Title

Class-Separation-Based Rotation Forest for Hyperspectral Image Classification

Permalink

https://escholarship.org/uc/item/7z8330dv

Journal

IEEE Geoscience and Remote Sensing Letters, 13(4)

ISSN

1545-598X

Authors

Xia, Junshi
Falco, Nicola
Benediktsson, Jón Atli
et al.

Publication Date

2016-04-01

DOI

10.1109/lgrs.2016.2528043

Peer reviewed

Class-Separation-Based Rotation Forest for Hyperspectral Image Classification

View Document

10
Paper
Citations

253
Full
Text Views

Related Articles

[(Semi-) Supervised Probabilistic Principal Component Analysis for Hyperspectral ...]

[Hyperspectral Image Spectral-Spatial Feature Extraction via Tensor Principal Com...]

[Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classif...]

View All

5
Author(s)
Junshi Xia ; Nicola Falco ; Jón Atli Benediktsson ; Jocelyn Chanussot ; Peijun Du
View All Authors

Abstract
Authors
Figures
References
Citations
Keywords
Metrics
Media

Abstract:

In this letter, we propose a new version of the rotation forest (RoF) method for the pixelwise classification of hyperspectral images. RoF, which is an ensemble of decision tree classifiers, uses random feature selection and data transformation techniques (i.e., principal component analysis) to improve both the accuracy of base classifiers and the diversity within the ensemble. Traditional RoF performs data transformation on the training samples of each subset. In order to further improve the performance of RoF, the data transformation is separately performed on each class, extracting sets of transformation matrices that are strictly dependent on the training samples of each single class. The approach, namely, class-separation-based RoF (RoF$_{\mathrm{CS}}$), is experimentally investigated on a hyperspectral image collected by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. Experimental results demonstrate that the proposed methodology achieves excellent performances, in comparison with random forest and RoF classifiers.

ISSN Information:

INSPEC Accession Number: 15871650

DOI: 10.1109/LGRS.2016.2528043

- Download PDF
- Download Citation
- View References
- Email

SECTION I.

Introduction

In THE last few decades, hyperspectral image classification has been an incredibly active research topic with widespread applications [1]. However, classification of hyperspectral data is a challenge due to issues such as the high ratio of feature (spectral bands) to instance (training samples) and the redundant information in the feature set [2], [3]. In the past two decades, researchers have investigated a variety of approaches to alleviate such issues [4], [5].

Recently, multiple-classifier systems (MCSs), which combine different classification algorithms or variants of the same classifier, have shown excellent performances in hyperspectral image classification compared to a single-classifier case [6]–[7][8]. Rotation forest (RoF), is a leading technique in MCSs, which aims at constructing multiple decision trees built on different sets of extracted features [9], [10]. More specifically, in RoF, the feature set is randomly split into several disjoint subsets. Principal component analysis (PCA) is then applied to each subset. Furthermore, new training data are formed by concatenating the linear extracted features contained in each subset and then used in an individual decision tree (DT) classifier. A series of individual classifiers is generated by repeating the aforementioned steps several times, fusing the results according to a majority decision. Studies on the use of RoF dealing with hyperspectral classification problems have been recently published [11]–[12][13][14]. RoF has proven to be effective not only in hyperspectral data analysis but also in very high-resolution image analysis [15], where object-based classification was investigated, and in synthetic aperture radar (SAR) image analysis [16], where RoF was applied to SAR images by integrating spatial and polarimetric features. Here, RoF provided better results in comparison to those obtained by exploiting support vector machines (SVMs) and random forest (RF). In general, accuracy of the base classifiers and diversity within the ensemble represent two important aspects that need to be taken in consideration when designing an MCS [6]. Diversity, in particular, can be improved by splitting the input feature space, where different splits on the feature space lead to different extracted features [9], [10]. Another strategy is represented by the employment of feature extraction techniques, which are used to transform the original feature space into another one in order to extract more representative information. In the case of the RoF classifier, data transformation is performed on the whole training set of each subset.

In order to further improve the diversity within the ensemble, we propose a new strategy based on the use of RoF, which is called class-separation-based

RoF ($\mathrm{RoF}_{\mathrm{CS}}$). Following the strategy described in [17] and [18], where class-specific independent components were extracted to address the dimensionality reduction task, we perform data transformation to each class, extracting sets of transformation matrices strictly dependent on the training samples of each single class. Here, PCA is used as the data transformation technique. We would like to emphasize that, in this work, we focus on pixelwise classification, although RoF can be combined with spatial information, such as Markov random fields [12]. The experimental analysis, including a comparison with RF and RoF, is carried out on the Indian Pine test site.

