Evaluation and Selection of the Optimal Normalization Scheme for Hyperspectral Feature Selection Based on Mutual Information

Bhagyashree Chopade
Bhagyashree Chopade
Department of Electronics and Communication,
Technocrats Institute of Technology, Bhopal, Madhya Pradesh, India

Vikas Gupta
Department of Electronics and Communication,
Technocrats Institute of Technology, Bhopal, Madhya Pradesh, India

Abstract- Hyperspectral remote sensing is significantly utilized in numerous applications due to the higher spectral coverage of hyperspectral sensors as compared to multispectral sensors. However, due to the higher dimensionality of hyperspectral data, the computational costs are significantly high which is often detrimental in near real-time or large-scale applications. Thus, feature selection techniques are often used to select the most useful information available from the hyperspectral data. Subsequently, there is a wide range of feature selection methods available in the literature, and several of them utilize concepts based on normalized mutual information. In this paper, we review the various normalization and adjustment schemes for deriving the normalized mutual information. We analyze the potential of these schemes for the ranking of the hyperspectral bands. The selected bands are classified using random forest technique, and the obtained accuracy is analyzed to determine the optimal normalization scheme. The experiments for the proposed study are carried out using three standard hyperspectral datasets that are widely used in the literature. It is observed that the sum normalization scheme shows the highest mean accuracy and is recommended for future application of normalized mutual information for feature selection in hyperspectral data.

Keywords – Feature Selection, Hyperspectral Band Selection, Mutual Information, Supervised Classification, Random Forest

I. INTRODUCTION

Hyperspectral sensors have a significant edge over multispectral sensors considering the larger spectral coverage and are thus, used in a wide range of applications in natural sciences. Some typical applications of hyperspectral data include material or surface characterization [1] and land use land cover classification [2, 3]. The larger dimensionality of the hyperspectral is a primary issue restricting its application from the perspective of higher computational complexity [4]. Thus, dimensionality reduction techniques are widely sought categorized mainly into feature extraction and feature selection. The former approached transforms the original image into a lower-dimensional space, while the latter is based on the selection of the most informative bands [5].

Information similarity measures have been used widely for the selection of best bands in hyperspectral data. The most commonly used measures include Kullback Leibler divergence (KLD) and mutual information (MI) [6–9]. Bajcsy and Groves investigated several statistical and information theory-based parameters for supervised and unsupervised hyperspectral band selection [8]. In an approach by Sarhrouni et al. [10], the inequality of Fano is used to derive the error probability for band selection using MI between the hyperspectral bands and the ground truth. Wang et al. [11] defined a spatial entropy-based approach for hyperspectral band selection. Varade et al. [7] investigated supervised and unsupervised frameworks utilizing MI for the selection of best bands in hyperspectral data. For the supervised case, the MI was determined between the hyperspectral bands and the ground truth, while for the unsupervised case between the hyperspectral bands and the first principal component of the hyperspectral bands. Uso et al. [6] utilized Ward's Linkage strategy using Mutual Information (WaLuMI) and Ward's strategy using KLD (WaLuDi) for hyperspectral band selection. Zhang et al. [12] proposed utilizing the nonlinear correlation coefficient (NCC) to replace some one-dimensional MI components, including the conditional ones for hyperspectral band selection. Amankwah [13] proposed the spatial MI for the selection of best bands. The spatial MI is based on a combination of mutual information and a weighting function based on a dissimilarity metric and hierarchical

clustering. Varade et al. [14] utilized the denoising error response of the hyperspectral bands and the first principal component for the selection of best bands. In another approach, Varade et al. [15] proposed utilizing the Normalized MI (NMI) based on the weighted entropy of hyperspectral band clusters and the first principal component.

In this paper, we assess the different normalization schemes in the literature for the determination of most informative bands in hyperspectral data. Additionally, we also investigate the potential of the various mutual information adjustment schemes for the dimensionality reduction of hyperspectral data. The proposed approach utilizes global mutual information in contrast to the conventional local mutual information. Thus, a new simple strategy is defined for deriving the expectation of mutual information required in the adjustment schemes. The experiments to evaluate the various normalization and adjustment schemes are performed using four state of the art hyperspectral datasets that are widely used in the literature. The potential of the schemes for dimensionality reduction is investigated using supervised classification of the selected bands using random forest technique.

