COMPARISON OF DIFFERENT SIMILARITY MEASURES FOR SELECTION OF OPTIMAL, INFORMATION-CENTRIC BANDS OF HYPERSPECTRAL IMAGES

Munmun Baisantry
School of Computing and Electrical Engineering
Dericks P Shukla
School of Engineering
Indian Institute of Technology Mandi
Himachal Pradesh, India 175005
munmun_baisantry@students.iitmandi.ac.in
dericks@iitmandi.ac.in

ABSTRACT

Hyperspectral images consisting of large number of spectral bands suffer from limitations like High data redundancy, curse of dimensionality (insufficient training samples), and high computational complexity. Therefore, dimensionality reduction & band-selection has become a common practice in the field of hyperspectral image processing. Graph-based band selection is a well-known technique which is based on spectral clustering on similarity matrix to select the optimal band set. Thus, choice of similarity/ affinity matrix is a vital decision in these methods. We have conducted a comparison of some well-known affinity matrices used in spectral clustering to divide graph to smaller sub-graphs (indicating subsets of bands). Comparison was done using various types of metrics like ACC, AIE and ARE.

KEYWORDS: Hyperspectral, AVIRIS, Indian Pines, Band selection, Similarity matrices, Dimensionality reduction.

INTRODUCTION

Hyperspectral sensors, with their nanometric spectral resolution and large number of continuous bands, have proved to be an asset to the remote sensing community, gaining prominence in myriad applications like target detection, classification etc. Subtle spectral divergences between objects can be easily identified using hyperspectral images, helping to distinguish between them (Vorovencii, 2009), (Mohan et. al., 2015). For example, deep-water bodies, roads or shadows are often considered as same in multispectral images based on reflectance values but they can be very easily distinguished in hyperspectral images. This property is very useful in detecting targets such as minerals, small vehicles, military camps, buildings,and oil tanks etc. in images where there are a lot of background classes.

With increasing dimensionality, the size of training samples required to estimate parameters also increases exponentially, a term characterized as curse of dimensionality (Kouiroukidis et. al., 2011).In high dimensional space, data is mostly clustered in small low-dimensional subspaces & structures instead of being uniformly distributed across the space. Most of the high dimensional space is sparse and data is far apart from each other (Li et. al.,2011).

Thus, conventionally used distance measures are not indicative of the true distance between the data points and may not give factual clustering results. Moreover, hyperspectral bands are highly correlated so band selection and dimensionality reduction methods are applied to reduce redundancy and complexity associated with high dimensionality in various applications.

Since hyperspectral bands are continuous in nature, it may seem that adjacent, contagious bands can be combined as one. However, band selection and dimensionality reduction is a data-dependent process and the assumption that the band similarity is location-dependent phenomena may not hold true in many cases.

In our paper, we have discussed the popular, graph-based band selection method and demonstrated the effects of different similarity measures available on it. The band selection & dimensionality reduction of hyperspectral image using graph-based approach consists of threemajor steps:

1) Learning a similarity matrix
2) Estimating the number of optimal number of bands.
3) Clustering the similarity matrix using spectral clustering.

DIFFERENT MEASURES OF SIMILARITY

The similarity between two objects is a numeral measure of the degree to which the two objects are alike.(Cha,2007) . Consequently, similarities are higher for pairs of objects that are more alike. Choosing the right similarity measure is of fundamental importance to applications like segmentation, pattern recognition, and content-based image retrieval systems but can take substantial efforts due to presence of plethora of measures available. Choice of similarity measure is also dependent on the representation of the objects which could be in probabilistic or vector form as well as in numeric or binary form.

Higher-dimensional hyperspectral data can be represented as a union of low-dimensional subspaces/ clusters having high intra-cluster correlation and low inter-cluster correlation (Sun et. al., 2015). Relationship between hyperspectral bands can be explained in terms of a graph wherein each band can be represented as a graph node. The adjacency matrix of the graph, thus, will represent the pair-wise adjacency between each band as quantified by the similarity measures. Hyperspectral bands, when grouped together, can result in a block-diagonal, sparse similarity matrix and could benefit the correct segmentation of all data points into separate subspaces. Such methods of band selection require that the distance metric & similarity score which should have higher values for bands within the same cluster and lower values for bands belonging to different clusters. Many band selection algorithms like the sparse representation based band selection (SpaBS) algorithm (Jia , 2012), the sparse nonnegative matrix factorization (Li et al, 2011)] algorithm, and the sparse support vector machine (SSVM)algorithm are based on the graph-partitioning principle using similarity matrices. Some other interesting approach to measure similarity between bands include frequency spectrum or fourier transform based similarity measures (Wang et al, 2016) and shape-based similarity measures (Shijin et al. 2014) . With this view, we have performed a comparative analysis of several conventionally used similarity measures in their contribution to improve the performance of band selection methods.

