Scholar Publishing Group
International Journal of Social Sciences and Economic Management
https://doi.org/10.38007/IJSSEM.2024.050104
ISSN 2789-715X Vol. 5, Issue 1: 23-31

Low Rank Representation Subspace Clustering Algorithm Based on Hessian Regularization and Non Negative Constraints

Chunzhong Li[1,a*]

[1]*College of Statistics and Applied Mathematics, Anhui University of Finance & Economics, Bengbu 233030, Anhui, China*

[a]*120120038@aufe.edu.cn*

[*]*corresponding author*

Keywords: Hessian Regularization, Non Negative Constraints, Low Rank Representation, Subspace Clustering, Local Structure

Abstract: Existing low rank representation methods do not fully utilize the local structural features of data, resulting in problems such as loss of local similarity during the learning process. This paper proposed to use the low rank representation subspace clustering algorithm based on Hessian regularization and non-negative constraint (LRR-HN), to explore the overall and local structures of data. Firstly, the high predictability of Hessian regularization was fully utilized to preserve the local manifold structure of the data, thereby improving the description of the local topological structure of the data. Secondly, in view of the fact that the obtained coefficient matrix is often positive or negative, and negative values often have no practical significance, this article intended to introduce non negative constraints to ensure the correctness of the model solution and better characterize the local structure of the data. The NMI (Normalized Mutual Information) of Ncut (Normalized cut, Ncut) was 23.3%, and the AC (Accuracy) was 34.6%. The NMI of PCA (Principal Component Analysis) was 25.9%, and the AC was 45.3%. The NMI of LRR-HN was 89.9%, and AC was 93.2%. Experimental results showed that LRR-HN outperformed existing algorithms in areas such as AC and NMI, and had good clustering performance.

1. Introduction

In today's society, high-dimensional data is increasing and its structure is becoming increasingly complex. How to perform clustering analysis on it is an urgent problem to be solved. At present,

researchers generally assume that high-dimensional data is in a common low dimensional subspace. Therefore, this method has been widely applied. In recent years, subspace clustering has become an effective clustering algorithm in multiple disciplines such as computer vision, pattern recognition, and machine learning.

On this basis, drawing on the ideas of manifold learning, a new low rank subspace clustering method based on Hessian regularization method is studied. Firstly, a kernel norm based method is used to mine the overall structure of the data, thereby clustering the same class of data with high correlation. Secondly, using the Hessian regularization term and nearest neighbor sampling to linearly express the data enhances the local dependency relationship between the data. During the solving process, non negative constraints are introduced to better characterize the local structure of the data. On this basis, this article finally utilizes a linear transformation method based on adaptive penalty function, and verifies the feasibility of the method used in this article through typical sample testing.

2. Related Work

In recent years, multi view subspace clustering has become a hot topic, and methods based on low rank tensors have received widespread attention. In order to better explore the high-order correlation between different views, Liu Yunxiang focused on the common problems in current multi view subspace clustering research, namely complementary information and perspective noise, and he attached great importance to and optimized them. His research has made the implicit representation of subspaces more precise [1]. Li Huan used the latest tensor kernel norm and the coefficient matrix kernel norm and Frobenius norm as regularization terms to effectively explore his scheme [2]. In order to better describe the noise and low rank characteristics of data, Tu Zhihui planned to study subspace clustering methods based on two types of norms, and conducted convergence analysis on them [3]. In response to the current deep multi view subspace clustering methods, which do not fully consider the constraints of low rank representation of self expression matrices, resulting in models that are not robust enough, Yan Jintao planned to study a deep multi view subspace clustering method based on deep multi view subspace clustering. The method he used improved accuracy and normalized mutual information by 0.097 and 0.103, respectively [4]. The existing multi view clustering algorithms only utilize the overall low rank structure during learning, ignoring its local features, and are easily affected by noise in high-dimensional situations. Li Li planned to study a new hidden multi view subspace clustering algorithm based on tensor learning. This algorithm constructed a new sparse dictionary by mapping multi view data to a low dimensional vector space, effectively removing redundant information and noise [5]. The above subspace clustering methods have contributed to mapping high-dimensional nonlinear structured data to linear feature spaces, but there are still some issues. For example, the existing single kernel and multi kernel subspace clustering methods do not consider the importance of different rank components in the matrix during the kernel mapping process, and cannot approach the rank function more accurately, which cannot guarantee the low rank structure of feature space data and thus affect clustering performance. In addition, most of the above models mainly optimize the processing of nonlinear structural data, without considering both the optimization data term and the regularization term simultaneously. Appropriate data items can be used to model errors based on the statistical distribution of noise in actual data, in order to improve the robustness of the model.

