The CSI Journal on
Computer Science and Engineering
Vol. 11, No. 2 & 4 (b), 2013
Pages 31-42
Regular Paper

Spectral-Spatial Graph Kernel Machines in the Context of Hyper spectral Remote Sensing Image Classification

Mostafa Borhani Hassan Ghassemian

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

Abstract

The concept of spectral-spatial graph kernel machinesis proposed in this paper to expand capabilities of the currently implemented classification of hyper spectral images. The frame work is investigated to handle one of the most important challenges of hyper spectral remotely sensed classification especially when the performance is affected by Hughes phenomenon i.e. the curse of dimensionality will be arises when using traditional procedure to classify hyper spectral data.This novel method is concentrated on spatial graph kernel machines, grouping methods and non-uniform distribution informationfor hyper spectral image classification. The innovation of this work consists in: 1) introducing novel weighted spectral-spatial kernel, 2) computing graph kernel by taking into account the non-uniform distribution of the spatial-spectral information, 3) extending some grouped multivariate analysis methods to nonlinear kernel based version and 4) clarifying simultaneous spectral spatial graph kernels theoretical relationships.From empirical results, we conclude that the novel proposed grouped kernel approach meaningfuly enhances the classification performance, it greatly improves the classification overall and per class accuracies and it also provides classification maps with more homogeneous regions particularly in terms of limited training samples.

Keywords: Spatial Graph Kernel, Spectral-Spatial Hyper Spectral Classification, Probabilistic SVM, Non-Uniform Distribution in Formation, Multivariate Discriminate Analysis, Grouping Methods.

1. Introduction

Hyper spectral sensors instantaneously capture hundreds of thin spectral images from a comprehensive electromagnetic range. This high dimensionality data represents several major challenges in image classification [1]. The dimensionality of hyper spectral image intensely affects performance of classic classification methods [2] [3]. Tolerating the Hughes phenomenon requires the cautious design of novel algorithms that are intelligent to handle hundreds of such spectral bands to minimize the effects from the “curse of dimensionality”. Integration of spatial and spectral information may improve classification results. Lately, many methods are focussed on using spatial information to improve spectral-based classifiers, in the remotely sensed data [4] [5]. Adding information about spatial structures into classifier improves the classification and simultaneously increases the dimension of feature vector.

Nonlinear approaches are less sensitive to the data’s dimensionality and have higher accuracies in many machine learning applications. Spectral-spatial kernels for hyper spectral image classification, e.g., generalized composite [6] [7], morphological [8], and graph kernels [9] [10], were introduced newly for the upgrading of the SVM classifier. The kernel machines have shown good results in terms of accuracies for classifying hyper spectral images.

This paper extends linear or nonlinear feature extraction and dimension reduction methods to grouped kernel form in hyper dimensional data analysis. We also study modern methods to make kernel based grouped Multi Variate discriminative Analysis (MVA) [11] more appropriate to hyper spectral image classification [12].

One of the contributions of this paper is focused on methods which decrease over-segmentation in a hyper spectral image. It is realized by habitually “true labelling”, the expressive spatial structures before execution on outstanding-controlled segmentation. A significant influence contains in investigating probabilistic classification results for choosing the most consistently classified pixels as repress enter samples of spatial areas. Several repress enter selection methods are planned, using either individual classifiers, or a multiple classifier (MC) system. Then, different tactics for outstanding-based classifier, using either probabilistic SVM [13] [14] or Multiple Spatial Spectral Classifiers (MSSC) followed with Minimum Spanning Forest (MSF) [15] as a spatial graph kernel. The proposed classification using two planned repress enter selection processes and spatial graph kernel followed by uncompelled majority voting as decision fusion is discussed in section 3.The details of the grouped kernel based method for spectral-spatial classification is described using non-uniform distributed information that is obtainable in last part of this section. Further in the rest of this paper, section 4 argues weighted spatial-spectral kernels, then, proposed classification methods are examined and presentedby real remote sensing dataset in section 5. Finally, conclusions are drawn in section 6.

2. Kernel Grouped Multivariate Analysis

In this section, a grouping method is projected as a dimension reduction approach and then, it uses to extend the linear Canonical Correlation Analysis (CCA) to kernel based grouped CCA as a sample of kernel based Grouped MVA methods. There are so many grouping methods but we follow the easiest way in this paper. For a given dataset $\{(x_i, y_i)\ i = 1{:}N\}$, the grouping algorithm computes the mean such as Equ. (1) and the covariance matrix such as Equ. (2), where Tmeans the transpose of a vector.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (1)$$

$$\widehat{\Sigma}_x = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\,(x_i - \bar{x})^T \quad (2)$$

The entries are sorted by mean values and collected in H groups. Then, the process leads to calculate the mean and weighted covariance matrix of grouped data as Equ. (3) and Equ. (4), respectively, where n_h is the number of elements in h-thgroup.

$$\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_i \quad (3)$$

$$\widehat{\Sigma}_W = \frac{n_h}{N}\sum_{h=1}^{H}(\bar{x}_h - \bar{x})\,(\bar{x}_h - \bar{x})^T \quad (4)$$

The last covariance is derived form the mean of groups and the total mean of elements. The rest of algorithms are alike. Using of unbiased covariance formula is straight forward.

CCA is frequently used for two underlying correlated datasets. Consider two independent and identically distributed random variables of input data, x_1 and x_2. Classical CCA discovers the linear combination of the variables which maximize correlation between the collections. Let

$$y_1 = w_1 x_1 = \sum_j w_{1j}\, x_{1j} \quad (5)$$

$$y_2 = w_2 x_2 = \sum_j w_{2j}\, x_{2j} \quad (6)$$

The CCA solves challenge of finding values of w_1 and w_2 which maximize the correlation between $y_1$and y_2, with constraint the solutions be a finite solution.Let has mean μ_1 and μ_2 are the mean of x_1 and x_2 and $\widehat{\Sigma}_{11}$, $\widehat{\Sigma}_{22}$, $\widehat{\Sigma}_{12}$ are meaning of auto covariance of x_1, auto covariance of x_2 and covariance of x_1 and x_2 respectively. The standard statistical method leads toEqu. (7). Grouped CCA (GCCA) customs the Equ. (4) for computing the covariance of grouped data and K is intended as Equ. (8).

$$K = \widehat{\Sigma}_{11}^{-\frac{1}{2}}\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-\frac{1}{2}} \quad (7)$$

$$K = \widehat{\Sigma}_{W11}^{-\frac{1}{2}}\widehat{\Sigma}_{W12}\widehat{\Sigma}_{W22}^{-\frac{1}{2}} \quad (8)$$

Grouped CCA then makes a Singular Value Decomposition (SVD) of K

$$K = (\alpha_1, \alpha_2, \dots, \alpha_k)D(\beta_1, \beta_2, \dots, \beta_k) \quad (9)$$

where α_i and β_i are the eigenvectors of Karush–Kuhn–Tucker (KKT) conditionsand Kuhn–Tucker (KT) conditions respectively and D is the diagonal matrix of eigenvalues. The canonical correlation vectors of CCA are given by Equ. (10) and Equ. (11) and also the canonical correlation vectors of GCCA are derived from Equ. (12) and Equ. (13).

$$w_1 = \widehat{\Sigma}_{11}^{-\frac{1}{2}}\alpha_1 \quad (10)$$

$$w_2 = \widehat{\Sigma}_{22}^{-\frac{1}{2}}\beta_1 \quad (11)$$

$$w_1 = \widehat{\Sigma}_{W11}^{-\frac{1}{2}}\alpha_1 \quad (12)$$

$$w_2 = \widehat{\Sigma}_{W22}^{-\frac{1}{2}}\beta_1 \quad (13)$$

The proposed GCCA can be transformed to the feature space by nonlinear kernel machines. Kernel methods are the modern novelty approaches. Kernel based Support Vector Machines (SVMs) makes a nonlinear mapping of the data set into high dimensional feature space (possibly infinite dimension). The covariance matrices in feature space can be defined by Equ. (14) for $i = 1, 2$ and covariance matrices of grouped data can be calculated by Equ. (15) where $\varphi(.)$ is the nonlinear one-to-one, onto, and inverse functions.

$$\widehat{\Sigma}_{\varphi ij} = \frac{1}{N}\sum_{i=1}^{N}\big(\varphi(x_i) - \varphi(\bar{x})\big)\left(\varphi(x_j) - \varphi(\bar{x})\right)^T \quad (14)$$

$$\widehat{\Sigma}_{W\varphi ij} = \frac{n_h}{N}\sum_{h=1}^{H}\big(\varphi(\bar{x}_{ih}) - \varphi(\bar{x})\big)\left(\varphi(\bar{x}_{jh}) - \varphi(\bar{x})\right)^T \quad (15)$$

w_1 and w_2 can be expressed in the feature space as

$$w_1 = \sum_{i=1}^{2}\sum_{j=1}^{M}\alpha_{ij}\,\varphi(x_{ij}) \quad (16)$$

$$w_2 = \sum_{i=1}^{2}\sum_{j=1}^{M}\beta_{ij}\,\varphi(x_{ij}) \quad (17)$$

where α_i and β_i are the eigenvectors of $K = \hat{\Sigma}_{W\varphi 11}^{-\frac{1}{2}} \hat{\Sigma}_{W\varphi 12} \hat{\Sigma}_{W\varphi 22}^{-\frac{1}{2}}$ with KKT and KT conditions respectively for Kernel Grouped CCA (KGCCA).The rest of KGCCA procedure is alike KCCA.