II. MATERIALS AND METHODS

2.1 Experimental datasets

The experiments to investigate the potential of the different normalization and adjustment schemes for mutual information in the context of hyperspectral band selection were carried out using four state of the art datasets described as follows.

- Indian Pines dataset: This dataset comprises 200 bands with a spatial coverage of 145x145 pixels per band at a spatial resolution of approximately 6m in the spectral range of 400 - 2500nm using the Airborne Visible/Infrared Imaging Spectrometer over the Indian Pines test site in North-West Indiana, U.S.A [16].
- Dhundi dataset: This dataset comprises a Fast Line-of-sight Atmospheric Analysis of Hypercubes (FLAASH) corrected Hyperion dataset of 196 bands at 30 m spatial resolution with dimensions of 200x200 per band. Dhundi is located in the lower Indian Himalayas, and subsequently, the Dhundi datasets cover land cover corresponding to the high mountain regions [15].
- Pavia University dataset: This dataset covers semi-urban classes of the Pavia University area from the Reflective Optics System Imaging Spectrometer (ROSIS-3) airborne sensor. The spectral range of ROSIS-3 is between 430 to 860 nm with 115 bands. However, the Pavia University dataset comprises 103 bands as 12 noisy bands were removed. The spatial resolution of this dataset is about 1.3m, and the dimensions per band are 610x340.
- Salinas dataset: This dataset consists of 204 bands at a high spatial resolution of 3.7 m pixels and 512 x 217 pixels per band over Salinas Valley, California acquired by the AVIRIS sensor.

2.2 Background

The mutual information of two random variables is defined using the individual entropies of the random variables and their joint entropy. Consider a discrete random variable A such that it takes on values {a1, a2, a3…., an} with a probability distribution of $P(A)$, then the entropy of A is given as follows.

$$H(A) = \sum_{j=1}^{n} p(a)\ ln\big(p(a)\big) \qquad (1)$$

Now, let us consider another discrete random variable B, defined by the probability distribution of $P(B)$ taking on values {b1, b2, b3…., bn}. If the variables A and B are jointly distributed, their joint probability distribution will be defined as $P(A, B)$ and the joint entropy will be defined as follows.

$$H(A,B)\ = \sum_{j=1}^{n} p(a,b)\ ln\big(p(a,b)\big) \qquad (2)$$

The information similarity or the independence of the variables A and B can be measured by evaluating the MI between these variables as follows.

$$\left.\begin{aligned} I(A,B) &= \sum_{j=1}^{n} p(a,b)\ ln\left(\frac{p(a,b)}{p(a)p(b)}\right) \\ &= H(A)+H(B)-H(A,B) \end{aligned}\right\} \quad (3)$$

where I indicates the MI between the two random variables A and B.

For any two random variables that are statistically dependent, MI>0 and when they are statistically independent, the MI=0. However, MI is not by itself a suitable measure of statistical dependency or independency between two random variables. For example, MI can be low when either A and B present a weak relation, or their entropies are small. Thus, it is a common practice to normalize the mutual information (NMI) [6, 17].

2.3. Normalization and adjustment schemes for mutual information

The normalization of a similarity metric or distance metric is carried out to restrict its range within a fixed interval, typically [0, 1]. There are different normalization schemes possible to constrain the range of MI, such that 1 represents the strongest relationship between the two random variables, and 0 represents their independency. Additionally, the metrics should exhibit the constant baseline property such that its expected values at random samples should remain constant. Typically, the baseline value should remain zero. However, seldom any metrics follow the constant baseline property. Thus, an adjustment scheme was proposed by Vinh et al. [18] to account for this issue. The various schemes for normalized MI (NMI) and corresponding adjustment MI (AMI) are shown in Table 1 [18].

