

文章编号: 1004—5570(2017)06—0074—03

基于 PARAFAC2 分解的手写数字识别算法

卢琳璋^{1,2}, 陈震^{1*}, 徐海利³

(1. 贵州师范大学 数学科学学院, 贵州 贵阳 550025; 2. 厦门大学 数学科学学院, 福建 厦门 361005;
3. 贵州大学附属中学, 贵州 贵阳 550025)

摘要: 提出了基于平行因子2 (PARAFAC2) 分解的手写数字识别算法, 与基于高阶奇异值分解的算法相比较, 新方法的识别率没有降低, 但是效率有所提高, 更重要的是, 新方法可以处理手写数字维数互异的情形。实验结果验证了算法的有效性。

关键词: 高阶奇异值分解 HOSVD; PARAFAC2 分解; 交替最小二乘法; 手写数字

中图分类号: O246 文献标识码: A

DOI:10.16614/j.cnki.issn1004-5570.2017.06.007

Handwritten digit classification algorithm based on PARAFAC2 decomposition

LU Linzhang^{1,2}, CHEN Zhen¹, XU Haili³

(1. School of Mathematics Science, Guizhou Normal University, Guiyang, Guizhou 550025, China; 2. School of Mathematics Science, Xiamen University, Xiamen, Fujian 361005, China; 3. The Middle School attached to the Guizhou University, Guiyang, Guizhou 550025, China)

Abstract: In this paper, we propose an algorithm for handwritten digit recognition based on the parallel factor 2 (PARAFAC2) decomposition. Comparing with the algorithm based on higher order singular value decomposition, the recognition rate of our method does not decrease, however, the efficiency is increased. Moreover, the new method can handle the case in which dimensions of handwritten digits are different. Numerical results validate the efficiency of our algorithm.

Key words: higher order singular value decomposition HOSVD; PARAFAC2 decomposition; alternating least square method; handwritten digit

0 引言

手写数字识别一直是图像处理、模式识别等领域的研究热点,在邮政编码数字自动分拣信件系统、统计报表、财务报表、银行票据等方面得到广泛应用。手写数字识别的方法很多,如:主成份分析

法、最近邻方法、统计方法、神经网络方法、高阶奇异值分解(HOSVD)等^[1-3]。传统的方法通常是在标准的数据库(手写数字的维数是统一的,例如,16×16像素)基础上进行实验,因为这些方法都是将手写数字向量化后构成矩阵或直接堆积为张量(也称为高维数组),然后进行特征提取。一般来说,一个识别率高的算法自然原理复杂,计算量

收稿日期: 2017-08-20

基金项目: 贵州省科学技术基金项目(黔科合J字[2013]2207号); 贵州省教育厅自然科学基金项目(黔教合KY字[2015]352号)

作者简介: 卢琳璋(1958-),男,博士,教授,博士生导师,研究方向: 应用数值代数, E-mail: llz@gznu.edu.cn.

* 通讯作者: 陈震(1979-),男,博士,教授,博士生导师,研究方向: 应用数值代数, E-mail: zchen@gznu.edu.cn.

大^[3]。本文基于 PARAFAC2 分解,提出一种简单、有效的新方法,与基于高阶奇异值分解的方法相比较,识别率保持一致,但效率提高,更重要的是,新方法不仅可以处理手写数字维数相同的情形,还可以处理维数互异的情形。

1 基于 HOSVD 的手写数字识别

HOSVD 于 2000 年首次提出^[4],是矩阵奇异值分解的推广,目前在信号处理、模式识别中有较为广泛的应用。简要叙述如下:

定理 1 任意一个三阶张量 $\bar{X} \in \mathbf{R}^{I \times J \times K}$ 部可以分解为如下的乘积形式:

$$\bar{X} = \bar{Y} \times_1 U \times_2 V \times_3 W,$$

其中,符号 \times_i 表示矩阵沿着第 i 个方向乘张量,详见文献[5], $U \in \mathbf{R}^{I \times I}$, $V \in \mathbf{R}^{J \times J}$, $W \in \mathbf{R}^{K \times K}$ 都是正交矩阵, \bar{Y} 与 \bar{X} 维数相同,满足

