Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220022

Adaptive Subspace Clustering with Sparse Constraints

Hao WANG and Zhiqiang ZENG[1]
College of Computer and Information Engineering, Xiamen University of Technology, Xiamen, China

Abstract. In many practical applications, data are represented by high-dimensional features. Although the traditional K-means algorithm is simple, it usually gets the approximation solution by eigenvalue decomposition, this method led to the model less efficient. In addition, their loss functions are sensitive to data distribution. In this paper, a clustering model of adaptive K-means with sparse constraints is proposed. The proposed method is designed by combining the dimension reduction with sparse constraints and adaptive clustering. It provides a flexible computational framework for subspace clustering and is suitable for different distribution data sets. Besides, the sparsely constraint in our method can remove redundant features and retain useful information. We develop an effective alternative optimization algorithm to solve our model. Finally, the extended experiments on several benchmark datasets demonstrate the advantages of our method over other clustering algorithms

Keywords. Feature selection, K-means, discriminative embedded clustering, adaptive learning

1. Introduction

Clustering and dimension reduction are widely used in machine learning, PCA and K-means (KM) are frequently used due to their simplicity and efficiency. KM first randomly initializes the cluster center, then determines the similarity of the data based on the distance between the data and its cluster center. After that, KM assigns the data to the nearest cluster center, and then updates its cluster center for each cluster of data.

However, due to the different distribution of data, a fixed norm may not suitable for all data. Typically, as shown in Fig.1(a), KM uses the L_2-norm to calculate the loss value, which assigns a larger weight to the farther data. Therefore, KM is sensitive to outliers. Different from KM, the $L_{2,1}$-norm used in Du et al. [1] proposed RMKKM in Fig.1(b) values higher density areas. However, this weight strategy may be overestimated, lead to misjudgement.

As an essential means to process high-dimensional data, dimensional reduce technology can not only reduce the computational time and storage space requirement, but also reduce the impact of noise and redundant features. At present, in the field of machine learning, the conventional dimensional reduce algorithm relies on the matrix decomposition, mutual information [2,3], regression-based feature selection, and so on. In addition, with the development of sparse representation and low rank representation [4-8], researchers have proposed many efficient and robust dimension reduction methods

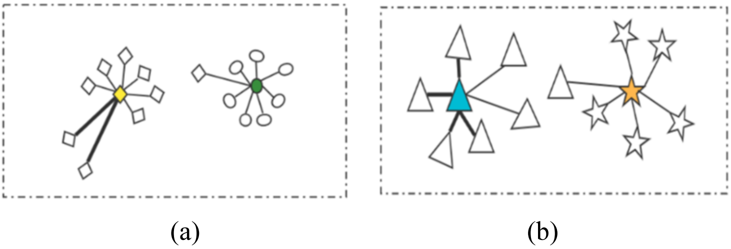
[1] Corresponding Author. E-mail: zqzeng@xmut.edu.cn.

such as SPCA [4], CSPCA [5]. These methods have achieved satisfactory results in practical applications. These sparse dimensional reduction algorithms not only improve the readability of the main components, but also preserve as much raw data structure information as possible while reducing data redundancy information. In this paper, the PCA is imposed to obtain a sparse projection matrix. Because of the diversity of sample distribution, a more flexible norm is needed to calculate the amount of data loss information.

In the field of machine learning, clustering, as a basic technique, has been widely discuss. In the past few decades, we have seen its remarkable effects in image, voice, word, and other information processing. KM is a commonly used clustering method, which is widely used in many applications because of its efficiency and simplicity. However, KM are very sensitive to initialization, the results are often unstable, especially when large amounts of noise and related characteristics are included in the data. In this case, the precision of the clustering algorithm will be directly affected.