Table 1. The different types of NMI and corresponding AMI variants

	Normalization scheme	Relation
NMI	Joint entropy	$NMI_{jt} = \frac{I(A,B)}{H(A,B)}$
	Maximum of individual entropies	$NMI_{\max} = \frac{I(A,B)}{\max\{H(A),H(B)\}}$
	Sum of individual entropies	$NMI_{sum} = \frac{2I(A,B)}{H(A)+H(B)}$
	Square root of the product of individual entropies	$NMI_{sqrt} = \frac{I(A,B)}{\sqrt{H(A)H(B)}}$
	Minimum of individual entropies	$NMI_{\min} = \frac{I(A,B)}{\min\{H(A),H(B)\}}$
AMI	Maximum of individual entropies	$NMI_{\max} = \frac{I(A,B)-E\{I(A,B)\}}{\max\{H(A),H(B)\}-E\{I(A,B)\}}$
	Sum of individual entropies	$NMI_{sum} = \frac{I(A,B)-E\{I(A,B)\}}{\frac{1}{2}[H(A)+H(B)]-E\{I(A,B)\}}$
	Square root of the product of individual entropies	$NMI_{sqrt} = \frac{I(A,B)-E\{I(A,B)\}}{\sqrt{H(A)H(B)}-E\{I(A,B)\}}$
	Minimum of individual entropies	$NMI_{\min} = \frac{I(A,B)-E\{I(A,B)\}}{\min\{H(A),H(B)\}-E\{I(A,B)\}}$

2.4. Methods

As evident from the previous background, the proposed strategy for investigating the various normalization and adjustment schemes for hyperspectral band selection first requires the determination of the individual and the joint entropies. Varade et al. [7] indicated an alternative approach to the forward search strategy for band selection by utilizing the global mutual information computed between each of the hyperspectral bands and a reference that is the ground truth for the supervised selection and the first principal component in the case of unsupervised selection. Further, in Varade et al. [15], the first principal component was shown to produce relatively better classification results as compared to the best band of hyperspectral datasets. In the proposed approach, as shown in Figure 1, we follow a similar strategy. Thus, the individual entropy for a particular band (X) amongst the N bands of hyperspectral

data (X) and the joint entropy of X with the first principal component PC1 which is selected as reference R. Then the mutual information between the band X and the reference R is shown in equation 4.

$$\left.\begin{aligned} I(X,R) &= \sum_{j=1}^{N} p(x,r)\ ln\left(\frac{p(x,r)}{p(x)p(r)}\right) \\ &= H(X)+H(R)-H(X,R) \end{aligned}\right\}; \quad (4)$$

$$H(X) = -\sum_{j=1}^{N} p(x)\ ln\big(p(x)\big),\ H(X,R) = -\sum_{j=1}^{N} p(x,r)\ ln\big(p(x,r)\big)$$

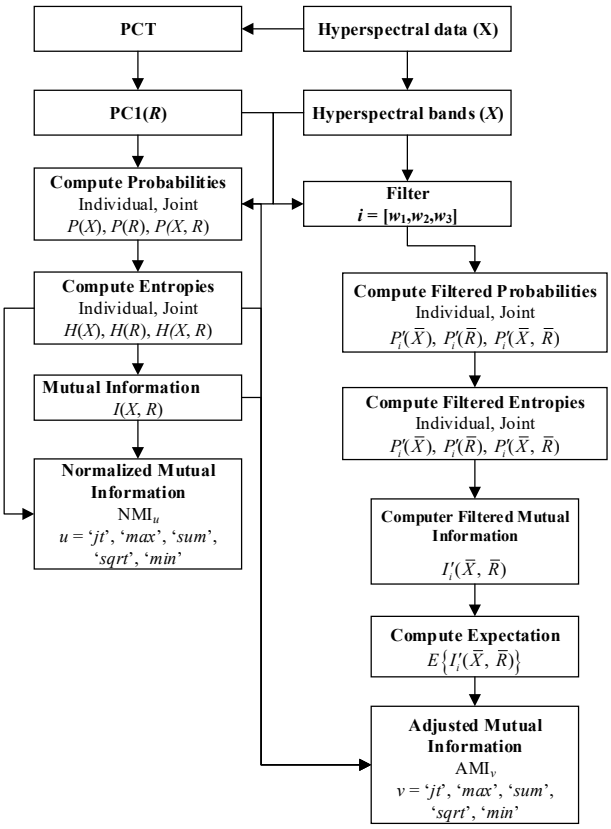


Figure 1. Workflow for the proposed approach for the computation of the normalized mutual information and the adjusted mutual information.

