SUBSPACE CLUSTERING VIA INDEPENDENT SUBSPACE ANALYSIS NETWORK

Chunchen Su[1,2], *Zongze Wu*[2], *Ming Yin*[2*], *KaiXin Li*[1], *Weijun Sun*[2]

[1] South China University of Technology, Guangzhou 510641, China
[2] School of Automation,Guangdong University of Technology, Guangzhou 510006, China
763140581@qq.com; zzwu@scut.edu.cn; yiming@gdut.edu.cn; 494113821@qq.com; 14341569@qq.com

ABSTRACT

Previous work on image clustering focused on seeking a low-dimensional structure from the high-dimensional image data by a shallow linear model, such as sparse subspace clustering (SSC) or low-rank representation (LRR). The recent advance of deep learning shows its superiority via handling data with nonlinear structure, i.e., sparse auto-encoder and independent subspace analysis(ISA), etc. However, most of this type of methods may ignore lots of useful information embedded in the original data. To this end, we propose a novel unsupervised learning algorithm via ISA incorporating the subspace structure within data. Specifically, we adopt the ISA to learn local translation invariant feature from data and integrate a *prior* subspace information into the output of the network simultaneously. This method performs an impressive powerful ability to learn the nature of data. By evaluating on public databases, CMU-PIE and ORL, the experimental results show that the proposed approach achieves better clustering results compared with the state-of-the-art ones.

Index Terms— Subspace clustering,Independent subspace analysis, Prior, Sparse representation

1. INTRODUCTION

In recent years, deep networks [3, 7], have been attracting more and more attentions from the communities of machine learning and computer vision, which have achieved considerable superior performance in face recognition [15], image understanding and natural language processing. Due to their powerful representation learning, many derivative algorithms on deep learning have been successfully developed for the practical tasks [12].

Subspace clustering aims at grouping the data into their intrinsic subspaces by uncovering their low-dimensional structures embedded in high-dimensional space [19][5]. In this context, image clustering is an important branch of subspace clustering, which tried to identify the groups of similar image primitives [17]. Roughly, subspace clustering methods can be grouped into four types, i.e., algebraic methods, iterative methods, statistical methods and spectral clustering-based methods [13]. Among them, spectral clustering-based approaches have been demonstrated to perform very well for some applications in the patter recognition. Actually, the key issue of this type of methods is to seek the similarity among data points, which is often measured in the raw space of data. In particular, this similarity is recently computed by the sparse or low-rank representations of data points [8], by exploiting the so-called *self-expressive* property of the data. In other words, these methods are regarded as a shallow linear model, which have an impressive ability of capturing linear structure of data. Unfortunately, they may fail in handling data with nonlinear structure [18]. While for deep learning, it learns features directly from data and consequently is more generalizable. More importantly, deep learning can handle data with significant non-linearity well [3]. However, deep learning mainly focuses on learning the nonlinear transformations ignoring the subspace structure within data [11].

To overcome this drawback, in this paper, we propose a nonlinear unsupervised learning method by using the Independent Subspace Analysis(ISA) [1] network and incorporating the subspace information of data simultaneously. In particular, we integrate the *prior* subspace information to enforce that the output of ISA network has the same subspace structure with the original data. Compared to the conventional unsupervised learning [2], our method discards the reconstruction procedure during learning the nonlinear structure within data. As such, a better low-dimensional representation will be learned and the clustering result can be improved.

The organization of this paper is as follows. In Section 2, the related works on image clustering are briefly reviewed. Subsequently, we elaborate our algorithm on how to integrate prior subspace information into the ISA network in Section 3. The experimental results are reported to show the efficacy of our proposed method in Section 4. Finally, Section 5 draws a conclusion of this paper.

2. RELATED WORKS

In this section, we briefly review some existing works on image clustering. **Subspace Clustering** : Recently, lots of subspace clustering algorithms [9, 2] have been developed, of

*Corresponding author. The Project was supported by NSFC (No.61673126),in part by the Guangdong Natural Science Foundation under Grant (No.2014A030313511).

