High-dimensional Teaching Data Clustering in Sparse Subspaces Based on Information Entropy

Huiyan Liu[1]

Artificial Intelligence College, Shenyang Normal University
Shenyang, 110034 China

Received May. 25, 2025; Revised and Accepted June. 25, 2025

Abstract. Due to the large scale and high dimension of teaching data, the using of traditional clustering algorithms has problems such as high computational complexity and low accuracy. Therefore, this paper proposes a weighted block sparse subspace clustering algorithm based on information entropy. The introduction of information entropy weight and block diagonal constraints can obtain the prior probability that two pixels belong to the same category before the simulation experiment, thereby positively intervening that the solutions solved by the model tend to be the optimal approximate solutions of the block diagonal structure. It can enable the model to obtain the performance against noise and outliers, and thereby improving the discriminative ability of the model classification. The experimental results show that the average probability Rand index of the proposed method is 0.86, which is higher than that of other algorithms. The average information change index of the proposed method is 1.55, which is lower than that of other algorithms, proving its strong robustness. On different datasets, the misclassification rates of the design method are 1.2% and 0.9% respectively, which proves that its classification accuracy is relatively high. The proposed method has high reliability in processing high-dimensional teaching data. It can play an important role in the field of educational data analysis and provide strong support for intelligent teaching.

Keywords: Intelligent teaching, Sparse subspace clustering, information entropy, high-dimensional.

1. Introduction

In high-dimensional clustering algorithms, if the data is distributed in the union of linear or affine subspaces, the subspace clustering algorithm is an effective way to achieve high-dimensional data clustering [1]. The subspace clustering algorithm assigns points in the same subspace to the corresponding subspace according to a certain classification to achieve the classification effect [2]. The sparse subspace clustering (SSC) algorithm [3] and the low-rank representation (LRR) algorithm [4] are classic subspace clustering algorithms. Although both the LRR algorithm and the SSC algorithm are subspace clustering algorithms based on spectral clustering [5], there are essential differences between the two algorithms in the sparsity constraints of the model. The LRR algorithm is a subspace clustering algorithm based on two-dimensional sparsity of data and with low-rank global constraints. However, when encountering noisy data, the sparsity of the low-rank representation coefficients is poor. The SSC algorithm is a subspace clustering algorithm based on one-dimensional sparsity of data. It constructs a similarity matrix using the sparse representation coefficients of the data and applies it to the spectral clustering method to obtain the subspace clustering results of the data [6]. Since SSC only utilizes the spectral information of each pixel point during the clustering process and does not consider the spatial context information, the connectivity of the adjacency matrix of the graph is reduced, and even a large amount of salt-and-pepper noise may appear in its final clustered image [7].

Since in practical applications, the data itself has noise and outliers, which cannot meet the assumption of subspace independence, the matrix structure of self-expression is disrupted, affecting the clustering results. Lu et al. [8] proposed a subspace clustering algorithm based on block diagonal representation, imposing block diagonal constraints on the representation matrix to enable the representation matrix to have a good block diagonal structure. If the representation matrix had a block diagonal structure, then the coefficients corresponding to pixels that did not belong to the same category are zero. Then the representation matrix with a block diagonal structure had a good grouping effect, and when it was input into spectral clustering, good clustering results could be obtained. Teaching data has high dimensionality and noise. Traditional subspace clustering algorithms are vulnerable to noise and have relatively low classification accuracy.

2. Sparse Subspace Clustering Algorithm

The SSC algorithm is accomplished based on spectral clustering. The basic idea is that pixel data in a high-dimensional space can be linearly represented in a low-dimensional space. The sparse representation matrix obtained by solving the SSC algorithm can better reflect the attributes of the pixel data subspace and has sparsity. Finally, it is applied to the spectral clustering algorithm to obtain the clustering results.

The sparse model is as follows.

$$\min_{c} ||C||_1 + \lambda||X - XC||_F^2, s.t.diag(C) = 0. \tag{1}$$

In the formula, $C \in R^{MN\times NM}$ is a sparse coefficient matrix. $X \in R^{D\times MN}$ is the data matrix. When imposing sparse constraints on the representation matrix, if equation (1) is an l_0-norm constraint, it is a non-convex optimization problem. The general processing method is to transform the optimization objective into the l_1-norm problem of convex optimization, and then use convex programming tools to obtain sparse solutions.

