A STUDY OF UNSUPERVISED CLASSIFICATION TECHNIQUES FOR HYPERSPECTRAL DATASETS

Himanshi Yadav, Alberto Candela and David Wettergreen

The Robotics Institute, School of Computer Science, Carnegie Mellon University

ABSTRACT

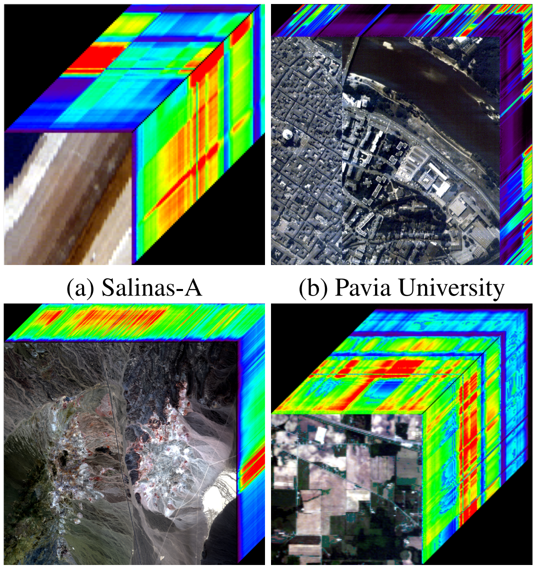
This work extensively studies and analyses several unsupervised clustering methods for hyperspectral data. We look at unsupervised classification solutions that accomplish adaptive cluster formation in anticipation for new data discoveries. We provide qualitative and quantitative answers to significant problems like high-dimensionality of hyperspectral datasets, multiple sources and relative amounts of existing noise in data and low class separability. The effectiveness of various clustering techniques is illustrated on diverse hyperspectral datasets by intensive experimentation, comparison between techniques and analysis.

Index Terms— Unsupervised, classification, hyperspectral, diffusion, learning

1. INTRODUCTION

Most traditional classification methods used by the imaging spectroscopy community compare features present in each pixel with spectral signatures of materials from a known or likely to be present exhaustive library of all materials discovered so far [1]. This poses a problem for scenarios like planetary exploration where specific materials have not been conclusively discovered and their proportions are yet to be determined. Unsupervised classification is therefore an obvious solution where test pixels are assigned to clusters without using expert labels. Clustering enables the discovery of new materials and their signatures in unexplored environments. Our paper extensively analyses such methods for available low-resolution and high-resolution hyperspectral datasets.

Classification of hyperspectral images is a challenging task due to the large size of each of these images. Each pixel in a hyperspectral image is a high-dimensional vector and any dataset will at least have several thousand pixels. With the advancement in imaging spectroscopy and measurement devices, we now have hyperspectral images with a higher spectral and spatial resolution [2]. These high-resolution high-dimensional hyperspectral images necessitate computationally faster and memory efficient algorithms. There is also a lack of generalizable algorithms that deal with the inherent noise and non-linearities present in hyperspectral datasets.



(a) Salinas-A (b) Pavia University

(c) Cuprite (d) Indian Pines

Fig. 1. Visualization of a high-dimensional hyperspectral image in the form of a hyperspectral cube for the (a) Salinas-A, (b) Pavia University, (c) Cuprite and (d) Indian Pines datasets.

Clustering techniques focus on grouping pixels by iteratively updating the centroid of a group, where the centroid represents the mean spectral signature of all the pixels in a group. These techniques perform equally well if the assumption that most of the pixels are governed by a single endmember holds true i.e. are pure pixels [3]. Such a scenario arises often in high-resolution hyperspectral images.

There are several machine learning methods for clustering [4]. Many of these methods have been applied on hyperspectral images with varying success [3, 5]. Some approaches reduce high-dimensional hyperspectral images to a low-dimensional coordinate space using methods like Principal Component Analysis (PCA), and Independent Component Analysis (ICA) before clustering [6, 7]. Deep learning methods have also been employed. Supervised deep classification

This research was supported by the National Science Foundation National Robotics Initiative Grant #IIS-1526667. We acknowledge support of the NASA Earth Science Division for the use of AVIRIS-NG data. We thank Dr. David R Thompson for his assistance with the AVIRIS-NG data. We also thank Dr. James M Murphy for his suggestions and expert advice on diffusion learning and clustering algorithms.

978-1-5386-9154-0/19/$31.00 ©2019 IEEE

of hyperspectral images [8, 9] produces almost 100 percent accurate results for most datasets. Several semi-supervised deep learning classification algorithms use auto-encoders [10, 11] to achieve high performances.