In equation 4, x and r represent the pixels values taken by the particular hyperspectral band and the first principal component. Based on the individual probabilities $H(X)$ and $H(R)$, the joint probability $H(X, R)$, and the mutual information $I(X, R)$, the various normalized variants of mutual information defined in Table 1 are computed. For the computation of the adjusted mutual information, we define a new strategy for the determination of the expectation of mutual information. Typically, the expectation of mutual information should be computed from the random samples collected from X, and R. Such a strategy does not work when we consider classes in the ground truth corresponding to very few pixels. This creates a problem in maintaining constant baseline property. Since the constant baseline property should be localized, considering the various classes, we define a simple strategy utilizing a moving average filter with three different window sizes. The objective here is to compute multiple instances of mutual information for a local mean filtered image with different neighbourhood sizes. The same process is applied to the reference image, and then the mutual information is computed for each of the neighbourhood sizes. The mean of the derived multiple mutual information is used as the expectation of mutual information. For simplicity and lower run time, we restrict the number of filter iterations to 3 with neighbourhood sizes 3, 5 and 7 corresponding to $w1$, $w2$ and $w3$.

III. Results and Discussion

The different variants of the adjusted mutual information are computed using the expressions shown in Table 1. The different variants of NMI and AMI are used separately to rank bands such that the highest value indicates the best

band in each of the variants. The subsequent best bands are selected by identifying the local maxima of a particular NMI/AMI variant for that variant. To evaluate the potential of the different variants of NMI/AMI for the selection of the best bands, we perform supervised classification using a random forest classifier with 100 trees and 20% training. A comparison for each of the case is carried out by evaluating the classification accuracy determined by the Kappa coefficient, as shown in Figures 2 and 3.

