Improved feature selection based on a mutual information measure for hyperspectral image classification

Conference Paper · July 2012

DOI: 10.1109/IGARSS.2012.6350780

Dr Md Ali Hossain
UNSW Australia

IMPROVED FEATURE SELECTION BASED ON A MUTUAL INFORMATION MEASURE FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Md. Ali Hossain, Xiuping Jia, Mark Pickering

School of Engineering and Information Technology, University of New South Wales, Canberra, Australia

ABSTRACT

Hyperspectral images contain a large amount of information which presents a major challenge for efficient classification. In this paper the information content of each spectral band is analyzed and an improved feature selection technique is proposed for the minimization of dependent information while maximizing the relevancy based on normalized mutual information (NMI). Experimental results are provided for comparisons among some relevant and recent methods for hyperspectral feature selection in terms of their classification accuracy using real hyperspectral images.

Index Terms— Hyperspectral image, feature selection, normalized mutual information, curse of dimensionality.

1. INTRODUCTION

Hyperspectral imaging consists of acquiring a scene with numerous and contiguous spectral bands centered on uniformly distributed wavelengths in the visible to near infrared spectrum [1]. For instance, the AVIRIS sensor simultaneously measures 224 bands with a fine resolution of $0.01\mu m$. However, the availability of this large amount of data presents some complex methodological problems for supervised image classification procedures. For example, the classification cost increases with the number of features used to describe the pixel vectors in hyperspectral space. On the other hand, features are highly correlated and much of the data does not add inherent information content for a particular application. Ideally each feature (spectral band) used in the classification process should add an independent set of information to provide accurate image classification. Another common problem often noted in classification accuracy of hyperspectral data is the Hughes phenomenon [2]. The key characteristics of the phenomenon is, assuming a fixed set of training samples, the rate of increase in accuracy declines and eventually accuracy starts to decrease as more features are added. Various approaches could be adopted for the efficient classification of hyperspectral data such as effectively increasing the training set size by the application of some dimensionality reduction procedure prior to the classification analysis [1, 3]. Feature reduction for hyperspectral data can be achieved by feature extraction or feature selection which could be linear or nonlinear and supervised or unsupervised [4, 5]. Feature extraction aims to project the original data set onto adequate subspaces chosen for their ability to explain the data. Sometimes feature extraction techniques such as PCA [6], LDA [7] etc. are avoided so as to retain the original hyperspectral wavelength information. Ranking and selecting the most informative features from the original feature space is called feature selection. Usually feature selection is easy to implement but the selection criteria is critical. Some popular feature selection criteria such as the Bhattacharya and J-M distance measures have their limitations and drawbacks including strong dependence on the training data, ineffective class pair wise treatment and only being reliable for normally distributed class data [8]. Alternatively mutual information is a non-parametric measure, that works for non Gaussian data and is effective for handling multi-class cases without the need for individual class pair evaluations [9]. In [10] MI is used between the input class data and spectral images as a feature selection criterion. However this paper did not consider the interaction among the selected and candidate features. The MIFS [11] and mRmR [12] approaches have been introduced for multiple feature selection. However, the selected features suffer from several weaknesses. Because the MI measured between the two images depends strongly on the original entropy of the images, it is not possible to properly rank the features based solely on the relative MI values. So an improvement based on normalized MI is proposed in this paper in order to make sure only informative features are selected.

This paper introduces a modified mutual information based feature selection method (m-MIFS) which uses normalized MI instead of conventional MI. An adaptive weighting factor and a minimum relevancy criteria for the feature independence check are added to remove the risk of noisy features being selected.

2. MUTUAL INFORMATION BASED CRITERION

2.1. Mutual information

The mutual information (MI) between two input variables **X** and **C** measures the general correlation between them and can

978-1-4673-1159-5/12/$31.00 ©2012 IEEE

be defined as:

$$I(\mathbf{X},\mathbf{C}) = \sum_{c\epsilon\mathbf{C}}\sum_{x\epsilon\mathbf{X}} p(x,c) log \frac{p(x,c)}{p(x)p(c)} \quad (1)$$

Where $p(x)$, $p(c)$ are the marginal and $p(x,c)$ is the joint probability of x and c. The MI value in eq. (1) is used as a selection criterion to identify the most relevant features where $\mathbf{X}$ is the input spectral image and $\mathbf{C}$ is the input class data. The candidate feature $\mathbf{X}_i$ having the highest value of MI is selected as the best feature for classification [13].