The sparse coefficient matrix $C = [c_1, C_2 \cdots, c_{MN}]$ of the pixel points is obtained by equation (1) through the Alternating direction method of multipliers (ADMM) [10]. Then it standardizes each column vector $c_i = \frac{c_i}{||c_i||_\infty}$ of the sparse coefficient matrix. Then, by solving the obtained coefficient matrix C, it calculates the similarity matrix $G = (C + C^T)/2$. Finally, it is applied to the spectral clustering of standardized segmentation to obtain the clustering results of all pixels [5]. Since the SSC model is sensitive to noise, Li et al. [9] proposed Gaussian weighted sparse subspace clustering (GSSC), introducing weights with sparse constraints to make the data linearly represented by data points in the same subspace as much as possible. The weights of the two data points were determined by the Gaussian similarity function, and the subspace representation model was as follows.

$$\min_{C} \sum_{j\neq i} \frac{1}{W_{ij}} |C_{ij}| + \lambda||X - XC||_F^2, s.t.diag(C) = 0. \tag{2}$$

Where $W_{ij} = exp(-\frac{||x_i - x_j||_2^2}{\sigma^2})$ denotes the Gaussian similarity between x_i and x_j.

3. LRR Subspace Clustering with Structural and Symmetry Constraints

In the study of subspace clustering, imposing constraints on the structure of low-rank representation solutions can obtain better clustering results. Therefore, this paper proposes a low-rank representation subspace clustering method with structural constraints, introducing structural constraints and symmetric constraints into the solutions of low-rank representations to construct a weighted sparse and symmetric low-rank affinity graph. Here, the low-rank constraint is used to capture the global structure of the data, the structural constraint is used to capture the local linear structure of the data, and the symmetric constraint can ensure the consistency of the weights between each data point. In fact, structural constraints, namely weighted sparse representation, can reveal the strong affinity among samples of the same class and the strong separability among samples of different classes, that is, the strong affinity within classes and the strong separability between classes [10-13].

To obtain the representation model from the structure of the data, constraint terms can be imposed on the structure of the solution of the LRR model. In this paper, the structure of the solution is restricted by adding the $\sum_{i,j} R_{ij}|z_{ij}|$ constraint and the $z_{ij} = z_{ji}$ constraint in the objective function (see Equation (3)). Compared with the objective function that only considers the kernel norm, this can not only improve the rank of the solution, but also retain the intrinsic geometric structure between data points, achieving a better subspace clustering effect.

$$\min_{Z} ||Z||_* + \beta||R \odot Z||_1, s.t.X = ZA + E, Z = Z^T. \tag{3}$$

To make the obtained Z more robust to noise and avoid the NP problem, a structurally constrained symmetric low-rank representation (SCSLR) model is constructed, as shown in equation (4):

$$\min_{Z} ||Z||_* + \beta||R \odot Z||_1 + \lambda||Z||_{2,1}. \tag{4}$$

In fact, when the data is labeled, SCSLR can be regarded as a semi-supervised LRRSC [14]. For data without labels, the structure of the data can be utilized to construct the weight R, that is, the weight is determined by the Angle. This means that the smaller the Angle between data points from the same category, the smaller the sample weight; conversely, the larger it is. Through data standardization processing, after calculating the absolute value of the inner product between data points, the ideal weight matrix R is constructed as follows:

$$R_{ij} = 1 - exp(-\frac{1 - |x_i^{*T} x_j^*|}{\sigma}). \tag{5}$$

Where, x_i^* and x_j^* are the normalized data points of x_i and x_j. σ is the mean value of B element ($B_{ij} = 1 - |x_i^{*T} x_j^*|$). Usually, the weights between data points from the same category are relatively small, while the weights between data points from different categories are relatively large. The following text will construct R in this way.

3.1. Model Optimization

In this paper, the alternating minimization method is used to solve the objective function in equation (5). It introduces the auxiliary variables J and L to transform the model as shown in equation (6).

$$\min_{Z,E} ||J||_* + \beta||R \odot L||_1 + \lambda||E||_{2,1}, s.t. X = AZ + E, Z = J, Z = L, J = J^T. \tag{6}$$

After obtaining the weighted sparse symmetric low-rank representation matrix Z^*, an affinity graph $G = (V, E)$ is constructed using Z^*. Where $V = v_{i}{}_{i=1}^{n}$ is the point set and $E = e_{i}{}_{i=1}^{n}$ is the edge set. When the data set is given, the problem of graph construction depends on the weight matrix $W = W_{ij}$. Since each data point can be represented by a linear combination of other data points, the contribution of other data points to the reconstruction of x_i is represented by the z_i^*-th column of Z^*.