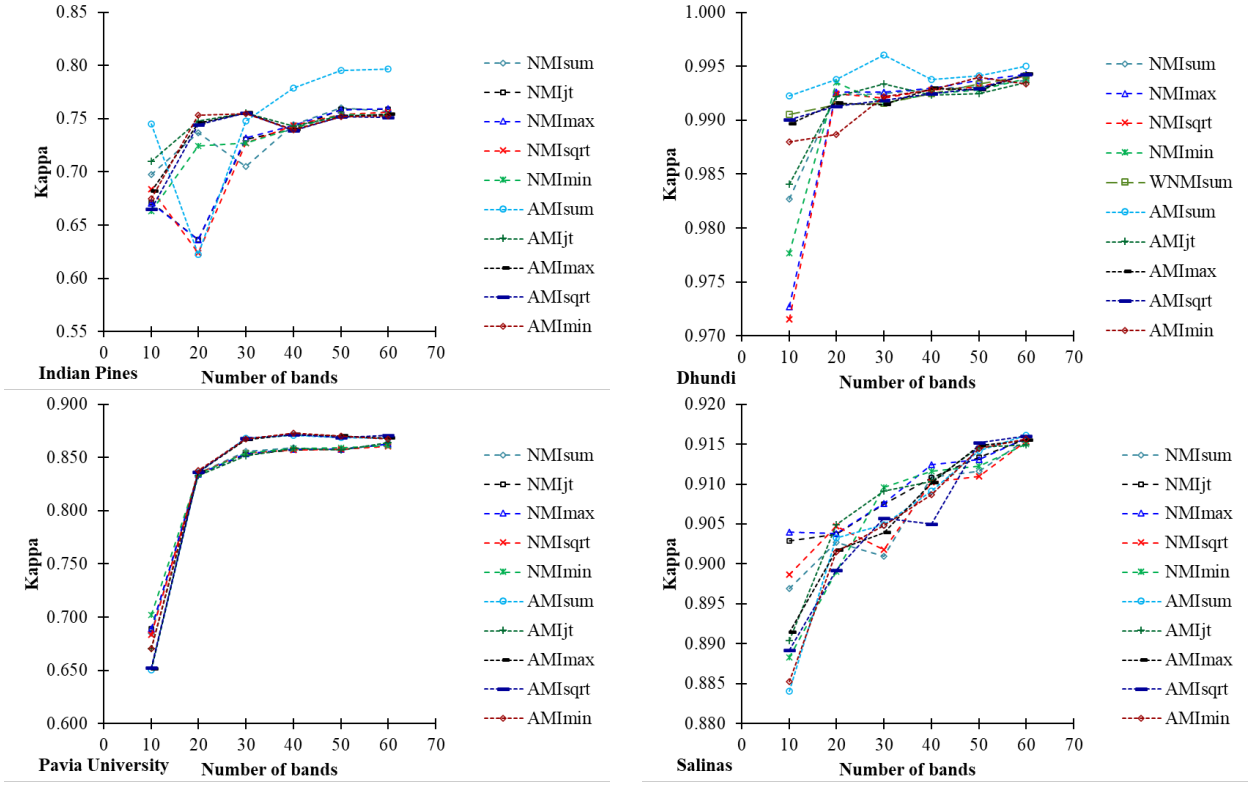


Figure 2. Classification accuracy for the different variants of mutual information with respect to the number of bands for the four test datasets at 20% training data supplied to the random forest classifier.

In Figure 2, it is observed that for each of the variants, the classification accuracy increases as the number of selected bands used are increased. Similarly, in Figure 3, expectedly, the classification accuracy increases as the percentage of training samples increase. It is worth mentioning that the training samples for classification are selected randomly, and the Kappa coefficient for each of the cases referring to the number of bands and the training sample volume is computed as the mean of Kappa coefficient values observed for four iterations. In general, the AMI based variants result in significantly better classification accuracy as compared to the NMI variants. Figure 4 (a) and (b) illustrates the mean accuracy for the two cases shown in Figures 2 and 3. It is observed that the mean Kappa coefficient for 10, 20, 30, 40, 50 and 60 bands at 20% training is the highest for the AMI_{min} followed by AMI_{jt} which is 0.9109 and 0.9107, respectively. For the case of 20 bands with training sample volumes varying between 15, 30, 45, 60, 75 and 90, the mean Kappa coefficient of 0.8665 for the AMI_{sum} is the best. For this case, the AMI_{min} and the AMI_{jt} follow next in terms of the relatively higher Kappa coefficient. Overall, it is observed that the AMI_{min} shows the best accuracy over the mean of the two cases, as mentioned above, as shown in Figure 5. Table 2 summarizes the mean Kappa coefficient values for the two cases. Overall from the analysis, it was observed that the AMI_{min} and the AMI_{jt} not only showed relatively higher performance for band selection but were also consistent as they observed the best mean Kappa for two datasets each. In case of the variants of the NMI, the best normalization scheme was observed to be NMI_{sum}. This is observed in agreement with the normalization scheme used for band selection in Uso et al. [6] and Varade et al. [14, 15].