$$I(\mathbf{X}_i,\mathbf{C}) > I(\mathbf{X}_j,\mathbf{C}) \quad for \;\; j \neq i, \;\; \{i = 1, 2, ...N\}$$

For multiple feature selection Battiti [11] developed a greedy search strategy called MIFS with the adoption of one dimensional MI evaluation to maximize the relevance and minimize the redundancy between selected features. This method considers a set of already selected features and chooses the next feature as the one that maximizes the mutual information between the input class data and candidate features after correction by subtracting a quantity proportional to the average MI of the candidate feature with the already selected features. Let $\mathbf{S}$ be the set of k already selected features. The selection of the $(k+1)^{th}$ feature is based on the following measure:

$$D(\mathbf{X}_i,\mathbf{C}) = I(\mathbf{X}_i,\mathbf{C}) - \beta \sum_{\mathbf{s}\epsilon\mathbf{S}} I(\mathbf{X}_i,\mathbf{s}_j), \, \mathbf{X}_i \notin \mathbf{S} \quad (2)$$

Where the first term measures the degree of information contained in $\mathbf{X}_i$ with reference to the input class data and the second term measures the redundancy between the candidate feature and the already selected features. β is a user-defined parameter that regulates the relative importance of the redundancy. If $\beta = 1$ then the average redundancy is eliminated. But if $\beta = 0$ then only the 1^{st} term will be used and highly dependent features will be selected and the class discrimination power will not be changed.

3. PROPOSED FEATURE SELECTION APPROACH

3.1. Normalized mutual information

Mutual information I has a limitation when used as a feature selection criterion. This limitation arises because I can be low due to the two features $\mathbf{X}_i$ and $\mathbf{X}_j$ having a weak relationship or because the marginal entropies of these two variables are high. Thus it is convenient to define a measure which is independent from the marginal entropies. Figure-1 illustrates this limitation of the MI measure where the MI value is measured between the input class data and all the spectral bands. It can be seen that MI could be any value, so it is difficult to apply as an absolute metric for similarity or dissimilarity. However, normalized MI [14, 15] can be used to solve this problem by normalizing the range of values to the range [0,1]. Normalized MI is defined as follows:

$$\hat{I}(\mathbf{X},\mathbf{C}) = \frac{I(\mathbf{X},\mathbf{C})}{\sqrt{I(\mathbf{X},\mathbf{X})}\sqrt{I(\mathbf{C},\mathbf{C})}} \quad (3)$$

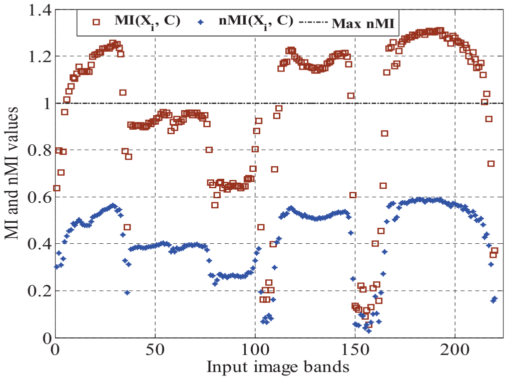


Fig. 1. MI and normalized MI values between the class data and the original image bands

This limitation of MI makes the performance of MIFS in eq. (2) unreliable for feature selection and normalized MI is proposed as a reliable feature selection criterion in this paper. After normalized MI is introduced the modified MIFS method becomes:

$$\hat{D}(\mathbf{X}_i,\mathbf{C}) = \hat{I}(\mathbf{X}_i,\mathbf{C}) - \frac{\alpha}{|S|} \sum_{\mathbf{s}\epsilon\mathbf{S}} \hat{I}(\mathbf{X}_i,\mathbf{s}_j), \, \mathbf{X}_i \notin \mathbf{S} \quad (4)$$

The value of the first term in eq. (4) is in the range of [0, 1] measuring the relevancy of the candidate feature and the value of the second term is also in the same range, measuring the average redundancy of the candidate feature. The candidate feature which gives the highest difference is selected as the next best feature. A different weighting factor α for the 2^{nd} term is also used in each feature selection step which influences the already selected feature when choosing the next best candidate feature.