The similarity measures compared in this study are discussed below:

Euclidean distance

Euclidean Distance between two points **p,q** is a Minkowski distance metric defined as (1)

$$\mathbf{euc(p,q)}=\sqrt{\left(p_{1}- q_{1}\right)^{2} + \left(p_{2} -q_{2}\right)^{2}+\ldots+ \left(p_{n} - q_{n}\right)^{2}} \quad (1)$$

Where n is the number of dimensions. Euclidean distance is only appropriate for data measured on the same scale and for other cases, Mahalanobis distance is considered to be a better measure.

Pearson Correlation

Pearson Correlation coefficient (Basseville , 1957) is a measure of how well two sets of data fit on a straight line. Correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship. If the correlation is 0, then there is no linear relationship between the attributes of the two data objects. However, the two data objects might have non-linear relationships.

Pearson correlation between two vectors **p,q**is defined by the following equation (2)

$$\mathbf{corr(p,q)}=\frac{\text{cov}(\mathbf{p,q})}{\text{stand.dev.}(\mathbf{p})*\text{stand.dev.}(\mathbf{q})} \quad (2)$$

Tanimoto Coefficient

Tanimoto coefficient, also known as extended Jaccard coefficient (Tanimoto, 1957) is used for handling the similarity of document data in text mining. In the case of binary attributes, it reduces to the Jaccard coefficient. Tanimoto coefficient between vectors **p,q** is defined by equation (3):

$$\boldsymbol{Tan(p,q)} = \frac{\boldsymbol{p.q}}{\|\boldsymbol{p}\|^2 + \|\boldsymbol{q}\|^2 \quad \boldsymbol{p.q}} \quad (3)$$

Cosine Similarity

The cosine similarity (Elhamifar et al. 2009)is a measure of similarity of two non-binary vectors. The cosine similarity ignores 0-0 matches like the Jaccard measure. The cosine similarity is defined by the equation (4):

$$\cos(\mathbf{p},\mathbf{q}) = \frac{\mathbf{pXq}}{\|\boldsymbol{p}\|\|\boldsymbol{q}\|} \qquad (4)$$

Spectral Angle Mapper

SAM is a spectral classifier that is able to determine thespectral similarity between image spectra and reference spectra by calculating the angle between thespectra, treating them as vectors in a space with dimensionality equal to the number of bands used eachtime .In SAM, only the angular information is used for identifying pixel spectra, where small angles between the two spectrums indicatehigh similarity and high angles indicate low similarity, whereas pixels with an angle larger than thetolerance level the specified maximum angle threshold are not classified. It can be defined as:

$$\mathrm{SAM}(\mathbf{p},\mathbf{q}) = cos^{-1}\frac{dot(\boldsymbol{p},\boldsymbol{q})}{norm(\boldsymbol{p})\quad norm(\boldsymbol{q})} \qquad (5)$$

Spectral Information Divergence

SID, also known as the Kullback–Leibler information measure, directed divergence, or cross-entropy is a symmetric hyperspectral measurethat can be used to measure the spectral similarity between two pixels samples p & q. SID offers a new look at spectral similarity by making use of relative entropy to account for the spectral information provided by each pixel. SID can be defined as:

$$\mathrm{SID}(\mathbf{p},\mathbf{q}) = \mathrm{D}(\mathbf{p}/\mathbf{q}) + \mathrm{D}(\mathbf{q}/\mathbf{p}) \qquad (6)$$

$$\text{Where}\mathrm{D}(\mathbf{p}/\mathbf{q}) = \sum_{i=1}^{N} p_i \log\frac{p_i}{q_i} \qquad (7)$$