Several MVA methods such as Principal Component Analysis (PCA), Partial Least Squares (PLS), Orthogonal Partial Least Squares (OPLS), CCA, NMF (Non-Negative Matrix Factorization) and Entropy Component Analysis (ECA) in linear, kernel and kernel grouped manners are examined in this paper.

Figure 1 shows the projections obtained in the toy problem by linear and proposed methods. Input data was normalized to zero mean and unit variance. An artificial two-class problem with the RBF kernel is used for this simulation.

The performance of kernel based grouped MVA methods for classification of hyper spectral images is examined by some experiments and their results are showed in figure 2 and figure 3 for various numbers of train samples. We use class 2 and 3 of Indian Pines dataset [16] as input data. Classification among these major classes is very difficult, which has made the scene a challenging benchmark to validate classification precision of hyper spectral imaging algorithms. Overall accuracy is depicted vs.

number of prediction for various feature extraction methods such as PCA, PLS, OPLS, CCA, MNF, KGPCA, KGPLS, KGOPLS, KGCCA, KGMNF and KGECA. Simulations results verify that the utilizing of proposed approach improves the overall accuracy especially in kernel grouped CCA in spite of CCA. Average accuracy of different classification approaches with various MVA approaches, for multiclass Indian Pines dataset [16] summarized in figure 4.

3. Spatial Graph Kernel Machine

The incorporation extra spatial knowledge into spectral information in kernel machines improves hyper spectral classification accuracy. We propose the use of outstanding points and graph kernels to obtain an accurate spectral-spatial classification map. The proposed approach that selects an indicator for each spatial object in the image is shown in figure 5 essentially it has three significant blocks included outstanding point selection, construction of a spatial graph kernel and optional decision fusion approach.

3.1. Outstanding Points' Selection

The outstanding points' selection strategy is presented in this section. The proposed approach uses probabilistic SVM outcomes [14] to select the most reliable pixels as repress enter points. The selection approach can be represented as followed steps for hyper spectral image:

The first step contains of the use of a pixel wise SVM classifier which is extremely well suited to classify hyper spectral images. This step results a classification map (where each pixel contains a label of the class) and a probability map (if a particular pixel is assigned to the class).

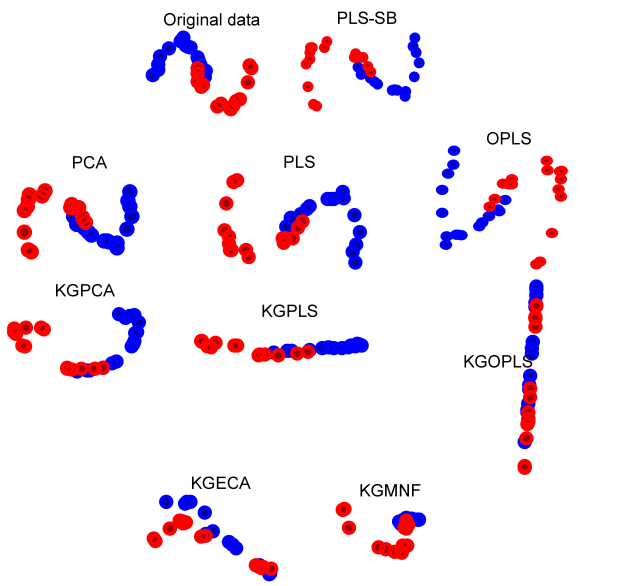


Figure 1. Score of various linear multivariate discriminative analyses, kernel based grouped multivariate analysis

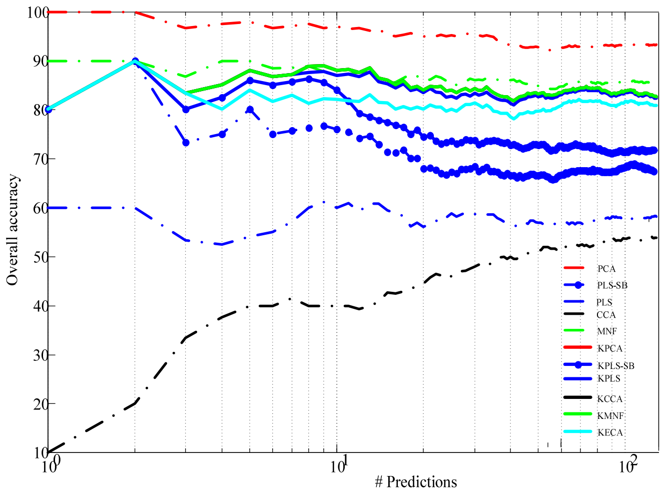


Figure 2. Feature extraction: PCA, PLS, OPLS, CCA, MNF, KGPCA, KGPLS, KGOPLS, KGCCA, KGMNF, KGECA, Train Sample=16;

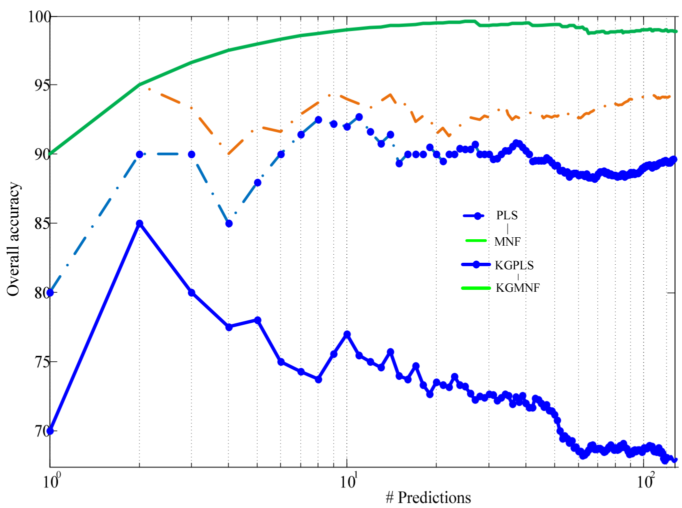


Figure 3. Feature extraction methods: PLS, MNF, KGPLS, and KGMNF, Train Sample=144