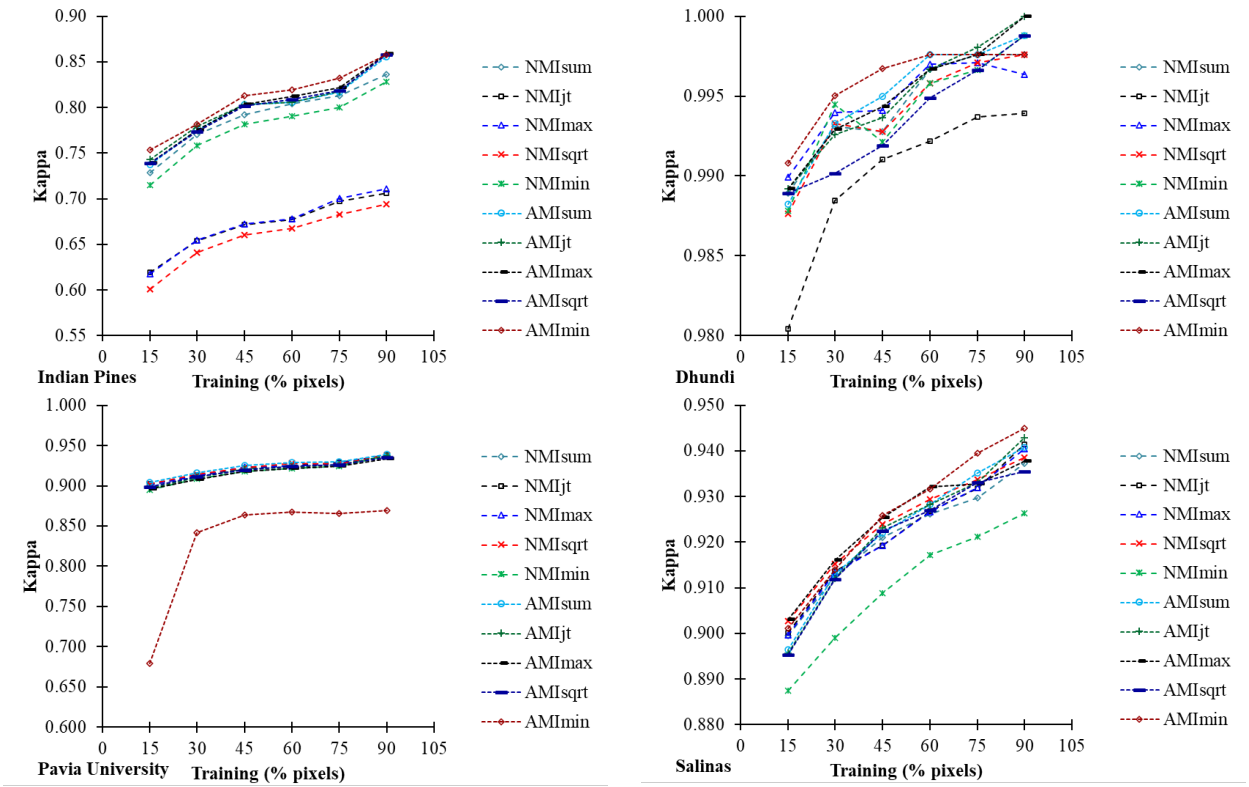
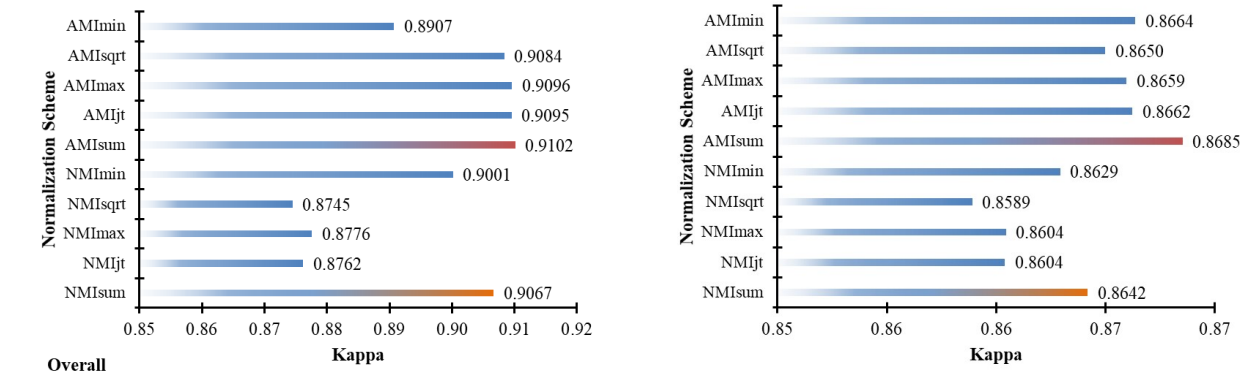


Figure 3. Classification accuracy for the different variants of mutual information with respect to the volume of training data for the four test datasets for 20 selected bands using the random forest classifier.



(a) (b)

Figure 4. Mean Kappa coefficients for the four test datasets for the different variants of mutual information with respect to the number of selected bands and the volume of training data in (a) and (b). The mean Kappa coefficient from (a) and (b) is shown in (c).