It is impractical to deal directly with the unnecessary noise of these high-dimensional data and the associated characteristics it contains. A simple method to handle this problem is to reduce the dimension of the data, such as the Principal Component Analysis (PCA), Linear Discrimination Analysis (LDA), and then preform clustering [9,10]. For example, a PCA and KM-based clustering algorithm called PCAKM was introduced by Hou C et al [9]. However, one of the main disadvantages of these clustering is that they perform subspace learning and clustering independently, but the subspace obtained by dimensionality reduction may not be the best result for subsequent clustering tasks. In order to effectively improve the accuracy of subspace clustering, the researchers found that joint subspace learning and clustering are beneficial for clustering [11,16]. For example, LDA and KM are adaptively applied to the Joint Framework (LDAKM) proposed by Ding C [12]. In this framework, clustering is combined with subspace learning adaptively to make the treatment of related features more effective. To better explore the identifying information in high-dimensional data, Hou et al. [13] have proposed a common framework for both PCA and KM, called Discriminative Embedded Clustering (DEC). However, these subspace clustering is generally limited to data processing for a specific distribution. When the data distribution changes, its accuracy often has some effect.

To solve the above problem, following the idea of adaptive learning in Nie et al. [17] and Wang et al. [18], this paper proposes a flexible dimension framework (named ASPCAKM). Besides, to adapt the data of different distributions, we also use adaptive paradigm to make the data similarity judgment in subspace clustering. At the same time, sparse constraints are added to the projection matrix to reduce data redundancy while preserving as much raw data structure information as possible. This improves the clustering accuracy.



(a) (b)

Figure 1. KM and RKM. (a) KM under the influence of outliers, (b) RKM under the influence of over-penalizing

2. Related work

This section mainly introduces some close related subspace clustering methods. We first familiarize some the notations.

2.1. Notations

Let X = $\{x_i \in R^d | i = 1,2...n\}$ be the input dataset with d features. Subspace clustering embeds data in feature space through the linear transformation matrix W= $\{w_i \in R^r | i = 1,2...d\}$. It obtains the prediction Y= $\{y_i \in R^r | i = 1,2...n\}$ after dimension reduction, and divides the data into class c groups. Tr(.) is the symbol of the trace operation. Table 1 introduces the other notations.

TABLE 1. Important notations

Notation	**Description**
d	Dimensionality of samples;
r	Dimensionality of embedded subspace;
c	Number of classes;
λ	Balance parameter;
σ	Adaptive parameter;
$t1_i$, $t2_i$	The i-th diagonal elements of T1, T2 respectively
T1, T2	diagonal matrix;
$x_i \in R^d$	The i-th sample;
$X = [x_1, x_2 x_3 ... x_4]^T$	Data matrix of the whole samples;
$B \in R^{dxr}$	Linear transformation matrix;
$G \in R^{dxr}$	Cluster centroid matrix;
$F \in R^{nxc}$	Cluster indicator matrix;
$A \in R^{dxr}$	Linear transformation matrix;

2.2. Robust K-means clustering

The purpose of KM clustering is to divide dataset X into c clusters, and for the distance between data, KM calculates the Euclidean distance (the L_2-norm of the data vector), representing the clustering center matrix of the data in F. Here, F_{ij} = 1 means that the i-th data belongs to cluster j. The target function for KM clustering can be expressed as:

$$\min_F \sum_{i=1}^{n} \sum_{k=1}^{c} F_{ik} \ ||x_i - \bar{x}_k||_2^2 \tag{1}$$

if the i-th data assigned to cluster k F_{ik} = 1, $\bar{x}_k$ is the k-th cluster center.

Despite the K-means algorithm is simple and efficient for clustering, Eq. (1) uses L_2-norm to calculate loss values, which are more sensitive to outliers because they assign a larger weight to distant points.

For the solution to this problem, Du et al. [1] proposed a more robust $L_{2,1}$-norm- based KM. Its loss function is as follows:

$$min_F \sum_{i=1}^{n} \sqrt{\sum_{k=1}^{c} F_{ik} \ ||x_i - \bar{x}_k||_2^2} \tag{2}$$

Such a robust loss function makes RKM reduce the impact of large outliers, but it is sensitive to small loss values and tends to overestimate the weight of data areas with higher density.

2.3. Sparse Principal Component Analysis

PCA is one of the most common used dimensional reduction algorithms by maximizing projection variance. Its loss function is as follows:

$$\min_{W} -tr(W^TXX^TW) \quad s.t. \quad W^T W = I \tag{3}$$

We obtain Eq. (4) by calculating derivative of Eq. (3) with respect to B

$$XX^TW = \lambda W \tag{4}$$

Then, we can choose the feature vector according to the feature value of the covariance matrix of X, and combine it into the projection matrix W.