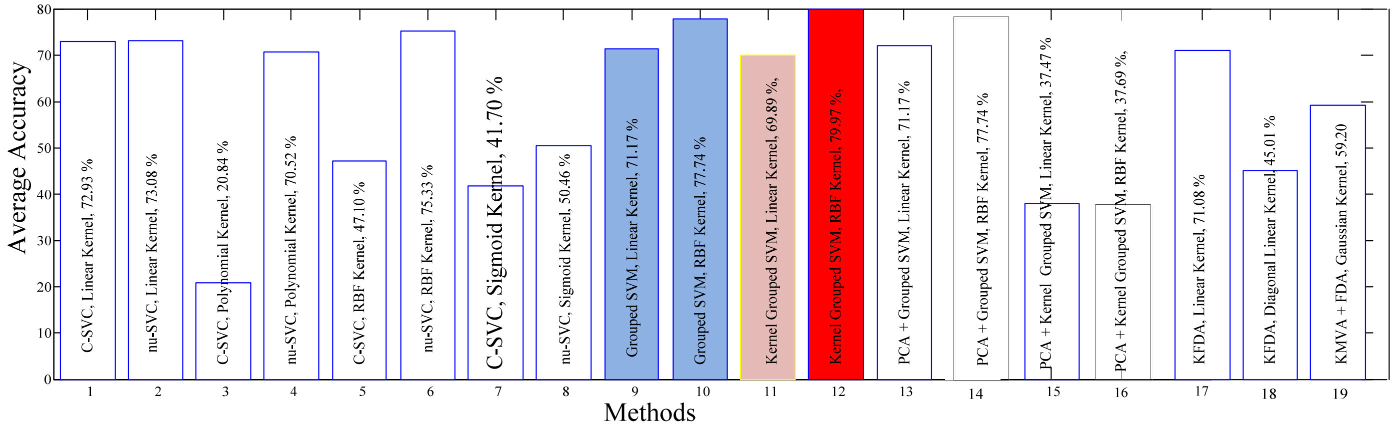


Figure 4. Average accuracy of different classification approaches, Indian Pines dataset, 10 classes, 64 train samples. 1. C-SVC, Linear Kernel, 72.93%, 2. nu-SVC, Linear Kernel, 73.08%, 3. C-SVC, Polynomial Kernel, 20.84%, 4.nu-SVC, Polynomial Kernel, 70.52%, 5. C-SVC, RBF Kernel, 47.10%, 6.nu-SVC, RBF Kernel, 75.33 %, 7. C-SVC, Sigmoid Kernel, 41.70 %, 8.nu-SVC, Sigmoid Kernel, 50.46 %, 9. Grouped SVM, Linear Kernel, 71.17 %, 10. Grouped SVM, RBF Kernel, 77.74 %, 11. Kernel Grouped SVM, Linear Kernel, 69.89 %, 12. Kernel Grouped SVM, RBF Kernel, 79.97 %, 13. PCA+Grouped SVM, Linear Kernel, 71.17 %, 14. PCA+Grouped SVM, RBF Kernel, 77.74 %, 15. PCA+Kernel Grouped SVM, Linear Kernel, 37.47 %, 16. PCA+Kernel Grouped SVM, RBF Kernel, 37.69 %, 17. KFDA, Linear Kernel, 71.08 %, 18. KFDA, Diagonal Linear Kernel, 45.01 %, 19. KMVA+FDA, Gaussian Kernel, 59.20%

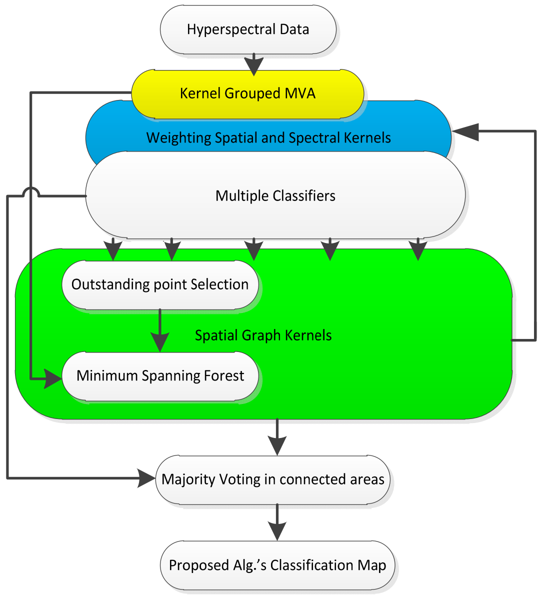


Figure 5. The block diagram of proposed approach

k, the probability map is a probability estimate for this pixel to actually belong to the class k). There are some parts of the image with a high probability of the correct classification and other parts where classification results are less reliable. Analysis of the obtained classification and probability maps is investigated to for choose the most reliable classified pixels to define suitable outstanding points as follow:

- First, a connected components labelling of the classification map is done.
- Then, each connected component is studied:
  - If it is big, it is considered as a relevant region and some indicators can be assigned in this component with the highest probabilities.
  - If it is small, we investigate if its pixels are classified with probability estimation higher than a threshold or not. If yes, the area represents an indicator. The method is robust to thresholds' selection.

The disadvantage of the presented technique is that the selection of outstanding points strongly depends on the performances of the selected probabilistic pixel wise classifier. Our next objective is to moderate this dependency. This can be realized by using not a single classification procedure for repress enter choosing, but rather multiple classifiers. Thus, we propose a new repress enter selection method based on a multiple classifiers, consists of two steps:

- First, several independent classifiers are used.
- Then, an indicator map is created: For each pixel, if all the classifiers reach agreement, the pixel is considered as reliably classified and it is reserved in the map of outstanding points.

Furthermore, to include spatial information in the repress enter selection procedure, the suggested method consists of the succeeding steps:

- First, unsupervised image segmentation is done. Segmentation approaches based on different principles must be taken. We investigate the use of the three procedures: watershed segmentation [19], segmentation by Expectation Maximization (EM) [20] and segmentation using the Hierarchical SEGmentation (HSEG) [21]. Then, Pixel wise classification is applied.
- Then, each of the obtained unsupervised segmentation maps is fussed with the pixel wise classification map using the majority voting. For every area in the segmentation map, all the pixels are assigned to the most common class inside this area. Different segmentation approaches based on different principles lead to diverse classification results. The spectral-spatial classifier produces more accurate classification map, when compared to pixel wise techniques.
- Finally, we make a map of outstanding points by choosing pixels assigned by all the classifiers to the same class. The next step is to use the acquired map of outstanding points for constructing spatial graph kernels.

3.2. Construction of a Spatial Graph Kernel

Here, we discuss the graph-based approach, which involves the building of a Minimum Spanning Forest (MSF), where each tree is rooted on one outstanding point.

First, we map an image onto a graph.

- Each pixel is considered as a vertex of an undirected graph. Each edge of this graph connects a couple of vertices corresponding to the closest samples.
- Furthermore, a weight is allocated to each edge, which shows the degree of unlikeness between two pixels connected by this edge.

And since we have a map of outstanding points, each recognized pixel is associated with the corresponding true marked pixel.

In this paper we use the MSF as a spatial graph kernel. Given a G graph, a MSF rooted on m vertices is:

- non-connected graph without cycles,
- contains all the vertices of G,
- consists of connected sub graphs, each sub graph (a tree) contains one root
- The sum of the edges' weights of this forest is minimal.

In order to obtain the MSF rooted on our outstanding points, m additional vertices corresponding to m outstanding points are introduced (one extra vertex for one agent).

The method of the construction of a MSF is a region growing process, which contains of the following phases:

- First, m roots are selectedas the nodes of forest.
- Then, at each iteration, an edge of adapted graph with the minimal weight is chosen such that one vertex adjacent to this edge is in the forest, and another is not.
- So this pixel and edge would be added to the forest (In other words, at each iteration a new pixel is added to the segmentation map, so the dissimilarity criterion between this pixel and one of the pixels already belonging to the map is minimal.).
- The procedure goes to the next iteration until all the vertices belong to the forest.
- Finally, we assign a class of each labelled point with high accuracy potential to all the pixels grown from this true labelled.

3.3. Post-Processing

An optional post-processing step is proposed to make the classification scheme more robust. The obtained spectral-spatial classification map is labelled by using four-proximity area connectivity to get a homogenous map, then, for each connected component, all the pixels are assigned to the most frequent class when investigating a pixel wise classification map. This post-processing scheme is shown in figure 5 as majority voting option. The achieved classification maps by this method contain more homogeneous regions, when compared to a pixel wise classification map.

4. Weighted Spatial-Spectral Kernels

Hyper spectral sensors achieve signal in a widespread range, and it is expected that different parts of the spectrum have different capability to separate features of interest. In some parts of the spectrum, the echo spectrum may be distinctive from other parts of the spectral band. In addition, the complex conditions of transfer such as water and CO2 in the atmosphere play a major role in this phenomenon.

To isolate the features, some statistical metrics such as mean and standard deviation for each class are investigated in different frequency bands. If we ignore the secondary and higher statistical criteria (use the mean difference between the two classes), it can be seen by taking out the meaning leads to losing large amounts of information, therefore, the better measurement of statistical methods to isolate that, is known as the Bhattacharya distance (use the secondary statistical measure such as variance) is used.

Some set of spectral bands may have more useful information for classification than other spectral bands. If separable criterion be considered as an estimation of the probability of correct classification, in this case it will be expected that the classifier accuracy can be improved output by focusing on the gangs that have further information.

This part of paper focuses on the classifier design based on prior knowledge integration. The classification based on SVM can be directly modifying by assigning different weights to different features, the weights can be adaptive and depended to classified information on each bond. In this section, we offer the weighted linear combination of spatial-spectral kernels for hyper spectral images' classifiers.