This paper addresses the need for automated analysis of spectral data and provides analysis of unsupervised solutions for clustering hyperspectral datasets into constituent classes.

2. METHODS

All methods considered in this paper are provided with an input image $I \in R^{h \times w \times d}$, where h and w are the height and width of the spatial dimensions, respectively, and d is the number of spectral bands. Let k be the number of classes present in I for every dataset considered.

In this work, we evaluate several clustering techniques on multiple datasets. We look at k-means clustering and multiple variants of k-means clustering. All methods which are based on k-means clustering should fail if the k-means assumptions are not valid. Most noticeable failures occur when the variance of the distribution of each variable is not spherical. Methods also do not fair well due to the inherent noise and curse of dimensionality. The techniques analyzed in this work which are direct variants of k-means mainly employ a two-stage process where first, the data dimensionality is reduced followed by k-means clustering in the low-dimensional feature space. PCA followed by k-means performs linear dimensionality reduction on the input image whereas auto-encoder followed by k-means clustering and Deep Embedded Clustering (DEC) [12] perform non-linear dimensionality reduction.

Auto-encoder is a deep learning network that learns a low-dimensional representation or encoding of a given image data by trying to reconstruct it in the output stage [13] and is a useful deep dimensionality reduction tool. Spectral Clustering (SC) [14] uses spectral techniques to learn a low dimensional representation of the image data. It does so by constructing similarities between individual data points and formalising them as a weight matrix. Then, eigen values of the weight matrix are used for dimensionality reduction, followed by clustering in thus obtained low-dimensional space. However, spectral clustering is known to be sensitive to irrelavant data dimensions and noise. DEC [12] defines a clustering loss which updates parameters of the deep dimensionality reduction network and the cluster centers simultaneously. This deep method outperforms traditional clustering methods for high-dimensional RGB datasets but shows limited performance when tested on high-dimensional hyperspectral data.

The state-of-the-art clustering methods considered are sparse manifold clustering and embedding (SMCE) [15], non-negative matrix factorization (HNMF) [3], fast search and find of density peaks clustering (FSFDPC) [16] algorithm, nonlocal total variation (NLTV) [17] , diffusion learning (DL) and spatial-spectral diffusion learning (DLSS) [18] algorithms. SMCE works somewhat like SC but it computes the neighborhood graph and the weights for the weight matrix simultaneously by solving a sparse optimization problem. FSFDPC makes the assumption that centroids of clusters have higher density than data points in the cluster and they are also quite distant from other high density centroids. This assumption is used again in the DL and DLSS algorithms. DL and DLSS algorithms use this assumption in a low dimensional embedding of the input data to then successfully find centroids or modes of data clusters whereas FSFDPC does so in the high dimensional space.

3. EXPERIMENTS AND RESULTS

3.1. Hyperspectral Datasets

The clustering methods were tested on the following datasets:

- Salinas-A dataset: The Salinas-A dataset was acquired by the AVIRIS sensor 204 bands and has 86 samples and 83 lines with 6 unique classes.

- Pavia University dataset: The Pavia University dataset was collected by the Reflective Optics System Imaging Spectrometer (ROSIS-3) sensor with 9 material classes. This dataset contains 115 bands and 103 clean bands (used for clustering) with a spatial resolution of 1.3 m. The dataset size is 1096 samples by 715 lines. In this work, we use a subset of the Pavia University dataset with 6 unique classes, as in [18], for better comparative analysis.

- Cuprite dataset: The Cuprite dataset was collected by AVIRIS-Next Generation (NG) sensor and is the most diverse geologic dataset with more than 200 mineral classes. It has more than 200 bands and we used 97 bands for clustering. We employ a subset of the Cuprite dataset for our experiments with the 10 most dominant classes.

- Indian Pines dataset: The Indian Pines dataset was acquired by AVIRIS sensor with 200 clean spectral bands and 16 classes. The dataset has 145 samples by 145 lines.

