一种新的广义鲁棒主成分分析(GRPCA)算法研究及应用

侯旭珂[1] 杨宏伟[2] 马 方[1] 赵丽娜[1*]
(北京化工大学 1. 理学院; 2. 信息中心, 北京 100029)

摘 要: 为恢复被混合噪声污染的低秩矩阵,提出了一种新的广义鲁棒主成分分析(GRPCA)算法。它通过最小化核范数、1 范数和 2,1 范数的组合问题,从观测矩阵中分离出低秩部分和混合噪声部分,并用随机排序的交替方向乘子法求解。利用本文方法进行垃圾邮件分类的实验结果表明,与经典的主成分分析(PCA)和鲁棒主成分分析(RPCA)算法相比,本文方法可以有效提高垃圾邮件分类的精确度和稳定性。

关键词: 广义鲁棒主成分分析(GRPCA); 降维; k 近邻(kNN); 支持向量机(SVM)
中图分类号: TP391.41 **DOI**: 10.13543/j.bhxbzr.2018.04.015

引言

在过去的十多年间,数据分析行业发生了革命性的变化。目前,用户需要处理的数据在尺寸、大小和复杂性上都呈爆炸式增长趋势[1],所以如何通过识别低维结构从海量数据中提取有用的信息,已成为一个迫切需要解决的问题。

主成分分析(PCA)是一种用于发现数据中低维结构的通用技术,它假设数据是从一个单一的低维数据中提取的高维空间的仿射子空间。PCA 作为最简单、最流行的降维工具,在计算机视觉与图像处理等许多领域得到广泛的应用[2-4]。然而,PCA 对于含野点和稀疏大噪声的数据非常敏感,为解决此问题,研究者们提出了鲁棒主成分分析(RPCA)方法[5]。RPCA 方法通过最小化核范数和 1 范数的组合问题,将观测矩阵分离为低秩部分和稠密小噪声部分。RPCA 方法已广泛应用于图像去噪、视频处理、网页搜索和生物信息等领域[6-8],但当数据被两种以上混合噪声污染时,RPCA 的表现不太理想,故本文提出一种可处理含有混合噪声的模型算法即广义鲁棒主成分分析(GRPCA),通过最小化核范数、1 范数和 2,1 范数的组合问题,从观测矩阵中分离出低秩部分和混合噪声部分,采用随机排序的交替方向乘子法(RPADMM)对 GRPCA 求解,并选取机器学习常用网站 UCI 中的一个垃圾邮件数据集(spambase)[9-10]来验证 GRPCA 算法的性能。

收稿日期: 2017-09-06
基金项目: 国家自然科学基金(11301021/11571031)
第一作者: 女,1993 年生,硕士生
*通信联系人
E-mail:zhaoln@mail.buct.edu.cn

1 鲁棒主成分分析

广义鲁棒主成分分析是由 RPCA 衍生出的一种新型降维算法,RPCA 由 PCA 发展而来,可以用于低秩矩阵恢复,即从一个被稀疏大噪声污染的低秩矩阵中恢复出低秩部分,并分离出噪声。

PCA 可以通过约束优化模型(1)求出低秩矩阵 $\boldsymbol{A}$ 的最优估计

$$\begin{cases}\min\limits_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{E}\|_F \\ \text{s.t. } \operatorname{rank}(\boldsymbol{A}) \leqslant r, \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E}\end{cases} \tag{1}$$

式中,$\|\cdot\|_F$ 为 Frobenius 范数。设 $\boldsymbol{X}$ 是一个 n 维方阵,则 $\|\boldsymbol{X}\|_F=\left(\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(a_{i,j})^2\right)^{\frac{1}{2}}$ 为矩阵元素的平方和的算术平方根。对观测矩阵 $\boldsymbol{D}$ 进行奇异值分解(SVD)即可求出最优解。

当 $\boldsymbol{D}$ 受到稀疏大噪声的干扰时,PCA 的假设条件不满足,无法正常工作。此时,恢复出低秩矩阵 $\boldsymbol{A}$ 就变成如式(2)的非凸优化问题

$$\begin{cases}\min\limits_{\boldsymbol{A},\boldsymbol{E}} \operatorname{rank}(\boldsymbol{A})+\lambda\|\boldsymbol{E}\|_0 \\ \text{s.t. } \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E}\end{cases} \tag{2}$$

式中,$\|\cdot\|_0$ 为 0 范数,表示矩阵中非 0 元素的个数。式(2)的优化问题经松弛变为一个新的凸优化问题[5],即 RPCA

$$\begin{cases}\min\limits_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 \\ \text{s.t. } \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E}\end{cases} \tag{3}$$

