第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)04—0066—03 文献标识码：A 中图分类号：TP311.12

一种增量式非负矩阵分解算法

郭 立，张守志，汪 卫，施伯乐
(复旦大学计算机科学技术学院，上海 200433)

摘 要：针对现有的非负矩阵分解算法在应用于问题规模逐渐增大的情形时，运算规模随之增大、空间和时间效率不高的情况，提出一种增量式非负矩阵分解算法，使用分块矩阵的思想降低运算规模，利用上一步的分解结果参与运算从而避免重复运算。实验结果表明，该算法对节约计算资源是有效的。

关键词：非负矩阵分解；矩阵分解；增量式算法

Incremental Non-negative Matrix Factorization Algorithm

GUO Li, ZHANG Shou-zhi, WANG Wei, SHI Bai-le
(School of Computer Science and Technology, Fudan University, Shanghai 200433)

【Abstract】 When existing Non-negative Matrix Factorization(NMF) algorithm is applied to a problem of incremental scale, the consumption of space and time behaves inefficiency. This paper proposes an Incremental Nonnegative Matrix Factorization(INMF) algorithm, which uses partitioned matrix theory to reduce the computing scale, and uses decomposition results already derived to avoid re-calculating every time. Experimental results show that the algorithm performs efficiently for saving computing resources.

【Key words】 Non-negative Matrix Factorization(NMF); matrix factorization; incremental algorithm

1 概述

非负矩阵分解(Nonnegative Matrix Factorization, NMF)一直是很有意义的研究问题，文献[1]阐述了其基本思想，让这一问题为更多人所关注，文献[2]提出一种迭代算法，使得相关研究蓬勃展开。近几年中，NMF 成功地应用于图像、语音等多媒体信息的分析与处理，文本聚类与挖掘，生物医学、遗传学和化学的各种研究，以及环境数据处理、信号分析与目标识别等诸多方面[3-7]，并表现出很好的效果。但当前的各种应用数据量显著变大，像网络数据以及大型的图像数据库所形成的矩阵规模都非常大，在问题规模逐渐增大的情形时，直接使用现有的算法每次都重新运算全部数据，处理的矩阵规模太大，在保证效果的同时效率比较低下。本文提出了一种增量式非负矩阵分解(Incremental Nonnegative Matrix Factorization, INMF)算法，使得非负矩阵分解算法在应用于逐渐增大的矩阵时能利用上一步的分解结果，从而达到节省内存资源和时间消耗的效果。

2 非负矩阵分解算法

非负矩阵分解的思想可以描述为给定一个非负矩阵 $\boldsymbol{V}$，非负矩阵分解算法寻找非负矩阵 $\boldsymbol{W}$ 和非负矩阵 $\boldsymbol{H}$，使之满足：

$$\boldsymbol{V}\approx\boldsymbol{WH} \tag{1}$$

这种分解可以解释为，初始矩阵 $\boldsymbol{V}$ 中的每一列向量为对左矩阵 $\boldsymbol{W}$ 中所有基向量(行向量)的加权和，其权重系数为右矩阵 $\boldsymbol{H}$ 中对应列向量中的元素，因此，$\boldsymbol{W}$ 称作基矩阵，$\boldsymbol{H}$ 称作系数矩阵。

式(1)表明，非负矩阵分解形成的只是一种近似分解，并不要求 $\boldsymbol{V}$ 严格等于 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的乘积，为此有必要建立一个损失函数来评价分解前后近似程度。这样 NMF 问题的目标就变为找到 2 个非负矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{H}$，使得损失函数 $f(\boldsymbol{W},\boldsymbol{H})$ 达到最小。可用的损失函数有很多，常用的如 Euclidean 距离：

$$f(\boldsymbol{W},\boldsymbol{H})=\frac{1}{2}\|\boldsymbol{V}-\boldsymbol{WH}\|_F^2=\frac{1}{2}\sum_{i,j}(\boldsymbol{V}_{ij}-(\boldsymbol{WH})_{ij})^2 \tag{2}$$

范数为 F-范数。对应的迭代算法[2]如下：

算法 1 非负矩阵分解算法

对给定的非负矩阵 $\boldsymbol{V}\in\boldsymbol{R}^{m\times n}$ 和正整数 $r<\min(m,n)$，找到非负矩阵 $\boldsymbol{W}\in\boldsymbol{R}^{m\times r}$ 和 $\boldsymbol{H}\in\boldsymbol{R}^{r\times n}$，使得损失函数 $f(\boldsymbol{W},\boldsymbol{H})$ 达到最小。

