ASSESSMENT OF NORMALIZATION TECHNIQUES ON THE ACCURACY OF HYPERSPECTRAL DATA CLUSTERING

A. Alizadeh Naeini [a, *], M. Babadi [a], S.Homayouni [b]

[a]Dept. of Geomatics Eng., Faculty of Civil Engineering and Transportation, U. of Isfahan, Iran
(a.alizadeh@eng.ui.ac.ir, masoud.babadi92@trn.ui.ac.ir)
[b]Dept. of Geography, Environment, and Geomatics, U. of Ottawa, Canada
(saeid.homayouni@uOttawa.ca)

KEYWORDS: K-means clustering, normalization techniques, density based initialization, hyperspectral data

ABSTRACT:

Partitioning clustering algorithms, such as k-means, is the most widely used clustering algorithms in the remote sensing community. They are the process of identifying clusters within multidimensional data based on some similarity measures (SM). SMs assign more weights to features with large ranges than those with small ranges. In this way, small-range features are suppressed by large-range features so that they cannot have any effect during clustering procedure. This problem deteriorates for the high-dimensional data such as hyperspectral remotely sensed images. To address this problem, the feature normalization (FN) can be used. However, since different FN methods have different performances, in this study, the effects of ten FN methods on hyperspectral data clustering were studied. The proposed method was implemented on both real and synthetic hyperspectral datasets. The evaluations demonstrated that FN could lead to better results than the case that FN is not performed. More importantly, obtained results showed that the rank-based FN with 15.7% and 12.8% improvement, respectively, in the synthetic and real datasets can be considered as the best FN method for hyperspectral data clustering.

1. INTRODUCTION

Classification can be categorized into two main groups of supervised and unsupervised classification methods. Although supervised methods lead to the better results, unsupervised or clustering techniques, have been attracted many attentions because they do not need any training data and assumption about data (Melgani and Pasolli, 2013).

Among different clustering algorithms, partitional methods are one of the best techniques for high-dimensional data, e.g., hyperspectral data. This is mainly because, they have lower complexity (Celebi et al., 2013). The k-means algorithm is undoubtedly the most widely used partitional clustering algorithm (Jain, 2010). However, k-means has two significant disadvantages. The first is its sensitivity to the range of image features. To address this drawback, the feature normalization (FN) methods can be used. The second disadvantage is its sensitivity to the selection of the initial clusters. To tackle this problem, either deterministic (Celebi et al., 2013) or heuristic methods (Abraham et al., 2008) can be used. Regarding the second problem, various solutions are proposed (Bradley and Fayyad, 1998, Khan and Ahmad, 2004). However, to the best of our knowledge, the first problem and its solutions (i.e. FN) have not been addressed in the literature. Accordingly, in this paper, we evaluated the performance of different FN methods and their effects on hyperspectral data clustering. Due to the impact of initialization on our aim, Distance-Based Neighborhood Density Initialization (DBNDI) (Zhang and Jiang, 2010), as an initialization method for high dimensional data, are considered here to alleviate the second problem.

Different classifiers have been used FN methods. In (Smits and Annoni, 2000), authors have used FN in order to do change detection in a better way. Their results show that FN methods are necessary in cases that distance measures are used among different physical features. Manian et al. (2000) have used FN to classify texture features. Their results indicate that FN can improve the classification. Zhang et al. (2015) have investigated on the influence of FN on the fusion of optical and SAR data for land cover classification. Their results show that distribution-dependent classifiers are independent of normalization. Li et al. (2015) have applied FN for hyperspectral image classification and have expressed that it is a necessary preprocessing for hyperspectral image analysis. In (Clausi and Deng, 2005), authors have normalized texture features by scaling each feature dimension to the range of [0, 1].

The focus of this study is on the evaluation and comparison of FN methods for clustering of hyperspectral data. To do this, ten FN methods were used. In this study, we aim at answering the following questions:

1) What is the best FN method for hyperspectral data clustering?
2) To what extent, FN can improve the accuracy of hyperspectral data clustering?

