Cross Correlation Based Clustering for Feature Selection in Hyperspectral Imagery

Hüseyin Çukur, Hamidullah Binol, Faruk Sukru Uslu, Yusuf Kalaycı, Abdullah Bal

Department of Electronics and Communications Engineering
Yildiz Technical University, 34220 İstanbul, Türkiye
{hcukur, hbinol, bal}@yildiz.edu.tr, fuslu@tekok.edu.tr, yusuf.kalayci@stu.khas.edu.tr

Abstract

One of the main problems with hyperspectral image processing is to be contained large amount of data. Furthermore, pattern recognition methods are highly sensitive to problems related to high dimensional feature spaces. Therefore, feature selection in hyperspectral remote sensing data is investigated by researchers. This paper propose a clustering strategy that divides a feature set into subsets within which features are closely related to each other by means of cross correlation between all spectral bands. After that a band selection strategy based on Minimum Redundancy Maximum Relevance (mRMR) eliminates redundant bands into band clusters. The effectiveness of the proposed method is carried out on a real hyperspectral data set. The obtained results clearly affirm the superiority of the proposed method.

1. INTRODUCTION

Nowadays, with the evolution of the remote sensing technology and hyperspectral sensors, the utilization of hyperspectral image is gradually spread [1]. It has hundreds of narrow and contiguous bands. Although these bands ensure greater detail and more information about different objects, computational cost of data processing is affected negatively. Moreover, higher correlation amid relative bands enhances the redundancy amid them. This phenomenon is called "curse of dimensionality" [2]. In this case, the classification accuracy first increases and then decreases with the increase of spectral bands , whereas training samples are remained the same. In order to struggle these issues, a tolerable large sample size is adjusted via feature reduction.

Feature reduction is carried out by two disparate techniques that are feature selection and feature extraction. Contrary to feature extraction, feature selection techniques preserve the physical meaning of feature unchanged. Even though a large training sample is also available, feature selection may still be more advantageous. In literature, a broad assortment of feature selection methods has been handled to hyperspectral data [3, 4]. In this study, minimum Redundancy Maximum Relevance (mRMR) is chosen as a feature selection technique. The mRMR is a principally fast feature selection method for detecting a set of both related and complementary features. It is a similarity based method that manages mutual information (MI) to define the relevance between the features [5]. Although mRMR has high speed and excessive achievement, dependency between bands decreases the performance of mRMR. Therefore, a clustering between bands adapted to mRMR to have better feature selection.

Clustering techniques allocate a feature set into subdivisions within which features are firmly linked to each other. In case that the collected features can be divided into such groups, similar characteristics are existed in the same clusters and hereby we need to keep only one feature in each group. In recent works, Conditional Entropy, Mutual Information, Euclidian Distance, Maximum Absolute Distance, and Centered Euclidian Distance are studied by researchers [6]. In this experiment, we have used the cross correlation as a clustering algorithm. Cross correlation is utilized to evaluate how much similar each of two random variables. If the correlation between two bands is high, it means that they are independent and they have to be in the same clusters. In this study we have proposed a new technique for bands regroup by finding the highly correlated groups of bands in the hyperspectral data cube based on cross correlation matrix. In this manner, we run the mRMR into each cluster one by one and we expect the more effective performance than conventional technique.

The paper is outlined as follows. Section 2 ensures mathematical description that describe the mRMR and cross correlation algorithms, as well as the implementation of these two methods together for dimensionality reduction. Section 3 describes, carries out, and analyzes the results of classification tests that illustrate the effectiveness of the methods presented in this study. Finally, conclusion is drawn in Sec. 4.

2. METHODS

2.1. mRMR

Minimum Redundancy Maximum Relevance (mRMR) is a similarity based method that handles mutual information to define the relevance between the features [5]. Mutual information is a similarity measurement of how much one random variable tells us about another random variable. Given two random variables X and Y with marginal probability distribution $p_1(x)$ and $p_2(y)$ and joint probability distribution $p(x,y)$, $x \in X$, $y \in Y$, MI between X and Y defined as:

$$I(X;Y) = \sum \sum p(x,y) \log\left(\frac{p(x,y)}{p_1(x)*p_2(y)}\right) \quad (1)$$

The principle of mRMR is similar to the MI. It selects the features that are independent from each other and provides greatest dependency on the target class. This method select a feature f_i amongst not selected features f_s that maximizes $(u_i - r_i)$, where u_i is the relevance of f_i to the class c alone and

r_i is the mean redundancy of f_i to each of the already selected features. u_i and r_i can be defined as via MI:

$$u_i = \frac{1}{|f|} \sum_{f_i \in f} I(f_i; c) \quad (2)$$

$$r_i = \frac{1}{|f|^2} \sum_{f_j \in f} I(f_i; f_j) \quad (3)$$

The mRMR combines the two criteria and tries to maximize $(u_i - r_i)$ to attain maximal relevance and minimal redundancy. In this study we use mRMR to decide the maximal relevance and minimal redundancy between hyperspectral bands. As a result to dimension reduction, the performance of mRMR is observed and noted via classification method. Additionally, after second part we utilize mRMR again and we expect the higher classification score than single mRMR.