(1)随机初始化 $\boldsymbol{W}_{ia}^1>0,\boldsymbol{H}_{bj}^1>0,\forall i,a,b,j$。

(2)对 k=1,2,…依照以下规则迭代：

$$\boldsymbol{H}_{bj}^{k+1}=\boldsymbol{H}_{bj}^{k}\frac{((\boldsymbol{W}^k)^{\mathrm{T}}\boldsymbol{V})_{bj}}{((\boldsymbol{W}^k)^{\mathrm{T}}\boldsymbol{W}^k\boldsymbol{H}^k)_{bj}},\forall b,j$$

$$\boldsymbol{W}_{ia}^{k+1}=\boldsymbol{W}_{ia}^{k}\frac{(\boldsymbol{V}(\boldsymbol{H}^{k+1})^{\mathrm{T}})_{ia}}{(\boldsymbol{W}^k\boldsymbol{H}^{k+1}(\boldsymbol{H}^{k+1})^{\mathrm{T}})_{ia}},\forall i,a \tag{3}$$

直到满足 $f(\boldsymbol{W},\boldsymbol{H})$ 收敛。

另一个常见的损失函数为 KL-Divergence(也称为 I-Divergence)[8]：

$$\boldsymbol{D}_{KL}(\boldsymbol{V}\|\boldsymbol{WH})=\sum_{i,j}(\boldsymbol{V}_{ij}\lg\frac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}}-\boldsymbol{V}_{ij}+(\boldsymbol{WH})_{ij}) \tag{4}$$

3 增量式非负矩阵分解算法

设矩阵 $\boldsymbol{A}$ 是 $m\times n$ 的矩阵，$\boldsymbol{B}$ 是 $m\times p$ 的矩阵，皆为非负，其中，$\boldsymbol{A}$ 矩阵已经通过 NMF 算法分解完毕，即 $\boldsymbol{A}\approx\boldsymbol{W}_1\times\boldsymbol{H}_1$，而 $\boldsymbol{B}$ 矩阵为增量部分矩阵，记增量后的矩阵为 $\boldsymbol{C}$，则 $\boldsymbol{C}=[\boldsymbol{A},\boldsymbol{B}]$，这里使用 Matlab 的语法，意为 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 矩阵的连接。若对增量后的 $\boldsymbol{C}$ 矩阵重新使用算法 1，则浪费了已有的 $\boldsymbol{A}$ 矩阵分解结果，提出一种增量式的非负矩阵分解算法，通过对 $\boldsymbol{A}$ 分解

作者简介：郭 立(1983－)，男，硕士研究生，主研方向：数据分析，数据挖掘；张守志，副教授；汪 卫、施伯乐，教授

收稿日期：2009-09-12 **E-mail：**062021125@fudan.edu.cn

后的基矩阵 W_1 和增量矩阵 B 连接形成的矩阵做非负矩阵分解求得新的基矩阵，并通过求 A 的系数矩阵 H_1 在新的基矩阵下的系数矩阵，得到 C 的一个非负矩阵分解结果。

算法 2 增量式非负矩阵分解算法

已知 NMF 分解结果 $A \approx W_1 \times H_1$，其中，$W_1 \in R^{m \times r}$，$H_1 \in R^{r \times n}$，正整数 $r<\min(m,n)$，对给定的增量非负矩阵 $B \in R^{m \times p}$，找到非负矩阵 $W \in R^{m \times r}$ 和 $H \in R^{r \times n}$，使得损失函数 $f(W,H)=\frac{1}{2}\|[A,B]-WH\|^2$ 达到最小。

(1)用 NMF 算法对$[W_1,B]$矩阵进行参数为 r 的分解，记 $D=[W_1,B]$。

(2)对分解结果 $D \approx W \times Y$ 中的 Y 矩阵按列分块，前 r 列组成的矩阵记为 Y_1，后 p 列组成的矩阵记为 Y_2，则 $D \approx W \times [Y_1,Y_2]$，其中，$W \in R^{m \times r}$，$Y_1 \in R^{r \times r}$，$Y_2 \in R^{r \times p}$，得到的 Y_1 矩阵称为转换矩阵。

