Coefficient optimization model for a class of robust principal component analysis algorithms

Songyi Wu, Qinghuai Liu
School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

Abstract: There are more and more large-scale data models with high dimensions, but these large-scale data often have strong noise and sparse, which is troubled by information loss, noise influence and small sample. Nowadays, there is a demand for extracting effective content from chaotic information in many fields, such as pattern recognition, machine learning and data mining, among which robust principal component analysis is a common method to separate effective information from these raw data. Aiming at the traditional algorithm of robust principal component analysis, this paper establishes a new optimization model by assigning new coefficients to the low-rank matrix, which has a better correlation to the original matrix and improves the accuracy problem without changing the solving speed.

Key words: Principal component analysis Robust principal component analysis soft threshold operator coordinate axis descent method

1. Introduction

Principal component analysis (PCA) is a common step for preprocessing and denoising, in which a low-order approximation is made to an input matrix, such as a covariance matrix. While PCA is easy to implement through feature decomposition, it is sensitive to the presence of outliers because it attempts to “force fit” outliers into low-rank approximations. To overcome this problem, the concept of robust PCA (RPCA) was proposed with the goal of removing sparse missing from the input matrix and obtaining a low-rank approximation.

At present, the principal component analysis of component data has been deeply studied in the selection of data processing methods. It is common to carry out principal component analysis on the data after logarithmic ratio conversion, and obtain the corresponding data processing method, so as to establish a complete analysis system, and widely used in the analysis of component data.

For the most common improved model of PCA, the robust Principal component analysis model (RPCA) has become a research hotspot in recent years. However, the traditional RPCA model ignores the influence of samples with large reconstruction errors and damages the effective information of these samples in the principal component space. The first problem is that such information will reduce the ability of PCA to extract principal components of data. Therefore, this paper improves the traditional RPCA model in order to obtain an RPCA model with less losses.

2. Preparation Knowledge

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a common data analysis method. Its main functions are as follows: one is to extract the main feature components of the data; the other is to reduce dimension data, which is commonly used in high-dimensional matrix processing. The main means to achieve this algorithm is the mapping of different dimensions, for example, the original matrix features are N-dimensional, the original data features are mapped to m dimension, the m dimension obtained is the mapped new orthogonal feature, the reconstructed M-dimensional feature is called the principal component obtained on the basis of the original feature. Therefore, the means of dimensionality reduction of PCA is to retain the features containing most of the variance from the original space, and on this premise, to find each group of orthogonal coordinate axes sequentially, and ignore the noisy part with almost zero variance which has little influence on the experiment, so as to achieve dimensionality reduction of the data.

2.2 Robust Principal Component Analysis (RPCA)

In order to isolate low-rank structures from raw data with sparse large noise, Chandrasekaran et al and Wright et al independently proposed robust principal component analysis (RPCA) models. The main principle of RPCA is to represent the data contained in the original matrix as two parts, that is, the sparse invalid part, that is, the noise; And the effective low-rank part, which refers to the information space of the data. Therefore, the original matrix can be separated by using the rank and L-0norm of the matrix, and the desired low-rank matrix and sparse matrix can be constrained respectively. However, traditional solving methods are difficult to deal with NP-hard and non-convex models, and the rank function and L-0norm of the required matrix belong to such problems. Therefore, in order to effectively optimize the model, Candes et al. gave a theoretical proof that in order to accurately separate the low-rank space and sparse noise of the original data under certain conditions, this requirement can be achieved by optimizing the kernel norm of the low-rank matrix and the L-0norm of the sparse matrix. Then the problem of optimizing the nuclear norm and L-0norm is usually called the robust principal component analysis model.

Assuming that the given original data matrix D□Rm*n has A structure space of low dimension, then the matrix can be represented as the sum of two matrices, that is, D=A+E, where the matrix A□Rm*n is of low rank. The purpose of principal component analysis (PCA) is to find A low-rank matrix A that has the least error from the matrix E=D-A. We can set up the following optimization problem:

$$\min_{A,E} \| E \|_F \quad s.t. rank(A) \le r, D = A + E \tag{2-1}$$

Where, is the Frobenius norm of the matrix, representing the maximum rank of the matrix. $\| E \|_F$ E r A

3 Model Construction

3.1 Improvements to RPCA

This paper attempts to improve the traditional RPCA model by adding coefficients to the low-rank matrix A obtained from the decomposition of the original matrix as weight coefficients in the calculation process, so as to ensure that the decomposed matrix retains more components of the original matrix. $1-\lambda$ $0<\lambda<1$

The initial optimization problem of D=A+E (2-1) has been obtained in the above article. Then, in order to further optimize the matrix D, the singular value decomposition is performed, and the feature space of the new matrix obtained after the singular value decomposition is analyzed according to the required conditions. The required solution can be obtained by retaining the feature space corresponding to the first r maximum singular values required. According to the above principles, the RPCA algorithm can be described as follows: given the original matrix D=A+E, where A is low-rank, E is sparse, and its element values can be arbitrarily large, try to restore matrix A. In order to achieve this goal, the lowest rank matrix A and the matrix E with the fewest non-zero elements can be searched. Based on this principle, the following optimization problem is established:

$$\min_{A,E}(rank(A) \mid E \mid_0)$$

$$s.t. D = A + E$$

(3, 1)

Representing the rank function of the matrix in this biobjective optimization problem, representing the norm of the matrix. $rank(A)$ A $\| E \|_0$ E l_0 In order to transform the biobjective optimization problem into a simpler single objective optimization problem, the following convex optimization problem is obtained by introducing the equilibrium parameters: $0<\lambda<1$

$$\min_{L,S}(1-\lambda) \mid A \mid_* + \lambda \mid E \mid_1 \quad (3\text{ -}2)$$

$$s.t. D = A + E$$

Then model (3-2) is the improved robust principal component analysis model. It is then proved that in most scenarios, the recovery of low-rank matrix A and sparse matrix E only needs to consider solving the convex optimization problem (3-2) and optimizing its obtained results.