- 1) 完全正交性:
 - $\langle \bar{Y}(i, :, :), \bar{Y}(j, :, :)\rangle = 0, i \neq j,$
 - $\langle \bar{Y}(:, i, :), \bar{Y}(:, j, :)\rangle = 0, i \neq j,$
 - $\langle \bar{Y}(:, :, i), \bar{Y}(:, :, j)\rangle = 0, i \neq j.$

2) 奇异值的有序性:无论沿着哪个方向展开,每一片的模是有序的,例如,沿第一个方向的模满足: $\|\bar{Y}(1, :, :)\| \geq \|\bar{Y}(2, :, :)\| \geq \dots \geq 0.$

所谓手写数字识别,即是先将训练集中的数字人工分为 10 类,代表数字 0-9。然后,利用具体的算法判断测试集中的每个数字属于哪一类。最后,计算各个数字的识别正确率和总体识别正确率。利用 HOSVD 进行手写数字识别的理论推导见文[3]。具体算法如下:

算法 1: 基于 HOSVD 的手写数字识别算法 训练阶段

- 1) 将表示相同数字的 k 个矩阵存放在同一个张量中,得到 10 个张量 $\bar{X}^\mu, \mu = 0, \dots, 9;$
 - 2) 计算每个张量的 HOSVD;
 - 3) 计算并存储单位化的基矩阵 $T^\mu = (A_i^\mu)_{i=1}^k$
- 其中 $A_i^\mu = \bar{Y}^\mu(:, :, i) \times_1 U^\mu \times_2 V^\mu, k$ 表示基的个数,一般远远小于 K ,本文取 $k = 10$ 。

测试阶段

- 1) 将待识别数字 D 单位化;
- 2) 计算 $R(\mu) = 1 - \sum_{i=1}^k [D A_i^\mu]_1;$
- 3) 确定使 $R(\mu)$ 最小的 μ ,将数字 D 归为第 μ 类。

因为 HOSVD 需要将张量沿不同方向展开后做奇异值分解,所以算法 1 中最大的工作量是计算每个张量的 HOSVD。

2 基于 PARAFAC2 分解的手写数字识别

平行因子(PARAFAC)分解是主成分分析的一种高维推广,最早出现在计量化学领域。以三阶张量为例,所谓 PARAFAC 分解,即是将张量 $\bar{X} \in \mathbf{R}^{I \times J \times K}$ 近似分解为如下形式:

$$\bar{X} \approx \sum_{i=1}^R a_i \otimes b_i \otimes c_i$$

本文将采用标准的交替最小二乘法对张量进行 PARAFAC 分解,算法详见文[5]。

PARAFAC2 严格来说不是张量分解,但是可以看成 PARAFAC 的变形,它作用的对象不是张量,而是一系列具有相同列,可以是不同行的矩阵。设矩阵 $X_k \in \mathbf{R}^{I_k \times J}, k = 1, \dots, K,$ PARAFAC2 分解即是将这些矩阵近似分解为如下形式:

$$X_k \approx U_k H S_k V^T, k = 1, \dots, K,$$

其中 $U_k \in \mathbf{R}^{I_k \times R}$ 是正交矩阵, S_k 是对角矩阵, H 是一个不随 k 变化的 $R \times R$ 的矩阵, V 也是一个不随 k 变化的 $J \times R$ 的矩阵,如图 1 所示。

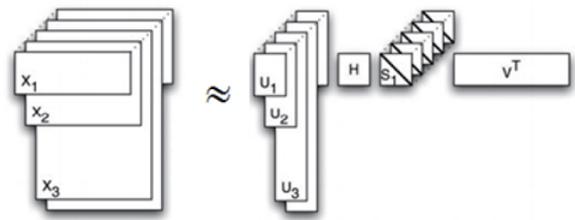


图 1 PARAFAC2 分解示意图

Figure 1 Diagrammatic sketch of PARAFAC2 decomposition

与 PARAFAC 比较,PARAFAC2 的使用范围更加广泛,后者可以处理矩阵列数相同而行数不相同的情形,在本文的实验中, X_k 表示同一类数字构成的矩阵,其维数是数字像素乘以数字个数($I_k \times J$),其中每一列由代表手写数字的矩阵向量化后得到,其长度即为行数。如果每一类手写数字像素大小不同,则行数不同,此时 PARAFAC 模型是不适用的。而 PARAFAC2 则是一个有效的模型。