The remainder of this letter is structured as follows: Section IIintroduces the proposed class-separation-based RoF ($\mathrm{RoF}_{\mathrm{CS}}$). Section III presents the results obtained by the experimental analysis and a comparison with other state-of-the-art classifiers. Conclusions are drawn in Section IV.

SECTION II.

Class-Separation-Based RoF ($\mathrm{RoF}_{\mathrm{CS}}$)

Let $\{\mathbf{X},\mathbf{Y}\}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$ be a set of training samples, where $\mathbf{x}_i=[x_{i,1},\ldots,x_{i,D}]\in\mathbb{R}^D$ is a pixel vector, and $y_i\in\{1,\ldots,C\}$ denotes the label information, where C is the total number of classes, and $\mathbb{F}$ be the set of D features in the training set. Class c has n_c samples with $\sum_{c=1}^{C} n_c=n$. The training set is ordered according to their labels, with the matrix $\mathbf{X}_c=\{\mathbf{x}_{c1},\mathbf{x}_{c2},\ldots,\mathbf{x}_{cn_c}\}$, where $\mathbf{x}_{ci}$ is the ith sample in class c. Thus, $\mathbf{X}$ can be expressed as $\mathbf{X}=\{\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_C\}$.

The detailed steps of $\mathrm{RoF}_{\mathrm{CS}}$ are summarized in Algorithm 1. In the training phase, the first step consists in splitting the feature space of training set $\mathbb{F}$ into K disjoint subsets, each containing M features. In the second step, PCA is applied to each class, to create a projection matrix that is specifically suitable to represent each specific class. The size of the projection matrix is then reduced by selecting those components that better represent each single class. This is usually achieved by selecting the principal components correspondent to the largest eigenvalues. However, in this letter, in order to provide a more general framework, we adopt the reconstruction error as measure of class information associated with a single component [17], which allows the use of other data transformation within this methodology.

SECTION Algorithm 1

$\mathrm{RoF}_{\mathrm{CS}}$

Training phase

Input: $\{\mathbf{X},\mathbf{Y}\}=\{\mathbf{x}_i,y_i\}_{i=1}^{n}$: training samples, T: number of classifiers, K: number of subsets, M: number of features extracted in a subset, L: base classifier. The ensemble $>=\emptyset$. $\mathbb{F}$: Feature set. l: number of retained components in each class

Output: The ensemble $>$

```
for $i=1: T$ do
    randomly split the features $\mathbb{F}$ into $K$ subsets $\mathbb{F}_{ij}$
    for $j=1: K$ do
        form the new training set $\mathbf{X}_{i,j}$ with $\mathbb{F}_{ij}$
        calculate the optimal matrix $\mathbf{W}_{opti,j}$ using (1)–(3)
    end for
    the features extracted will be given by: $\mathbb{F}_{newi}=[\mathbf{W}_{opti,1}\mathbf{X}_{i,1},\ldots,\mathbf{W}_{opti,K}\mathbf{X}_{i,K}]$
    train an DT classifier $L_{i}$ using $\{\mathbb{F}_{i}^{\mathrm{new}},\textbf{Y}\}$
    Add the classifier to the current ensemble, $\mathcal{L}=\mathcal{L}\cup L_{i}$.
end for
```

