样本权重的影响已被默认为一致,导致关键区域缺损较大的样本对分解后的基图像产生较大的干扰。显然,应当相对降低关键区域缺损较大的样本的权重。

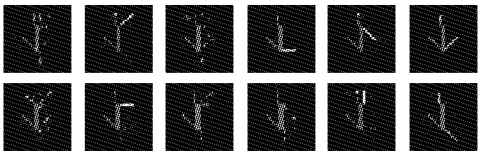


图 8　直接计算缺损度后分解产生的基图像

接下来按区域权重来计算样本的缺损度,计算公式如下:

$$\theta = \frac{0.1}{688} \cdot x + \frac{0.9}{366} \cdot y \tag{14}$$

其中:x 为背景区域缺损的元素个数;y 为关键区域缺损的元素个数。

再根据式(13)的权重计算公式得出各样本权重,则样本矩阵随机修复后分解得到的基图像如图 9。

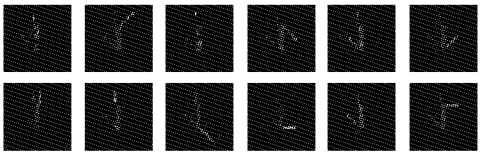


图 9　根据区域权重分解产生的基图像

从图 9 中可以看出提取的信息有所减少,即四肢和躯干相对模糊,但基图像的关键区域受到缺损数据的干扰变小。这是由于关键区域重要性较背景区域高,后 10 个图像的权重较前 10 张图像的权重低,因此分解产生的基图像受关键区域数据缺损的干扰降低。

2.5　算法设计

根据前面的分析,本文提出的 NMFI 算法的具体步骤如下。

1)扫描样本矩阵,根据区域重要性计算各样本的缺损度 θ 及高完整数 β。

2)计算各样本的权重 p。

3)采用随机修复的方法对不完备样本进行修复。

4)对基矩阵 $\boldsymbol{W}$ 和编码矩阵 $\boldsymbol{H}$ 随机初始化,并将 $\boldsymbol{H}$ 赋予 $\boldsymbol{H}_{\text{temp}}$。

5)将 $\boldsymbol{H}_{\text{temp}}$ 赋予 $\boldsymbol{H}$,使用式(12)乘法更新规则更新 $\boldsymbol{H}$,并将新的编码矩阵赋予 $\boldsymbol{H}_{\text{temp}}$。

6)使用式(9)乘法更新规则更新 $\boldsymbol{W}$,其中 $\boldsymbol{H}' = [\boldsymbol{H}_k \quad p_{k+1}\boldsymbol{h}_{k+1} \quad p_{k+2}\boldsymbol{h}_{k+2} \quad \cdots \quad p_n\boldsymbol{h}_n]$,即 $\boldsymbol{H}'$ 通过编码向量加权得到。

7)符合迭代中止条件,算法结束,否则转 5)。

3　Triangle 数据集的实验结果

在文献[11]中 Triangle 数据集被用于评价算法识别数据集生成因子的能力,且各张图像均有关键区域和背景区域。该数据集的生成因子是三张 32×32 的灰度图,每张图像包含一个大小不同的三角形,如图 10。

图 11 是根据生成因子产生的该数据集的一部分图像。

实验时,从 Triangle 数据集中取出 75 张图像,每张图像为一个样本,即样本矩阵的一列。保留前 25 张图像的完整性,其余 50 张图像随机去除元素产生不完备样本,基图像个数取为 3。则采用单一值 0 对不完备样本矩阵进行修复时,NMF 分解产生的基图像如图 12。

图 10　Triangle 数据集生成因子

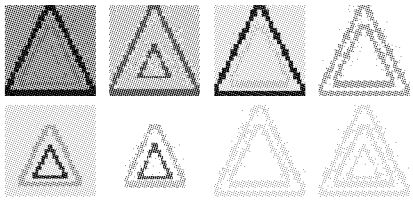
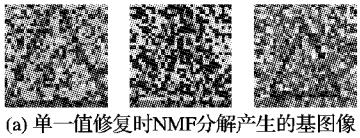
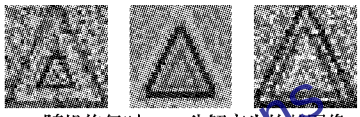