3.2 Coordinate axis descent method

Firstly, the concept of soft-thresholding operator is given:

Define 3.1 If a function is of the shape

$$soft(x,t) = \begin{cases} x+t, & x \le -t \\ 0, & |x| \le t \\ x-t, & x \ge t \end{cases}$$

Where,, then it is called a soft threshold operator. $x \in R$ $t>0$

The above definition of the soft threshold operator can also be generalized to matrix form.

Lemma 3.1 Let the singular value of the matrix be decomposed as, the rank of the matrix is, where, is singular, then there is a closed solution to the kernel norm minimization problem: $W = U\Sigma V^T$ r $\Sigma = diag(\sigma_1,\sigma_2,\cdots\sigma_r)$ $\sigma_1,\sigma_2,\cdots\sigma_r$

$$G_z[W] = Usoft(\sum \varepsilon)V^t = \arg\min \varepsilon \| X \| + \frac{1}{2}\| X - W \|^2$$

For the solution of the new RPCA, it can be calculated using the augmented Lagrange multiplier method (ALM) :

Firstly, the Lagrange multiplier matrix Y□Rm*n is introduced to construct the Lagrange function

$$L(A,E,Y) = (1-\lambda)\| A \|_* + \lambda \| E \|_1 + < Y, D - A - E >$$

Add a positive scalar as a penalty term μ

$$L(A,E,Y,\mu) = (1-\lambda)\| A \|_* + \lambda \| E \|_1 + < Y, D - A - E > + \frac{\mu}{2}\| D - A - E \|_F^2$$

Using the descending method of axes, fix the other axes and find an extreme value for one axis

Fix, solve, have A, Y, μ E

$$\min_E L(A,E,Y) = (1-\lambda)\| A \|_* + \lambda \| E \|_1 + < Y, D - A - E > + \frac{\mu}{2}\| D - A - E \|_F^2$$

Equivalent to solving

$$\min_E \mid E \mid_1 + \frac{\mu}{2\lambda} \mid E - (D - A + \frac{Y}{\mu}) \mid_F^2$$

The iterative formula that can be obtained from lemma 3.1: E_k

$$E_{k+1} = \arg\min_{E} \| E \|_1 + \frac{\mu_k}{2\lambda} \| E - (D - A_k + \frac{Y_k}{\mu_k}) \|_F^2$$

$$= soft(D - A_k + \frac{Y_k}{\mu_k}, \frac{\lambda}{\mu_k})$$

The same goes for

$$A_{k+1} = \arg\min_{E} \| A \|_1 + \frac{\mu_k}{2(1-\lambda)} \| A_k - (D - E_k + \frac{Y_k}{\mu_k}) \|_F^2$$

$$= soft(D - E_k + \frac{Y_k}{\mu_k}, \frac{1-\lambda}{\mu_k})$$

The iterative formula of the available matrix obtained from the above formula,, A_{k+1} E_{k+1} Y

$$Y_{k+1} = Y_k + \mu_k (D - A_{k+1} - E_{k+1})$$

Finally set the constant, the updated formula of the available parameters: $\rho > 1$ μ

$$\mu_{k+1} = \rho \mu_k$$

3.3 Algorithm steps:

Enter the initial matrix and balance parameters $D = A_0 + E_0$ $(D \in R^{m \times n})$ $0 < \lambda < 1$

Give the positive scalar and Lagrange multiplier μ, ρ Y_0

Ream, repeat the following iterations until convergence: $k = 0$

$$A_{k+1} = soft(D - E_k + \frac{Y_k}{\mu_k}, \frac{1-\lambda}{\mu_k})$$

$$E_{k+1} = soft(D - A_k + \frac{Y_k}{\mu_k}, \frac{\lambda}{\mu_k})$$

$$Y_{k+1} = Y_k + \mu_k (D - A_{k+1} - E_{k+1})$$

$$\mu_{k+1} = \rho \mu_k$$

Output the desired, matrix. A_k E_k

4 Summary

In this paper, by changing the coefficient of low-rank matrix A in RPCA algorithm and introducing the equilibrium parameter, a new formula of RPCA algorithm is obtained. The matrix decomposed by this method has a higher correlation with the original data, and tries to get a better convergence effect under the condition that the convergence speed is consistent.

References:

[1] Xue Wang,Miao Xie,Lingfei Zhou et al. Research on Principal Component Analysis based on Component Data processing [J]. Science and Technology Innovation,2023(18):94-98.

[2]ZHOU Z, LI X,WRIGHT J, et al. Stable principal component pursuit[C]. 2010 IEEE International Symposium on Information Theory. IEEE 20 10.

[3] Haipeng Wang,Ailian Jian,Pengxiang Li. Newton-soft Threshold iterative robust Principal Component Analysis Algorithm [J]. Journal of Computer Appli cations,20,40(11):3133-3138.

[4] Chong Zhang. The robust principal component analysis and its application [D]. Xi ‘an university of electronic science and technology, 2020. The DOI: 10.27389 /, dc nki. Gxadu. 2019.002480.