(3)用转换矩阵 Y_1 求 A 的系数矩阵 H_1 在新的基矩阵 W 下的系数矩阵 $Z=Y_1 \times H_1$，拼接 Z 和 Y_2 矩阵得到 $H=[Z,Y_2]$，则得到 C 矩阵的一个非负矩阵分解结果：$C=W \times H$。

算法 2 的流程如图 1 所示。算法 2 是一个通用的增量式非负矩阵分解算法构架，损失函数可以换用 KL 等其他函数，此外，步骤 1 和步骤 2 中对非负矩阵分解算法并无指定，可以与现有的各种非负矩阵分解衍生算法配合使用。

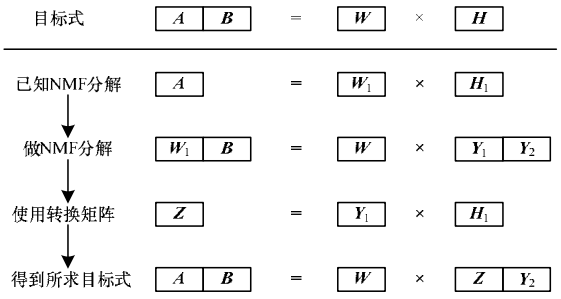


图 1　增量式非负矩阵分解算法流程

4　算法 2 的收敛性

为证明算法 2 的收敛性，首先介绍一个引理。

引理 对给定的非负矩阵 $V \in R^{m \times n}$ 和正整数 $r<\min(m,n)$，算法 1 得到的非负矩阵 $W \in R^{m \times r}$ 和 $H \in R^{r \times n}$，使得损失函数 $f(W,H)=\frac{1}{2}\|V-WH\|_F^2$ 收敛。

此引理证明的详细讨论参见文献[2, 9-10]。

以欧式范数对应的算法 2 给出如下定理并进行证明。

定理 给定的非负矩阵 $A \in R^{m \times n}$，增量非负矩阵 $B \in R^{m \times p}$ 和正整数 $r<\min(m,n)$，记连接后的矩阵为 $C=[A,B]$，算法 2 得到非负矩阵 $W \in R^{m \times r}$ 和 $H \in R^{r \times (n+p)}$，使得损失函数 $f(W,H)=\frac{1}{2}\|C-WH\|_F^2$ 收敛。

证明：由算法 2 过程和引理可知：

$$f_1(W_1,H_1)=\frac{1}{2}\|A-W_1H_1\|_F^2$$

收敛，设收敛到驻点 ε_1；根据 F-范数定义易验证：

$$\|D-WY\|_F^2=\|[W_1,B]-W\times[Y_1,Y_2]\|_F^2=$$
$$\|W_1-WY_1\|_F^2+\|B-WY_2\|_F^2$$

由引理可知 $f_2(W,Y)=\frac{1}{2}\|D-WY\|_F^2$ 收敛，设收敛到驻点为 ε_2，因此 $\|B-WY_2\|_F^2$ 和 $\|W_1-WY_1\|_F^2$ 也被 ε_2 控制，设 $WY_1=W_1+E$，其中，误差矩阵 $\|E_{m\times r}\|_F^2 \le \varepsilon_2$。

考察算法 2 的损失函数：

$$2f(W,H)=\|C-WH\|_F^2=\|[A,B]-WH\|_F^2=$$
$$\|A-WZ\|_F^2+\|B-WY_2\|_F^2=$$
$$\|A-WY_1H_1\|_F^2+\|B-WY_2\|_F^2=$$
$$\|A-W_1H_1+EH_1\|_F^2+\|B-WY_2\|_F^2$$

由 F-范数的相容性，上式中，

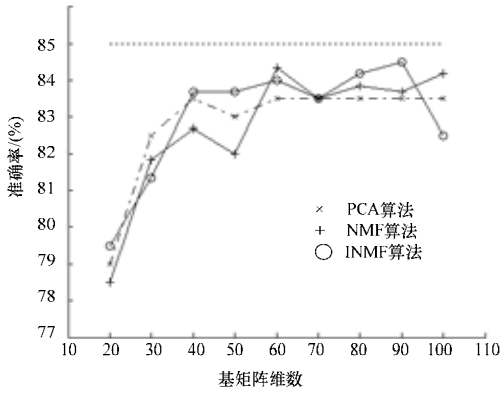
$$\|A-W_1H_1+EH_1\|_F^2=\|A-W_1H_1\|_F^2+\|EH_1\|_F^2 \le \varepsilon_1+\varepsilon_2\|H_1\|_F^2$$