Prediction phase

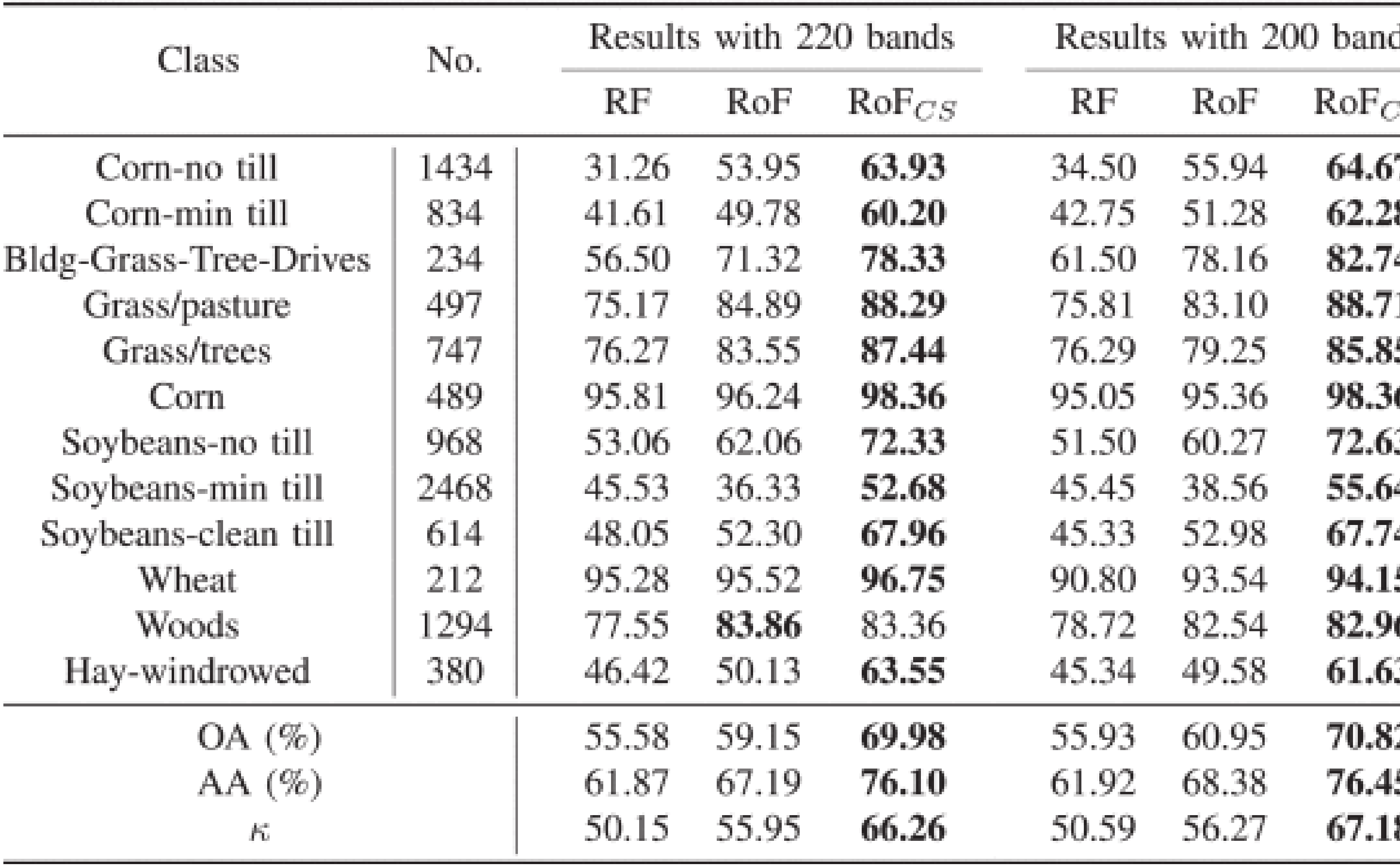
```
Input: The ensemble $\mathcal{L}={\{L_{i}\}}_{i}^{T}$. A new sample $\textbf{x}^{\ast}$. Transformation matrix: $\textbf{W}$.
Output: class label $y^{\ast}$
for $i=1$: $T$
    for $j=1$: $K$
        generate the test features of $\textbf{x}^{\ast}$, $\mathbb{F}_{i}^{\mathrm{test}}=[\textbf{W}_{i,1}^{\mathrm{opt}}\textbf{x}_{i,1}^{\ast},\ldots,\textbf{W}_{i,K}^{\mathrm{opt}}\textbf{x}_{i,K}^{\ast}]$
    end for
    run the DT classifier $L_{i}$ using $\mathbb{F}_{i}^{\mathrm{test}}$ as input
end for

calculate the confidence $\textbf{x}^{\ast}$ for each class and assign the class label $y^{\ast}$ to the class with the largest confidence.
```

PCA is a linear orthogonal data transformation technique that aims at projecting the original features $\textbf{X}$ into another space, where the transformed features $\textbf{Z}$ are linearly uncorrelated, which are called principal components. Following the linear decomposition model $\textbf{Z}=\textbf{W}\textbf{X}$, where $\textbf{W}$ represents the unmixing matrix, the unmixing matrix $\textbf{W}$ and the principal components $\textbf{Z}$ can be estimated by solving an eigenvalue decomposition problem. Reformulating the linear mixing model as $\textbf{X}=\textbf{A}\textbf{Z}$, where $\textbf{A}=\textbf{W}^{-1}$ represents the unknown mixing matrix, it is possible to compute the reconstruction error. Considering the PCA applied to each specific class, the reconstruction error e_{c}, which is estimated by computing the Frobenius norm $(\|\cdot\|_{F}^{2})$ between the original feature space and the back projection of the extracted components, is given by

$$e_{c}={\|\textbf{X}^{c}-\textbf{A}^{c}\textbf{Z}^{c}\|}_{F}^{2}=\left\|\textbf{X}^{c}-\sum_{i=1}^{n}\textbf{a}_{i}\textbf{z}_{i}^{T}\right\|_{F}^{2}\tag{1}$$