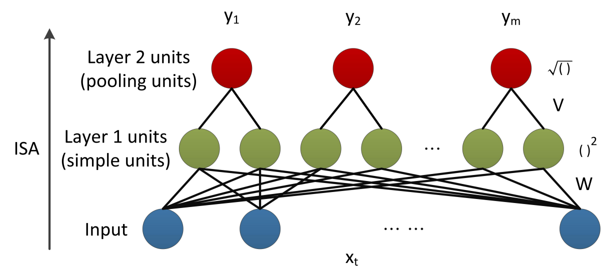


Fig. 1. The neural network architecture of an ISA network. [6].

which the major concern is the way to learn an affinity matrix [16]. These methods impose different constraints on the arrangement of subspaces and the distribution of data, and succeed in recovering the desired low-dimensional structure, i.e., the similarity within data. However, these approaches mostly belong to the linear models, and are not be suitable for the data with nonlinear structure. Thus, they may fail to perform the clustering tasks well when used in the real scenario. To address this issue, some kernel methods have been proposed such as kernel SSC [10] and kernel LRR [14]. However, it is not easy to choose a suitable kernel function, which depends on the experience in most cases.

Deep Learning : Deep learning has achieved promising success in many areas, especially in the facial recognition, and demonstrated the powerful nonlinear representation ability [4]. Nevertheless, there are still some open problems on applying deep learning to clustering task. In work [12], the authors adopted the auto-encoder network to clustering. Specifically, Tian *et al.* [12] proposed a novel graph clustering approach in the sparse auto-encoder framework. Furthermore, Peng *et al.* [11] presented a deeP subspAce clusteRing with sparsiTY prior, termed as PARTY, by combining the deep neural network and sparsity information of original data to perform subspace clustering. This framework achieved a satisfactory performance while extracting low-dimensional feature in the unsupervised learning.

ISA : ISA is usually regarded as an extension of Independent Component Analysis(ICA), which can be depicted as a two-layer network (illustrated in Fig.1), with square and square-root active functions in the first and second layer respectively. In Fig.1, the first layer connection is weighted by W learned from data, and the second layer's weight is denoted by V that is fixed. Moreover, each of the hidden units in the second layer connects a small number of neighbor units from the first layer. Based on this understanding, the units in the first and second layer are named as simple and pooling units respectively.

Given $\mathbf{x}_t$ as the input of the network, the output is $y_l(\mathbf{x}_t; W, V) = \sqrt{\sum_{j=1}^{k} V_{lj} \left(\sum_{i=1}^{n} W_{ji} x_i\right)^2}$, and x_i denotes the element at position i of the input vector $\mathbf{x}_t$. ISA learns the network parameters W through finding sparse feature representations in the second layer, by solving an optimization problem as follows.

$$\min_{W} \sum_{t=1}^{N} \sum_{l=1}^{m} y_l(\mathrm{x}_t; W, V), \text{s.t. } WW^T = \mathbf{I}. \tag{1}$$

where $\{\mathbf{x}_t\}_{t=1}^{N}$ are input data. Here, $W \in \mathbb{R}^{k\times n}$ and $V \in \mathbb{R}^{m\times k}$ denote the weights connecting in the first and second layer of ISA respectively. n, k, m are the input dimension, number of simple units and pooling units respectively. $\mathbf{I}$ is the identity matrix with suitable dimension.

Although ISA has an advantage that it learns features that are robust to the local translation, it may not be sufficient to represent the data feature so as to not have enough superiority to be super enough in subspace learning.

3. SUBSPACE CLUSTERING VIA ISA NETWORK

In this section, we will elaborate on our method for image clustering. Firstly, we attain the *prior* sparsity subspace representation of data using SSC algorithm, and then learn the subspace feature through ISA network. Finally, the low-dimensional feature is utilized to cluster the data into multiple classes.

3.1. ISA with Subspace Prior

Let $\mathbf{z}_t^{(i)}$ be the weighted input to the neurons in i-th layer corresponding to the t-th sample, and $f(x) = x^2$ and $g(x) = \sqrt{x}$ as the active function of the first and second layer of ISA network respectively. That is, the computation of $\mathbf{z}_t^{(i)}$ in each layer is given by,

$$\mathbf{z}_t^{(1)} = W\mathbf{x}_t,\ \mathbf{z}_t^{(2)} = Vf(\mathbf{z}_t^{(1)}). \tag{2}$$

Given $\mathbf{y}_t = [y_1, y_2, \cdots, y_m]^T \in \mathbb{R}^m$ the output of the second layer and

$$\mathbf{y}_t = g(\mathbf{z}_t^{(2)}) \in \mathbb{R}^m. \tag{3}$$

where $t = 1, 2, \cdots, N$ indexs the sample and m denotes the dimension of the output at the second layer of ISA. For a collection of N given samples $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}^{n\times N}$, the corresponding outputs of the network is $Y = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]$.