由此可得 $f(W,H)=\frac{1}{2}\|C-WH\|_F^2$ 收敛。证毕。

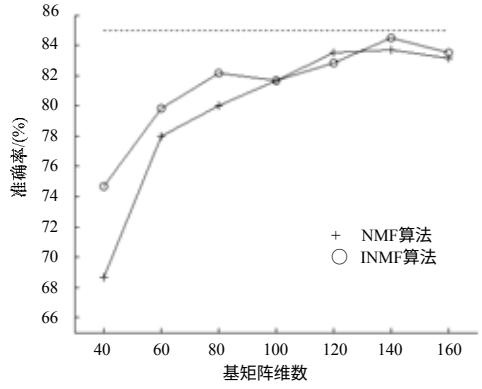
5　实验

使用如下实验场景，以观察增量式非负矩阵分解算法的特点。这是一个人脸识别环境，这个系统实时接收若干幅待检测的图片，判断这些人脸在已有的数据库中是否已经存在，并把这些图片加入到该数据库中，作为下一轮检测时的比照图片。这里使用 ORL 数据库[11]进行测试，ORL 数据库包含 40 个人，每个人 10 幅照片，使用每个人的前 5 幅做训练样本，后 5 幅用最小距离分类器进行测试，这样形成的原有矩阵和增量矩阵的规模都是 200 幅图片的数据。若对数据不降维直接分类，得到的准确率为 85%(在图 2 中用虚线标出)，使用非负矩阵分解降维后，通过提高迭代次数、增加基的数量等方法，可以接近这个准确率的上界，实验中使用了不同的基的数量和迭代次数，用随机初始化的方法，通过多次运算得到平均表现，此外还与用 PCA 方法分类的结果进行了比较，PCA 方法变化的参数为特征脸子空间的个数。运行环境为 Pentium D 3.20 GHz，1 GB 内存，Matlab 7.0。

图 2 是聚类准确率结果对比，图 2(a)固定迭代 140 次，变化基矩阵维数 r 的取值；图 2(b)固定基矩阵维数 r 为 90，使用不同的迭代次数。图 2 显示 2 种算法聚类准确率的差异在可以接受的范围内。



(a)固定迭代次数时的聚类准确率



(b)固定基矩阵维数时的聚类准确率

图 2　聚类准确率对比

设基础矩阵规模为 $m\times n$，分解的基矩阵为 r 维，增量矩阵规模为 $m\times p$，由于通常 $r<\min(m,n)$，则使用增量式矩阵分解算法，参与运算的矩阵规模从 $m\times(n+p)$降为 $m\times(r+p)$。

表 1 和表 2 分别固定了迭代次数和基矩阵维数时的运行结果对比。

表 1　迭代 140 次时不同基维数对应的时间消耗

基矩阵维数 r	NMF 时间消耗/s	INMF 时间消耗/s
20	121.1	68.6
30	137.9	82.3
40	141.1	86.4
50	158.0	105.3
60	168.9	113.4
70	192.3	137.4
80	198.4	143.3
90	220.9	167.6
100	233.1	177.9

表 2　基矩阵维数为 90 时不同迭代次数下的时间消耗

迭代次数	NMF 时间消耗/s	INMF 时间消耗/s
40	67.4	51.9
60	100.8	76.7
80	131.0	99.8
100	166.0	123.5
120	189.0	144.0
140	220.9	167.6
160	250.0	190.5

从表 1 和表 2 的结果可以看出使用增量式非负矩阵分解算法后，时间消耗比传统方法有明显改善。而从内存占用方面看，算法 1 需存储的最大矩阵规模为 $m\times(n+p)$，而算法 2 存储的最大矩阵为 $m\times(r+p)$，这样也节省了运算时的存贮资源。

从上面的分析和表 1 的结果也可以得到，在分解的基的维数 r 取值较小时，算法 2 对时空效率提升的效果更加显著，如表 1 中当 r 取 40 时，时间效率提高了约 38.9%。事实上通常人们使用 NMF 算法时，为了得到较好的降维效率，也总是期望使用较小的 r 值。

6　结束语

本文提出了一种增量式非负矩阵分解算法，在应用于实际问题时表现出对增量情景的良好支持，从实验部分可以看出，计算资源特别是存储空间和时间的节约效果都很显著。

由于这种增量式算法也可以看作分块式 NMF 算法的一种特例，研究更有效的分块式算法降低每次迭代时的运算规模，可以使 NMF 算法处理更大规模的问题，以及适用于分布式条件下的应用。这也是有待进一步继续研究的问题。

参考文献

[1] Lee D, Seung H. Learning the Parts of Objects by Non-negative Matrix Factorization[J]. Nature, 1999, 401(6755): 788-791.