式中 $\|\cdot\|_1$ 为 1 范数，则 $\|\boldsymbol{X}\|_1 = \left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}|x_{i,j}|\right)$ 表示矩阵各元素绝对值之和。式(3)的 RPCA 模型可用迭代阈值算法(IT)、交替方向乘子法(ADMM)[6]或近端梯度算法(APG)[7]等算法求解。

由于污染矩阵的噪声可能不唯一，RPCA 算法无法处理被两种不同噪声污染的情况，所以本文在 RPCA 算法基础上加入新的噪声矩阵 $\boldsymbol{H}$，并采用 RPADMM 对 GRPCA 求解，该方法的收敛性已被证明[11]。

2　广义鲁棒主成分分析(GRPCA)及其算法

2.1　模型及求解

当 $\boldsymbol{D}$ 同时被稀疏大噪声(如椒盐噪声)和稠密小噪声(如高斯噪声)干扰时，RPCA 模型无法识别稠密小噪声，因此本文在模型(3)中引入了第三项 $l_{2,1}$ 范数来表示稠密的小噪声。$l_{2,1}$ 范数定义如下。

设 $\boldsymbol{X}$ 是一个 n 维方阵，且 $\|\boldsymbol{X}\|_{2,1} = \sum\limits_{i=1}^{n}\sqrt{\sum\limits_{j=1}^{n}x_{i,j}^2}$，此时 RPCA 模型可改写为

$$\begin{cases}\min\limits_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + \gamma\|\boldsymbol{H}\|_{2,1} \\ \text{s.t. } \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E} + \boldsymbol{H}\end{cases} \tag{4}$$

模型(4)的增广拉格朗日函数为

$$\begin{cases}L(\boldsymbol{A},\boldsymbol{E},\boldsymbol{H},\boldsymbol{Y},\mu) = \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + \\ \qquad \gamma\|\boldsymbol{H}\|_{2,1} + <\boldsymbol{Y},\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}-\boldsymbol{H}> + \\ \qquad \mu\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}-\boldsymbol{H}\|_F^2\end{cases} \tag{5}$$

由于式(5)含有 3 个变量$(\boldsymbol{A},\boldsymbol{E},\boldsymbol{H})$，ADMM 求解不再收敛，故本文采用 RPADMM 对其求解。RPADMM 与 ADMM 的区别在于，每一步迭代的顺序是不确定的，取决于迭代前产生的一组随机数即(1,2,3)的随机排序。随机数为(2,3,1)表示先求第 2 个变量，再求第 3 个变量，最后求第 1 个变量，即先固定 $\boldsymbol{A}$、$\boldsymbol{H}$ 同时更新 $\boldsymbol{E}$，再固定 $\boldsymbol{A}$、$\boldsymbol{E}$ 同时更新 $\boldsymbol{H}$，最后固定 $\boldsymbol{E}$、$\boldsymbol{H}$ 同时更新 $\boldsymbol{A}$。

GRPCA 求解迭代格式如下。

固定 $\boldsymbol{E}$、$\boldsymbol{H}$，更新 $\boldsymbol{A}$ 时，有

$$\boldsymbol{A}_{k+1} = \boldsymbol{D}_{\mu^{-1}}(\boldsymbol{D}-\boldsymbol{E}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y})$$

固定 $\boldsymbol{A}$、$\boldsymbol{H}$，更新 $\boldsymbol{E}$ 时，有

$$\boldsymbol{E}_{k+1} = \arg\min_{\boldsymbol{E}}\lambda\|\boldsymbol{E}\|_1 + \frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y}\|_F^2 = S_{\frac{\lambda}{\mu}}(\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y})$$

固定 $\boldsymbol{A}$、$\boldsymbol{E}$，更新 $\boldsymbol{H}$ 时，有

$$\boldsymbol{H}_{k+1} = \arg\min_{\boldsymbol{F}}\gamma\|\boldsymbol{H}\|_{2,1} + \frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y}\|_F^2 = (\gamma\boldsymbol{G}+\mu\boldsymbol{I})^{-1}(\boldsymbol{Y}+\mu(\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}))$$

式中 $\boldsymbol{G} = \text{diag}\left(\frac{1}{\|\boldsymbol{H}^i\|_2}\right)$ 是由 $\boldsymbol{H}$ 决定的对角矩阵。

GRPCA 算法步骤如下。

1)　初始化参数 $Y_0 = 0, \mu_0 > 0, \rho > 0, \gamma = 0.5$，$k = 0, \boldsymbol{A}_0 = 0, \boldsymbol{E}_0 = 0, \boldsymbol{H}_0 = 0$

2)　if　收敛条件满足，输出结果

3)　else

　　for $k = 0$ 到最大迭代次数

　　生成(1,2,3)的一个随机排序 $\boldsymbol{\sigma}$

4)　　遍历 $\boldsymbol{\sigma}$

5)　　if $\boldsymbol{\sigma}(i) = 1$

　　$\boldsymbol{A}_{k+1} = \boldsymbol{D}_{\mu^{-1}}(\boldsymbol{D}-\boldsymbol{E}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y})$