Hybrid similarity

Hybrid metric is a combination of SAM and SID, both of which are spectral measures. SAM is a deterministic method that looks for an exact pixel match and weighs the differences of the same. SID is a probabilistic method that allows for variations in pixel measurements, where probability is measured from zero to a user-defined threshold. Smaller the values of SAM, greater is the similarity. Similarly, smaller the value of divergence, the more likely the pixels are similar. Using this concept, hybrid metric suggested in (Kumar et al. 2011) is defined as:

$$SIDSAM_{tan} = SID(s_i, s_j) \times \tan(SAM(s_i, s_j)) \qquad (8)$$

BAND CLUSTER USING SPECTRAL CLUSTERING

Spectral clustering techniques (Von Luxburg, 2007), (Honarkhah et al. 2010) make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform band clustering. The premise is to find a partition of thegraph such that the edges between different groups have very low weights and the edges within a group have high weights. For clustering, laplacian of the similarity matrix is calculated and k-means is performed along the direction of the first k eigen vectors. It is very essential that the similarity matrix is chosen wisely such that its Laplacian is as block-sparse as possible for correct clustering of the graph, each block representing a sub-graph.

Generally, affinity/similarity matrices are not sparse block diagonal in nature. The coefficient values which represent affinities between bands belonging to same cluster are not strong enough to show a stark difference from bands of different clusters. To successfully meet our objectives, we need to define some mathematical operation on the similarity matrix such that we get block structure with strongly weighed affinities for same clusters and reduce weak inter-cluster affinities further.

BAND SELECTION METHODOLOGY

The complete workflow of graph-based methodology used for band selection is shown in Figure 1. Here, the similarity matrix will be used to divide/ cluster the graph into smaller sub-graphs representing low-dimensional subspaces. Clustering is done on similarity matrix using the K-means algorithm. One representative band from each cluster combines to form the final subset.

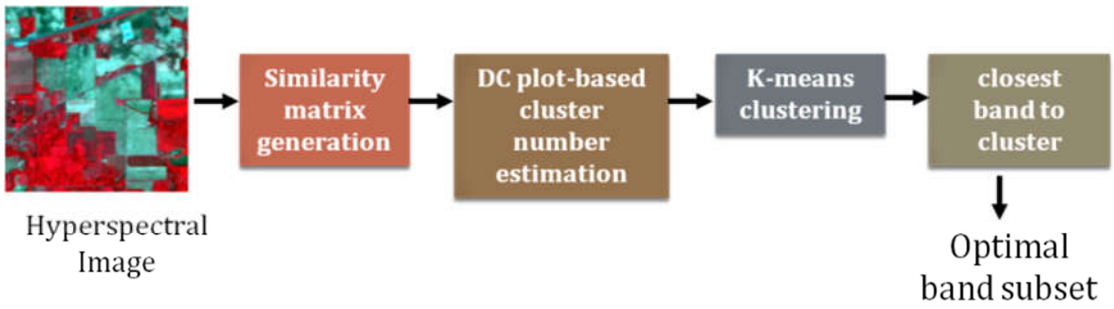


Figure 1: Workflow for graph-based band selection

Here, the number of sub-graphs or subspaces is estimated using DC (Distribution Compactness) method (Sun et. al., 2015) which uses an Eigen value decomposition of the similarity matrix to estimate the total compactness of all clusters. Compactness is measured by low intra-cluster variance and high inter-cluster variance. Number of Eigen vectors which contribute to an increase in total inter-cluster variance and decrease in total intra-cluster variance significantly is the estimated number of clusters. This is shown in Figure 2 where total count of all the Eigen vectors whose contribution to compactness is greater than 2 percent is the number of clusters.