However, PCA has one obvious disadvantage that each main component analysis is a linear combination of all variables. This makes it often difficult to interpret the main components of the data. In order to solve this problem, a new method is proposed to estimate the main component with sparse load, called sparse main component analysis (SPCA). SPCA is based on the fact that the PCA can be written as a regression optimization problem with secondary penalties, and lasso penalties (via elastic nets) can be integrated directly into the regression criteria to obtain sparsely loaded PCA. Its objective function is as follows:

$$(\hat{A}, \ \hat{B}) = \sum_{i=1}^{n} ||X_i - AB^TX_i||^2 \ + \alpha||B||_2^2 \tag{5}$$

where A ∈ R^{dxr} is an orthogonal matrix, B ∈ R^{dxr} is the projection matrix and the rank of both A and B are k, α is the regularization parameter;

2.4. Fast Adaptive KM Subspace Clustering

FAKM use an adaptive loss function for subspace clustering. Its objective function is,

$$\max_{F,G,W} Tr\ (W^TS_tW) - \lambda||X^TW - FG^T||_\sigma \tag{6}$$

where $W \in R^{dxr}$, $F \in R^{nxc}$ is the cluster indicator matrix, $G \in R^{rxc}$ is the cluster center matrix, λ is the balance parameter, σ is the adaptive parameter

It applies the adaptive norm defined in [17].

$$||A||_\sigma \ = \ \sum_i \frac{(1+\sigma)||a_i||_2^2}{||a_i||_2 + \sigma} \tag{7}$$

Where $\sigma > 0$ is an adaptive parameter.

By analyzing the Eq. (7), one can find that when $\sigma \rightarrow 0$, $||A||_\sigma$ tends to become the $L_{2,1}$-norm of A; when $\sigma \rightarrow \infty$, $||A||_\sigma$ is actually Frobenius norm.

Although FAKM uses adaptive strategy for clustering to cope with the distribution of different samples, it selects the PCA method with the largest variance for dimension reduction.

3. Proposed method

3.1. Adaptive discriminative clustering with sparse PCA

For compressing the projection matrix and further reducing redundancy information. we replace PCA with SPCA in subspace clustering. Combining with the SPCA in Eq. (4), we first proposed the subspace clustering with sparse constraints method.

$$\min_{B,F,G} ||X - XBA^T||_F^2 + \alpha||B||_2^2 + \lambda||XB - FG||_F^2 \tag{8}$$

In order to cope with data in different distributions, refer to the recent discriminant subspace clustering [18], we add the adaptive loss function in subspace clustering and get the objective function

$$\min_{B,F,G} ||X - XBA^T||_\sigma + \alpha||B||_2^2 + \lambda||XB - FG||_\sigma \tag{9}$$

where A ∈ $R^{dxr}$is an orthogonal matrix, B ∈ R^{dxr} is the projection matrix, α is the regularization parameter, λ is the balance parameter σ is the adaptive parameter.

3.2. Optimization

We proposed a simple iterative algorithm to solve the Eq. (8). When we update a variable, we keep the other's fixed. According to the theory in [17], Eq. (8) can be translated into the following optimization problem:

$$\min_{B,F,G\ t1,t2} \sum_{i=1}^{n} t1_i||x_i - AB^T x_i||_2^2 + \alpha||B||_2^2 + \lambda \sum_{i=1}^{n} t2_i||B^T x_i - G^T f_i||_2^2 \tag{10}$$

Where $t1_i = (1+\sigma)\frac{||x_i - AB^T x_i||_2 + 2\sigma}{2(||x_i - AB^T x_i||_2 + \sigma)^2}$ $t2_i = (1+\sigma)\frac{||B^T x_i - G^T f_i||_2 + 2\sigma}{2(||B^T x_i - G^T f_i||_2 + \sigma)^2}$

Note $U1 = X - XBA^T$, $U2 = XB - FG$; T1, T2 are diagonal matrix; diagonal element are $t1_i$ $t2_i$ respectively, We can obtain

$$\min_{B,F,G\ T1,T2} Tr(U1^T\ T1\ U1) + \lambda\ Tr(U2^T T2\ U2) + \alpha||B||_2^2 \tag{11}$$