2.2. Cross Correlation

In statistics, the correlation expresses the strength and direction of a relationship between two random variables. [7]. The mathematical formula of cross correlation between two variables *x* and *y*:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i.y_i - nxy)}{\sqrt{\sum_{i=1}^{n}(x_i^2 - nx^2)}\sqrt{\sum_{i=1}^{n}(y_i^2 - ny^2)}} \quad (4)$$

The value of r_{xy} is such that $-1 < r < +1$. The + and – signs are applied for positive correlations and negative correlations, respectively. If x and y have a strong positive correlation, r is close to $+1$. An r value of exactly $+1$ indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase. If x and y have a strong negative correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease. If there is no correlation or a weak correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables. Correlation values greater than 0.8 are generally described as strong, whereas correlation values less than 0.5 as weak.

3. RESULTS

In this section, we submit and compare the experimental results acquired by applying our proposed techniques, and then commentate its effectiveness.

3.1. Experimental Setup

3.1.1. Hyperspectral Data

Kennedy Space Center Data (KSC) is one of the most widely used hyperspectral image in the literature was first applied in this study. This data set was gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Kennedy Space Center (KSC), Florida, on March 23, 1996 [8]. The data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. Also it consists of 512x614 pixels and 224 spectral reflectance bands in the wavelength range of 0.4–2.5 μm. After removing water absorption and low signal to noise (SNR) bands, 176 bands were utilized for the analysis. A three band image and online available ground-truth map are given in Fig. 1.

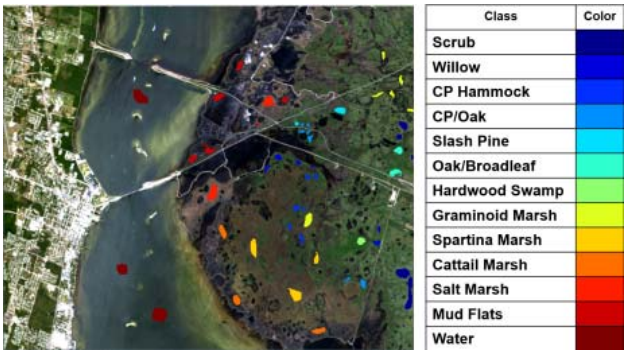
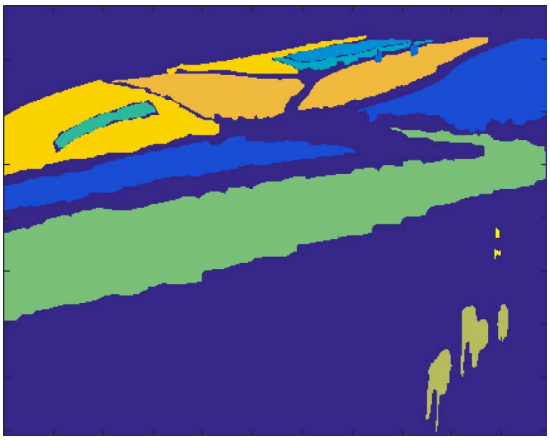


Fig. 1. Colored ground reference of KSC

The second data set is Catalca data gathered by the SPECIM sensor over the Catalca, Tekirdag, Turkey on June 7, 2015. It consists of 810x1091 pixels and 196 spectral reflectance bands in the wavelength range of 0.4–1 μm. The RGB image and ground-truth map are given in Fig. 2.

(a)



(b)

Fig. 2. A part of Catalca01 data (a) RGB Color image, (b) Ground truth

3.1.2. Classification Method

SVM [9] is selected as classifier method in this study. SVM is a modern classifier that applies kernels to construct linear classification boundaries in higher dimensional spaces. It classifies data into two groups by building a hyperplane. We practise the LIBSVM package [10], which supports both 2-class and multiclass classification. The radial basis function (RBF) is also used as the kernel function. The pixels from every 13 classes on KSC and every 10 classes on Catalca01 are randomly separated into 10% and 90% as the training and testing data, respectively. Number of training and test samples are listed in Table 1.