In recent years, subspace clustering algorithms based on low rank representation have received attention due to their ability to effectively handle high-dimensional data [6]. Rank, as a measure of sparsity in a matrix, can better grasp the overall structure of data. How to replace matrix rank with better modulus to obtain a more accurate low rank structure is a highly challenging problem [7].

Currently, clustering algorithms based on low rank subspaces have been widely used in image segmentation, facial recognition, speech emotion recognition, and more.

3. Methods

3.1 Subspace Clustering Algorithm

In the era of big data, the scale of data is showing explosive growth, and high-dimensional data often reaches tens of thousands of dimensions. How to extract beneficial information for humans from a large amount of high-dimensional data is a current research challenge. Clustering algorithm is an effective method for solving such problems. Cluster analysis is an unsupervised learning behavior that does not require labeling and is of great significance in the context of big data. Traditional clustering algorithms are distance based and suitable for low dimensional clustering. However, in high-dimensional data, due to the redundancy of the data, the probability of clustering is very small, so clustering cannot be constructed based on distance. High dimensional data has significant applications in various fields such as information security and finance, and is a challenge in clustering analysis. In response to the complexity of high-dimensional data, parameterized modeling methods are generally used for modeling, and subspace clustering algorithms based on self-expression are one type. This algorithm assumes the union of high-dimensional data in multiple low dimensional subspaces, and then divides the high-dimensional subspaces in each subspace into corresponding low dimensional subspaces. The research on subspace clustering can not only provide new methods for processing high-dimensional data, but also make important progress in data classification, action segmentation, facial clustering, text clustering, heterogeneous data analysis, subspace learning, hybrid system identification in the field of control, and community segmentation in social networks.

In recent years, researchers have utilized different subspace clustering algorithms. The existing research methods can be classified into five categories based on their expression mechanisms: matrix factorization, Gaussian mixture, algebra, spectral clustering, and deep learning. On this basis, a new matrix factorization based algorithm is utilized, which has strong sensitivity to noise and outliers. The Gaussian mixture model assumes that each observation point is an independent sampling point after Gaussian distribution mixing, and uses maximum likelihood estimation to perform clustering analysis on it. These types of algorithms are generally sensitive to outliers and initial states. Algebraic based methods that utilize mathematical techniques to process data points, such as fitting data points with quadratic or higher-order polynomials. However, existing algorithms are costly, especially when dealing with high-dimensional data, and are sensitive to noise and outliers. In recent years, the use of spectral clustering technology to solve optimal problems has been widely applied due to its ease of solving and strong computational efficiency.

On the basis of spectral clustering, a similarity based clustering algorithm can be utilized, which can effectively extract low dimensional steganography of data, and then use the k-means method to obtain clustering results. The difference between various spectral clustering algorithms is the construction of affinity matrices. The data in the similarity matrix represents the similarity between data points. In an ideal situation, when the similarity matrix is a block diagonal structure, that is, when the similarity between all classes is 0, spectral clustering technology can achieve the desired clustering effect. Traditional Gaussian kernel functions or other affinity matrices constructed based on local information (such as local subspace similarity) may not necessarily point to an ideal subspace clustering algorithm. Therefore, this article utilizes a new subspace clustering method based on global information, which adopts regularization method and applies it to subspace clustering.

3.2 LRR Algorithm

Unlike traditional algorithms, the goal of low rank representation algorithms is to find the minimum rank of the data. This approach can better understand the overall and inherent information of the data space [8-9]. However, existing low rank representation methods cannot fully utilize the local linear relationships between data, and the constructed similarity matrix is often too dense, and its negative values have little impact on constructing affinity [10]. In addition, since the samples used to represent cannot represent a dictionary, the established dictionary is also very robust to noise.

LRR is another mainstream algorithm based on spectral clustering, which has been highly valued since its utilization because it can effectively mine low dimensional subspace structures hidden in data, and has better stability and overall structure than SSC (Sparse subspace clustering) [11-12]. It is precisely because of this advantage that scholars have made significant improvements and improvements to low rank representation, resulting in a number of excellent low rank representation algorithms [13]. Before exploring these methods, it is necessary to briefly explain their basic principles, namely LRR. LRR is built on the premise of union sampling on multiple low rank linear subsets, and LRR also has the property of self-expression, that is, each sample is expressed in its own form. Unlike SSC, which selects the least number of features from multiple expressions, LRR aims to find the minimum rank of all data [14-15]. Therefore, compared with SSC, LRR can better control the overall structure of sample points, that is, by applying low rank constraints to the expression matrix, it aggregates data points with high correlation together, thereby achieving overall clustering.