To provide weighted spatial-spectral kernel, firstly we give some equations related to SVM. Suppose x_i is one of N pixel vector in hyper spectral data and index i means the i-thpixel of train samples. SVM classifier can be presented as follows:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{M} y_i\, \alpha_i K(x_i, x) + b\right) \tag{18}$$

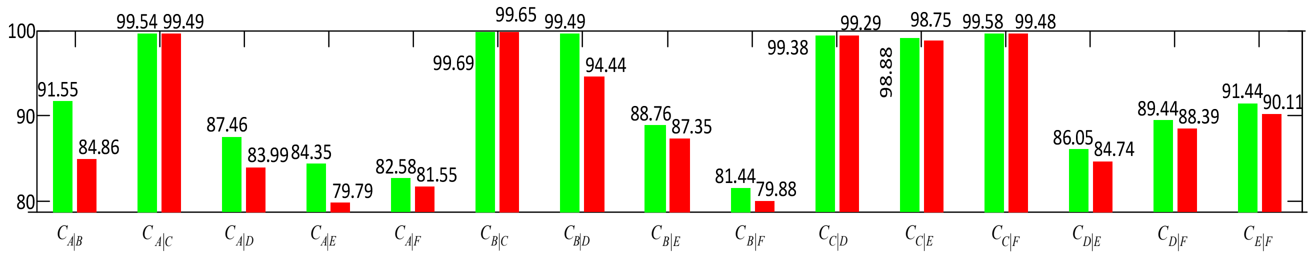


Figure 6. The comparison of classification accuracy of the proposed weighing method with pair wise SVM classifiers $C_{A|B}$,$C_{A|C}$,$C_{A|D}$,$C_{A|E}$,$C_{A|F}$,$C_{B|C}$,$C_{B|D}$,$C_{B|E}$,$C_{B|F}$,$C_{C|D}$,$C_{C|F}$,$C_{D|E}$,$C_{D|F}$,$C_{E|F}$, (Green: Grouped kernel spatial graph kernel machine, Red: Weighted grouped kernel spatial graph kernel machine)

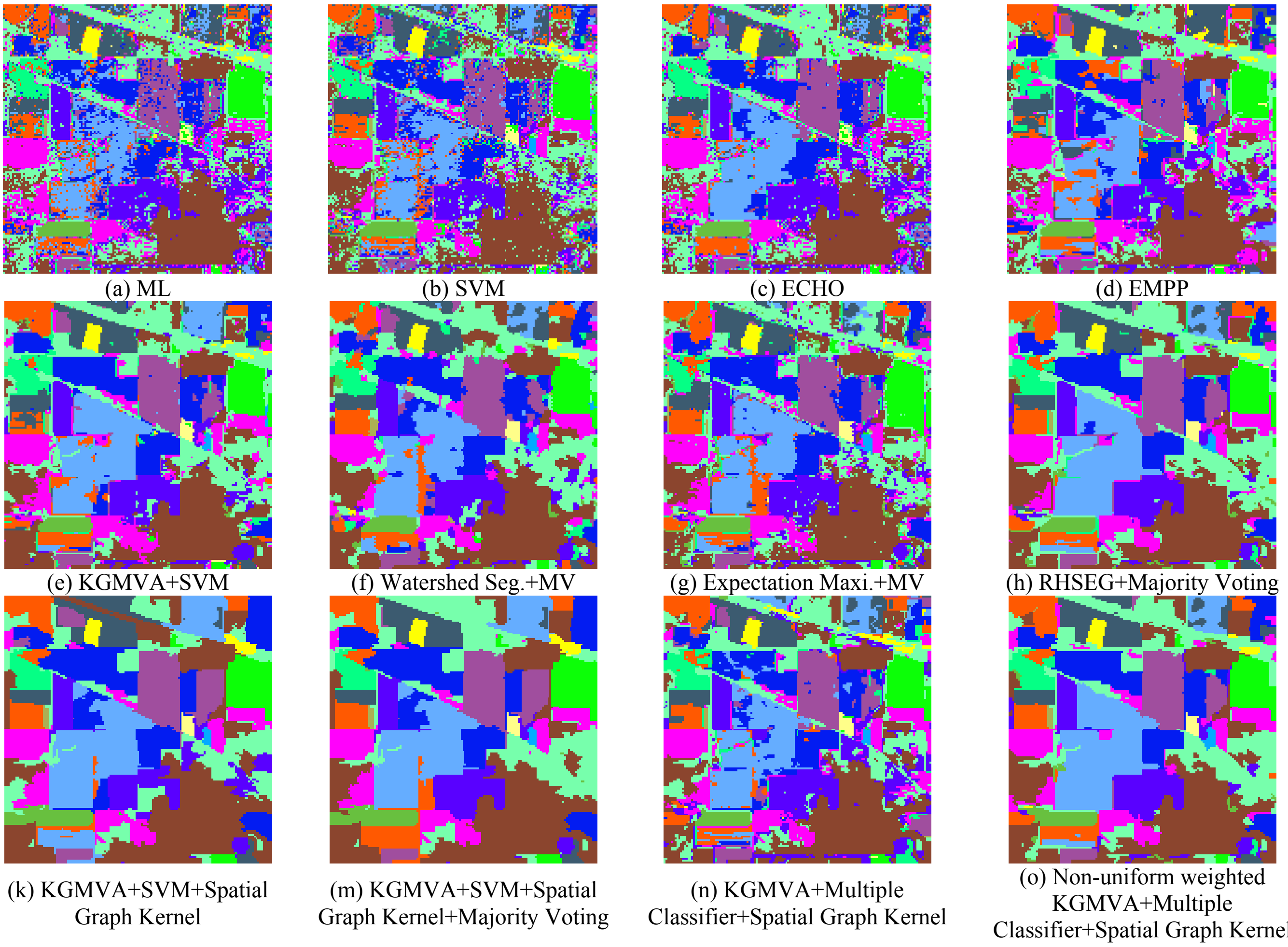


Figure 7. Classification maps for the Indian Pines image (a) ML (b) SVM (c) ECHO [17] (d) EMPP [18] (e) KGMVA+SVM (f) Watershed Segmentation+Majority Voting (g) EM+Majority Voting (h) RHSEG+Majority Voting (k) KGMVA+SVM+Spatial Graph Kernel (m) KGMVA+SVM+Spatial Graph Kernel+MV (n) KGMVA+Multiple Classifier+Spatial Graph Kernel (o) Non-uniform weighted KGMVA+Multiple Classifier+Spatial Graph Kernel

where $y_i \in \{-1,1\}$ is classified label and $\alpha_1, \alpha_2, \dots, \alpha_M$ are Lagrange coefficients and M is the number of test points and b is the threshold level. Moreover, $K(x,\acute{x}) = \varphi(x)\varphi^T(\acute{x})$ is the kernel function that is usually a polynomial or Gaussian or neural network kernel.

The kernel deals identically with each element x_i^j in feature space. Let's hyper spectral data vector is $x_i = [x_i^1, x_i^2, \dots, x_i^B]$ and x_i^j represents j-th band of the i-th training sample, where $1 \le j \le B$ and B is the number of bands., if x_i^j indicates the spectral value or band (spectral feature) of the two classes which are quite distinct.

In the proposed method, a larger share of the kernel is given by the weighted spectral features, and this proposed approach improves efficiency of class. Modified kernel function for implementing the outlined requirements, a weighting vector $s = [s_1, s_2, \dots, s_N]$ associated with the frequency bands is considered to weight each feature vector in hyper spectral data. The diagonal matrix $S = diag(s_1, s_2, \dots, s_N)$ is introduced simply; the polynomial, Gaussian and neural networks kernels can be rewritten as follows as the weighted spatial-spectral kernel:

$$K(x,\acute{x}) = (x^T C_{ss} \acute{x} + 1)^d \tag{19}$$

$$K(x,\acute{x}) = \exp\left(-\frac{\|S(x-\acute{x})\|^2}{2\sigma^2}\right) \tag{20}$$

$$K(x,\acute{x}) = \tanh(b_1 x^T C_{ss} \acute{x} + b_2) \tag{21}$$

where $C_{ss} = S^T S = diag(s_1^2, s_2^2, \dots, s_N^2)$, d is the degree of polynomial and σ is the characteristic parameter of the Gaussian kernel width and b_1 and are b_2 free parameters of the neural network kernel. Necessary and sufficient condition for the latter functions to be a kernel functions is Mercer theory.