Step 1: Solving F while fixing B, G, $T1$, and $T2$

When B, G, T1, and T2 fixed, the optimization problem in Eq. (11) become

$$\min_F \sum_{i=1}^{n} ||B^T x_i - G^T f_i||_2^2 = \min_F \sum_{i=1}^{n} \sum_{k=1}^{c} ||B^T x_i - g_k||_2^2\ F_{ik} \tag{12}$$

By assign data to its nearest cluster center Minimize Eq. (12)

$$F_{ij} = \begin{cases} 1 & j = argmin_k\ ||B^T x_i - g_k||_2^2 \\ 0 & otherwise \end{cases} \tag{13}$$

Step 2: Solving B while fixing G, F, $T1$, $T2$

We obtain Eq. (14) by calculating derivative of Eq. (11) with respect to B

$$B = (\alpha E + X^T T1 X + \lambda X^T T2 X)^{-1}(X^T T1 X A + \lambda X^T T2 F G) \tag{14}$$

Where E is the unit matrix,

Update the A matrix after getting B

$$X^T X B = U D V^T; A = U V^T \tag{15}$$

Step 3: Solving G while fixing B, F, $T1$, $T2$

We obtain Eq. (16) by calculating derivative of Eq. (11) with respect to G

$$G = (F^TF)^{-1} F^TXB \tag{16}$$

Step 4: Updating T1 and T2 by calculating its i-th element as:

$$t1_i = (1+\sigma)\,\frac{||x_i - AB^Tx_i||_2 + 2\sigma}{2(||x_i - AB^Tx_i||_2+\sigma)^2} \tag{17}$$

$$t2_i = (1+\sigma)\,\frac{||B^Tx_i - G^Tf_i||_2 + 2\sigma}{2(||B^Tx_i - G^Tf_i||_2+\sigma)^2} \tag{18}$$

Eq. (13) is sensitive to the initialization of F. Concretely, it first initializes F, which is used to solve W, G. However, the next time when you need to update, the initial cluster center is calculated by the previous F, finally lead to unstable results. Hence, when updating F, we initialize F several times, and choose to the one makes Eq (13) smallest.

we sum up the iteration process of optimization in **Algorithm 1.**

Algorithm 1 Algorithm to solve the problem Eq. (10).

```
Input
      The input data $X \in R^{nxd}$
      The number of clusters c, the reduced dimension number r, regularization
parameter λ,$\alpha$ and adaptive parameter σ
1 Initialize T1, T2  as an identity matrix,  randomly initialize B G
2 repeat
3         Update F by Eq. (13)
4         Update B by Eq. (14)
5         Update A by Eq. (15)
6         Update G by Eq. (16)
7         Update T1, T2 Eq. (17)(18)
8   until Convergence
Output: Selection matrix B, Cluster indicator matrix F, Cluster centroid matrix
G
```

4. Experiments

We analyzed five different public data sets to evaluate the algorithm performance. The details of these datasets are summarized in Table 2.

4.1. Experiment setup

We have conducted comparative experiment on five public datasets to evaluate the algorithm performance. Before executing the algorithm, we perform data normalize. These datasets include: one face image datasets (Yale), three UCI datasets (wine, Breast and lung), one text dataset (WebKB),

Detailed information of these datasets is summarized in Table 2. We compared ASPCAKM with FAKM and KM, briefly described as follows:

KM is the traditional K-means algorithm.

FAKM is a novel subspace clustering with adaptive loss function.

To make comparison fair, we record the best results of Comparison algorithm. Using Accuracy (ACC) to measure clustering performance. A higher value indicates better performance. [19] [20] have explanations of ACC. Table 3 show the results of clustering.

Table 2. A brief description of the datasets.