6)　　else if $\boldsymbol{\sigma}(i) = 2$

　　$\boldsymbol{E}_{k+1} = S_{\frac{\lambda}{\mu}}(\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{H}+\mu^{-1}\boldsymbol{Y})$

7)　　else $\boldsymbol{\sigma}(i) = 3$

　　$\boldsymbol{H}_{k+1} = (\gamma\boldsymbol{G}+\mu\boldsymbol{I})^{-1}(\boldsymbol{Y}+\mu(\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}))$

8)　　end if

9)　　$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k + \mu_k(\boldsymbol{D}-\boldsymbol{A}_k-\boldsymbol{E}_k-\boldsymbol{H}_k)$

　　$\mu_{k+1} = \rho\mu_k, k = k+1$

10)　end for

　　end if

11)　输出结果$(\boldsymbol{A}^*,\boldsymbol{E}^*,\boldsymbol{H}^*)$

2.2　算法复杂度分析

在求解模型(4)的 GRPCA 算法中，运算量主要集中在奇异值分解和矩阵求逆中。假设输入矩阵的维度为 $R^{m\times n}$，每一次循环中，步骤 5)和步骤 6)奇异值分解的运算量为 $O((m+n)^3)$，步骤 7)矩阵求逆的运算量为 $O((m+n)^3)$。此外 GRPCA 算法循环内的多次乘法加法和截断阈值的运算量为 $O(5n(m+n)+14n)$。因此，GRPCA 算法的总运算量为 $O((m+n)^3 + I(5n(m+n)+14n))$，其中 I 为循环的次数。

3　实验验证及结果分析

所有实验运行环境均为 Matlab R2016b 版，机

器配置为 ThinkCentre M8400t-N000,内存 4 GB,频率 3.4 GHz,操作系统为 Windows 7。

3.1 数据来源

本文数据集来自 UCI 中的 Spambase Data Set。Spambase 是一个 4 601 ×58 的矩阵,包含 4 601 个邮件,每一行代表一个邮件。其中前 57 列为邮件的属性,第 58 列为邮件的标签(1 表示垃圾邮件 0 表示非垃圾邮件)。本文采用分类精确度(accuracy)来度量实验的有效性。

3.2 实验设置

实验分为两大部分。

(1) 将本文提出的 GRPCA 与 kNN 算法结合构成垃圾邮件分类的一个新型算法:先用 GRPCA 对邮件属性进行去冗余预处理,再用 kNN 分类。并与经典的降维算法 PCA、RPCA 和 kNN 结合所形成的垃圾邮件分类算法比较。

(2)将 GRPCA 与 SVM 算法结合构成垃圾邮件分类的一个新型算法:先用 GRPCA 算法对邮件属性进行去冗余预处理,再用 SVM 算法分类。并与经典的降维算法 PCA、RPCA 和 SVM 结合所形成的垃圾邮件分类算法比较。实验流程如图 1。

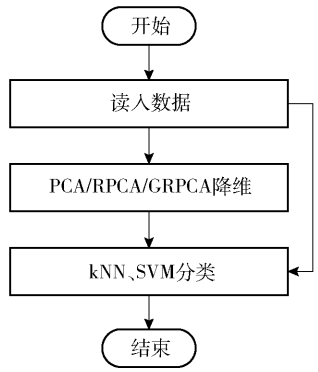

图 1 实验流程图

Fig. 1 Experimental flowchart

3.3 结果分析

3.2 节中两部分实验所得结果分别如表 1、图 2 所示。从表 1 可以看出,当固定 kNN 算法的 k 值时,经过 GRPCA 去冗余后再分类的精确度均达到了 85.5% 以上,比直接使用 kNN 的精确度要高出 2.91% ~4.79%;相对经典降维算法 PCA 和 RPCA 预处理后的结果的精确度分别高出 2.91% ~4.79%和 1.22% ~1.85%,并且 3 种方法所用时间几乎无差别。从图 2 可以看出,在 SVM 算法中,经过 GRPCA 去冗余处理后的精确度高达 89.52%,比直接使用 SVM 的精确度要高出 2.76%,比 PCA 和 RPCA 预处理后的精确度分别高出 2.76% 和 3.89%。由于3 种方法所需时间均较长(大于 6 h),故不作时间对比。

表 1 kNN 算法精确度及时间对比

Table 1 A accuracy and time comparison of the kNN algorithm

方法	精确度/%				时间/s			
	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
kNN	83.047 2	80.938 9	81.482 3	80.721 6	53.029 8	52.863 9	53.013 1	53.122 8
PCA + kNN	83.068 9	80.938 9	81.482 3	80.721 6	54.316 5	54.256 5	53.547 6	53.497 2
RPCA + kNN	84.742 4	84.112 1	84.459 9	83.655 7	53.194 2	52.490 3	52.630 7	52.435 6
GRPCA + kNN	85.959 6	85.720 5	86.176 9	85.503 2	54.188 1	54.919 4	52.573 9	52.632 4