Assuming a set of data:

$$S_J = \left[s_{1,}s_{2,}s_{3,\cdots,}s_{d\times n}\right](1)$$

Here, d represents the dimension of each data sample, and n represents the number of samples for all data. Because each sample can be linearly represented by other samples in the same subspace, a self representation model can be defined as follows:

$$\mathrm{X} = \mathrm{z_d B}(2)$$

Among them, $\mathrm{z_d}$ is the dictionary, and B represents the coefficient matrix.

In the LRR method, if the dataset X itself is viewed as a data dictionary, then this model can be represented as:

$$\mathrm{minrank(B)s.t.X = XB}(3)$$

However, in real life, data nodes are often noisy or damaged. Moreover, given that the kernel norm can be used to implement low rank constraints, the canonical function can be rewritten as follows:

$$\mathrm{minrank(B)s.t.X_S = XB + \lambda E}(4)$$

Among them, norm, $\lambda > 0$. After obtaining B, it is defined as:

$$\mathrm{G = [B + B'\ ]/2}(5)$$

The final image is obtained using the HN method. With the development of society and the continuous progress of human civilization, high-performance computers are no longer a new thing. The ways to obtain data have also greatly developed from manual statistics to various sensors such as fingerprints, portraits, sounds, etc. In recent years, due to the diversification of data collection methods, data has shown an explosive development trend. With the continuous progress of

instrument technology, the quality of collected data is also getting better and better, thereby greatly improving the effective utilization of data. This provides opportunities for technological research, but also brings new challenges: how to discover and extract useful information from complex big data. The clustering algorithm has achieved good results in solving such problems.

3.3 Objective Function of LRR-HN

Usually, the objective function of LRR-HN is negative, and in real life, negative coefficients often become unreasonable or even meaningless. This article intends to use $B \geqslant 0$ as a constraint condition to ensure that all data points fall within the convex hull of their adjacent points, thereby better reflecting the correlation between data and better characterizing local structures. Furthermore, according to the manifold assumption, when the geometric structures between two data points are similar, their embedding and projection in the new space are also similar. Therefore, considering the excellent characteristics of Hessian energy, this paper intends to introduce Hessian regularization terms into low rank models based on Hessian regularization terms to better characterize the correlation between local data. On this basis, the Hessian regularization term is used to replace the coefficient matrix B, and the objective function of LRR-HN is obtained as follows:

$$\mathrm{minrank}(B)\mathrm{s.t.}X_H = E\| B\| + \lambda 2\| E\| + \lambda 3 (6)$$

4. Results and Discussion

4.1 Evaluation Indicators

On this basis, this article intends to quantitatively evaluate the algorithm using accuracy (AC) and normalized mutual information (NMI) as indicators. Then, the AC method is defined as:

$$\mathrm{AC} = \frac{b+\sqrt{g_1+g_2}}{2a} (7)$$

g_1 represents the true category of the sample, and g_2 represents the clustering category of the sample.

There are different clustering results J_1 and J_2, and the NMI method is:

$$\mathrm{NMI} = \sum_{k=0}^{n} (J_1 + J_2) x^k a^{n-k} (8)$$

4.2 DBI Evaluation

This article uses the Davies-Bouldin Index (DBI) as an evaluation metric to measure the compactness and separability of clustering results. The number of categories and DBI under Hessian regularization and low rank representation are shown in Figure 1. The DBI of dataset A is 0.85, and the DBI of dataset B is 0.88.