PARAFAC2 分解的算法有很多,这里采用直接拟合的算法,具体见文[6]。注意到,在 PARAFAC2 中, U_k 也是正交矩阵,类似于 HOSVD, U_k 可以看作是 X_k 的一组正交基,因此,基于 PARAFAC2 分解的手写数字识别算法可以归结如下:

算法 2: 基于 PARAFAC2 的手写数字识别算法
训练阶段

1) 用 $\sum_{k=1}^K X_k^T X_k$ 的前 R 个主特征向量初始化 V , 用单位矩阵 I_R 初始化 H, S_1, \dots, S_K ;

2) 对 $Z_k = HS_k V^T X_k^T$ 作 SVD 分解 $P_k \sum_k Q_k^T$, 首先通过计算 $Z_k Z_k^T$ 的前 R 个主特征向量得到 P_k , 然后对 $Z_k^T P_k$ 的列单位化得到 Q_k , 最后用 $Q_k P_k^T$ 更新 $U_k, k = 1, \dots, K$;

3) 将 $U_k^T X_k$ 作为正面切片堆积成张量 \bar{T} , 利用交替最小二乘算法对张量 \bar{T} 进行 PARAFAC 分解, 从而更新 H, V, S_k ;

4) 重复 2) ~ 3) 直至达到最大迭代次数或误差 $\sum_{k=1}^R \|X_k - U_k H S_k V^T\|$ 不再下降;

5) 存储 2) ~ 4) 中输出的 U_k , 作为每一类数字矩阵的正交基矩阵。

测试阶段

1) 将待识别数字拉成向量 $d_p = \text{vec}(D_p)$;

2) 计算 $R(k) = \|d_p - U_k U_k^T d_p\|, k = 0, \dots, 9$;

3) 确定使 $R(k)$ 最小的 k , 将数字 d_p 归为第 k 类。

算法 2 的好处在于避免计算每一类数字构成的张量的 HOSVD, 只需要计算 $R \times R$ 矩阵的特征向量, 以及张量的 PARAFAC 分解。更重要的是, 相比 HOSVD 和 PARAFAC, 算法 2 的应用范围更加广泛。

3 数值试验

下面的试验是在计算机内存为 4GB, 处理器为 Inter(R) Core(TM) i5 - 6500 CPU 3.20 GHz, 操作系统为 Windows 7 的台式机上使用 MATLAB 2016a 编写的, 涉及到张量部分的程序使用了 Bader 和 Kolda 开发的 Matlab Tensor Toolbox [7]。

试验中使用的 USPS 数据库是美国邮政信封上扫描得到的数据, 可以在网上免费下载, 详见 [8], 该数据库的训练集包含了 1707 个数据, 测试集包含了 2007 个数据, 具体分布情况如表 1 所示:

表 1 USPS 数据库中的手写数字分布

Tab. 1 The digit distribution in the US Postal Service data sets

	0	1	2	3	4	5	6	7	8	9	共计
训练集	319	252	202	131	122	88	151	166	144	132	1707
测试集	359	264	198	166	200	160	170	147	166	177	2007

例 1 为了比较算法 1 和算法 2, 首先对训练集中的每一类数字取 88 个样本(因为由表 1 可知, 数字 5 的样本数最少, 只有 88 个)。对于算法 1, 将每一类数字构成 $16 \times 16 \times 88$ 的三阶张量 \bar{X}^u ; 对于算法 2, 将每个样本向量化构成 256×88 的矩阵

X_k , 然后执行算法 1 和算法 2。其中, 在算法 2 训练阶段的第 3 步中进行 PARAFAC 的交替最小二乘算法时, 我们取张量的秩 R 取为 48, 迭代次数取为 3, 第 4 步中的最大迭代次数取为 15, 容许误差取为 0.5×10^{-5} , 运行结果见表 2。

表 2 算法 1 和算法 2 的识别率

Tab. 2 Correct classification results of two Algorithms

数字	0	1	2	3	4	5	6	7	8	9	平均	时间
识别率(算法 1)	0.986	0.985	0.889	0.892	0.910	0.893	0.941	0.938	0.891	0.909	0.931	2.16s
识别率(算法 2)	0.981	0.974	0.889	0.898	0.930	0.869	0.924	0.918	0.898	0.944	0.930	1.58s