View Source where $\textbf{a}_{i}$ is a column vector of the mixing matrix $\textbf{A}^{c}$, and $\textbf{z}_{i}^{T}$ is a row vector of estimated components. Furthermore, the pairs ($\textbf{a}_{i}$, $\textbf{z}_{i}^{T}$) are ranked based on their relative contribution to the minimization of the reconstruction error.

TABLE I Classification Results Obtained for the Indian Pines Image Using 240 Training Samples (20 Samples Per Class). For Each Method, "OA (%)," "AA (%)," "κ," and "CA (%)" are Reported. No. Means the Total Number of Samples for Each Class in Reference Map



Class	No.	Results with 220 bands			Results with 200 band		
		RF	RoF	RoF$_{CS}$	RF	RoF	RoF$_{C}$
Corn-no till	1434	31.26	53.95	**63.93**	34.50	55.94	**64.6**
Corn-min till	834	41.61	49.78	**60.20**	42.75	51.28	**62.2**
Bldg-Grass-Tree-Drives	234	56.50	71.32	**78.33**	61.50	78.16	**82.7**
Grass/pasture	497	75.17	84.89	**88.29**	75.81	83.10	**88.7**
Grass/trees	747	76.27	83.55	**87.44**	76.29	79.25	**85.8**
Corn	489	95.81	96.24	**98.36**	95.05	95.36	**98.3**
Soybeans-no till	968	53.06	62.06	**72.33**	51.50	60.27	**72.6**
Soybeans-min till	2468	45.53	36.33	**52.68**	45.45	38.56	**55.6**
Soybeans-clean till	614	48.05	52.30	**67.96**	45.33	52.98	**67.7**
Wheat	212	95.28	95.52	**96.75**	90.80	93.54	**94.1**
Woods	1294	77.55	**83.86**	83.36	78.72	82.54	**82.9**
Hay-windrowed	380	46.42	50.13	**63.55**	45.34	49.58	**61.6**
OA (%)		55.58	59.15	**69.98**	55.93	60.95	**70.8**
AA (%)		61.87	67.19	**76.10**	61.92	68.38	**76.4**
κ		50.15	55.95	**66.26**	50.59	56.27	**67.1**

The following iterative procedure is used to perform the ranking and identifies the lth couple with the smallest reconstruction error [17]:

$$\begin{align*} idx=&\,\arg\min_{i}\mathrm{err}(i)=\left\{i|\min_{i}\left\|\textbf{X}_{l}-\textbf{a}_{i}\textbf{z}_{i}^{T}\right\|_{F}^{2}\right\}\tag{2}\\ \textbf{X}_{l+1}\leftarrow &\, \textbf{X}_{l}-\textbf{a}_{idx}\textbf{z}_{idx}^{T}\tag{3} \end{align*}$$

View Source where idx is the index of the chosen lth pair at the lth iteration. $\textbf{X}_{l}$ is initially set as $\textbf{X}^{c}$ and updated at each iteration by subtracting the contribution provided by $\textbf{a}_{idx}\textbf{z}_{idx}^{T}$ computed at the previous iteration [see (3)]. The tuning parameter l, which represents the number of pairs to retain after the ranking, is the only parameter required in the procedure [17]. For each class c, a matrix $\textbf{A}^{c}$, which is composed of the best elements $[\textbf{a}_{1},\ldots,\textbf{a}_{l}]$, is defined. It is worth noting that the aforementioned steps are applied to each specific class, and thus, the process of extraction can be done in parallel fashion, decreasing the computational time, which can be approximated to the one of a single-class PCA. The final mixing matrix $\textbf{A}_{\mathrm{opt}}$, which integrates all the specific class information, is represented by $\textbf{A}^{\mathrm{opt}}=[{(\textbf{A}^{1})}^{\prime},\ldots,{(\textbf{A}^{C})}^{\prime}]$. The obtained $\textbf{A}^{\mathrm{opt}}$ is an $M\times(C\times l)$ matrix. The unmixing matrix $\textbf{W}^{\mathrm{opt}}$ is obtained by $(\textbf{A}^{\mathrm{opt}})^{-1}$.