The Pavia University and the Salinas-A datasets have very distinct spectral classes that are spread out homogeneously in the hyperspectral image. These datasets also have comparatively lesser number of classes. On the other hand, the Indian Pines and the Cuprite datasets have more number of classes (although only 10 dominant classes of the Cuprite dataset were used for experimentation). There is considerable amount of overlap between classes in these datasets. It is immensely more difficult to cluster the latter two datasets than the former which can clearly be seen in Table 1. Figure 2 depicts the differences between the former and the latter type

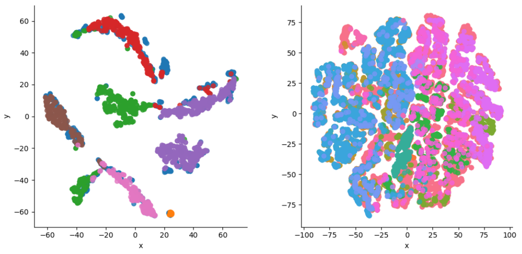
Table 1. Comparison of overall clustering accuracy in percentages for each algorithm implemented on different hyperspectral datasets

Datasets	Number of Classes	Clustering Accuracy (OA) (in percentages)								
		k-means	PCA + k-means	Auto + k-means	DEC	SMCE	HNMF	FSFDPC	DL	DLSS
Salinas-A	6	62.5	62.5	60.67	40.6	46.62	63.20	63.22	83.13	84.76
Pavia	6	77.6	77.55	70.36	63.54	83.52	72.17	77.83	84.9	93.6
Cuprite	10	24.42	24.39	24.00	23.55	20.98	40.65	25.48	29.30	29.12
Indian Pines	16	39.6	39.42	42.24	46.74	33.89	36.36	39.16	35.78	41.82

Table 2. Comparison of run time in seconds for each algorithm implemented on different hyperspectral datasets

Datasets	Time (in seconds)								
	k-means	PCA + k-means	Auto + k-means	DEC	SMCE	HNMF	FSFDPC	DL	DLSS
Salinas-A	0.69	0.01	16.37	31.34	180.86	0.45	3.42	4.44	6.11
Pavia	2.71	0.01	50.01	946.15	313.60	0.53	10.74	14.76	30.69
Cuprite	0.99	0.35	22.02	19.00	93.44	0.48	1.56	1.85	4.93
Indian Pines	27.59	0.01	15.00	6.06	270.56	1.29	28.79	49.84	41.82

of datasets where the former has fewer number of classes and lower amounts of classes overlap.



(a) Salinas-A dataset (b) Indian Pines dataset

Fig. 2. The t-SNE representation of the Salinas-A dataset and the Indian Pines datasets along with the ground truth classes, which depicts the differences in the number of classes and class separability that in turn affect the clustering accuracies.

3.2. Implementation

The algorithms implemented in this work use the hyperparameters and parameters as stated in their original works. Specifically for autoencoder along with k-means and DEC methods, we employ the autoencoder used for pretraining in [12]. For the rest of the techniques, the hyperparameters stated in [18] were used.

3.3. Analysis of Performance of Hyperspectral Datasets on Various Clustering Algorithms

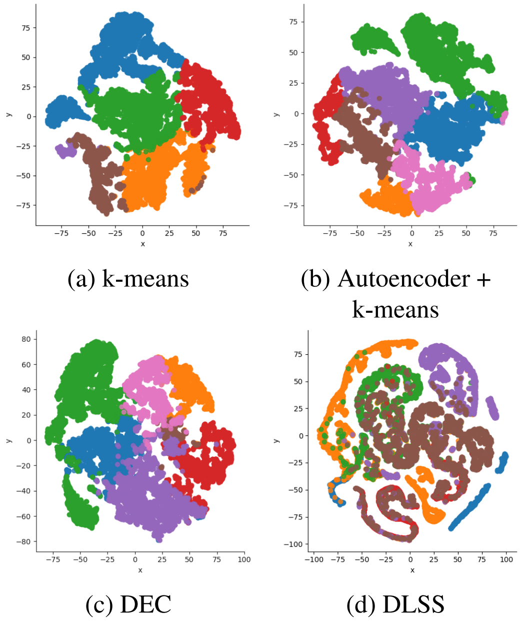
From Table 1, we see that DEC, SMCE, and DLSS perform equally well, however DLSS gives more consistent results for all the datasets. DEC falls short for Salinas-A and Cuprite datasets and this can be attributed to DEC using a feed-forward artificial neural network instead of a convolutional neural network which is also noted by the authors in [19](for RGB datasets). SMCE does not perform well for the Cuprite and the Indian Pines dataset and takes longer times to converge to a solution. DEC and SMCE also take longer to compute results, and on being compared to DLSS, the two fall short in computation time.

As is in shown in [17], the number of clusters affects the clustering accuracy negatively. Therefore, Indian Pines and Cuprite datasets show worse accuracies compared to Pavia and Salinas-A datasets due to having more number of classes.