3.2. Minimum relevancy criterion

Although the method described in eq. (4) improves the feature selection performance over the MIFS approach, these two methods still have a risk of selecting noisy features when the redundancy dominates the relevancy. This may also be the results of two small values when the selected features are weakly related to the target and the already selected features. Therefore a constraints needs to be imposed so that only the candidate features having high minimum relevancy can compete for selection. The following constraint is applied to remove the risk of selecting noisy features.

$$\hat{I}(\mathbf{X}_i,\mathbf{C}) < T \quad remove \, \mathbf{X}_i \quad \{i = 1, 2, ...N\}$$

where, T is a user defined threshold, corresponding to the minimum relevancy criterion. This criterion also significantly reduces the feature search space by discarding the irrelevant candidate features which fails to satisfy the minimum relevancy criterion.

4. RESULTS

4.1. Experimental procedure

To assess the performance of the proposed approach a comparison has been made, using real hyperspectral images, between the proposed method and two other relevant methods taken from the recent literature. The proposed normalized MI is measured for two cases; one is without interaction between the candidate and the already selected features (NMI-WtFI) and the other is with interaction to avoid redundant features being selected. An adaptive weight is also used in each step of the feature selection step to provide a more optimal feature set. The new order of the selected features is shown in table-1 for each algorithm compared. For each method the top five ranked features were used for classification using a maximum likelihood classifier.

Table 1. NMI-WtFI, MIFS and m-MIFS ranked features

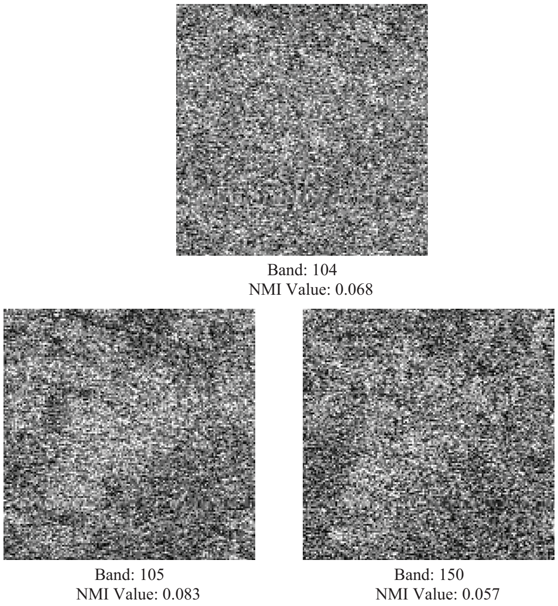
Methods	Order of selected feature				
	1st	2nd	3rd	4th	5th
NMI-WtFI	184	183	187	182	191
MIFS	193	65	31	170	2
m-MIFS	184	29	169	67	139

Table 2. Training and Testing samples in each class

Class name	Training Samples	Testing samples
Hay-windrowed	165	150
Soybean-notill	180	150
Woods	279	248
Wheat	63	63
Grass-trees	96	88
Soybean-min	204	204
Grass-pasture	108	90
Corn-notill	130	104

4.2. Performance evaluation

The proposed m-MIFS method is applied to a real hyperspectral image which was acquired on June 12, 1992 by the AVIRIS sensor in Northwestern Indiana over the Indian Pines test site [16]. The performance of the m-MIFS is compared with two other methods NMI-WtFI and MIFS described in eq. (2). For each method the overall classification accuracy is plotted in figure-3 with respect to the subset of k features selected. Table-2 lists all the input classes with the number of training and testing samples used in this experiment.


Band: 104
NMI Value: 0.068

Band: 105
NMI Value: 0.083

Band: 150
NMI Value: 0.057

Fig. 2. Noisy image bands identified after applying the minimum relevancy criterion