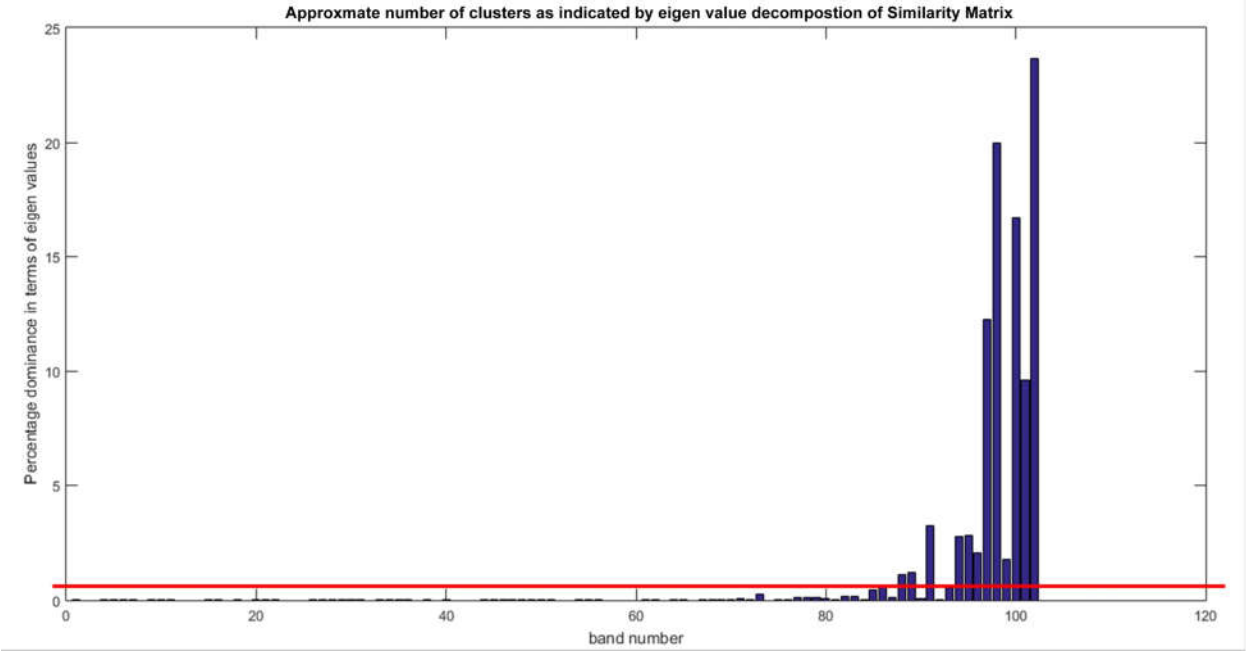


Figure 2:DC (Distribution Compactness) plot for estimation of cluster number

RESULTS & DISCUSSION

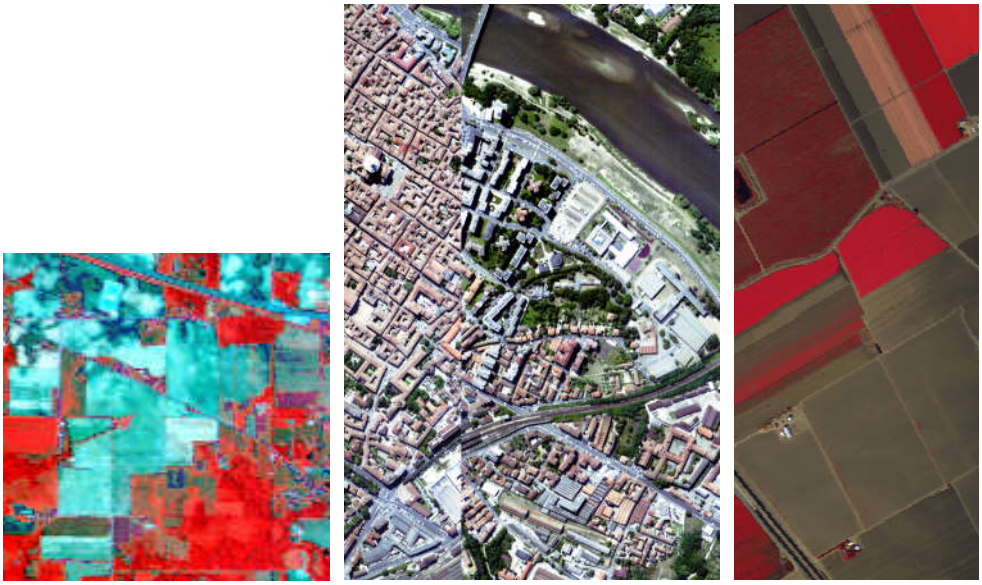
The experiment was conducted on three well known AVIRIS hyperspectral datasets available at <http://www.ehu.eus/ccwintco/index.php?titleHyperspectral_Remote_Sensing_Scenes >[15]. First set is called Indian-pines which is an AVIRIS sensor image with 20 m spatial resolutions bands (Figure. 3(a)). The dataset consists of 224 spectral reflectance bands in the wavelength range 0.4–2.5 10-6 meters. The datasets have been preprocessed for

radiometric corrections, bad band and water absorption bands removal. This dataset contains classes of agriculture, forest or other natural perennial vegetation. There are total 16 classes such as corn, soybeans, wheat, hay etc which are highly similar and mutually non-exclusive resulting in poor discrimination amongst the classes and poor classification accuracies.