The weighted kernels with prior knowledge (i.e., non-uniformly distributed data) can be used in the SVM training process [22]. For minimization of the generalization error of SVM, we can determine weights that lead to better accuracy. This process affects computational complexity of proposed classification as well as Hughes phenomena affect [2], and thus, the classification accuracy for high dimensional data with limited number of training samples can be treated. Kernel weighting method can be rewritten as follows:

$$K(x_i, x_j) = \exp\left(-\sum_{p=1}^{N} \frac{\left\| s_p^2 \left(x_i^p - x_j^p\right)\right\|^2}{2\sigma^2}\right) \quad (22)$$

Where i and j are indexes of the training samples and $p = 1, 2, ..., N$ is the number of bands. Bands with low variance lead to higher separation values (good bands lead smaller classification error). The above equation shows that the spectral weight s_p changes proportionally with the "similarity" in the proposed classifier. Thus, obtaining less generalization error is derived by weighting spectral features with "similarity".

Among the various measures of "similarity", we can use the mutual information in each feature. The advantage of using MI is closely related to the effective implementation and the bayes classification error. Let's consider spectral bands pixel as well as samples of a continuous random variable A with values $a \in R$ and set of related classes B and their labels as discrete variables them $a \in \{w_1, w_2, ..., w_N\}$. The mutual information between A and B, can be calculate as follows:

$$I(A, B) = -\int_a p(a) \mathrm{Log}(p(a)) da - \sum_{b \in B} p(b) \mathrm{Log}(p(b)) + \sum_{b \in B} \int_a p(a, b) \mathrm{Log}(p(a, b)) da \quad (23)$$

Using the above equation, the mutual information between each of the bands and the reference map of Indian Pines [16] data set has been calculated. The bands with the highest similarity with the reference class maps have also the highest MI. MI curves reflect the effects of atmospheric water absorption bands with low values of MI in 104 to 108 and 150 to 163. In Indian Pinescase [16], the MI curve changes in various characteristics such as changes in the bands of Bhattacharya distance.

It is suggested that MI can be used as an effective measure of "similarity" features in terms of classification and can be considered as the feature weight in proposed approach. We use spatial Markov Random Fields (MRF) energy function to measure the spatial distances [23].

According to the concept of "similarity" and the proposed weighted kernels, and effects of mutual information, we proposed multi-class SVM with better overall accuracy as in the following:

To determine the influence of each band proportional to the amount of its mutual information, we consider the weighting vector as $S_{n,m} = \frac{MI_{n,m}}{\left\| MI_{n,m} \right\|}$. This means that the Gaussian kernel for class C_n and C_m for one against one strategy in multi-class SVM classification can be rewritten as follows:

$$k_{n,m}(x_i, x_j) = \exp\left(-\sum_{p=1}^{N} \frac{\left\| s_{n,m,p}^2 \left(x_i^p - x_j^p\right)\right\|^2}{2\sigma_{n,m}^2}\right) \quad (24)$$

Or the polynomial kernel and the neural network kernel can be written as follows:

$$k_{n,m}(x, \acute{x}) = (x^T C_{n,m} \acute{x} + 1)^{d_{n,m}} \quad (25)$$

$$k_{n,m}(x, \acute{x}) = \tanh(k_{n,m,1} x^T C_{n,m} \acute{x} + k_{n,m,2}) \quad (26)$$

where $C_{n,m} = S_{n,m}^T S_{n,m}$. In this way, each classifier has its own free parameters and values $\sigma_{n,m}^2$, $d_{n,m}$, $k_{n,m,1}$ and $k_{n,m,2}$ are selected independently for each classifier. The Gaussian kernel for class n for one against the rest strategy in multi-class SVM classification can be rewritten as follows

$$k_{n,\bar{n}}(x_i, x_j) = \exp\left(-\sum_{p=1}^{N} \frac{\left\| s_{n,\bar{n},p}^2 \left(x_i^p - x_j^p\right)\right\|^2}{2\sigma_{n,\bar{n}}^2}\right) \quad (27)$$

or the polynomial kernel and the neural network kernel can be written as follows:

$$K_{n,\bar{n}}(x, \acute{x}) = (x^T C_{n,\bar{n}} \acute{x} + 1)^{d_{n,\bar{n}}} \quad (28)$$

$$K_{n,\bar{n}}(x_i, x_j) = \exp\left(-\sum_{p=1}^{N} \frac{\left\| s_{n,\bar{n},p}^2 \left(x_i^p - x_j^p\right)\right\|^2}{2\sigma_{n,\bar{n}}^2}\right) \quad (29)$$

as the expression $\bar{n}$means all classes except n class as the "others" classes and $C_{n,\bar{n}} = S_{n,\bar{n}}^T S_{n,\bar{n}}$. In this way, each classifier has its own free parameters and values$\sigma_{n,\bar{n}}^2$, $d_{n,\bar{n}}$, $k_{n,\bar{n},1}$ and $k_{n,\bar{n},2}$ are selected independently for each classifier.

The paper proposes using the Spectral Angle map per (SAM) for the mean vectors of the most similar pair of regions as a measure of dissimilarity. For$X_i = (X_{i\,1}, ..., X_{i\,B})^T$ and $X_j = (X_{j\,1}, ..., X_{j\,B})^T$ the SAM dissimilarity measure defines as:

$$\mathrm{SAM}(x_i, x_j) = \arccos\left(\frac{\sum_{b=1}^{B} x_{ib} x_{jb}}{\left(\sum_{b=1}^{B} {x_{ib}}^2\right)^{\frac{1}{2}} \left(\sum_{b=1}^{B} {x_{jb}}^2\right)^{\frac{1}{2}}}\right) \quad (30)$$

So far we have introduced the four major to measure the similarity and for each criterion, the weighted kernel was discussed: Bhattacharya distance, mutual information and spectral angle measures. We know that the linear combination of kernel is kernel with considering the Mercer limitations. Conditions can be considered that even non-linear combinations of kernels can be also a kernel.

It can briefly be noted that by applying multiple kernel, we expect the proposed approach archives better separation in kernel space and also higher overall accuracy for classification. From previous sections it can be summarized that the linear combination of various weighted kernels, including Gaussian kernel, polynomial, neural network and ... can be mathematically recommended with different scales such as Bhattacharya distance, mutual information, spatial MRF energy function and spectral angle.

Let's,x_i^k is the training data set (which x_i^k indicates the k-th training samples from i-th class.) and$K_{n,m,t,d}(x, \acute{x})$ is the basic kernel where n and $m$are classes that the classifier aimed to separate them, t represents the type of kernel (Gaussian, polynomial, neural, etc.) and d is weighting criterion (Bhattacharya distance, mutual information, spatial MRF energy function and spectral angle ...). For given positive definite kernels, our proposed kernel can be defined by the following combination:

$$K_{n,m} = \sum_t \sum_d \mu_{t,d} K_{n,m,t,d}(x, \acute{x}) \tag{31}$$

where $\mu_{t,d} > 0$ and $\sum_t \sum_d \mu_{t,d} = 1$. Another view of the proposed kernel can be combined such as following:

$$K_{n,m} = \sum_t \sum_d \mu'_{t,d} k_{n,m,t,d}(x, \acute{x}) \tag{32}$$

Where $\mu'_{t,d} > 0$ and $\sum_t \sum_d \acute{\mu}^2_{t,d} = 1$. This kernel converged in optimization process and the last proposed kernel has some privileges to the ones before, and its coefficients can be determined automatically. Also the proposed kernel can be investigated by non-linear combinations of kernel such as equation (33) where $a_{t,d} \in \mathbb{R}^+$and $a_{t,d} > 0$and$0 < \sum_t \sum_d a_{t,d} \leq d$and $\mu \geq 0$are used.

$$K_{n,m} = \sum_t \sum_d \mu K^{a_{t,d}}_{n,m,d,t}(x, \acute{x}) \tag{33}$$

The proposed kernel can be combined such as equation (34).

$$K_{C_n,C_m} = \sum_t \sum_d \mu^{a_{t,d}}_{t,d} K^{a_{t,d}}_{n,m,d,t}(x, \acute{x}) \tag{34}$$

Where $a_{t,d} \in \mathbb{R}^+$ and $\sum_t \sum_d a_{t,d} = d$ and $\mu_{t,d} \geq 0$. In experimental part of this paper, we have used only the linear combination of proposed kernel and $\mu_{t,d}$ values have been calculated from the following equation:

$$\mu_{t,d} = \frac{Var(S_{t,d,n,m})}{\sum_t \sum_d Var(S_{t,d,n,m})} \tag{35}$$

$S_{t,d,n,m}$ denotes the weight vector of the kernels t, d distance and n and m binary classifier. The more variance of feature distance measure leads to the greater weight of the kernel.