In this paper, the weight matrix W is constructed based on the structure of matrix Z^*. It decomposes the singular values of Z^* into $U^*\Sigma^*(V^*)^T$. The parameters U^* and V^* are respectively the orthogonal bases of the column and row vectors of the matrix Z^*. First, according to reference [15], it multiplies the weight of each column of U^* by $(\Sigma^*)^{0.5}$, and multiplies the weight of each row of $(V^*)^T$ by $(\Sigma^*)^{0.5}$. Then, by defining $M = U^*(\Sigma^*)^{0.5}$ and $N = V^*(\Sigma^*)^{0.5}$, let $Z^* : Z^* = MN$ be represented by the matrices M and N. That is, the weight matrix W of the affinity graph is defined by using the Angle information of all row vectors from matrix M or all column vectors from matrix N, as shown in equation (7). Among them, $m_i(n_i)$ and $m_j(n_j)$ are respectively the $i - th$ row and the $j - th$ row of the matrix $M(N)$, and the parameter $\alpha \in N$ is used to adjust the similarity between samples.

$$W_{ij} = (\frac{m_i^T m_j}{||m_i||_2||m_j||_2}). \tag{7}$$

On this basis, the NCuts algorithm is applied to segment the samples into the corresponding subspaces. Suppose the graph $G = (V, E, W)$ is divided into two parts, A and B. These two parts satisfy the conditions $A \cup B = V$ and $A \cap B = O$. Then the division formula is as follows:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)}. \tag{8}$$

$$assoc(A, V) = \sum_{u \in A, v \in V} w(u, v). \tag{9}$$

Equation (8) represents the inter-class similarity between parts A and B, and the smaller the value, the better. Equation (9) represents the sum of the weights of part A and the overall node V. Solving the minimum value of Ncut can lead to a better segmentation result.

4. Experimental Results and Analysis

To verify the performance of the proposed teaching data classification method that combines deep learning and sparse subspace clustering algorithm, simulation experiments are conducted on a computer with an Intel Core i7-10700 processor, 6GB of running memory, an independent graphics card configuration of NVIDIA GeForce GT930M, and Windows 10 system with the Matlab software. Firstly, the model is trained on the SCB-Dataset. The clustering accuracy and computing time of the designed method are calculated and it is compared with the traditional sparse subspace clustering (SSC) algorithm and the low-rank subspace clustering (LSC) algorithm. The results are shown in Figures 1 and 2.

It can be seen from Figure 1 that the clustering accuracy rates of the three algorithms increase with the increase of the number of iterations. When the number of iterations is 180, the clustering accuracy rates of the three algorithms are 91.7%, 78.3%, and 70.1% respectively. It can be seen from Figure 2 that the computing times of the three algorithms show an upward trend. When the number of classifications is 27, the computing times of the three algorithms are 11.8s,18.6s and 26.4s respectively. The clustering accuracy of the proposed algorithm is significantly higher than that of other algorithms, and its computing time is much lower than that of other algorithms, which proves the accuracy and efficiency of the proposed method in processing teaching data.

To verify the effect of the designed algorithm in practical applications, the study first introduces the normalized mutual information index, which is a common evaluation index for the effectiveness of clustering methods. The comparison results of normalized mutual information indicators of different algorithms are shown in table 1.

It can be seen from Table 1 that the average normalized mutual information value of the proposed algorithm is 0.89, the average normalized mutual information value of the sparse subspace clustering algorithm is 0.78, and the