The second dataset covers Pavia University, Northern Italy and is acquired using ROSIS sensor with 1.3 m spatial resolution and 115 bands. Both the dataset as well as the ground truth was taken from the Computational Intelligence Group in the Basque University. After removing low signal to noise ratio (SNR) bands, 103 bands were left for further analysis. The dataset, sized 610* 610, covers a complex urban scene consisting of 9 spectrally distinct classes like alphalt, bricks, meadows, trees, metallic sheets etc. The datset is shown in (Figure. 3(b)).

The third dataset was collected by the 224-band AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution (3.7-meter pixels). After discarding the water absorption bands, 204 bands were used in the experiment. The image sized 512 * 217, consists of 16 spectrally confusing and similar classes like vegetables, bare soils, and vineyard fields. The datset is shown in (Figure. 3(c)).

These are challenging datsets and thus, chosen to test the similarity measures.



(a) **(b)** **(c)**

Figure 3: Hyperspectral dataset (a) Indian pines (b) Pavia dataset (c) Salinas

Similarity measures determined from the methods discussed above are diagrammatically shown in Figure 4. While the bluish regions show low values of similarity, yellow-colored portions show high correlation appearing in blocks between neighboring sets of bands. Euclidean distance, due to weaker similarity affinities, does not offer stark discrimination between the clusters of similar bands, thus, number of clusters are more and loosely defined. This can also be seen in Table 1, where the number of clusters (in this case is 20) as estimated by DC plot algorithm is mentioned. The same behavior is also shown by Tanimoto matrix to some extent. However, its structure is more pronounced than Euclidean. The similarity matrices of these two measures are block-diagonal in nature, thus corroborating the assumption that neighboring bands are correlated. Similarity matrices seen in Figure 4 (b), (d) & (f) have fewer and larger blocks. Thus, estimated number of clusters is fewer in comparison as seen in Table 1. Further, the lack of block diagonal structure implies that correlation between the bands is not a localized phenomenon. Poor demarcation between clusters lead to incorrect estimation of clusters in SAM, as seen in Figure 4(e). From Table 1 and Figure 4, it is quite clear that the bands selected eventually by these methods are very close to the bands falling under the green portions of the graphs, green portions depicting band clusters. Here, final reduced band set consists of atleast one band from each cluster representative of the cluster (in terms of closeness to the cluster centroid) .