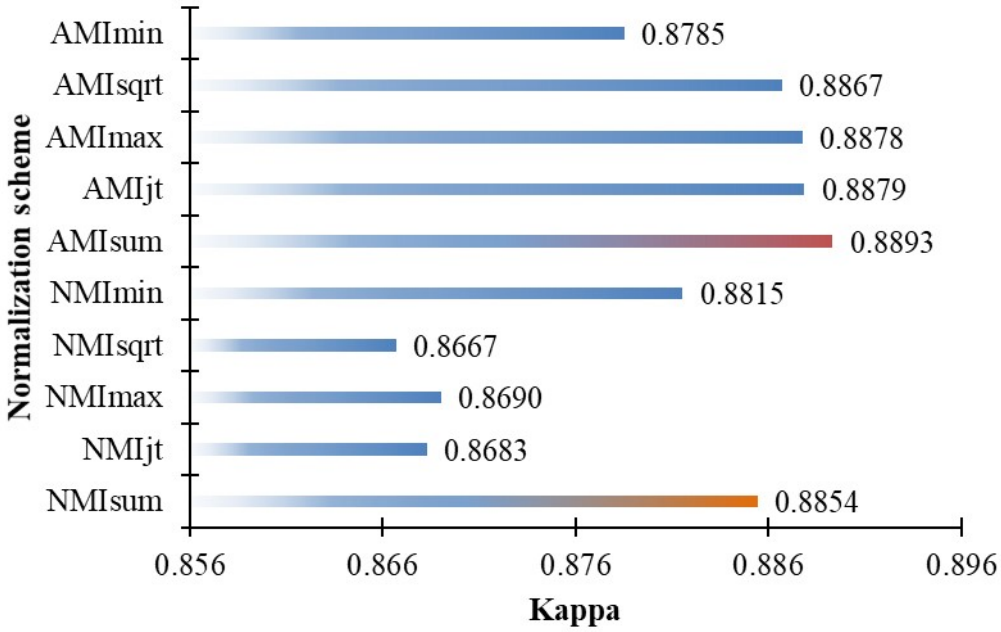


Figure 5. The mean Kappa coefficient from (a) and (b) is shown in (c).

Table 2. Kappa coefficients values corresponding to Figure 4 and 5.

	Fixed Training at 20%				**Fixed number of selected bands 20**				**Overall**
Scheme	Indian Pines	Dhundi	PaviaU	Salinas	Indian Pines	Dhundi	PaviaU	Salinas	Mean
NMI_{sum}	0.7906	0.9944	0.9206	0.9211	0.7335	0.9912	0.8255	0.9064	0.8854
NMI_{jt}	0.6709	0.9900	0.9217	0.9222	0.7159	0.9912	0.8256	0.9090	0.8683
NMI_{max}	0.6723	0.9947	0.9214	0.9219	0.7167	0.9898	0.8258	0.9095	0.8690
NMI_{sqrt}	0.6577	0.9940	0.9223	0.9239	0.7147	0.9893	0.8246	0.9070	0.8667
NMI_{min}	0.7789	0.9943	0.9175	0.9100	0.7272	0.9903	0.8282	0.9059	0.8815
AMI_{sum}	0.7991	0.9952	0.9211	0.9128	0.7416	0.9923	0.8270	0.9053	0.8868
AMI_{jt}	0.8014	0.9951	0.9239	0.9226	0.7439	0.9913	0.8224	0.9074	0.8885
AMI_{max}	0.8017	0.9950	0.9192	0.9225	0.7381	0.9921	0.8274	0.9062	0.8878
AMI_{sqrt}	0.7999	0.9951	0.9170	0.9245	0.7347	0.9921	0.8280	0.9051	0.8871
AMI_{min}	0.8098	0.9936	0.9192	0.9209	0.7378	0.9915	0.8310	0.9051	0.8886

IV. Conclusion

Mutual information and techniques based on mutual information have been widely used in the literature for feature selection in numerous areas, including medical imaging, land use land cover analysis, etc. Normalization of mutual information is significant to impart the essential qualities of a metric for the determination of the dependency between two random variables. In this study, we investigated the optimal normalization scheme for mutual information from the perspective of feature selection in hyperspectral data. We further examined the AMI as a band ranking parameter for hyperspectral feature selection. Towards this end, we performed experiments with four state of the art hyperspectral datasets. The experimental results revealed that the best bands selected using the different variants of AMI outperform those selected by NMI in terms of the classification accuracy. It was observed that the most appropriate information similarity measures for hyperspectral band selection include the AMI_{min} and the AMI_{jt}.