[2] Lee D, Seung H. Algorithms for Non-negative Matrix Factorization[C]//Proc. of Neural Information Processing Systems Conference. Vancouver, Canada: MIT Press, 2000: 556-562.

[3] Buciu I, Pitas I. Application of Non-negative and Local Non Negative Matrix Factorization to Facial Expression Recognition[C]//Proc. of the 17th International Conference on Pattern Recognition. Cambridge, UK: IEEE Press, 2004: 288-291.

[4] Xu Wei, Liu Xin, Gong Yihong. Document-clustering Based on Non-negative Matrix Factorization[C]//Proc. of the 26th ACM SIGIR'03. Toronto, Canada: ACM Press, 2003: 267-273.

[5] Okun O, Priisalu H. Fast Nonnegative Matrix Factorization and Its Application for Protein Fold Recognition[J]. EURASIP Journal on Applied Signal Processing, 2006, (16): 117-125.

[6] Wang Beiming, Plumbley M. Musical Audio Stream Separation by Non-negative Matrix Factorization[C]//Proc. of Digital Music Research Network Summer Conference. Glasgow, UK: [s. n.], 2005: 23-24.

[7] Berne O, Deville Y, Joblin C. Blind Signal Separation Methods for the Identification of Interstellar Carbonaceous Nanoparticles[C]// Proc. of ICA'07. London, UK: Springer, 2007: 681-688.

[8] Dhillon I, Sra S. Generalized Nonnegative Matrix Approximations with Bregman Divergences[C]//Proc. of Neural Information Processing Systems Conference. Vancouver, Canada: MIT Press, 2005: 283-290.

[9] Donoho D, Stodden V. When Does Non-negative Matrix Factorization Give a Correct Decomposition into Parts?[C]//Proc. of Neural Information Processing Systems Conference. Vancouver, Canada: MIT Press, 2003: 1141-1148.

[10] Berry M, Browne M, Langville A, et al. Algorithms and Applications for Approximation Nonnegative Matrix Factorization[J]. Computational Statistics and Data Analysis, 2007, 52(1): 155-173.

[11] Cambridge University. The ORL Face Database[EB/OL]. (2008-09-02). www.orl.co.uk/facedatabase.html.

编辑　任吉慧

(上接第 65 页)

5　结束语

本文算法能优化调谐时间，并使访问时间保持在一个合理范围内。实验结果表明，其性能较 VF 算法有很大改进。

参考文献

[1] Tsai C F, Tsai C W. A New Approach for Solving Large Traveling Salesman Problem Using Evolutionary Ant Rules[C]//Proceedings of International Joint Conference on Neural Network. New Jersey, USA: IEEE Press, 2002: 1540-1545.

[2] Tsakiridis F, Bozanis P, Katsaros D. Interpolating the Air for Optimizing Wireless Data Broadcast[C]//Proceedings of the 5th ACM International Workshop on Mobility Management and Wireless Access. [S. l.]: ACM Press, 2007: 112-119.

[3] Madhuka A, Alhajj R. An Adaptive Energy Effieient Cache Invalidation Seheme for Mobile DataBases[C]//Proceedings of the 2006 ACM Symposium on Aplied Computing. [S. l.]: ACM Press, 2006: 1122-1126.

[4] Chen Ming-Syan, Wu Kun-lung, Yu S P. Optimizing Index Allocation for Sequential Data Broadcasting in Wireless Mobile Computing[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(1): 161-173.

[5] Shen Junhong, Chang Ye-In. A Skewed Distributed Indexing for Skewed Access Patterns on the Wireless Broadcast[J]. Syst. Software, 2007, 80(5): 711-723.

[6] Acharya S, Alonso R, Franklin M, et al. Broadcast Disks: Data Management for Asymmetric Communications Environment[C]// Proceedings of ACM SIGMOD International Conference. San Jose, CA: [s. n.], 1995: 199-210.

编辑　陈　晖