The rest of the paper is organized as follows. Section 2 presents a summary of k-means algorithm and introduces feature selection methods. Section 3 describes the experimental setup and results. Lastly, Section 4 gives our conclusion.

2. THEORETICAL BACKGROUND

2.1 K-Means Clustering

K-means clustering (MacQueen, 1967) is a method commonly used to partition a dataset into K groups automatically. It proceeds by selecting K initial cluster centers and then iteratively refining them as follows. 1) First, each point is assigned to its closest cluster center. 2) Each cluster center Cj is updated to be the mean of its constituent points (Wagstaff et al., 2001). From the mathematical perspective, given dataset $X = \{x_1, x_2, \dots, x_N\}$

* Corresponding author

in $\mathbb{R}^D$, i.e. N points (vectors) each with D attributes (components), K-means algorithm divides X into K exhaustive and mutually exclusive clusters $P = \{p_1, p_2, \dots, p_K\}$, $\bigcup_{i=1}^{K} p_i = X$, $p_i \cap p_j = \emptyset$ for $1 \le i \ne j \le K$. This algorithm generates clusters by optimizing a criterion function. The most intuitive and frequently used criterion function is the Sum of Squared Error (SSE) given by:

$$SSE = \sum_{i=1}^{K} \sum_{x_j \in p_i} \|x_j - c_i\|_2^2 \quad (1)$$

where, $\|.\|_2$ denotes the Euclidean ($\mathcal{L}_2$) norm and $c_i = \frac{1}{|p_i|} \sum_{x_j \in p_i} x_j$ is the centroid of cluster p_i whose cardinality is $|p_i|$. The optimization of (1) is often referred to as the minimum SSE clustering (MSSC) problem (Celebi et al., 2013). To address sensitivity of k-means solutions to initial cluster centers, in this study, an initialization method for high dimensional data, namely, the Distance-Based Neighborhood Density Initialization (DBNDI) (Zhang and Jiang, 2010) is used. Furthermore, to investigate the influence of different FN methods, Euclidean distance (ED), as the most frequently used measures in the remote sensing literature is applied (Celik, 2009).

2.2 Feature Normalization:

FN methods aim at normalizing each feature of the image in different ranges. Normalization of image values is necessary, especially when distance classifiers are used. By normalizing features, equal weights are given to different features of an image on the one hand, and on the other computational burden is reduced.

2.2.1 Stretching based method (SBM):

2.2.1.1 Mapping between [0,1]: Given a lower bound l and an upper bound u for a feature component x_i, normalization can be done as follows:

$$xn_i = \frac{x_i - l}{u - l} \quad (2)$$

Where i is number of feature component and $xn_i$is normalized feature ranged from 0 to 1 (Aksoy and Haralick, 2001).

2.2.1.2 Trimming: In this approach, the value located at the point of the top 95% of the distribution is taken as the nominal maximum. All features greater than the nominal minimum in the feature space were clipped to the nominal maximal value, i.e. the top 5% of distribution are trimmed. Then all values are divided by the maximal values (Wei and Li and Wilson, 2005).

2.2.2 Statistical normalization methods: In this method of normalization, a feature component $x_i$is transformed to a random variable with zero mean, median or mode and unit variance as:

$$xn_i = \frac{x_i - \mu}{\sigma} \quad (3)$$

Where μ is mean, median or mode and σ is the sample standard deviation of that feature (Aksoy and Haralick, 2001).

2.2.3 Norm-based normalization (p-norm and infinity norm normalization): The p-norm of a 1-by-n or n-by-1 vector V is defined as follows (Xie and Tian and Zhang, 2013):

$$\|V\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \quad (4)$$

Where, x_i Is a feature component, and n is some components in each feature. If p tend to infinity, the infinity norm of a 1-by-n or n-by-1 vector V is defined as follows:

$$\|V\|_\infty = (|x_i|) \quad (5)$$

Then all values are divided by the p-norm of each feature:

$$xn_i = \frac{x_i}{\|V\|_p} \quad (6)$$

2.2.4 Rank normalization: With rank normalization, each feature component x_i In one array are replaced by its position (rank) in the ordered array counted from the smallest value divided by the total number of components, but if each feature contains a repetitive value, each value is divided by the maximum of the feature. Denote r_i as the rank of $x_i$in the array to which it belongs, the normalized component expressions are:

$$xn_i = \frac{r_i}{m} \quad (7)$$

Where m is the total number of components or maximum of feature (Qui and Wu and Hu, 2013).