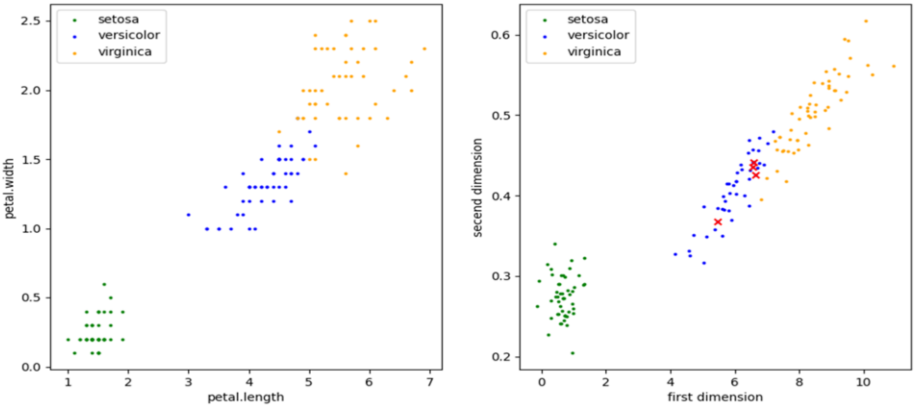
Datasets	Classes	# of instances	Dimensions	# of reduced dimensions
Breast	2	699	10	[3,4,5,6,7,8]
wine	3	178	13	[3,4,5,6,7,8,9,10]
Lung	3	32	56	[5,10,15,20,25,30,35,40,45,50]
Yale	15	165	1024	[100,200…900,1000]
WebKB	7	814	4029	[100,200 …2000,2100]

Table 3. Clustering results of compared algorithms (ACC%).

	Breast	Wine	Lung	Yale	WebKB
KM	60.09	70.22	53.88	40.3	51.61
FAKM	94.59	87.02	55.56	42.34	67.09
ASPCAKM	**95.57**	**95.34**	**59.75**	**50**	**69.01**

4.2. visualization

We perform clustering and visualization on iris dataset. The dataset includes versicolor, virginica and setosa three clusters, and has 4 dimensions. The original data with the ground truth label is shown in Fig. 2(a), where two characteristics are selected. We used ASPCAKM to subspace cluster the original 4-D data and reduced dimension to 2. ASPCAKM results are shown in Fig. 2(b). Misclassified data are marked with forks.



(a)

(b)

Figure 2. (a) Original data with ground truth labels. (b) Clustering result of ASPCA

4.3. Influence of dimension reduction

The purpose of next experiment is testing the effect of dimension reduction applied to ASPCAKM

We set the projection matrix B in Eq. (9) as a d-dimensional identity matrix, and generate Eq. (19)

$$\min_{F,G} \left|\left|X - FG\right|\right|_{\sigma} \tag{19}$$

Obviously, if we never perform dimensionality reduction, ASPCAKM become KM with an adaptive loss function, name it ASPCAKM-ND. Fig. 3 shows a comparison of two algorithm ACC on several data sets. The results show that, because of adaptive loss function, ASPCAKM-ND obtains higher accuracy than KM. ASPCAKM gets the highest accuracy. Therefore, we argue that it is beneficial for clustering to perform feature learning and adaptive clustering.

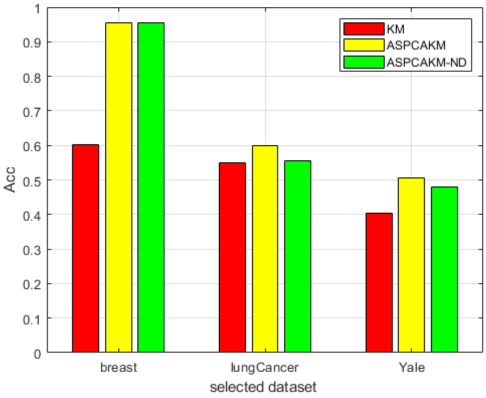


Figure 3. Influence of dimension reduction in ASPCAKM

5. Conclusion

In this paper, we proposed a sparse constraints subspace clustering with adaptive loss function to deal with the challenges of existing subspace clustering algorithms with K-means. We jointly integrated cluster and Sparse PCA into a framework. Additionally, to ease the affected by redundant features and outliers, we apply an adaptive loss function to calculate the distance data elastically. Based on this theory, we experimented on several data sets to demonstrate the advantages of the ASPCAKM algorithm. Indeed, you can set the dimension reduced to d-dimension and set the appropriate sparse constraint α to get the sparse projection matrix $B \in R^{dxd}$, and because the characteristics of the sparse PCA redundant features will be compressed to 0, which is similar to the best results of the algorithm. In the future we will extend it into a 2-D clustering framework.