Another notable anomaly observed is in the case of Salinas-A dataset where DEC performs the worst. A possible reason for this is the stopping criteria used in [12]. The authors stop the procedure when less than δ percentage of the points change cluster assignments between any two successive iterations. We also observe that for certain runs of the algorithm, DEC produces lesser number of clusters than that present in the ground truth data. This is seen for the Salinas-A, the Indian Pines and the Cuprite datasets.

Figure 3 shows the t-SNE representations [20] for the Pavia University dataset when clustered using k-means clustering, autoencoder along with k-means clustering, DEC and DLSS. The cluster labels obtained from each algorithm are depicted in color. For autoencoder along with k-means, we use the feature embedding obtained at the last iteration and visualize those using t-SNE. Similarly, for DEC we use the latent feature representation obtained during the last iteration. In the case of DLSS, we first obtained the weighted eigen vectors and then employ t-SNE.

Finally, we see that linear techniques like k-means and PCA along with k-means clustering take less computational time and memory than deep methods like autoencoder along with k-means and DEC. Furthermore, the deep methods implemented in this work do not provide better accuracies than that by the former two linear techniques.



(a) k-means

(b) Autoencoder + k-means

(c) DEC

(d) DLSS

Fig. 3. The t-SNE representations for the Pavia University dataset for (a) k-means clustering, (b) Autoencoder + k-means clustering, (c) DEC and (d) DLSS algorithms.

4. CONCLUSIONS

This work is a first hand examination of high-resolution Cuprite dataset. We implement and study how various state-of-the-art clustering algorithms perform on it. We also examine the performance of these state-of-the-art clustering techniques on multiple other hyperspectral datasets. Quantitative results show that the DLSS algorithm consistently outperforms other algorithms. We also observe and identify the shortcomings of the other techniques.

In future efforts, the authors would like to combine the feature representation learnt from DEC with the clustering technique from DLSS and also analyse the Cuprite dataset using the NLTV algorithm.

5. REFERENCES

[1] R. N. Clark, "Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems," *Journal of Geophysical Research*, vol. 108, no. E12, p. 5131, 2003.

[2] L. Hamlin *et al.*, "Imaging spectrometer science measurements for terrestrial ecology: AVIRIS and new developments," *IEEE Aerospace Conference Proceedings*, pp. 1–7, 2011.

[3] N. Gillis *et al.*, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," *CoRR*, vol. abs/1310.7441, 2013.

[4] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, Jun 2015.

[5] N. Acito *et al.*, "An unsupervised algorithm for hyperspectral image segmentation based on the gaussian mixture model," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, vol. 6, July 2003, pp. 3745–3747 vol.6.

[6] M. D. Farrell and R. M. Mersereau, "On the impact of pca dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 2, pp. 192–195, April 2005.

[7] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1586–1600, June 2006.

[8] M. Zhang *et al.*, "Diverse region-based cnn for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2623–2634, June 2018.

[9] H. Lee *et al.*, "Cross-domain CNN for hyperspectral image classification," *CoRR*, vol. abs/1802.00093, 2018.

[10] X. Cao *et al.*, "Hyperspectral image classification with markov random fields and a convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[11] C. Zhao *et al.*, "Spectral-spatial classification of hyperspectral images based on joint bilateral filter and stacked sparse autoencoder," in *2017 First International Conference on Electronics Instrumentation Information Systems (EIIS)*, June 2017, pp. 1–5.

[12] J. Xie *et al.*, "Unsupervised deep embedding for clustering analysis," *CoRR*, vol. abs/1511.06335, 2015.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] A. Y. Ng *et al.*, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, ser. NIPS'01. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856.

[15] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor *et al.*, Eds. Curran Associates, Inc., 2011, pp. 55–63.

[16] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[17] W. Zhu *et al.*, "Unsupervised classification in hyperspectral imagery with nonlocal total variation and primal-dual hybrid gradient algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2786–2798, May 2017.

[18] J. M. Murphy and M. Maggioni, "Unsupervised geometric learning of hyperspectral images," *CoRR*, vol. abs/1704.07961, 2017.

[19] E. Aljalbout *et al.*, "Clustering with deep learning: Taxonomy and new methods," *CoRR*, vol. abs/1801.07648, 2018.

[20] G. Hinton and Y. Bengio, "Visualizing data using t-sne," in *Cost-sensitive Machine Learning for Information Retrieval 33*.