In the third step, $\textbf{W}^{\mathrm{opt}}$ is obtained for each subset, and a new training set is constructed by concatenating the extracted features $\textbf{W}_{j}^{\mathrm{opt}}\textbf{X}_{j}$ ($j=1,\ldots,K$ is the index of subset), which are then used to train an individual DT classifier. The size of $\textbf{W}^{\mathrm{opt}}$ in each subset is $(C\times l)\times M$. Then, the final ensemble is produced by integrating the individual DT classifiers that are generated, by repeating the aforementioned steps T times.

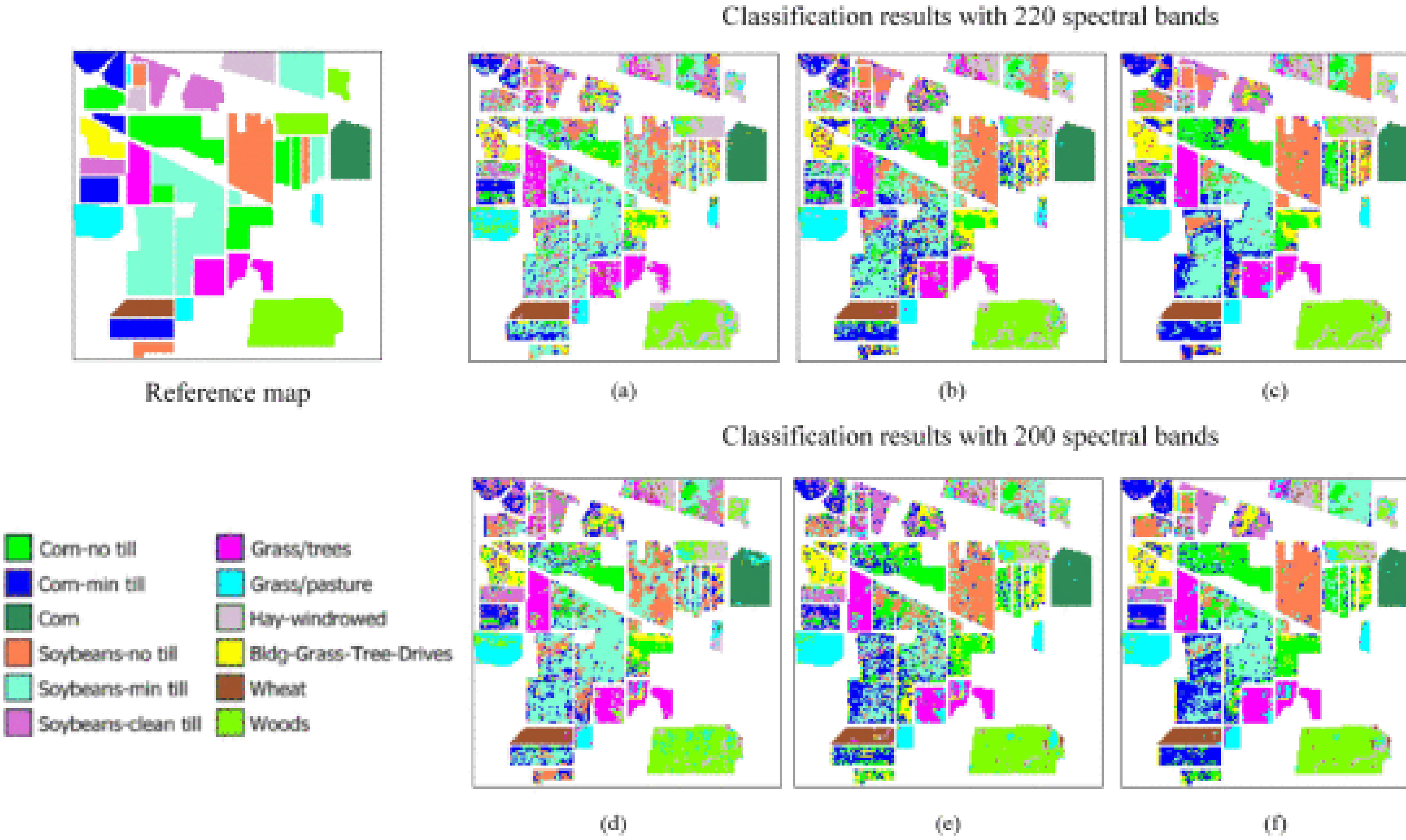
In the prediction phase, for a new sample $\textbf{x}^{\ast}$, the final result is generated by combining the results from individual DT classifiers in the ensemble based on the transformation matrix $\textbf{W}$ using a majority voting rule.