Both these measure were significantly better than NMI_{sum}, which was determined to be the best variant of NMI for hyperspectral band selection.

REFERENCES

[1] D. Varade, A. K. Maurya, and O. Dikshit, "Development of spectral indexes in hyperspectral imagery for land cover assessment," IETE Technical Review, pp.1–9, 2018.

[2] D. Varade, A. Sure, and O. Dikshit, "Potential of Landsat-8 and Sentinel-2A composite for land use land cover analysis," Geocarto International, Vol. 10, pp.1–16, 2018.

[3] A. Vali, S. Comai, and M. Matteucci, "Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review," Remote Sensing, Vol. 12, No. 15, pp.2495, 2020.

[4] D. Landgrebe, 2002, "Hyperspectral Image Data Analysis as a High Dimensional Signal Processing Problem," Special Issue of the IEEE Signal Processing Magazine, Vol. 19, pp.17–28.

[5] P. K. Varshney and M. K. Arora, Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data: SPRINGER, 2010.

[6] A. Martinez-Uso, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-Based Hyperspectral Band Selection Using Information Measures," IEEE Trans. Geosci. Remote Sensing, Vol. 45, No. 12, pp.4158–4171, 2007.

[7] D. Varade, A. K. Maurya, A. Sure, and O. Dikshit, "Supervised classification of snow cover using hyperspectral imagery," in 2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT), Dehradun, India, 2017, pp.1–7.

[8] P. Bajcsy and P. Groves, 2004, "Methodology for Hyperspectral Band Selection", Photogrammetric Engineering and Remote," Sensing Journal, Vol. 70, pp.793–802.

[9] B. Guo, S. R. Gunn, R. I. Damper, and J.D.B. Nelson, "Band Selection for Hyperspectral Image Classification Using Mutual Information," IEEE Geosci. Remote Sensing Lett., Vol. 3, No. 4, pp.522–526, 2006.

[10] E. Sarhrouni, A. Hammouch, and D. Aboutajdine, "Band selection and classification of hyperspectral images using mutual information: An algorithm based on minimizing the error probability using the inequality of Fano," IEEE International Conference on Multimedia Computing and Systems (ICMCS)., pp.155–159, 2012.

[11] B. Wang, X. Wang, and Z. Chen, 2012, "Spatial Entropy Based Mutual Information in Hyperspectral Band Selection for Supervised Classification," International Journal of Numerical Analysis & Modeling, Vol. 9, No. 2, pp.181–192.

[12] M. Zhang, Q. Wang, Y. Shen, and B. Zhang, "Hyperspectral Feature Selection Based on Mutual Information and Nonlinear Correlation Coefficient," in 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto, Japan, 2009, pp.965–968.

[13] A. Amankwah, "Spatial mutual information based hyperspectral band selection for classification," TheScientificWorldJournal, Vol. 2015, pp.630918, 2015.

[14] D. Varade, A. K. Maurya, and O. Dikshit, "Unsupervised hyperspectral band selection using ranking based on a denoising error matching approach," International Journal of Remote Sensing, 2019.

[15] D. Varade, A. K. Maurya, and O. Dikshit, "Unsupervised Band Selection of Hyperspectral Data Based on Mutual Information Derived from Weighted Cluster Entropy for Snow Classification," Geocarto International, 2019.

[16] M. Baumgardner, L. Biehl, and D. Landgrebe, "220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3," 2015.

[17] T. M. Cover and J. A. Thomas, Elements of information theory, 2nd ed. Hoboken, N.J.: Wiley; Chichester : John Wiley, 2006. [Online]. Available: http://www.loc.gov/catdir/enhancements/fy0624/2005047799-d.html

[18] N. X. Vinh, J. Epps, and J. Bailey, 2010, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," Journal of Machine Learning Research, Vol. 11, No. 95, pp.2837–2854. [Online]. Available: http://jmlr.org/papers/v11/vinh10a.html