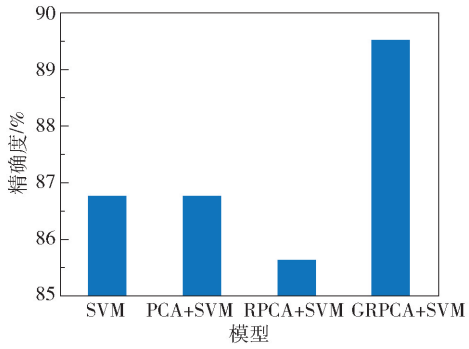

图 2 SVM 算法的精确度对比

Fig. 2 A accuracy of the SVM algorithm

此外,还可以看出,在 PCA 去冗余处理前后精确度几乎没有发生变化,这说明属性矩阵受到稀疏大噪声的干扰,使 PCA 无法正确降维,从而证明本文在邮件过滤时对属性矩阵进行去冗余处理是必要的。

4 结束语

本文提出了一种新的广义鲁棒主成分分析(GRPCA)算法,通过最小化核范数、1 范数和 2,1 范数的组合问题,可以分离出被混合噪声污染的低秩矩阵。选用垃圾邮件来检验算法,结果表明本文算法在提高垃圾邮件的精确度上是有效的。

参考文献：

[1] LU C Y, LIN Z C, YAN S C. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization [J]. IEEE Transactions on Image Processing, 2015, 24(2): 646-654.

[2] FU Y F, GAO J B, TIEN D, et al. Tensor LRR and sparse coding-based subspace clustering [J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(10): 2120-2133.

[3] LU C Y, LI H, LIN Z C. Optimized projections for compressed sensing via direct mutual coherence minimization [J]. Signal Processing, 2018, 151: 45-55.

[4] ZHAO Z Y, LIN Z C, WU Y. A fast alternating time-splitting approach for learning partial differential equations[J]. Neurocomputing, 2016, 185: 171-182.

[5] WRIGHT J, PENG Y G, MA Y. Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization[J]. Journal of the ACM, 2009, 87(4): 1-44.

[6] LIN Z C, CHEN M M, MA Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices [J/OL]. arxiv. org(2013-10-18). https://arxiv. org/abs/1009. 5055.

[7] LIU R S, ZHONG G Y, CAO J J, et al. Learning to diffuse: a new perspective to design PDEs for visual analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(12): 2457-2471.

[8] CANDÈS E J, LI X, MA Y, et al. Robust principal component analysis? [J]. Journal of the ACM, 2011, 58(3): 11.

[9] 杨雷,曹翠玲,孙建国,等. 改进的朴素贝叶斯算法在垃圾邮件过滤中的研究[J]. 通信学报,2017,38(4): 140-148.
YANG L, CAO C L, SUN J G, et al. Study on spam filtering based on improved naive Bayesian algorithm [J]. Journal on Communications, 2017, 38(4): 140-148. (in Chinese)

[10] KAMBLE M, MALIK L G. Detecting image spam using principal component analysis & SVM classifier [J]. International Journal of Computer Science and Information Technology & Security, 2012, 2(6): 1217-1220.

[11] 李吉,赵丽娜,侯旭珂. 通过随机排序的交替方向乘子法的矩阵恢复[J]. 北京化工大学学报(自然科学版), 2017, 44(3): 123-128.
LI J, ZHAO L N, HOU X K. Matrix recovery by randomly permuted alternating direction method of multipliers (ADMM) [J]. Journal of Beijing University Chemical Technolgy(Natural Science), 2017, 44(3): 123-128. (in Chinese)

A new generalized robust principal component analysis (GRPCA) algorithm

HOU XuKe[1]　YANG HongWei[2]　MA Fang[1]　ZHAO LiNa[1*]

(1. Faculty of Science; 2. Center for Information Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: A new generalized robust principal component analysis (GRPCA) algorithm is proposed in order to recover the low-rank matrix with mixed noise pollution. It separates the low-rank part and the mixed noise part from the observation matrix by minimizing the combination of the kernel norm, the 1 norm, and the 2,1 norm, and then solving by a randomly permuted alternating direction multiplier method. Using spam classification as an example and a comparison with the classic methods PCA and RPCA shows that this method can effectively improve the accuracy and robustness of spam classification.

Key words: generalized robust principal component analysis(GRPCA); dimensionality reduction; k-nearest neighber (kNN); support vector machines(SVM)

(责任编辑:汪　琴)