由表 2, 表 3 可以看出: 基于 PARAFAC2 分解的算法与基于 HOSVD 的算法相比较, 无论对具体数字的识别率, 还是总体识别率几乎都是一致的, 但是算法 2 的速度提高了 27%。例 2 首先在训练

集中对每一类数字随机抽取相同的样本, 然后对其像素进行处理, 把数字 0, 2, 7, 9 数字像素由 16×16 改变为 15×15 , 其余不变, 此时利用算法 2 同样具有很好的识别效果, 运行结果见表 3。

表 3 算法 2 对不同像素数字的识别率

Tab. 3 Correct classification results of Algorithm 2 on digits of different sizes

数字	0	1	2	3	4	5	6	7	8	9	平均	时间
准确率	0.983	0.973	0.878	0.891	0.905	0.862	0.929	0.932	0.891	0.932	0.926	1.52s

(下转第 88 页)

- World Scientific, 1994.
- [16]SANTALÓ L. Integral geometry and geometric probability [M]. Reading, MA: Addison-Wesley, 1976.
- [17]SCHNEIDER R. Convex Bodies: The Brunn-Minkowski Theory [M]. Cambridge: Cambridge Univ Press, 1993.
- [18]ZHANG G, ZHOU J. Containment measures in integral geometry, Integral geometry and convexity [M]. Singapore: World Sci Publ, 2006: 153-168.
- [19]ZHANG Z, LUO M, CHEN F. New convex bodies and the reverse Bonnesen-style inequalities in the plane [J]. Journal of Southwest China Normal University (Natural Science Edition), 2015, 40: 27-30.
- [20]ZHOU J, DU Y, CHENG F. Some Bonnesen-style inequalities for higher dimensions [J]. Acta Math Sinica, 2012, 28: 2561-2568.
- [21]ZHOU J, REN D. Geometric inequalities—from integral geometry point of view [J]. Acta Math Scientia, 2010, 30: 1322-1339.
- [22]ZHOU J. Integral geometry and isoperimetric inequality [J]. Journal of Guizhou Normal University(Natural Sciences Edition), 2002, 20: 1-3.

欧氏平面 R^2 中几个加强的 Minkowski 不等式

罗 森¹ 胡 娟² 杨 林³

(1. 贵州师范大学 数学科学学院, 贵州 贵阳 550025; 2. 贵州师范大学 地理与环境科学学院, 贵州 贵阳 550025; 3. 铜仁职业技术工程学院 信息工程学院, 贵州 铜仁 554300)

摘要: 运用积分几何中的平移包含测度, 估计对称混合等似亏格 $\Delta_2(K_0, K_1) = A_{01}^2 - A_0A_1$, 其中 A_{01} 是凸域 K_0 与 K_1 的混合面积。获得了欧氏平面 R^2 中一些加强的 Minkowski 不等式。

关键词: Minkowski 不等式; 加强的 Minkowski 不等式; 平移包含测度; 对称混合等似亏格

中图分类号: O186.1 文献标识码: A

(上接第 76 页)

参考文献:

- [1] DUDA R O, HART P E, STORK D G. Pattern classification [M]. Beijing: China Machine Press, 2004.
- [2] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning [M]. Springer: New York, USA, 2000.
- [3] SAVAS B, ELDEN L. Handwritten digit classification using higher order singular value decomposition [J]. Pattern Recognition, 2007, 40: 993-1003.
- [4] LATHAUWER L D E, MOOR B D E, VANDEWALLE J. A multilinear singular value decomposition [J]. SIAM J Matrix Anal Appl, 2000, 21(4): 1253-1278.
- [5] KOLDA T G, BADER B W. Tensor decompositions and applications [J]. SIAM Review 2009, 51(3): 455-500.
- [6] KIERS H A L, TEN BERGE J M F, BRO R. PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model [J]. Journal of Chemometrics, 2015, 13(3/4): 275-294.
- [7] BADER B W, KOLDA T G. https://gitlab.com/tensors/tensor_toolbox.
- [8] ELDEN L. Matrix Methods in Data Mining and Pattern Recognition [M]. New York: Society for Industrial and Applied Mathematics, 2007.