Pair classifiers' accuracy for classes from A to F for 25 Mount Carlo runs for the proposed weighted method and the grouped kernel spatial graph kernel machine were compared in the figure 6 as it can be seen in some experiments the weighted method is better than previous methods in all of the average accuracy measurement. We use a part of AVIRIS hyper spectral dataset in this experiment, called F210. It is illustrative of hyper spectral image analysis problem to determine land use. However the AVIRIS sensor [16] collects nominally 224 bands (or images) of data, four of these contain only zeros and so are discarded, leaving 220 bands in the 92AV3C dataset. At special frequencies, the spectral images are kenned to be adversely affected by atmospheric dihydrogen monoxide absorption. This affects about 20 bands. Each image size is 145*145 pixels. The dataset was collected over a test site called Indian Pinesin north-western Indiana [16]. The database is accompanied by a reference map; signify partial ground truth, whereby pixels are labeled as belonging to one of 16 classes of vegetation or other land types. Not all pixels are so labeled, presumably because they correspond to uninteresting regions or were too arduous to label. In this experiment, we use only class A to class F.

5. Performance Evaluation

The proposed method has been evaluated in terms of classification map and accuracy. For this purpose, we evaluate the Overall Accuracy (OA), which is the fraction of all correctly classified samples versus all samples and the Average Accuracy (AA) reflecting the average number of correctly classified samples of each class as well as Kappa factor (κ is the percentage of agreement, i.e., correctly classified pixels, corrected by the number of agreements that would be expected purely by chance [24]). In the following, we present results for the Indian Pines [16], centre of Pavia [25] beside university of Pavia [25] datasets.

Our proposed approaches were compared with two other spectral-spatial approaches. One of them is the Extraction and Classification of Homogeneous Objects (ECHO) [17] as a well-known spectral-spatial issue and the other is Extended Morphological Profiles with Partial reconstruction (EMPP) [18] as a state-of-the-art approach.

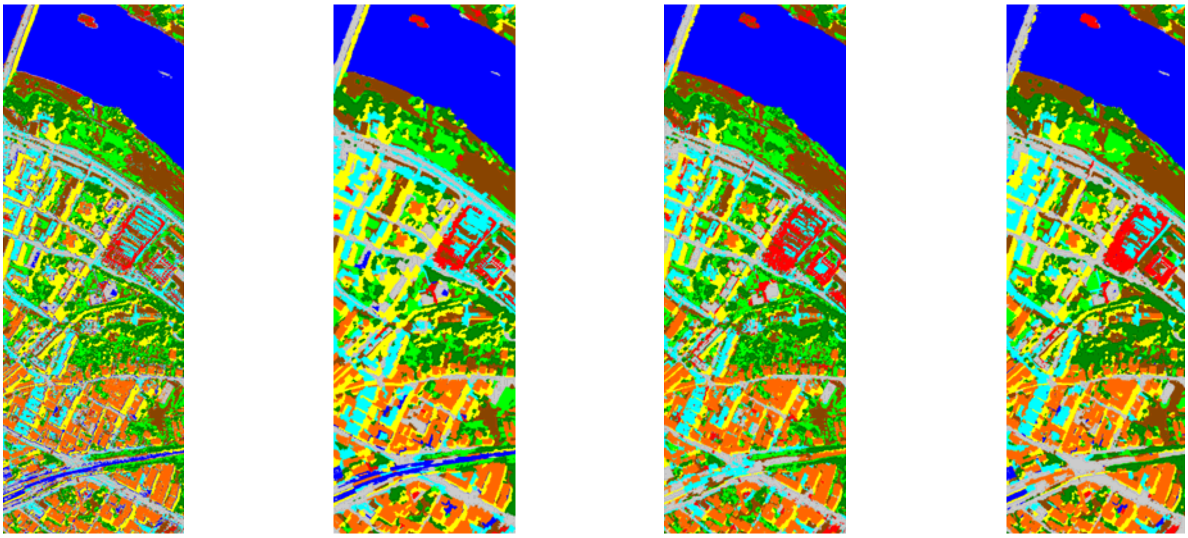
To demonstrate the proposed algorithm performance, let's take a look at classification accuracies. Our experiments are based on three benchmark hyper spectral remote sensed images.

Figure 7, figure 8 and figure 9 present classification maps for the AVIRIS Indian Pines dataset [16], university of Pavia [25] and centre of Pavia [25] images. For classification, we consider the k-nearest neighbor (kNN) [26] classifier and support vector machines (SVM) [27]. The experiments have been repeated ten times and the training samples have been randomly chosen. For Indian Pines, we consider 50 training samples for each class, for image of Pavia; we select30 training samples per class. During model selection, the parameters of the classifiers have been estimated, i.e. the number of neighbors k for kNN and the regularization C and the kernel bandwidth γ.

Figure 10 and figure 11 summarize the global and class-specific accuracies of the pixel wise SVM, ML, ECHO[17], EMPP [18], three segmentation techniques plus majority voting, and some proposed classification methods such as KGMVA plus SVM with spatial graph kernel, then continue experiments with multiple classifiers approach and at the end we use the non-uniform weighting approach, the results obtained using the construction of an MSF as a symbolic of spatial graph kernel from the probabilistic SVM-derived markers followed by majority voting within connected regions.

We can conclude that all our proposed approaches yield higher accuracies when compared to the pixel wise SVM or to the ECHO [17] or EMPP [18]classification results.

The proposed strategy gives the best classification results. In particular, it is a great interest to use non-uniform weighted KGMVA approach before selecting outstanding point from multiple classifiers probabilistic outputs. There are not any challenges for using the minimum spanning forest for spatial graph kernel-based region growing and majority voting as post processing block.