2.2.5 Cosine distance as normalization method: Cosine distance can be used as a normalization method. Cosine distance (or vector dot product), which is the sum of the product of each component from two vectors z_u and z_w, defined as:

$$< z_u, z_w >= \frac{\sum_{j=1}^{N_d} z_{u,j} z_{w,j}}{\|z_u\| \|z_w\|} \quad (8)$$

Where z_u and z_w are two vectors and $z_{u,j}$ and $z_{w,j}$ are components of the vector z_u and z_w, respectively, and also $<z_u, z_w> \epsilon$ [-1, 1].
The cosine distance is not a distance but rather a similarity metric. In other words, the cosine distance measures the difference in the angle between tow vectors not the difference in the magnitude of two vectors. The cosine distance is suitable for clustering data of high dimensionality (Omran, 2005).

3. DISCUSSION AND RESULTS

3.1 Hyperspectral Data

To assess the efficiency of different normalization methods on hyperspectral image clustering, one real dataset, and one simulated dataset were used. The synthetic dataset is one of the five well-known synthetic images in hyperspectral-processing space (Martin and Plaza, 2011, Plaza et al., 2012). An image window of 100×100 pixels was created and used to simulate the linear mixtures. Nine selected minerals from the U.S. Geological Survey (USGS) spectral library were used to simulate hyperspectral data. The real data set was acquired by the ROSIS sensor during a flight campaign in 2003 over the campus of Pavia University in the north of Italy. This data contains 610 by 340 pixels with 103 spectral bands. This dataset contains nine ground-truth classes, namely, Trees, Gravel, Meadows, Asphalt, Metal sheets, Bricks, Bitumen, Shadows and Bare soil. Figure 1 and figure 2 shows the ground-truth map and a color composite image of real and simulated data sets, respectively. Before using these

data sets, their background is ignored. This is because no information is available about these areas and using background only increases computing time.

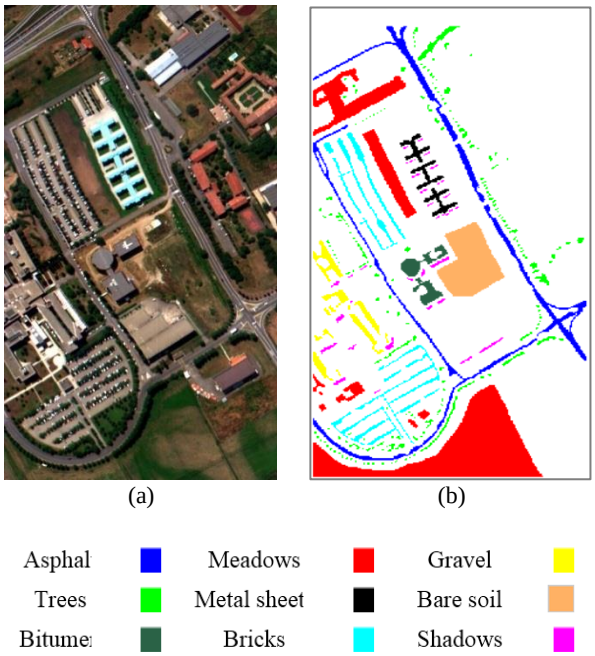

Figure 3. ROSIS hyperspectral dataset over Pavia University used in experiments: (a) color composition image (R: 60, G: 30, B: 10). (b) Ground truth map.