In order to preserve the *prior* sparsity subspace information within data, we propose to integrate the subspace information into the ISA. Mathematically, the objective of our model is to minimize the following problem.

$$\min_{W} \sum_{t=1}^{N} \sum_{l=1}^{m} y_l(\mathrm{x}_t; W, V) + \frac{\lambda}{2}\|Y - YC\|_F^2, \text{s.t. } WW^T = \mathbf{I}. \tag{4}$$

where C denotes the global subspace prior, and $\|\cdot\|_F$ is the Frobenius norm defined as $\|X\|_F^2 = \sum_i \sum_j |X_{ij}|^2$. λ is a tradeoff parameter.

In this objective function, the first term aims to learn invariant feature via minimizing the sum of all the network output. As for the second one, it is designed to preserve the affinity between the original data samples that is invariant to different feature spaces. This point is also motivated by the well-known manifold assumption. As well known, the orthogonal constraint $WW^T = \mathbf{I}$ aims to avoid the trivial solution, and then W can be set by computing $(WW^T)^{-\frac{1}{2}}W$. Once when $\lambda = 0$, our model will degrade into a ISA network. To some extent, we can see the proposed model is more general than ISA.

However, how to define the C is not a trivial issue. In this paper, we adopt the SSC to learn the subspace structure of data as the superiority of SSC in subspace learning, and incorporate into the ISA network. That is, C can be learned by solving several sub-problems as follows.

$$\min_{\mathbf{c}_i} \ \|\mathbf{x}_i - X\mathbf{c}_i\|_2^2 + \alpha\|\mathbf{c}_i\|_1, \text{ s.t. } c_{ii} = 0. \tag{5}$$

where $C = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_N]$ and $\|\cdot\|_1$ denote ℓ_1-norm that is usually used to achieve sparsity, c_{ii} denotes the i-th entry of the column vector $\mathbf{c}_i$,which is utilized to prevent degenerate solution of C, as a result the proposed model can hold the potential affinity among the data.

In our method, by making full use of ISA and the prior subspace information, the feature learned from data can capture the intrinsic structure well so as to be more powerful for the next clustering task.

3.2. Optimization

To optimize the proposed model, gradient descend method is usually adopted. For the sake of simplicity, we define $\boldsymbol{\varphi}_t = g(\mathbf{y}_t)$, and the $\Phi = [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_N] \in \mathbb{R}^{n\times N}$. Then, we can rewrite (4) as the following equivalent problem.

$$\min_W \mathcal{J} = \|\Phi\|_F^2 + \frac{\lambda}{2}\|Y - YC\|_F^2, \text{ s.t. } WW^T = \mathbf{I}. \tag{6}$$

That is,

$$\min_W \mathcal{J} = \sum_{t=1}^{N}\left(\|\boldsymbol{\varphi}_t\|_2^2 + \frac{\lambda}{2}\|\mathbf{y}_t - Y\mathbf{c}_t\|_2^2\right), \text{ s.t. } WW^T = \mathbf{I}. \tag{7}$$

According to the definition of $\mathbf{y}_t$ in (3). We can compute the gradient of (7) *w.r.t.* W as follows.

$$\begin{aligned}\nabla\mathcal{J}_W &= V^T\left\{[\boldsymbol{\varphi}_t \odot g'(\mathbf{y}_t) + \lambda(\mathbf{y}_t - Y\mathbf{c}_t)] \odot g'(\mathbf{z}_t^{(2)})\right\} \\ &\odot f'(\mathbf{z}_t^{(1)})(\mathbf{x}_t)^T \\ &= V^T\left[\frac{1}{2}g'(\mathbf{z}_t^{(2)}) + \lambda(\mathbf{y}_t - Y\mathbf{c}_t) \odot g'(\mathbf{z}_t^{(2)})\right] \\ &\odot f'(\mathbf{z}_t^{(1)})(\mathbf{x}_t)^T \end{aligned} \tag{8}$$

where $\odot$ denotes element-wise multiplication. Here $f'(\cdot)$ and $g'(\cdot)$ are the derivative of the activation $f(\cdot)$ and $g(\cdot)$ respectively.

Once obtaining the gradient, the weight W will be updated by,

$$W = W - \mu\nabla\mathcal{J}_W. \tag{9}$$

where $\mu > 0$ is the learning rate which is typically set to a small value such as 10^{-4} in our experiments. Then adding the orthogonal constraint until convergence. The weight V is only initialized and not updated in each iteration.

Algorithm 1 briefly describes the detailed procedure for optimizing our model.