Table 1. Training and testing set sizes for the SVM classification experiment

KSC			Catalca01		
Class	Train	Test	Class	Train	Test
C1	76	685	C1	6768	60920
C2	24	219	C2	49	445
C3	25	231	C3	413	3726
C4	25	227	C4	333	3002
C5	16	145	C5	451	4066
C6	22	207	C6	14512	130612
C7	10	95	C7	898	8083
C8	43	388	C8	3867	34806
C9	52	468	C9	4648	41833
C10	40	364	C10	25	233
C11	42	377			
C12	50	453			
C13	92	835			

3.2. Experimental Results

We first find the dissimilarity matrix of hyperspectral data sets using cross correlation. According to dissimilarity matrix, the bands which have higher correlation have to be in the same cluster. Based on this information we divide bands four clusters on KSC and two classes on Catalca01. Followed by the determination of clustering region, all clusters do not have the same band number. Therefore, we decide to define selection criteria according to clusters total bands. If any cluster has much more bands, it also has bigger selection criteria. It means bigger clusters are represented by much more individual. Band intervals and selection criteria based on total band number is presented in table 2 and table 3..

In Table 2 and Table 3 clusters obtained from two data are shown.

Table 2. Bands grouping and selection criteria on KSC

Group number	1	2	3	4
Bands interval	1 - 32	33 - 97	98 - 131	132 - 176
Total band	32	65	33	45
Selection criteria	$\frac{32}{176}$	$\frac{65}{176}$	$\frac{33}{176}$	$\frac{45}{176}$

Table 3. Bands grouping and selection criteria on Catalca01

Group number	1	2
Bands interval	1 - 105	105 - 196
Total band	105	91
Selection criteria	$\frac{105}{196}$	$\frac{91}{196}$

After the separate bands into clusters, mRMR algorithm is applied. To compare the effectiveness of proposed method, SVM has been used to obtain the classification accuracy for both proposed method and single mRMR. The results can be shown in Fig. 3 and 4. The classification results are given in table 4 and 5 for two datasets respectively. The results without feature selection are also given in table 4 and table 5. As shown in tables the classification accuracy is significantly improved based on our proposed feature selection technique. Additionally, we have seen that classification accuracy decreases over 40 bands for KSC and 20 bands for Catalca01 due to the Hughes phenomenon.

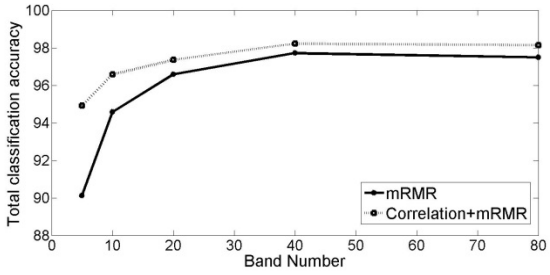


Fig. 3. Overall evaluation of classification performance for mRMR and proposed method on KSC

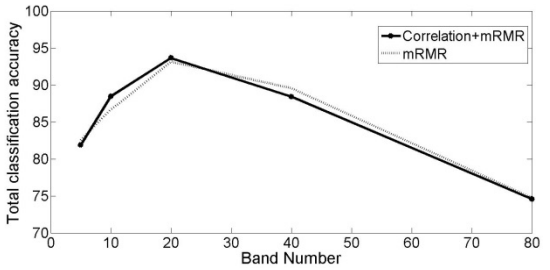


Fig. 4. Overall evaluation of classification performance for mRMR and proposed method on Catalca01

Table 4. SVM accuracy comparisons of all KSC classes between mRMR and cross correlation + mRMR

Class name	Full Band	mRMR					Cross Correlation + mRMR				
		Number of Bands					Number of Bands				
		5	*10*	*20*	*40*	*80*	*5*	*10*	*20*	*40*	*80*
Scrub	94,306	95,457	96,524	97,377	97,142	96,374	96,758	97,377	97,356	**97,697**	97,292
Willow swamp	97,505	92,962	93,261	98,401	98,273	98,486	97,867	98,401	**98,784**	98,571	98,656
Cabbage palm hammock	98,166	91,896	94,135	96,588	98,443	98,273	95,329	97,483	97,356	**98,955**	98,806
Cabbage palm/oak hammock	95,969	87,673	90,36	94,029	95,628	95,777	86,308	90,232	94,412	95,756	**96,353**
Slash pine	97,419	79,484	88,036	94,263	96,694	96,438	82,96	90,531	96,033	97,547	**97,654**
Oak/broadleaf hammock	95,948	84,069	90,531	93,517	95,884	95,649	90,872	93,901	94,86	95,628	**96,417**
Hardwood swamp	98,87	92,877	96,929	97,803	98,891	99,147	96,268	98,699	98,912	**99,083**	98,87
Graminioid marsh	94,263	90,254	95,884	95,628	96,652	96,055	91,256	94,434	95,202	**98,059**	97,867
Spartina marsh	95,756	94,562	95,415	95,948	97,014	97,057	94,093	96,247	96,225	98,059	**98,528**
Cattail marsh	93,367	82,214	89,934	96,737	96,972	96,289	93,965	95,074	97,419	**98,294**	97,505
Salt marsh	93,325	98,742	98,848	99,168	98,699	97,803	99,062	**99,552**	98,72	98,464	97,825
Mud flats	94,583	79,953	94,519	96,46	97,505	**97,867**	92,088	94,54	96,46	97,654	97,718
Water	99,787	90,68	95,479	96,78	99,744	99,936	**100**	99,957	99,893	99,979	**100**