图 11　Triangle 数据集部分图像



(a) 单一值修复时NMF分解产生的基图像



(b) 随机修复时NMF分解产生的基图像

(c) NMFI分解产生的基图像

图 12　单一值修复时 NMF 分解产生的基图像

在 Triangle 数据集中,若采用单值修复,会使修复后的样本矩阵与原始图像相比有较大误差,缺损数据对分解结果干扰会比较明显,因此从图 12(a)的分解结果中很难分辨出该数据集的生成因子。采用随机修复的方法对不完备样本进行修复后,NMF 分解产生的基图像如图 12(b)。从图 12(b)中可以看出,随机修复使分解误差降低,但由于各样本的权重已默认为相同,缺损数据对分解结果的干扰仍然较大。

若设单一样本中三角形所在区域与背景区域的重要程度之比为 9∶1,则采用区域权重的方法按 2.5 节中的算法步骤,对不完备数据矩阵进行分解后产生的基图像如图 12(c)。缺损度的计算公式如下:

$$\theta = \frac{0.1}{528} \cdot x + \frac{0.9}{496} \cdot y$$

其中:x 为背景区域缺损的元素个数;y 为关键区域缺损的元素个数。

从图 12(c)中可以看出,由于削弱了关键区域缺损度较高样本的影响,分解产生的基图像误差进一步降低。

4　结语

本文在 NMF 算法的基础上提出了一种基于加权的不完备非负矩阵分解算法 NMFI。实验表明该算法能有效提取未缺损数据中的信息,避免缺损数据对分解结果产生较大的干扰。该算法中权重的取值至关重要,合适的权重不但能降低不完备数据的干扰,而且能最大限度提取残存数据中的信息。关于权重的取值方法有待我们今后进一步的研究。此外,对于图像恢复,本文算法的思想也提供了一个新的思路,为今后在这方面的研究提供一定的借鉴。

(下转第 1286 页)

《计算机应用》Journal of Computer Applications

3 实验及其结果分析

文献[6]采用了 UCI 数据集[9]中 16 个属性的 Zoo 数据和 9 个属性的 Winsconsin breast cancer 数据，本文为了在更复杂的数据上比较 UpdateENBROD 和 ENBROD 算法，采用 UCI 数据集中的 Thyroid disease records 数据。该数据集中有 3 772 条记录，每个记录有 30 个属性值，被标记为 negative(3 540 个记录，占 93.8%)和 sick(232 个记录，占 6.2%)。实验中，在标记为 negative 的数据中选分别随机选出 100，200，…，600 条数据，求出每个数据对象的 S_{ROF}，然后在标记为 sick 的数据中随机选出几条数据插入到前面的数据集中作为异常数据记录，分别采用 UpdateENBROD 和 ENBROD 求出每个数据对象的 S_{ROF}；然后随机删除一些数据对象，再次采用 UpdateENBROD 和 ENBROD 分别求出每个数据对象的 S_{ROF}。实验在 CPU 为 Intel core 2 T5500 1.66 GHz，内存 2 GB 的 PC 机上进行，算法用C++实现。