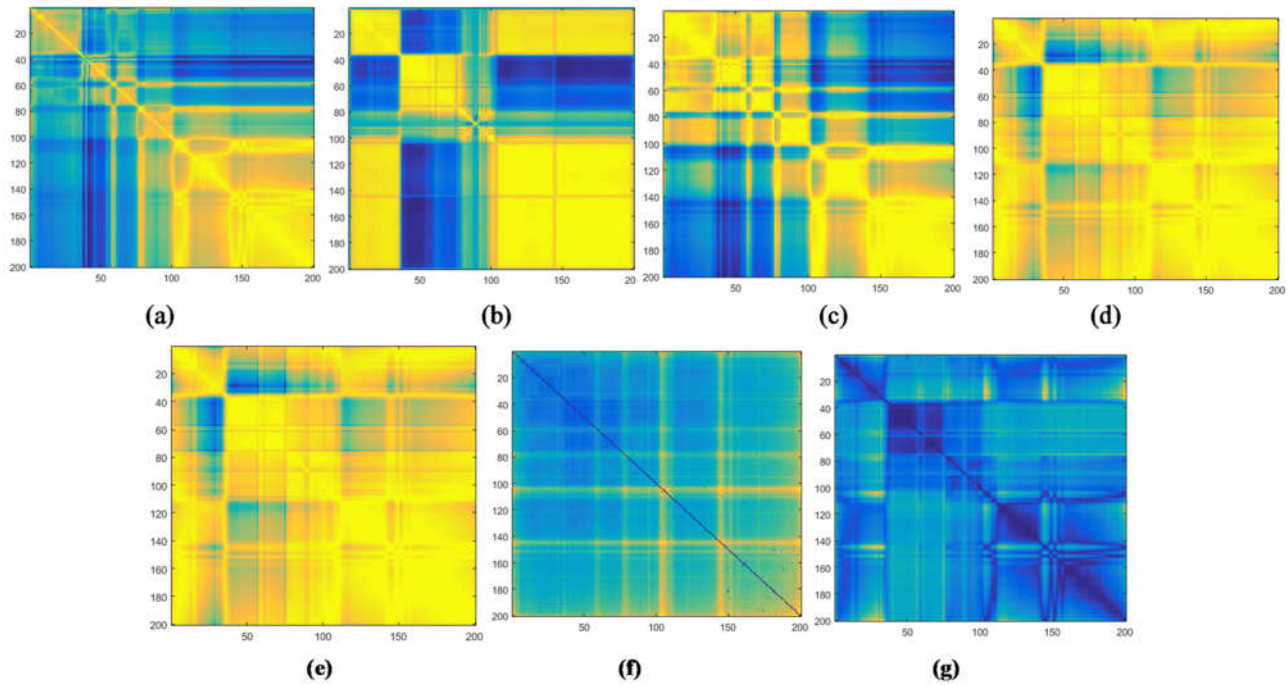


Figure 4: Similarity matrix of AVIRIS Indian Pines dataset using (a) Euclidean (b) Pearson(c) Tanimoto (d) Cosine (e) SAM (f) SID (g)Hybrid method

In the experiment, the appropriate size of band subset k (i.e., the number of bands in the subset) is estimated using the DC plot algorithm and is then set as the dimension of band subset for all the methods. This step of determining the virtual dimensionality is very crucial. If too few bands are taken, it can lead to loss of some subtle but vital information in the low entropy bands. On the other, too many bands also fail to serve the purpose of dimensionality reduction. More number of clusters are estimated if, the coefficient values, which represent affinities between bands belonging to same cluster, are not strong enough to show a stark difference from bands of different clusters. To support our argument, we can see in Table 1 that number of clusters in Euclidean, 20 and Tanimoto, 19 are higher in comparison to other measures.

Table 1.Comparison of similarity measures in terms of selected bands and classification accuracy

Similarity matrix used	Estimated no. of clusters	Bands selected	Classification Accuracy
Euclidean	20	5,17,27,31,45,50,51,55,80,,94,100,105,114,133,140,151,163 ,176,182,196	62.52
Pearson	14	1,7,15,69,76,84,92,98,112,119,135,167,174,193	58.74
Tanimoto	19	8,15,24,58,72,73,74,83,86,95,99,134,142,165,168,174,181,1 97	76.20
Cosine	16	1,7,15,69,76,84,92,98,112,119,135,144, 166,167,174,193	67.91
SAM	14	1,7,29,31,35,69,96,98,113,117,129,133,141,151	70.90
SID	11	56,67,71,77,83,99,116,127,134,151,188,195,196,200	61.06
Hybrid	13	9,20,35,56,67,77,99,116,123,133,140,163,196	67.14

Three metrics namely Average Information Entropy (AIE), Average Correlation Coefficient (ACC) and Average Relative Entropy (ARE) were used to evaluate the information carrying capacity and inter-band separability of the selected band sets [7]. AIE is used to measure the information amount and to evaluate the richness of spectrum information in the band subset, ACC gives the estimate of intra-band correlations in the band subset and ARE (also called average Kullback–Leibler divergence, AKLD) measures the inter-separabilities of selected bands and assess the distinguishing ability within the band subset for classification. These three quantitative metrics were used because they measure the three desired performance characteristic of selecting an appropriate band subset having high information

Table 2.Quantitative Evaluation of band selection method

Data Sets	Measure	ACC	AIE	ARE
Indian pines	**Cosine**	0.3703	3.1904	0.3559
	Euclidean	0.2422	3.4057	0.3164
	Tanimoto	**0.1846**	**4.0270**	0.4354
	Pearson	0.2311	3.6243	**0.4820**
	Hybrid	0.3010	3.8396	0.2268
	SAM	0.722	3.700	0.356
	SID	0.4839	3.0839	0.2918
Salinas	**Cosine**	0.3102	**7.3078**	1.1681
	Euclidean	0.3195	5.6815	1.2198
	Tanimoto	**0.2450**	6.7285	**1.2695**
	Pearson	0.3188	6.6501	1.2264
	Hybrid	0.3557	6.5279	1.0193
	SAM	0.367	6.3014	1.1866
	SID	0.3826	5.6342	1.1238
Pavia	**Cosine**	0.7504	**5.8444**	1.5582
	Euclidean	0.7352	5.8436	1.5771
	Tanimoto	**0.7059**	5.8085	**1.6187**
	Pearson	0.7150	5.8375	1.6039
	Hybrid	0.7184	5.7935	1.5504
	SAM	0.7158	5.8019	1.6011
	SID	0.8550	5.9022	1.5249

amount, low intra-band correlations, and high inter-separabilities in the band subset.