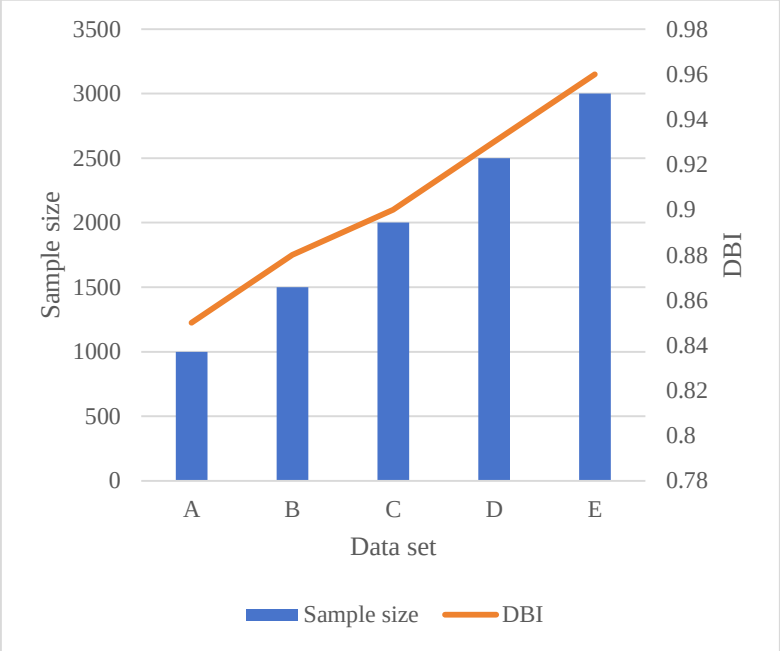

Figure 1. Number of categories and DBI under Hessian regularization and low rank representation

4.3 Evaluation of Algorithm Execution Time

This article intends to select large-scale samples to study the running speed of low rank representation subspace clustering methods under Hessian regularization and non negative constraints, and evaluate the computational efficiency of the algorithm.

The dimensions corresponding to different datasets are shown in Table 1. The dimension of dataset A is 50, and the dimension of dataset B is 100.

Table 1. Dimensions corresponding to different datasets

Data set	Dimension
A	50
B	100
C	200
D	300
E	400

The number of samples and execution time corresponding to different datasets are shown in Figure 2. Dataset A has a sample size of 10000 and an execution time of 32.5 seconds. The sample size of dataset B is 20000, and the execution time is 78.4 seconds.

Figure 2. Sample size and execution time corresponding to different datasets

4.4 Comparison of Clustering Effects under Different Hyperparameter Settings

This article intends to use the variable Hessian regularization and non negative regularization parameter setting methods (regularization parameters, non negative constraint strength, etc.) to study the clustering effect under different conditions, and optimize the optimal hyperparameter setting through comparative analysis.

The regularization parameters, non negative constraint strength, and DBI corresponding to different hyperparameter experimental groups are shown in Table 2. The regularization parameter of hyperparameter experimental group 1 is 0.1, with weak non negative constraint strength and a DBI index of 0.82. The regularization parameter of hyperparameter experimental group 2 is 0.5, with weak non negative constraint strength and a DBI index of 0.78.

Table 2. Regularization parameters corresponding to different hyperparameter experimental groups, non negative constraint strength and DBI

Hyperparameter experimental group	Regularization parameter	Non negative constraint strength	DBI
1	0.1	Weak	0.82
2	0.5	Weak	0.78
3	1.0	Weak	0.75
4	2.0	Weak	0.71

4.5 NMI and AC

The K-means algorithm is a distance based clustering method. This method randomly selects

several classes as the initial cluster centers, classifies the classes according to their distance from the cluster center, and updates the cluster center until convergence is achieved. NMF is an unsupervised learning method that enables data dimensionality reduction. On this basis, non negative matrix factorization is used to reduce dimensionality, and then K-means algorithm is used to perform clustering analysis on the samples. PCA is a widely used method for unsupervised data dimensionality reduction. This method first performs a linear transformation on the original data, making it linearly independent in all dimensions, which can be used for feature extraction and denoising. Ncut is a spectral clustering algorithm. This algorithm uses adjacent matrices to obtain eigenvalues and eigenvectors, normalizes them to form a new matrix, and then uses the K-means algorithm to achieve clustering.

The NMI and AC of different algorithms on the Yale dataset are shown in Figure 3. The NMI of Ncut is 23.3%, and the AC is 34.6%. The NMI of PCA is 25.9%, and the AC is 45.3%. The NMI of LRR-HN is 89.9%, and AC is 93.2%.