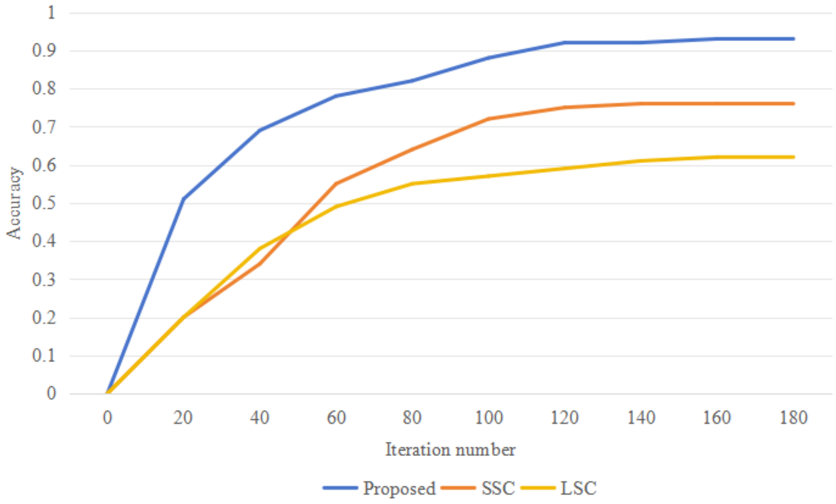

Fig. 1. The accuracy rates with different algorithms

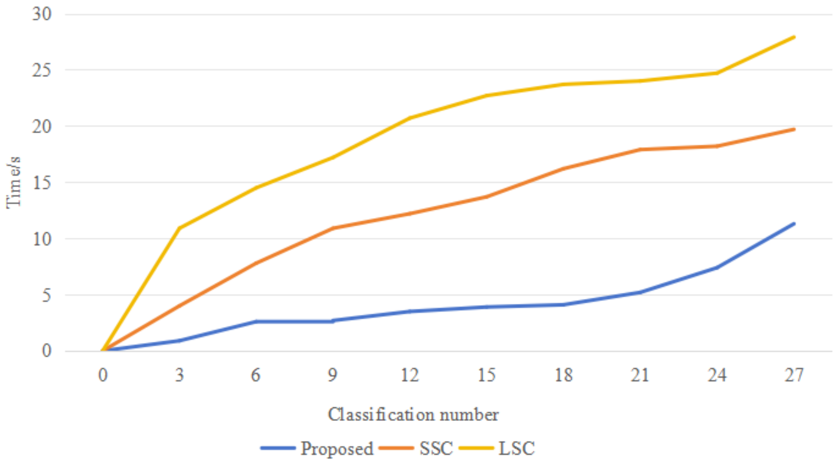

Fig. 2. The computing time with different algorithms

Table 1. Normalized mutual information indicators with different algorithms

Iteration number	Proposed	SSC	LSC
0	0.95	0.86	0.67
20	0.93	0.77	0.61
40	0.91	0.74	0.59
60	0.89	0.72	0.58
80	0.88	0.71	0.57
100	0.89	0.72	0.57
120	0.89	0.73	0.58
140	0.87	0.73	0.52
160	0.86	0.72	0.59
180	0.85	0.70	0.55

average normalized mutual information value of the low-rank subspace clustering algorithm is 0.61. The average normalized mutual information value of the proposed method is higher than that of the other two algorithms, which proves that its classification effect is better.

The next step is to select a group of noisy data to verify the processing effect of the proposed method on the noisy data. The evaluation is conducted through the probability rand index and the information change index. The larger the probability rand index, the better the processing effect; the smaller the information change index, the better the processing effect. The variation of the probability Rand index curve and the information change curve of different algorithms is shown in Figure 3. It can be seen from Figure 3 that the average probability Rand indices of the three algorithms are 0.85, 0.66, and 0.58 respectively. The average Rand index of the proposed algorithm is higher than that of other algorithms, which proves that the proposed algorithm has strong robustness.