SECTION III.

Experimental Results and Analysis

Here, the proposed RoF_{CS} is evaluated on the Indian Pines AVIRIS hyperspectral data, which were acquired over Northwestern Indiana, USA, in June of 1992. The AVIRIS image is composed of 145 $\!\times\!$ 145 pixels, with a spatial resolution of 20 m/pixel and 220 spectral channels in the spectral range from 400 to 2500 nm. The original reference data contain 16 classes, whereas in this letter, we kept 12 classes, which contain large numbers of labeled samples. After removing 20 noisy and water absorption bands, the final data set is composed of 200 bands. In order to investigate the performance of the proposed method in noisy environments, the full spectral image is also used. In order to provide a more exhaustive analysis, a comparison with the RF [19] and RoF [10] classifiers is also provided. The parameters M and Kneeded in RF, RoF, and RoF_{CS} are set to be 10, whereas the parameter l used in RoF_{CS} is set to 7. In this work, classification and regression tree (CART) is considered as the base classifier.

The classification results achieved by the different methods, considering both the cleaned and the full data sets, are presented in Table I. For each method, the table reports the percentage of correctly classified samples, i.e., “overall accuracy (OA)”; the average percentage of correctly classified samples for individual class, i.e., “average accuracy (AA)”; the percentage of correctly classified samples for each class, i.e., “class accuracy (CA)”; and the percentage agreement corrected by the level of agreement that could be expected to chance alone, i.e., “kappa coefficients (κ).” In order to test the method in a practical scenario, where limited training samples are available, we perform the analysis considering only 240 training samples (i.e., 20 samples per class). From the obtained results, we can see that the proposed RoF_{CS} can better exploit the information present in both the scenarios, providing the best results, even in very noisy conditions. Moreover, all the ensemble classifiers appear to be robust in noisy conditions. Fig. 1 shows the classification maps produced by the classifiers under the noisy condition (one of the ten Monte Carlo runs). As it can be seen, RoF_{CS} exhibits much less classification errors than RF and RoF. In addition, considering the low number of samples used for the training of the classifier, the obtained results indicate that RoF_{CS} can properly deal with the high ratio between high dimensionality and limited training samples. Moreover, the proposed method performs well also in the presence of mixed pixels, as it is in the Indian Pines scene.