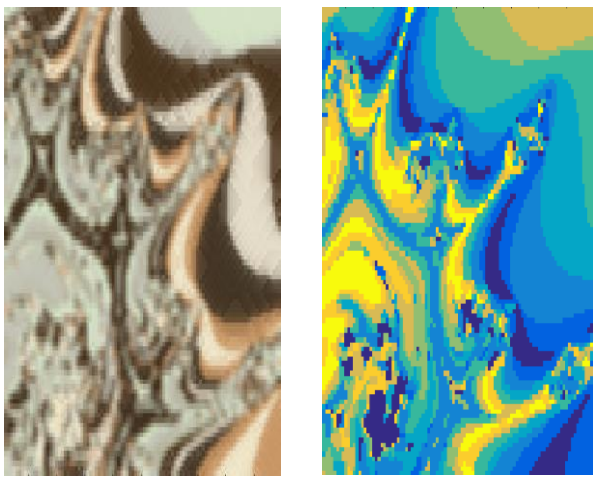

Figure 2. Simulated Plaza dataset: (a) color composition image (R: 60, G: 30, B: 10). (b) Ground truth map.

3.2 Experimental results

Kappa coefficients of k-means clustering for different normalization methods, are tabulated in Table 1 and illustrated in figure 3. As is obvious from the results, in the simulated dataset, norm2 and trimming methods led to the best and worst results, respectively, for k-means clustering. However, in real dataset, SBM (mapping between [0,1]) and Cosine distance method led to the best and worst results for clustering, respectively. In the simulated dataset, norm2 normalization method resulted in 17.8% improvement in results than clustering without normalization, and in real dataset, SBM (mapping between [0,1]) resulted in 14.5% improvement in results than clustering without normalization.

According to the results, different normalization methods, regardless of their methodology, almost led to improvement in clustering results though the amount of this improvement varies from normalization method to normalization method. Among normalization methods, rank normalization, had reasonable results in both data sets and norm infinity method, despite its application in classification papers ((Xie and Tian and Zhang, 2013), had the worst result in both data sets. Rank normalization method led to 15.7% and 12.8% improvement in results than clustering without normalization in the simulated and real dataset, respectively.

Table 1. Kappa coefficients of k-means clustering for different normalization methods.

Normalization method \ Data set		Simulated (Plaza)	Pavia University
Without normalization (WON)		0.2677	0.2757
SBM	mapping between [0,1]	0.3818	0.4209
	Trimming	0.2851	0.4145
Statistical	mean	0.4389	0.2673
	median	0.4389	0.2673
	mode	0.4389	0.2673
Norm based	Norm1	0.4038	0.3632
	Norm2	0.4458	0.2824
	Infinity	0.3839	0.2712
Rank normalization		0.4247	0.4037
Cosine distance		0.4251	0.1676

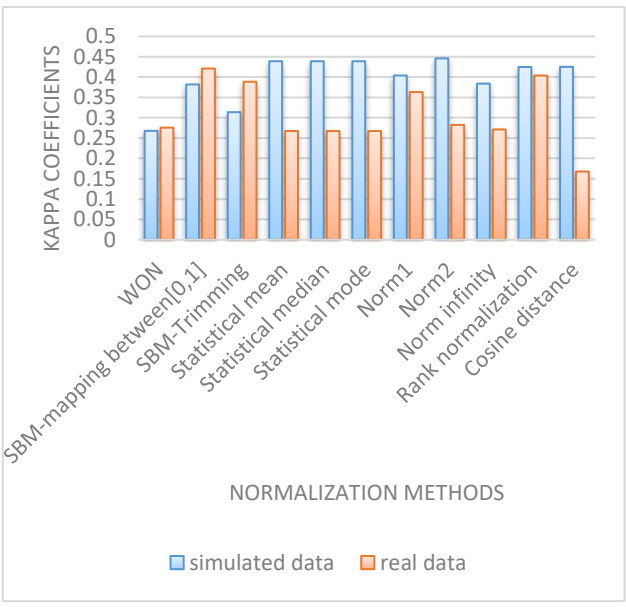

Figure 6. Kappa coefficients of k-means clustering for different normalization methods.