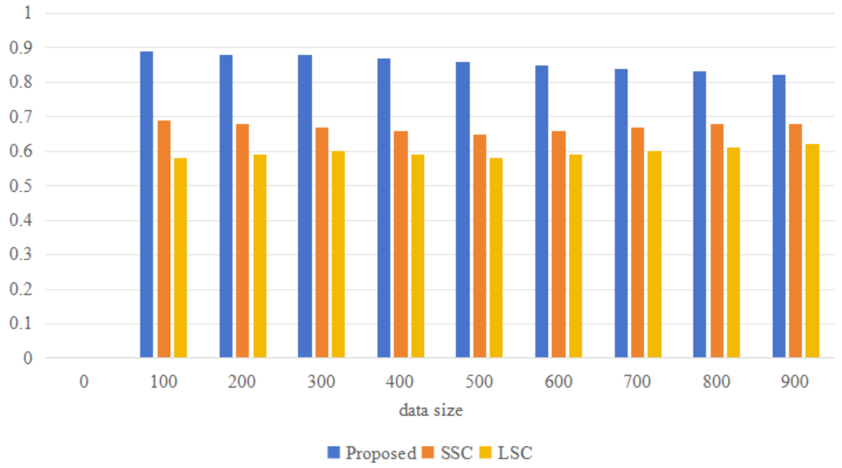

Fig. 3. Probabilistic Rand exponential curves of different algorithms

5. Conclusion

With the advent of the era of big data, the processing of high-dimensional data has become an important research direction in the field of data mining. Teaching data, as a type of high-dimensional data, its clustering processing is of great significance for the formulation of teaching plans. Based on the sparse subspace clustering algorithm, the research introduces unsupervised metric learning to preprocess the data to improve the classification effect, and designs a teaching data classification method based on the sparse subspace clustering algorithm and deep learning. The results show that in the comparison of clustering accuracy and computing time, the clustering accuracy of the designed algorithm is 91.7% and the computing time is 11.8s, which proves its accuracy and efficiency. The above results prove the effectiveness of the proposed algorithm in processing high-dimensional teaching data. In the future, it will be verified on more datasets to further improve the performance of the algorithm.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. None.

References

1. Kumar, Abhishek, et al. ”Entropy-weighted medoid shift: An automated clustering algorithm for high-dimensional data.”Applied Soft Computing 169 (2025): 112347.
2. Lehner, Sebastian, Katharina Enigl, and Matthias Schlgl. ”Derivation of characteristic physioclimatic regions through density-based spatial clustering of high-dimensional data.”Environmental Modelling & Software (2025): 106324.
3. Zhu, Yanjiao, et al. ”Robust and stochastic sparse subspace clustering.”Neurocomputing 611 (2025): 128703.

4. Leng, Chengcai, et al. ”Dual graph-regularized low-rank representation for hyperspectral image denoising.”Engineering Applications of Artificial Intelligence 139 (2025): 109659.
5. Lavindi, Eri Eli, and Nina Faoziyah. ”Adaptive Swarm Intelligence Algorithms for High-Dimensional Data Clustering in Big Data Analytics.”ALCOM: Journal of Algorithm and Computing 1.01 (2025): 23-32.
6. Singh, Vikas, and Nishchal K. Verma. ”Variable feature weighted fuzzy k-means algorithm for high dimensional data.”Multimedia Tools and Applications? (2025): 1-18.
7. Li, Yuchen, et al. ”Stacked Ensemble of Extremely Interpretable TakagiCSugenoCKang Fuzzy Classifiers for High-Dimensional Data.”IEEE Transactions on Systems, Man, and Cybernetics: Systems (2025).
8. Lu, Canyi, et al. ”Subspace clustering by block diagonal representation.”IEEE transactions on pattern analysis and machine intelligence? 41.2 (2018): 487-501.
9. Jiang, Kun, et al. ”Graph embedded subspace clustering with entropy-based feature weighting.”International Journal of Machine Learning and Cybernetics (2025): 1-19.
10. Chen, Xu, et al. ”Optimizing Block Skipping for High-Dimensional Data with Learned Adaptive Curve.”Proceedings of the ACM on Management of Data 3.1 (2025): 1-26.
11. Lu, Minyuan, and Bu Zhou. ”A one-way MANOVA test for high-dimensional data using clustering subspaces.”Statistics & Probability Letters 217 (2025): 110293.
12. Liu, Peng, et al. ”Comprehensive evaluation and practical guideline of gating methods for high-dimensional cytometry data: manual gating, unsupervised clustering, and auto-gating.”Briefings in Bioinformatics 26.1 (2025): bbae633.
13. Niu, Guo, et al. ”RKHS reconstruction based on manifold learning for high-dimensional data.”Applied Intelligence 55.2 (2025): 1-24.
14. Chen, Shuai, et al. ”RadioLLM: Introducing Large Language Model into Cognitive Radio via Hybrid Prompt and Token Reprogrammings.”arxiv preprint arxiv:2501.17888 (2025).
15. Zong, Wei, et al. ”Model-based multifacet clustering with high-dimensional omics applications.”Biostatistics 26.1 (2025): kxae020.