A good band selection method is judged by low values of ACC (due to high un-correlation of the selected bands) and high values of ARE & AIE (due to high information content of the selected bands). Based on this premise, Table 2 illustrates the quantitative evaluation results of all the methods on the chosen Hyperspectral Indian Pines datasets. It can be noticed that in all the datasets, Tanimoto measure has the lowest values of ACC, and in most of the cases, Tanimoto has highest values of ARE & AIE. This shows that in most of cases, the Tanimoto similarity measure outperforms the other measures as evident from high information content and low correlation of the selected bands.

We can notice from Table 1 that the classification accuracy for Indian Pines dataset is also highest for Tanimoto measure which further corroborates that this measure can divide the total number of bands into appreciable number of clusters and can also select the appropriate number of bands from each cluster. As expected, Euclidean distance and SID show lowest accuracies at 62.52 and 61.06 percent. Although more or less similar bands are chosen in Euclidean and cosine methods, there is a great difference in their accuracies.

The radiance curve of some classes of Indian Pines is shown in Figure 5.

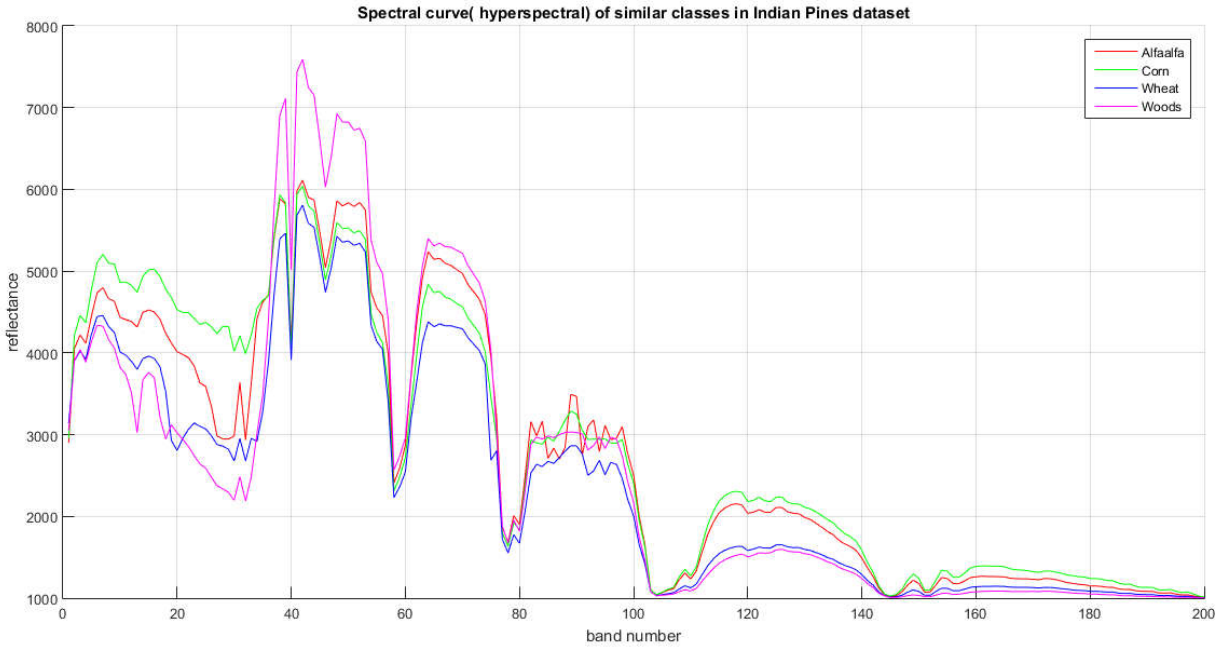


Figure 5: Spectral Reflectance curve of some classes of AVIRIS Indian Pines data set.