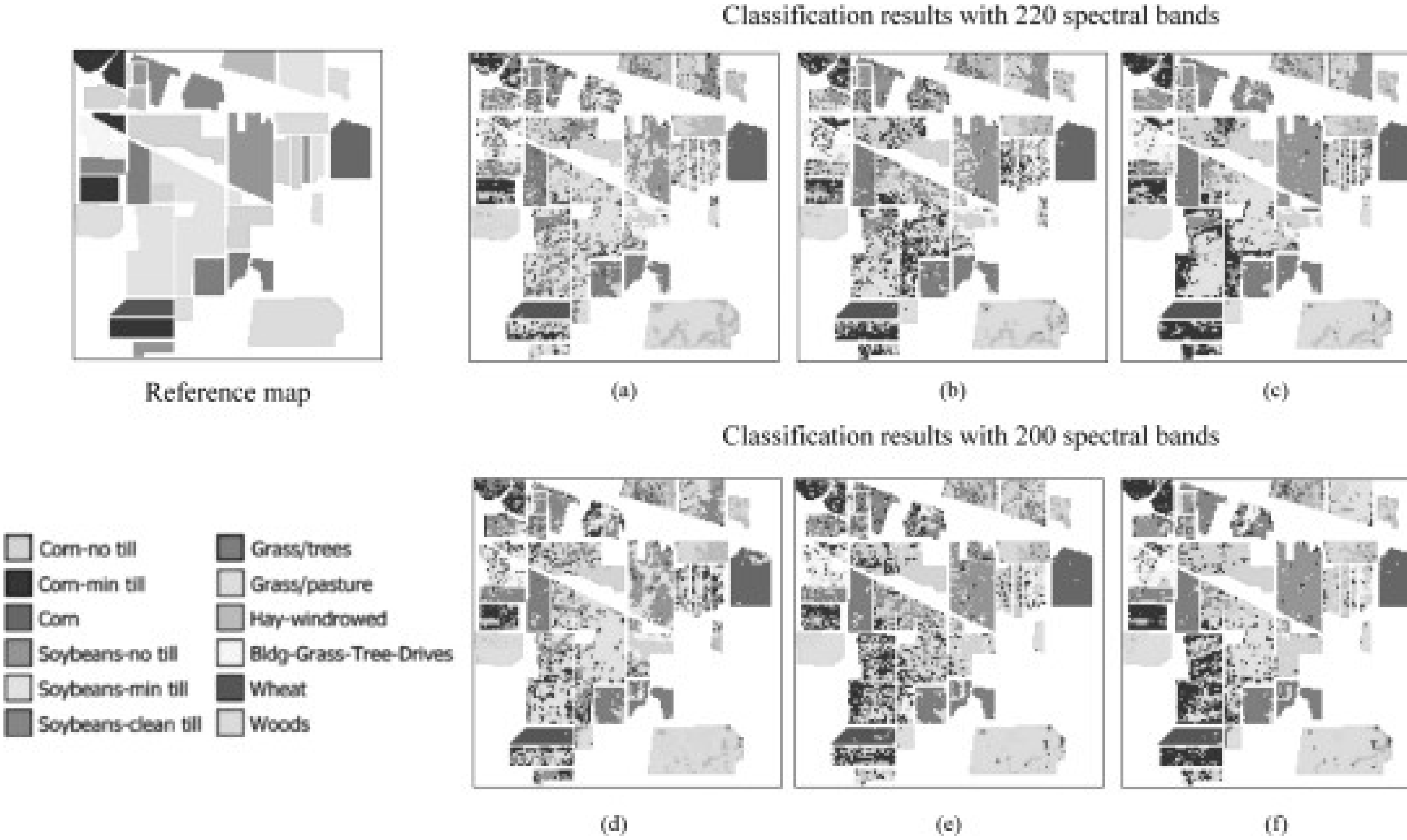
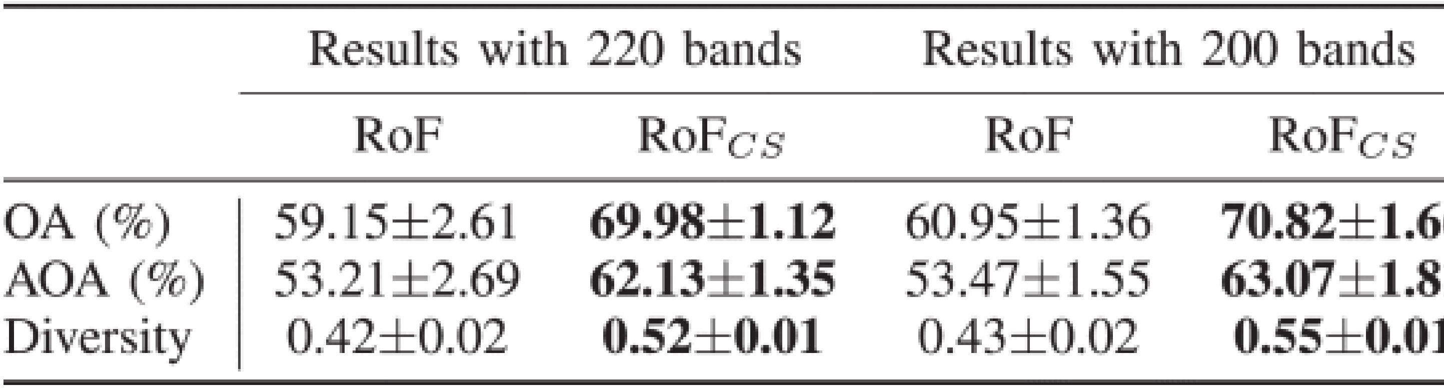


Fig. 1.
Classification maps of Indian Pines image obtained by the classifiers using 20 training samples per class. Results with 220 bands: (a) RF, $\text{OA}=55.19\%$; (b) RoF, $\text{OA}=58.31\%$; (c) RoF_{CS}, $\text{OA}=67.57\%$. Results with 200 bands: (d) RF, $\text{OA}=58.02\%$; (e) RoF, $\text{OA}=60.14\%$; (f) RoF_{CS}, $\text{OA}=69.23\%$.