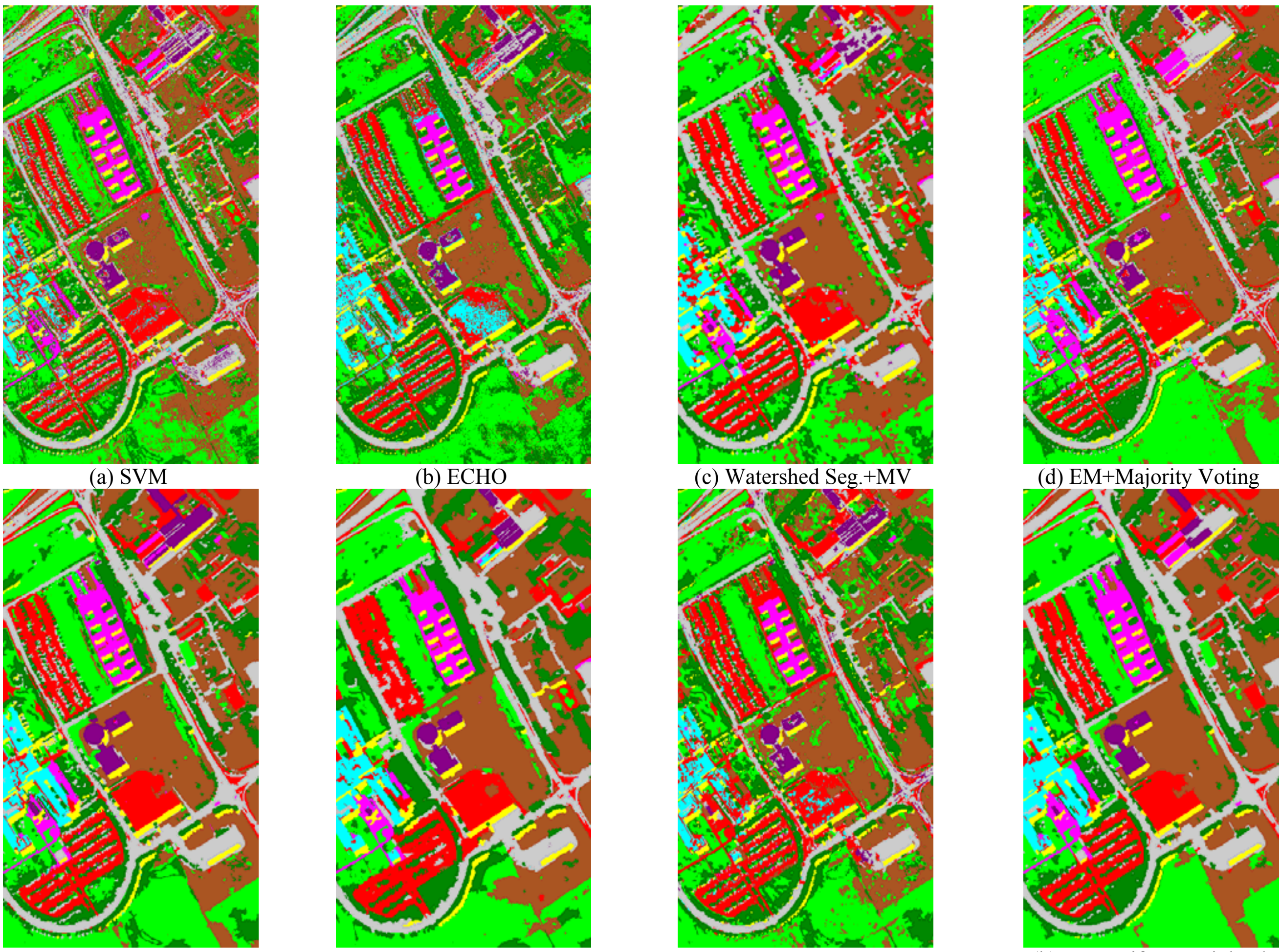
(a) SVM

(b) KGMVA+SVM+Spatial Graph Kernel+Majority Voting

(c) KGMVA+Multiple Classifier+Spatial Graph Kernel

(d) Non-uniform weighted KGMVA+Multiple Classifier+Spatial Graph Kernel

Figure 8. Classification maps for the Center of Pavia image (a) SVM (b) KGMVA+SVM+Spatial Graph Kernel+Majority Voting (c) KGMVA+Multiple Classifier+Spatial Graph Kernel (d) Non-uniform weighted KGMVA+Multiple Classifier+Spatial Graph Kernel



(a) SVM

(b) ECHO

(c) Watershed Seg.+MV

(d) EM+Majority Voting

(e) RHSEG+Majority Voting

(f) KGMVA+SVM+Spatial Graph Kernel+Majority Voting

(g) KGMVA+Multiple Classifier+Spatial Graph Kernel

(h) Non-uniform weighted KGMVA+MC+Spatial Graph Kernel

Figure 9. Classification maps for the University of Pavia image (a) SVM (b) ECHO [17] (c) Watershed Segmentation+Majority Voting (d) EM+Majority Voting (e) RHSEG+Majority Voting (f) KGMVA+SVM+Spatial Graph Kernel+Majority Voting (g) KGMVA+Multiple Classifier+Spatial Graph Kernel (h) Non-uniform weighted KGMVA+Multiple Classifier+Spatial Graph Kernel

We can see that all of classification maps of our proposed methods contain much more homogeneous regions when compared to a pixel wise map. The drawback of spectral-spatial classification approaches consists in the fact that they smooth a classification map, and in some parts of the image they may smooth it too much.

From experimental results, we can make conclusions that

- The proposed spectral-spatial classification approach uses automatically selected outstanding points:
  - Meaningfuly decreases over segmentation
  - improves classification accuracies
  - and provides classification maps with homogeneous regions
- For the repress enter selection step, its advantageous is to use an SVM classifier, spatial information and multi classifier approaches.
- The spatial graph kernel for agent-controlled region growing has proven to be an efficient and robust technique.

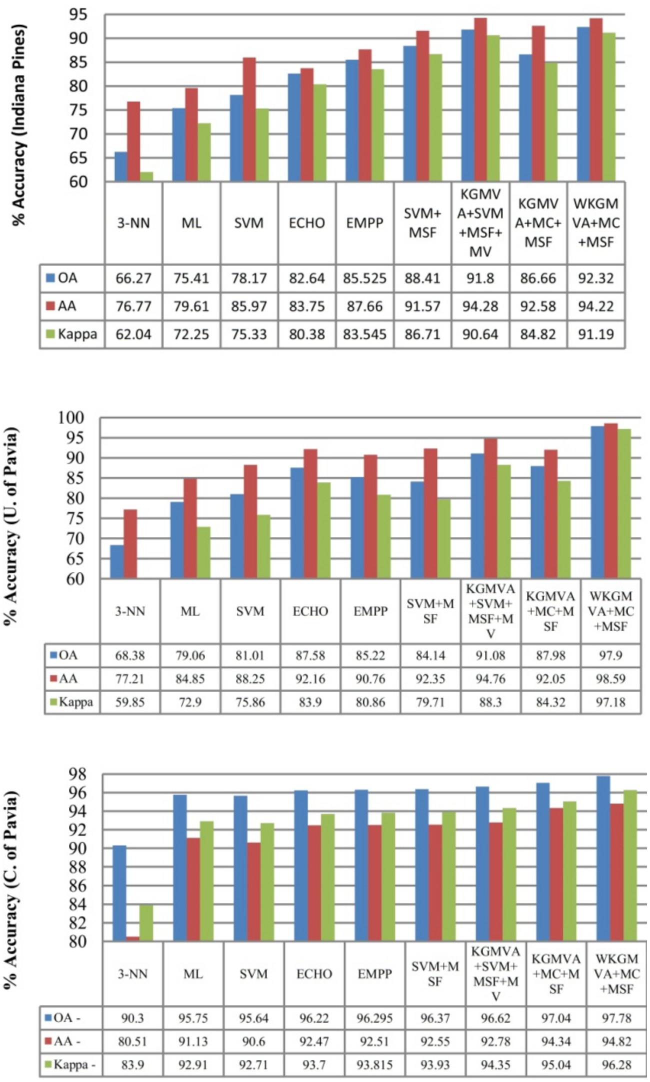


Figure 10. Classification accuracy for various images in terms of Overall accuracy (OA), Average Accuracy (AA) and Kappa factor for various approaches such as kNN, ML, SVM, ECHO [17], EMPP[18], SVM+Spatial Graph Kernel, KGMVA+SVM+Spatial Graph Kernel+Majority Voting, KGMVA+MC+Spatial Graph Kernel, Non-uniform weighted KGMVA+Multiple Classifiers+Spatial Graph Kernel

- Implementation results with Indian Pines hyper spectral image showed better performance for proposed approach in limited training samples compared to some pixel wise methods (3-NN, ML and SVM) and some spatial-spectral approaches (ECHO [17] and EMPP [18]).

In particular, we assess performance of several criteria to measure the non-uniform distribution of the spatial-spectral information, grouping based techniques of multivariate analysis, probabilistic SVMs and simultaneous spectral spatial graph kernels. To achieve these purposes, we focus on the average accuracy, overall accuracy and Kappa coefficient of methods when working in higher-dimensional data, and limited training sets. Experimental implementations with three benchmark hyper spectral datasets (Indian Pines, University of Pavia and Centre of Pavia) represent advantageous of the proposed framework in hyper spectral remote sensing applications.

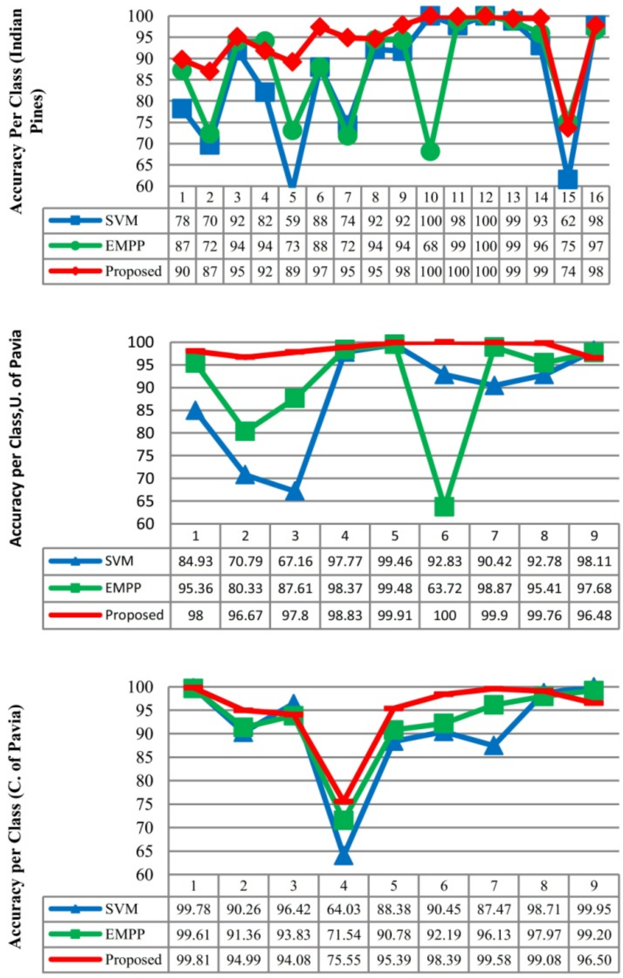


Figure 11. Accuracy per class for SVM, EMPP [18] and proposed approach in various dataset, for Indian Pines dataset: Class 1: Alfalfa, 2: Corn-no till, 3: Corn-min till, 4: Corn, 5: Grass-pasture, 6: Grass-trees, 7: Grass-pasture-mowed, 8: Hay-windrowed, 9: Oats, 10: Soybean-no till, 11: Soybean-min till, 12: Soybean-clean, 13: Wheat, 14: Woods, 15: Buildings-Grass-Trees-Drives, 16: Stone-Steel-Towers, for university and center of Pavia: Class 1: Water, 2: Trees, 3: Asphalt, 4: Self-Blocking Bricks, 5: Bitumen, 6: Tiles, 7: Shadows, 8: Meadows, 9: Bare Soil, we use Non-uniform weighted KGMVA+Multiple Classifier+Spatial Graph Kernel+Majority Voting for proposed approach