表 1 给出了算法 UpdateENBROD 和 ENBROD 在数据记录数为 600，k = 80 发现所有异常记录时排序后的最大记录数。

表 1 发现所有异常记录时的最大记录数

异常数据的条数	发现所有异常数据记录时的最大记录数	
	UpdateENBROD	ENBROD
3	3	3
5	5	5
10	11	11

图 3 给出了当数据集中有数据对象更新时，算法 UpdateENBROD 和 ENBROD 在不同数据记录数下的运行时间曲线图。

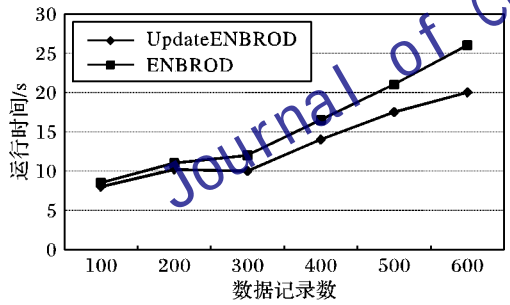


图 3 UpdateENBROD 和 ENBROD 的运行时间曲线图

由表 1 可以看出，当数据集里有 3 条和 5 条离群数据记录时，UpdateENBROD 与 ENBROD 都能准确地发现异常的数据，当数据集中插入了 10 条异常数据，两个算法都在前 11 条里面发现了所有的异常数据，所以 UpdateENBROD 与 ENBROD 都能较高效地发现离群数据的记录，检测离群数据的能力完全一样，从而说明了算法 UpdateENBROD 的正确性。当数据集中的数据发生动态更新的时候，从图 3 可以看出，当数据集的记录数不大的时候，两个算法的运行时间相差不大，随着数据集记录数的增加，UpdateENBROD 的运行时间小于 ENBROD。

4 结语

在研究 ENBROD 算法的基础上，本文提出了一种动态数据环境下的离群点检测算法，当数据集中的数据存在一些更新操作的时候，不必重新计算所有数据点的 ROF 值，而只需要计算受影响点的 ROF 值。实验表明，该算法在动态数据环境下的时间性能优于 ENBROD。

参考文献：

[1] HAWKINS D. Identification of outliers [M]. Berlin: Springer-Verlag, 1980.

[2] 薛安荣，鞠时光，何伟华，等. 局部离群点挖掘算法研究[J]. 计算机学报，2007，30(8)：1455－1463.

[3] BREUNIG M M, KRIEGEL H-P, NG T, *et al.* LOF: Identifying density-based local outliers [J]. ACM SIGMOD Record, 2000, 29(2): 93－104.

[4] CHAWLA S, SUN PEI. SLOM: A new measure for local spatial outliers [J]. Knowledge and Information Systems, 2006, 9(4): 412－429.

[5] BARBARÁ D, LI YI, COUTO J. Coolcat: An entropy-based algorithm for categorical clustering [C]// Proceedings of the 11th International Conference on Information and Knowledge Management. New York: ACM Press, 2002: 582－589.

[6] 于绍越，商琳. 基于信息熵的相对离群点的检测方法：ENBROD [J]. 南京大学学报：自然科学版，2008，44(2)：212－218.

[7] SHANNON C E. A mathematic theory of communication [J]. Bell System Technical Journal, 1948, XXVH(3): 379－423, 623－656.

[8] 杨风召，朱扬勇，施伯乐. IncLOF：动态环境下局部异常的增量挖掘算法[J]. 计算机研究与发展，2004，41(3)：477－484.

[9] MURPHY P M, AHA D W. UCI repository of machine learning database[EB/OL]. (1994－06) [2009－10－06]. http://www.ics.uci.edu/mlearn/MLRepository.html.

《计算机应用》Journal of Computer Applications

(上接第 1283 页)

参考文献：

[1] LEE D D, SEUNG H S. Learning the parts of objects by nonnegative matrix factorization [J]. Nature, 1999, 401(6755): 788－791.

[2] XU WEI, LIU XIN, GONG YIHONG. 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. New York: ACM Press, 2003: 267－273.

[3] GUILLAMET D, VITRIA J. Non-negative matrix factorization for face recognition [C]// Proceedings of the 5th Catalonian Conference on AI: Topics in Artificial Intelligence, LNCS 2504. Berlin: Springer-Verlag, 2002: 336－344.

[4] VIRTANEN T. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1066－1074.

[5] DONOHO D, STODDEN V. When does nonnegative matrix factorization give a correct decomposition into parts? [C]// Proceedings of the 17th Annual Conference Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2003: 1141－1148.

[6] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization[C]// Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2001, 13: 556－562.

[7] CAO BIN, SHEN DOU, SUN JIAN-TAO, *et al.* Detect and track latent factors with online nonnegative matrix factorization [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2007: 2689－2694.

[8] KLINGENBERG B, CURRY J, DOUGHERTY A. Non-negative matrix factorization: Ill-posedness and a geometric algorithm[J]. Pattern Recognition, 2009, 42(5): 918－928.