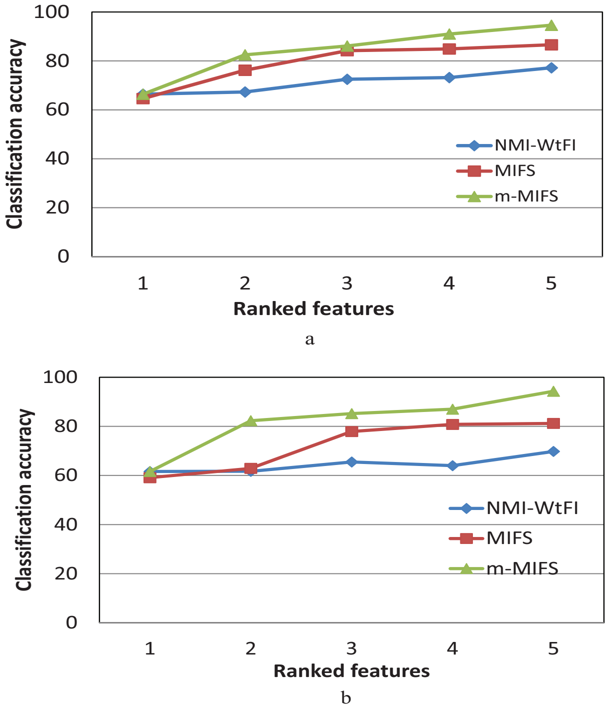


Fig. 3. Classification accuracy of a. Training data and b. Test data

It can be seen from figure-3 that the classification accuracy of m-MIFS increases the most as more features are added which is due to the improvement in class discrimination power. On the other hand NMI-WtFI gives the poorest performance because it selects features which are highly correlated and the discrimination power is not improved much. Figure-2 shows some noisy image bands which fail to achieve the minimum relevancy criterion and are removed by the proposed approach. A clear improvement of classification accuracy can be observed from the above result where the proposed m-MIFS approach performs much better in all cases than the MIFS and NMI-WtFI approaches. This is because the normalized MI measure has the ability to select features more reliably. Another reason is that the adaptive weighting is used in each feature selection step which enables the features which have already been selected to influence the selection of the next feature.

5. CONCLUSIONS

In this paper we have demonstrated the benefits of applying normalized MI rather than conventional MI for feature selection to provide improved classification accuracy. The proposed approach can identify features which have more relevance to the desired input classes. The proposed method can be used as a practical tool for supervised hyperspectral image classification.

6. REFERENCES

[1] J.A. Richards and Xiuping Jia, *Remote Sensing Digital Image Analysis, 4th Edition*, Springer-verlag Berlin Heidelberg, Germany, 2006.

[2] G.F. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Transactions on Information Theory*, vol. IT-14 no. 1, pp. 55–63, January 1968.

[3] P. F. Hsieh and D. A. Landgrebe, “Classification of high dimensional data,” *PhD Thesis and School of Electrical and Computer Engineering Technical Report*, vol. TR-ECE 98-4, pp. 40–42, May 1998.

[4] J. Yang, P. Yu, and B. Kuo, “A nonparametric feature extraction and its application to nearest neighbour classification for hyperspectral image data,” *IEEE Transaction on Geoscience and Remote Sensing*, vol. 48 no. 3, pp. 1279–1293, March 2010.

[5] S. Tajudin and D.A. Landgrebe, “Covariance estimation with limited training samples,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2113–2118, July 1999.

[6] L. Ying, G. Yanfeng, and Z. Ye, “Hyperspectral feature extraction using selective pca based on genetic algorithm with subgroups,” in *Innovative Computing, Information and Control, 2006. ICICIC ’06. First International Conference*, 2006, Pages 652 -656.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York: Academic, 1990.

[8] L. Bruzzone, F. Roli, and S. B. Serpico, “An extension of the jeffreys-matusita distance to multiclass cases for feature selection,” *IEEE Transaction on Geoscience and Remote Sensing*, vol. 33 no. 6, pp. 1318–13121, Nov. 1995.

[9] Cover T.M. and J.A. Thomas, *Elements of Information Theory, Second Edition*, 2006.

[10] C.Conese and F. Maselli, “Selection of optimum bands from tm scenes through mutual information analysis,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 48(3), pp. 2–11, 1993.

[11] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, September 1994.

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

[13] K. Torkkola, “Feature extraction by non-parametric mutual information maximization,” *Journal of Machine Learning Research 3*, vol. 3, pp. 1415–1438, March 2003.

[14] B. Guo, R.I. Damper, S. R. Gunn, and J.D.B. Nelson, “Normalized mutual information feature selection,” *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, February 2009.

[15] M. Fauvel, J. Chanussot, and J.A. Benediktsson, “Kernel principal component analysis for classification of hyperspectral remote sensing data over urban areas,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. Art. 783194, pp. 1–14, 2005.

[16] D. Landgrebe, “https://engineering.purdue.edu/ biehl/ multispec/hyperspectral.html,” .