Algorithm1 Independent Subspace Analysis with Sparsity Prior

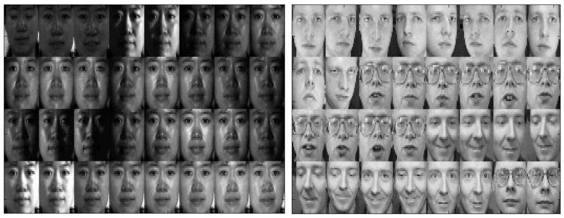
Input: A data X, and the tradeoff parameter λ.
Initializing W and V.
Compute the sparsity prior C over X via solving Eq.(5).
Do forward propagation to compute Y via Eqs.(2)- (3) .
while not converge **do**
  for $t = 1, 2, \cdots, N$ **do**
    Sequentially select a data point $\mathbf{x}_t$ as the input of nework,
    Computer $\mathbf{y}_t$ via Eq. (3),
    Caculate the gradient via Eq. (8),
    Update W using Eq. (9).
  end
end
Obtain the data segmentation by clustering based on Y.
Output: W and the clustering result.

4. EXPERIMENTAL RESULT

In this section, we conduct several experiments to evaluate the effectiveness of Independent Subspace Analysis with Sparsity Prior, termed as ISASP, and compare our algorithm against the state-of-the-art clustering algorithms. Specifically, we compare ISASP with K-means, SSC [2], and ISA [6]. Moreover, we investigate the performance of our approach with the post-processing using K-means and SSC respectively. Therefore, these two algorithms are called by ISASP-k and ISASPs respectively. Similarly, the ISA with the K-means (termed as ISAk) and post-processed by SSC (termed as ISAs) are also studied. In our experiments, we use the *theano*as the deep learning framework so that the computationally complexity can be reduced effectively.

4.1. Datasets

Two famous benchmark face datasets are utilized to evaluate our algorithm, CMU-PIE and ORL. The subset of CMU-PIE, termed as PIE_pose27, is used in our experiments, which contains 2,856 samples distributed over 68 volunteers. Each image of PIE_pose27 is with size of 32×32. The ORL consists of 400 samples from 40 individuals, where each image is with size of 92 $\times$ 112. We reshape the ORL images to 32 $\times$ 32. Some sample images of datasets are shown in Fig. 2.



(a) (b)

Fig. 2. Samples on the PIE_pose27 (a) and ORL (b) database.

4.2. Parameter settings

To effectively evaluate our algorithm, clustering accuracy and the normalized mutual information (NMI) are selected as the literature. Usually, we usually flatten an image into a vector as the input of the first layer of ISA. Thus, 1024 features will be input into ISA network (i.e., there are 1024 red nodes in Fig.1). In our experiments, we utilize 400 neurons in the first layer and the entries of the weight matrix V only can be defined by 0 or 1. Specifically, the value in a adjacent position is 1 and other position is 0 for each dimension of the matrix V, so that each of the hidden units in the second layer connects two neighbor units from the first layer. Thus, the second layer of the ISA contains only 200 neurons. To attain the best result, we experimentally choose λ in the experiments, similar to other compared methods.

4.3. Results

The performance of our method on the PIE_pose27 and ORL are reported in Table 1 and Table 2 respectively. For fair comparison, we report the best result of all the evaluated methods, which are achieved by their optimal parameters. It can be observed that the results in Tables 1 and 2 show the superiority of our method. For the PIE_pose27, the gains of ISASPs are 4.44% and 1.61% against SSC in terms of Accuracy and NMI respectively. Similarly, for the ORL, ISASPs also achieved the best results, of which the Accuracy is approximately 1.07% higher than the second best method. In addition, for the two datasets, the result of ISASP is better than that of ISA, which shows that the introduction of sparsity prior in our algorithm can obtain better performance. In most cases, ISASPs is better than ISASPk, which may owe to that the segmentation of the data using spectral clustering is more discriminative than the original space [11].

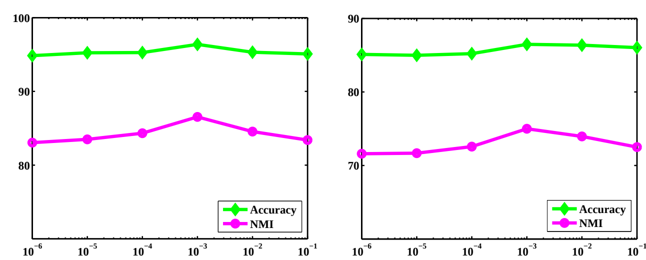
Next, we tested the effect of parameter λ in ISASP. Fig. 3 presents the clustering performance versus the varying of parameter λ on the PIE_pose27 and ORL, respectively. From the figure 3, it can be seen that the clustering scores increase as λ becomes larger, reaching peak value at about 10^{-3} and decreasing afterwards. This helps to determine the value of λ in our experiments.