4. CONCLUSION

Normalization of features is a necessary pre-processing step in hyperspectral clustering tasks. FN gives equal weight to different features of an image, especially when distance classifiers are used. In order to compare the performance of different FN

methods, in this study, the effects of ten FN methods on hyperspectral data clustering were discussed. Different FN methods were investigated on both real and synthetic hyperspectral datasets. Based on the results, different normalization methods, regardless of their methodology, almost could improve the clustering results. Although, the amount of this improvement varies for different FN methods. The results of simulated dataset showed that k-means clustering in the case of 2-norm normalization and SBM (trimming) methods led to the best and worst results, respectively. In the real dataset, SBM (mapping between [0,1]) and Cosine distance method led to the best and worst results, respectively. Among normalization methods, rank normalization led to convincing improvement on both datasets; 15.7% and 12.8% in the simulated and real datasets respectively, when compared to the clustering without normalization. On the other hand, Norm infinity method had the worst results in both data sets. After all, we can conclude that different FN methods on various datasets, lead to Different ranges of improvement and this fact should be considered in clustering and classification works.

In future works, we study the effects of different FN methods on different clustering methods and the effect of the various features such as object-based ones.

REFERENCES

Abraham, A., Das, S. and Roy, S., 2008. Swarm intelligence algorithms for data clustering. In Soft Computing for Knowledge discovery and data mining (pp. 279-313). Springer US.

Aksoy, S. and Haralick, R.M., 2001. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern recognition letters, 22(5), pp.563-582.

Bradley, P.S., and Fayyad, U.M., 1998, July. Refining Initial Points for K-Means Clustering. In ICML (Vol. 98, pp. 91-99).

Celebi, M.E., Kingravi, H.A. and Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40(1), pp.200-210.

Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geoscience and Remote Sensing Letters, 6(4), pp.772-776.

Clausi, D.A. and Deng, H., 2005. Design-based texture feature fusion using Gabor filters and co-occurrence probabilities. IEEE Transactions on Image Processing, 14(7), pp.925-936.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

Khan, S.S., and Ahmad, A., 2004. Cluster center initialization algorithm for K-means clustering. Pattern recognition letters, 25(11), pp.1293-1302.

Li, W., Chen, C., Su, H. and Du, Q., 2015. Local binary patterns and extreme learning machine for hyperspectral imagery classification. IEEE Transactions on Geoscience and Remote Sensing, 53(7), pp.3681-3693.

MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Manian, V., Vásquez, R. and Katiyar, P., 2000. Texture classification using logical operators. IEEE Transactions on image processing, 9(10), pp.1693-1703.

Melgani, F. and Pasolli, E., 2013. Multi-objective PSO for hyperspectral image clustering. In Computational Intelligence in Image Processing (pp. 265-280). Springer Berlin Heidelberg.

Omran, M.G., 2005. Particle swarm optimization methods for pattern recognition and image processing (Doctoral dissertation, University of Pretoria).

Qiu, X., Wu, H. and Hu, R., 2013. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC bioinformatics, 14(1), p.124.

Smits, P.C. and Annoni, A., 2000. Toward specification-driven change detection. IEEE Transactions on Geoscience and Remote Sensing, 38(3), pp.1484-1488.

Wagstaff, K., Cardie, C., Rogers, S. and Schrödl, S., 2001, June. Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).

Wei, C.H., Li, C.T. and Wilson, R., 2005, April. A general framework for content-based medical image retrieval with its application to mammograms. In Medical Imaging (pp. 134-143). International Society for Optics and Photonics.

Xie, L., Tian, Q. and Zhang, B., 2013, September. Feature normalization for part-based image classification. In Image Processing (ICIP), 2013 20th IEEE International Conference on (pp. 2607-2611). IEEE.

Zhang, H., Lin, H., and Li, Y., 2015. Impacts of feature normalization on optical and SAR data fusion for land use/land cover classification. IEEE Geoscience and Remote Sensing Letters, 12(5), pp.1061-1065.

Zhang, Y. and Jiang, Q., 2010, November. An Improved initialization method for clustering high-dimensional data. In Database Technology and Applications (DBTA), 2010 2nd International Workshop on (pp. 1-4). IEEE.