Since the classes are highly similar in Indian Pines dataset, as indicated in radiance curves of some of the classes like Corn, Alpha Alpha, Wheat, Woods (Figure 5) (Refer to the legend of the graph), even a small difference in the number and type of bands selected can create a significant difference in the classification accuracies. As most of the bands would give similar information but only a few bands can identify the subtle differences between curves. If any of these bands are not present in the reduced band set, it may affect the overall accuracy of classification. All the similarity matrices show only average performance for classification. The reason behind this is that classes of Indian pines are very extremely similar to each other (Figure 5). It can be seen from here that the classes are extremely similar to each other, thus if insufficient number of bands are selected for classification, subtle differences between these The lasses cannot be fully discovered. This could also be possible reason for poor accuracy for methods like SID & Hybrid. Even good band selection algorithms are not able to find specific bands which can help in distinguishing between these classes, resulting in poor discrimination between classes and lower classification accuracy.

CONCLUSION

This work was conducted to evaluate the performance of different similarity measures in graph-based band selection methods. Various metrics were used to review the performance like ACC, AIE, ARE to measure the information-content and uncorrelatedness of the selected bands. The Tanimoto index was identified to be the best method for similarity calculations because it is a measure of 'commonness', where 'commonness' would include both intra-cluster similarity and inter-cluster dissimilarity. This is very essential as coefficient values, which represent affinities between bands belonging to same cluster, should be strong enough to show a stark difference from bands of different clusters. A good similarity matrix, when used further in Spectral clustering would optimal results for clustering and thus, an optimal partition of high-dimensional space into smaller, low-dimensional subspaces.

REFERENCES

B. Krishna Mohan and Alok Porwal,"Hyperspectral image processing and analysis,"*Current Science,* vol. 108, no.5, 10 march 2015

E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.(CVPR'09)*, Miami, FL, USA, Jun. 20–26, 2009, pp. 2790–2797.

Iosif Vorovencii, "The Hyperspectral Sensors Used In Satellite And Aerial Remote Sensing," *Bulletin of the Transilvania University of Braşo,* vol. 2 (51) , pp.510-56, 2009 .

J. M. Li and Y. T. Qian, "Clustering-based hyperspectral band selection using sparse nonnegative matrix factorization," *J. Zhejiang Univ. Sci. C*, vol. 12, no. 7, pp. 542–549, 2011

M. Basseville "Distance measures for signal processing and pattern recognition",.*European Journal Signal Processing*, 18 (1989), pp. 349-369

M. Honarkhah and J. Caers, "Stochastic simulation of patterns using distance-based pattern modeling," *Math. Geosci*., vol. 42, no. 5, pp. 487-517, 2010

M. Naresh Kumar and M. V. R. Seshasai and K. S. Vara Prasad and V. Kamala and K. V. Ramana and R. S. Dwivedi and P. S. Roy,"A new hybrid spectral similarity measure for discrimination among Vignaspecies",*International Journal of Remote Sensing*,vol.32(14),pp.4041-4053,2011.

N. Kouiroukidis and G. Evangelidis, "*The Effects of Dimensionality Curse in High Dimensional kNN Search*," 2011 15th Panhellenic Conference on Informatics, Kastonia, 2011, pp. 41-45.

S. Jia, "Unsupervised band selection for hyperspectral imagery classification without manual band removal," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., Apr. 2012.

S. Li and H. Qi, "Sparse representation based band selection for hyperspectral images," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP),* Brussels, Belgium, Sep. 11–14, 2011, pp. 2693–2696.

S.-H. Cha."Comprehensive survey on distance/similarity measures between probability density functions". *International Journal of Mathematical Models and Methods in Applied Sciences,* 1 (2007), pp.300-307

Shijin Li, JianbinQiu, Xinxin Yang, Huan Liu, Dingsheng Wan, and Yuelong Zhu. 2014. A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search. *Eng. Appl. Artif. Intell*. 27, C (January 2014), 241-250.

Tanimoto, T.T. (1957) IBM Internal Report 17th Nov. 1957.

Xiao guang Jiang, Lingli Tang ,Changyao Wang and and Cheng Wang, "Spectral characteristics and feature selection of hyperspectral remote sensing data," *International Journal of Remote Sensing*, vol. 25(1),pp. 51–59, 2004.

VonLuxburg, Ulrike,"*A tutorial on spectral clustering"*,Statistics and Computing,vol .17, no.4, pp.395-416,2007.

Wang, Ke; Yong, Bin, "Application of the Frequency Spectrum to Spectral Similarity Measures."*Remote Sens*. 8,no. 4: 344, 2016.

W. Sun, L. Zhang, B. Du, W. Li and Y. Mark Lai, "Band Selection Using Improved Sparse Subspace Clustering for Hyperspectral Imagery Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2784-2797, June 2015.