Acknowledgement

This paper was supported by National Natural Science Foundation of China (Grant Nos.61871464), National Natural Science Foundation of Fujian Province (Grant Nos.2020J01266, 2021J011186), the "Climbing" Program of XMUT (Grant No. XPDKT20031), Scientific Research Fund of Fujian Provincial Education Department (Grant No. JAT200486), Program of XMUT for high-Level talents introduction plan (Grant No. YKJ19003R).

References

[1] Du L, Zhou P, Shi L, et al. *Robust multiple kernel K-means using 2;1 -norm*. AAAI Press, 2015.

[2] Wang X, He Y, Wang L, et al. An Unsupervised Feature Selection Method Based on Information Entropy // 2018 3rd International Conference on System Reliability and Safety (ICSRS). IEEE, 2018.

[3] Yan C, Kang X, Li M, et al. *A Novel Feature Selection Method on Mutual Information and Improved Gravitational Search Algorithm for High Dimensional Biomedical Data* // 2021 13th International Conference on Computer and Automation Engineering (ICCAE). 2021.

[4] Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. 2019, 13(2):1016-1042.

[5] Chang X, Nie F, Yang Y, et al. A Convex Sparse PCA for Feature Analysis. Computer Science, 2014.

[6] EJ Cande, Xiaodong Li Robust Principal Component Analysis. 2009.

[7] Lu H, Plataniotis K N, Venetsanopoulos A. *Multilinear Principal Component Analysis of Tensor Objects for Recognition* // International Conference on Pattern Recognition. IEEE, 2006.

[8] Zhang J, Zhao D, Gao, Group-based Sparse Representation for Image Restoration. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, **23** (2014), 8, 3336.

[9] Hou C, Nie F, Jiao Y, et al. *Learning a subspace for clustering via pattern shrinking*. Information Processing & Management, **49** (2013), 4,871-883.

[10] Yin X, Chen S, Hu E. *Regularized soft K-means for discriminant analysis*. Neurocomputing, **103** (2013), 29-42.

[11] Wang D, Nie F, Huang H. *Unsupervised Feature Selection via Unified Trace Ratio Formulation and K-means Clustering (TRACK)* // ECML/PKDD. 2014.

[12] Ding C, Berkeley L, Li T. *Adaptive dimension reduction using discriminant analysis and K-means clustering*. ICML '07: Proceedings of the 24th international conference on Machine learning ACM, 2007.

[13] Chenping, Hou, Fe iping, et al. *Discriminative embedded clustering: a framework for grouping high-dimensional data*. IEEE transactions on neural networks and learning systems, **26** (2015), 6, 1287-1299.

[14] Cnl A, Yhs A, Wjc B, et al. Generalized two-dimensional linear discriminant analysis with regularization. Neural Networks, **142** (2021), 73-91.

[15] Li, Yao L, Wang S, et al. *Adaptive Two-Dimensional Embedded Image Clustering*. Proceedings of the AAAI Conference on Artificial Intelligence, **34** (2020), 4, 4796-4803.

[16] Cnla B, Yhs A, Zhen W C, et al. *Robust bilateral Lp-norm two-dimensional linear discriminant analysis - ScienceDirect*. Information Sciences, **500** (2019), 274-297.

[17] Nie F, Wang H, Huang H, et al. *Adaptive loss minimization for semi-supervised elastic embedding* // Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013.

[18] Wang X D, Chen R C, F Yan, et al. *Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data*. IEEE Access, 2019, 42639-42651.

[19] Deng C, He X, Han J. *Document Clustering Using Locality Preserving Indexing*. IEEE Transactions on Knowledge and Data Engineering, **17** (2005), 12,1624-1637.

[20] Strehl, Alexander, Ghosh, et al. *Cluster Ensembles -- A Knowledge Reuse Framework for Combining Multiple Partitions*. Journal of Machine Learning Research, **3** (2002), 583-617.