Table 5. SVM accuracy comparisons of all Catalca01 classes between mRMR and cross correlation + mRMR

Class name	Full Band	mRMR					Cross Correlation + mRMR				
		Number of Bands					Number of Bands				
		5	*10*	*20*	*40*	*80*	*5*	*10*	*20*	*40*	*80*
Wheat straw	78,826	69,444	75,267	86,234	83,001	78,826	70,351	79,091	**88,289**	85,716	78,833
Vehicles	99,847	99,492	99,802	99,847	99,847	99,847	99,76	**99,857**	99,854	99,847	99,847
Processed wheat spike	98,71	98,112	98,689	99,597	99,791	98,707	97,959	99,75	**99,927**	99,861	99,625
Ploved wheat fields	98,961	88,591	98,137	99,034	98,957	98,957	94,779	97,539	**99,068**	98,957	98,957
Dry bushes	98,589	90,023	97,191	99,263	**99,555**	98,943	94,525	98,71	99,308	98,672	98,592
Wet bushes	55,546	87,732	90,83	**95,79**	91,991	62,266	81,826	90,333	95,634	87,148	61,963
Donkeys	97,191	98,981	99,433	**99,482**	98,564	97,57	98,905	99,346	99,013	97,667	97,191
Wheat bales	87,903	58,522	69,413	87,903	89,265	87,903	72,302	78,618	87,854	**89,46**	87,903
Green sunflower	85,459	99,552	99,016	96,437	87,246	85,483	**99,858**	99,461	97,327	90,59	85,459
Human	99,924	99,541	99,91	99,924	99,924	99,924	99,691	99,903	**99,924**	99,924	99,924

4. CONCLUSION

In this paper, we proposed a novel method combining mRMR with cross correlation for feature selection in hyperspectral imagery. We add the cross correlation as preprocessing step into mRMR algorithm to improve the feature selection performance. The proposed method initially can obtain an optimal subset of features via cross correlation. The optimal subset of features is then utilized for band selection. Selected bands used in both training and testing for optimal outcomes in the classification.. As a conculision, it has been shown that the novel proposed approach not only improves the classification accuracy and but also reduces the computation consumption.

ACKNOWLEDGMENT

This research was supported by a grant from The Scientific and Technological Research Council of Turkey (TUBITAK - 112E207).

The authors wish to thank members of the YTU-YAZGI Laboratory team for HSI Dataset.

REFERENCES

[1] P.K. Varshney, M.K. Arora, “Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data”, 2nd ed., Springer Verlag, Berlin, 2004.

[2] G. F. Hughes, “On the mean accuracy of statistical pattern recognizers,”IEEE Trans. Inf. Theory, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[3] T. Kavzoglu and P. M. Mather, “The role of feature selection in artificial neural network applications,” Int. J. Remote Sens., vol. 23, no. 15, pp. 2787–2803, Aug. 2002.

[4] S. B. Serpico and L. Bruzzone, “A new search algorithm for feature selection in hyperspectral remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[5] H. Peng, F. Long and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, August 2005.

[6] W. Wang, Z. Zhao, and H Zhu, "Hyperspectral Image Compression Method Based on Spectral Statistical Correlation," 2nd International Congress on Image and Signal Processing, Tianjin, china, pp. 1 – 5, Oct. 2009.

[7] D. Manolakis, R. Lockwood, and T. Cooley, "On the Spectral Correlation Structure of Hyperspectral Imaging Data," IEEE International Geoscience and Remote Sensing Symposium, Boston, MA USA, pp.581-584, July 2008.

[8] J. Ham, Y. Chen, M.M. Crawford, J. Ghosh, “Investigation of the random forest framework for classification of hyperspectral data”, IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 492–501, March 2005.

[9] J. A. Gualtieri and R. F. Cromp, “Support vector machines for hyperspectral remote sensing classification,” Proc. SPIE, vol. 3584, pp. 221–232, 1998.

[10] C.C. Chang, C.J. Lin. (2015, July 20). LIBSVM a library for support vector machines. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.