Table 1. Clustering results in terms of Accuracy (%) and NMI (%) on PIE_pose27 dataset(mean $\pm$ standard deviation).

Algorithm	Accuracy	NMI
K-means	18.33 ± 0.85	40.62 ± 0.79
SSC	82.10 ± 2.30	94.77 ± 0.61
ISAk	58.26 ± 2.78	74.43 ± 1.37
ISAs	84.72 ± 1.69	95.74 ± 0.60
ISASPk	59.68 ± 2.85	75.07 ± 0.89
ISASPs	$\mathbf{86.54 \pm 2.92}$	$\mathbf{96.38 \pm 0.77}$

Table 2. Clustering results in terms of Accuracy (%) and NMI (%) on ORL dataset(mean $\pm$ standard deviation).

Algorithm	Accuracy	NMI
K-means	58.25 ± 3.56	78.84 ± 1.69
SSC	73.93 ± 2.03	$\mathbf{88.09 \pm 0.61}$
ISAk	48.85 ± 2.39	68.86 ± 2.14
ISAs	72.53 ± 1.50	84.66 ± 0.45
ISASPk	52.43 ± 2.24	71.46 ± 1.58
ISASPs	$\mathbf{75.00 \pm 2.01}$	86.48 ± 0.67



(a) (b)

Fig. 3. Accuracy and NMI (%) (y-axis) of ISASP with different λ (x-axis) on PIE_pose27(a) and ORL(b) dataset.

5. CONCLUSION

In this paper, we presented a novel approach that learns features from original data using ISA network incorporated the sparsity subspace *prior*. By this, the segmentation of the data can be effectively performed. The experimental results, on two real world datasets, show that our method remarkably outperforms the state-of-the-art methods.

6. REFERENCES

[1] Pierre Comon. Supervised classification: a probabilistic approach. In *ESANN95-European Symposium on Artificial Neural Networks*, pages 111–128. University Press, 1995.

[2] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[3] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPSs*, pages 1097–1105, 2012.

[5] Zhihui Lai, Yong Xu, Qingcai Chen, Jian Yang, and David Zhang. Multilinear sparse principal component analysis. *IEEE transactions on neural networks and learning systems*, 25(10):1942–1950, 2014.

[6] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of CVPR*, pages 3361–3368. IEEE, 2011.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[8] Kun Li, Jingyu Yang, and Jianmin Jiang. Nonrigid structure from motion via sparse representation. *IEEE transactions on cybernetics*, 45(8):1401–1413, 2015.

[9] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machince Intelligence*, 35(1):171 – 184, Jan. 2013.

[10] Vishal M Patel and René Vidal. Kernel sparse subspace clustering. In *Proceedings of ICIP*, pages 2849–2853. IEEE, 2014.

[11] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Proceedings of IJCAI*, 2016.

[12] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. Learning deep representations for graph clustering. In *Proceedings of AAAI*, pages 1293–1299, 2014.

[13] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[14] Shijie Xiao, Mingkui Tan, Dong Xu, and Zhao Yang Dong. Robust kernel low-rank representation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2268–2281, 2016.

[15] Meng Yang, Pengfei Zhu, Feng Liu, and Linlin Shen. Joint representation and pattern learning for robust face recognition. *Neurocomputing*, 168:70–80, 2015.

[16] Ming Yin, Junbin Gao, and Zhouchen Lin. Laplacian regularized low-rank representation and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):504–517, 2016.

[17] Ming Yin, Junbin Gao, Zhouchen Lin, Qinfeng Shi, and Yi Guo. Dual graph regularized latent low-rank representation for subspace clustering. *IEEE Transactions on Image Processing*, 24(12):4918–4933, 2015.

[18] Ming Yin, Yi Guo, Junbin Gao, Zhaoshui He, and Shengli Xie. Kernel sparse subspace clustering on symmetric positive definite manifolds. In *Proceedings of CVPR*, pages 5157–5164, 2016.

[19] Zexuan Zhu, Sen Jia, Shan He, Yiwen Sun, Zhen Ji, and Linlin Shen. Three-dimensional gabor feature extraction for hyperspectral imagery classification using a memetic framework. *Information Sciences*, 298:274–287, 2015.