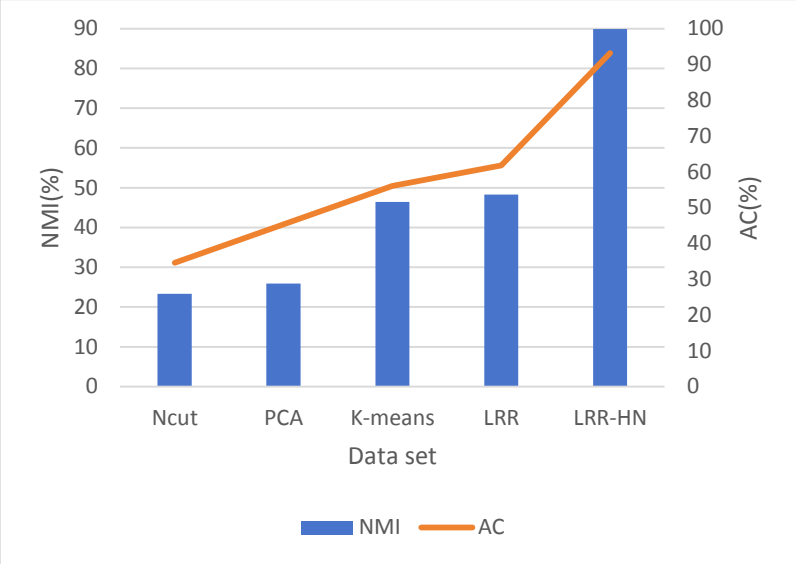

Figure 3. NMI and AC of different algorithms on the Yale dataset

5. Conclusions

In recent years, with the widespread application of computer vision, image processing and other fields, subspace clustering methods have received increasing attention. Among these methods, subspace clustering has attracted much attention for its simplicity and good performance. This article intended to use the minimum low rank representation subspace clustering algorithm to study the low rank nature of the overall structure of data and achieve effective clustering of high-dimensional data. On this basis, this article intended to design an efficient LRR-HN algorithm based on the adaptive penalty function linear alternating direction method, and evaluate the performance of the proposed algorithm based on actual data using methods such as accuracy and normalized mutual information. However, this method also has some drawbacks: firstly, the data dictionary needs to be denoised to ensure the accuracy of clustering, which can lead to an extension of calculation time; secondly, when the number of clusters is large, there are significant errors in the clustering results. Further research in the future can reduce computation time and clustering errors while ensuring accuracy.

Funding

This work was supported by the Anhui Provincial University Natural Science Foundation (No. KJ2021A0481)

References

[1] Liu Yunxiang, Wang Yibin. Adaptive weighted multi view subspace clustering algorithm based on latent representation. Computer Knowledge and Technology: Academic Edition, 2023, 19 (17): 10-15.
[2] Li Huan, Tang Kewei. Multi view subspace clustering based on low rank tensor representation. Theoretical Mathematics, 2023, 13 (10): 2877-2887.
[3] Tu Zhihui, Chen Long, Zhang Zichang, et al. Subspace clustering algorithm for joint Capped norm minimization. Journal of Gannan Normal University, 2020, 041 (006): 56-61.
[4] Yan Jintao, Li Zhongyu, Tang Qifan, et al. Deep low rank multi view subspace clustering. Journal of Xi'an Jiaotong University, 2021, 055 (011): 125-135.
[5] Li Li, Li Jinghao, Zhang Xiaoqian. Potential Multi View Subspace Clustering Based on Tensor Learning. Journal of Southwest University of Science and Technology, 2022, 37 (3): 52-59.
[6] Chen J, Yang S, Mao H, et al. Multiview subspace clustering using low-rank representation. IEEE Transactions on Cybernetics, 2021, 52(11): 12364-12378.
[7] Abhadiomhen S E, Wang Z Y, Shen X J. Coupled low rank representation and subspace clustering. Applied Intelligence, 2022, 52(1): 530-546.
[8] Abhadiomhen S E, Wang Z, Shen X, et al. Multiview common subspace clustering via coupled low rank representation. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(4): 1-25.
[9] Nie F, Chang W, Hu Z, et al. Robust subspace clustering with low-rank structure constraint. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(3): 1404-1415.
[10] Khan G A, Hu J, Li T, et al. Multi-view subspace clustering for learning joint representation via low-rank sparse representation. Applied Intelligence, 2023, 53(19): 22511-22530.
[11] Chen Y, Xiao X, Peng C, et al. Low-rank tensor graph learning for multi-view subspace clustering. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 92-104.
[12] Guo J, Sun Y, Gao J, et al. Rank consistency induced multiview subspace clustering via low-rank matrix factorization. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(7): 3157-3170.
[13] Sun W, Peng J, Yang G, et al. Fast and latent low-rank subspace clustering for hyperspectral band selection. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(6): 3906-3915.
[14] Peng X, Feng J, Zhou J T, et al. Deep subspace clustering. IEEE transactions on neural networks and learning systems, 2020, 31(12): 5509-5521.
[15] Sui J, Liu Z, Liu L, et al. Dynamic sparse subspace clustering for evolving high-dimensional data streams. IEEE Transactions on Cybernetics, 2020, 52(6): 4173-4186.