6. Conclusions

This section concludes by recalling the main contributions of this paper. The new strategy for spectral-spatial classification of hyper spectral data has been proposed and investigated. The proposed strategy has been concentrated on grouped MVA techniques in hyper spectral images. An important contribution consists in analyzing probabilistic classification results for selecting the most reliably classified pixels as outstanding points of spatial regions. We have concluded on the interest of using spatial information and multiple classifier approaches for repress enter selection, and using a spatial graph kernel-based approach for outstanding point-controlled region growing. This paper proposed a novel weighting spatial-spectral kernel by discovery and application of the prior spectral information contained in hyper spectral data with non-uniform weighting to increase the accuracy and reliability of classification. The used prior knowledge considers the non-uniform distribution of data in different spectral and spatial features. The various metrics such as spatial energy of MRFs, Bhattacharya distance function, mutual information and SAM were introduced and calculated. Linear or non-linear combinations were employed as a novel kernel for kernel maximum margin classifiers when the convex optimization was solved.

References

[1] B. C. Kuo, H. H. Ho, C. H. Li, C. C. Hung, and J. ShiuhTaur, "A Kernel-based Feature Selection Method for SVM with RBF Kernel for Hyper Spectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 317-326, 2014.

[2] G. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Trans. Information Theory,* vol. 14, no. 1, pp. 55-63, 1968.

[3] M. Borhani, and H. Ghassemian, "Novel Spatial Approaches for Classification of Hyper Spectral Remotely Sensed Landscapes," *Proc. IEEE Intl Conf. Artificial Intelligence and Signal Processing*, pp. 41-51, 2013.

[4] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial-Spectral Kernel Sparse Representation for Hyper Spectral Image Classification," *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2462-2471, 2013.

[5] M. Borhani, and H. Ghassemian, "Hyper Spectral Spatial-spectral Feature Classification based on Adequate Adaptive Segmentation," *Proc. IEEE Intl Conf. Intelligent Systems*, pp. 4-6, 2014.

[6] J. Li, P. Reddy Marpu, A. Plaza, J. M. Bioucas-Dias, and J. Atli Benediktsson, "Generalized Composite Kernel Framework for Hyper Spectral Image Classification," *IEEE Trans. Geoscience and Remote Sensing,* vol. 51, no. 9, pp. 4816-4829, 2013.

[7] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite Kernels for Hyper Spectral Image Classification," *IEEE Trans. Geoscience and Remote Sensing,* vol. 3, no. 1, pp. 93-97, 2006.

[8] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, "Spectral and Spatial Classification of Hyper Spectral Data using SVMs and Morphological Profiles," *IEEE Trans. Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804-3814, 2008.

[9] N. Hoai Ly, Q. Du, and J. E. Fowler, "Sparse Graph-based Discriminant Analysis for Hyper Spectral Imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 3872-3884, 2014.

[10] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt, "Spatio-spectral Remote Sensing Image Classification with Graph Kernels," *IEEE Trans. Geoscience and Remote Sensing*, vol. 7, no. 4, pp. 741-745, 2010.

[11] K. Brandt Petersen, G. Camps-Valls, L. Kai Hansen, and J. Arenas-García, "Kernel Multivariate Analysis Framework for Supervised Subspace Learning," *IEEE Signal Processing Magazine*, pp. 1-14, 2013.

[12] M. Borhani, and H. Ghassemian, "Kernel Grouped Multivariate Discriminant Analysis for Hyper Spectral Image Classification," *Journal of Artificial Intelligence and Signal Processing*, pp. 3-13, 2013.

[13] J. Platt, *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods*, Advances in Large Margin Classifiers, MIT, Cambridge, MA, 2000.

[14] J. L. Yang, G. Zhang, and B. F. Han-Xiong Li, "Probabilistic Support Vector Machines for Classification of Noise Affected Data," *International Journal of Information Sciences*, vol. 22, no. 1, pp. 60-71, 2013.

[15] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and Classification of Hyper Spectral Images using Minimum Spanning Forest Grown from Automatically Selected Markers," *IEEE Trans. Geoscience and Remote Sensing*, vol. 40, no. 5, pp. 1267-1279, 2010.

[16] AVIRIS Indian Pines Hyper Spectral Dataset, https://engineering.purdue.edu/~biehl/MultiSpec/hyper spectral.html, May 2008.

[17] R. L. Kettig, and D. A. Landgrebe, "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects," *IEEE Trans. Geoscience Electronics*, vol. 14, no. 1, pp. 19-26, 1976.

[18] R. Bellens, A. Pižurica, W. Philips, Y. Pi, and W. Liao, "Classification of Hyper Spectral Data over Urban Areas based on Extended Morphological Profile with Partial Reconstruction," *Lecture Notes in Computer Science*, vol. 75, no. 17, pp. 278-289, 2012.

[19] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and Classification of Hyper Spectral Images using Watershed Transformation," *Journal of Pattern Recognition*, vol. 43, no. 7, pp. 2367-2379, 2010.

[20] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial Classification of Hyper Spectral Imagery based on Partitional Clustering Techniques," *IEEE Trans. Geoscience and Remote Sensing*, vol. 47, no. 8, pp. 2973-2987, 2009.

[21] H. G. Akcay, and S. Aksoy, "Automatic Detection of Geospatial Objects using Multiple Hierarchical Segmentations," *IEEE Trans. Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2097–2111, 2008.

[22] M. Borhani, and H. Ghassemian, "Hyper Spectral Image Classification based on Non-uniform Spatial-spectral Kernels," *Proc. IEEE Intl Conf. Intelligent Systems*, pp. 1-6, 2014.

[23] M. Borhani, and H. Ghassemian, "Hyper Spectral Image Classification based on Spectral-spatial Features using Probabilistic SVM and Locally Weighted Markov Random Fields," *Proc. IEEE Intl Conf. Intelligent Systems*, pp. 1-6, 2014.

[24] J. A. Richards, and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag: Berlin Heidelberg, 2006.

[25] ROSIS Pavia University Hyper Spectral Dataset, http://www.ehu.es/ccwintco/index.php/Hyper, April 2009.

[26] T. Cover, and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 13, no. 1, pp. 21-27, 1967.

[27] C. Chang, and C. Lin, "LIBSVM: A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, June 2011.

Mostafa Borhani received the B.Sc. and M.Sc. Degrees in Electrical Engineering from Sharif University of Technology, Tehran, Iran in 2002 and 2004 respectively. He is currently working toward the Ph.D. degree in the department of Electrical and Computer Engineering at the Tarbiat Modares University, Tehran, Iran. His research interests include the hyper spectral image analysis, kernel machines and pattern recognition in remote sensing applications.

E-mail: m.borhani@modares.ac.ir

Hassan Ghassemian received the B.S.E.E. degree from Tehran College of Telecommunication in and the M.S.E.E. and Ph.D. degree from Purdue University, West Lafayette, USA in 1984 and 1988 respectively. Since 1988, he has been with Tarbiat Modares University in Tehran, Iran, where he is a Professor of Computer and Electrical Engineering. His research interests focus on Multi-Source Signal/Image Processing and Information Analysis and Remote Sensing

E-mail: ghassemi@modares.ac.ir

Paper Handling Data:

Submitted: 16. 06. 2014
Received in revised form: 22.01.2015
Accepted: 25.01.2015
Corresponding author: Mostafa Borhani,
Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.