View All

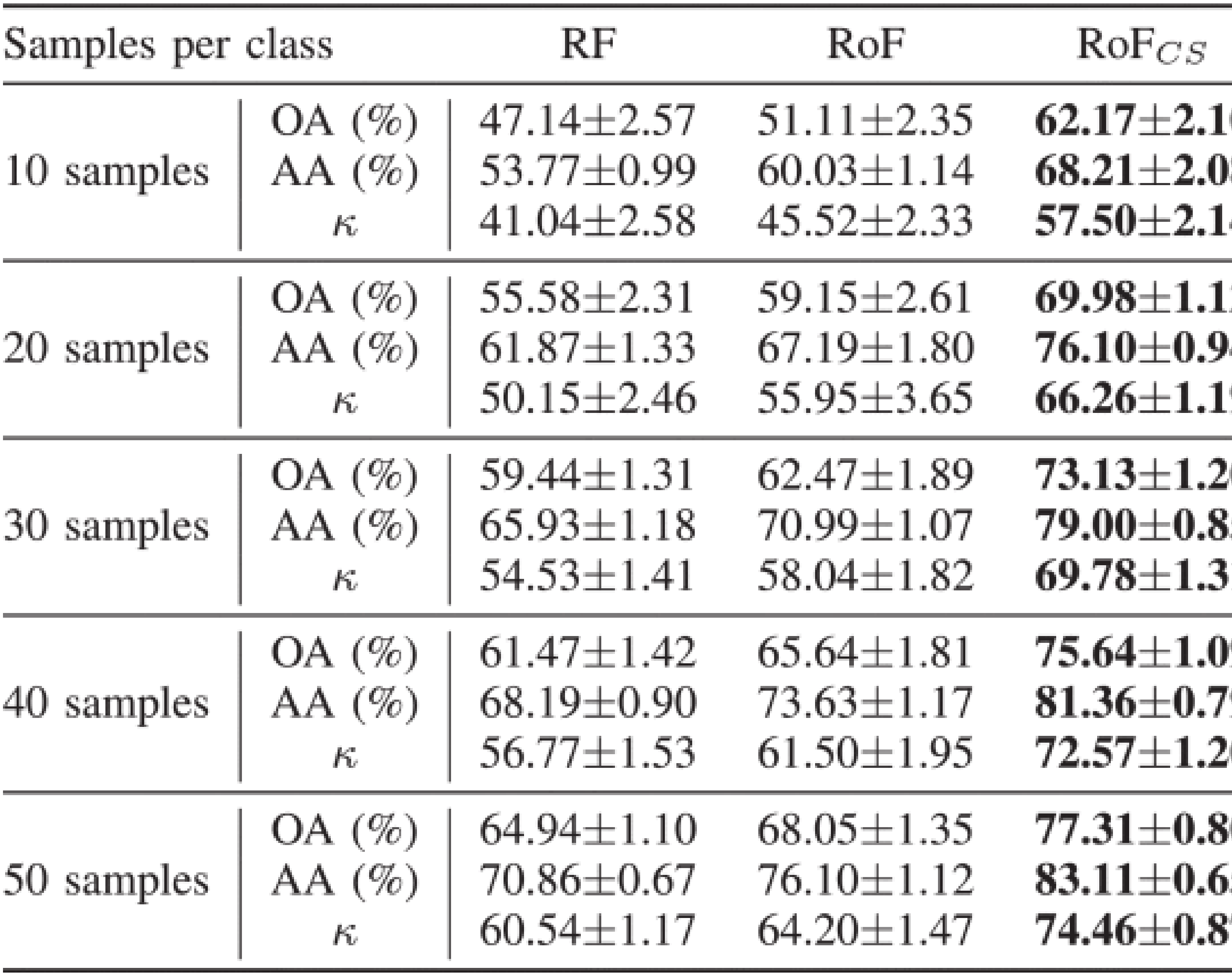
As stated in Section I, two important needed components to construct a strong ensemble are high accuracy of the base classifier and strong diversity within the ensemble. Here, to compare the two ensemble classifiers, i.e., RoF and RoF_{CS}, measures such as the “OA (%),” the percentage average OA of the individual DT classifier, i.e., “AOA (%),” and the *coincident failure diversity* (CFD) [20] are used. A higher value of CFD represents a stronger diversity. As shown in Table II, the proposed $\textrm{RoF}_{\mathrm{CS}}$produces higher values of AOAs and diversities than RoF, leading to better classification results.

TABLE II Comparison Between RoF and $\textrm{RoF}_{\mathrm{CS}}$. For Each Method, “OA (%),” “AOA (%),” and Diversities are Reported



	Results with 220 bands		Results with 200 bands	
	RoF	RoF$_{CS}$	RoF	RoF$_{CS}$
OA (%)	59.15±2.61	**69.98±1.12**	60.95±1.36	**70.82±1.6**
AOA (%)	53.21±2.69	**62.13±1.35**	53.47±1.55	**63.07±1.8**
Diversity	0.42±0.02	**0.52±0.01**	0.43±0.02	**0.55±0.01**

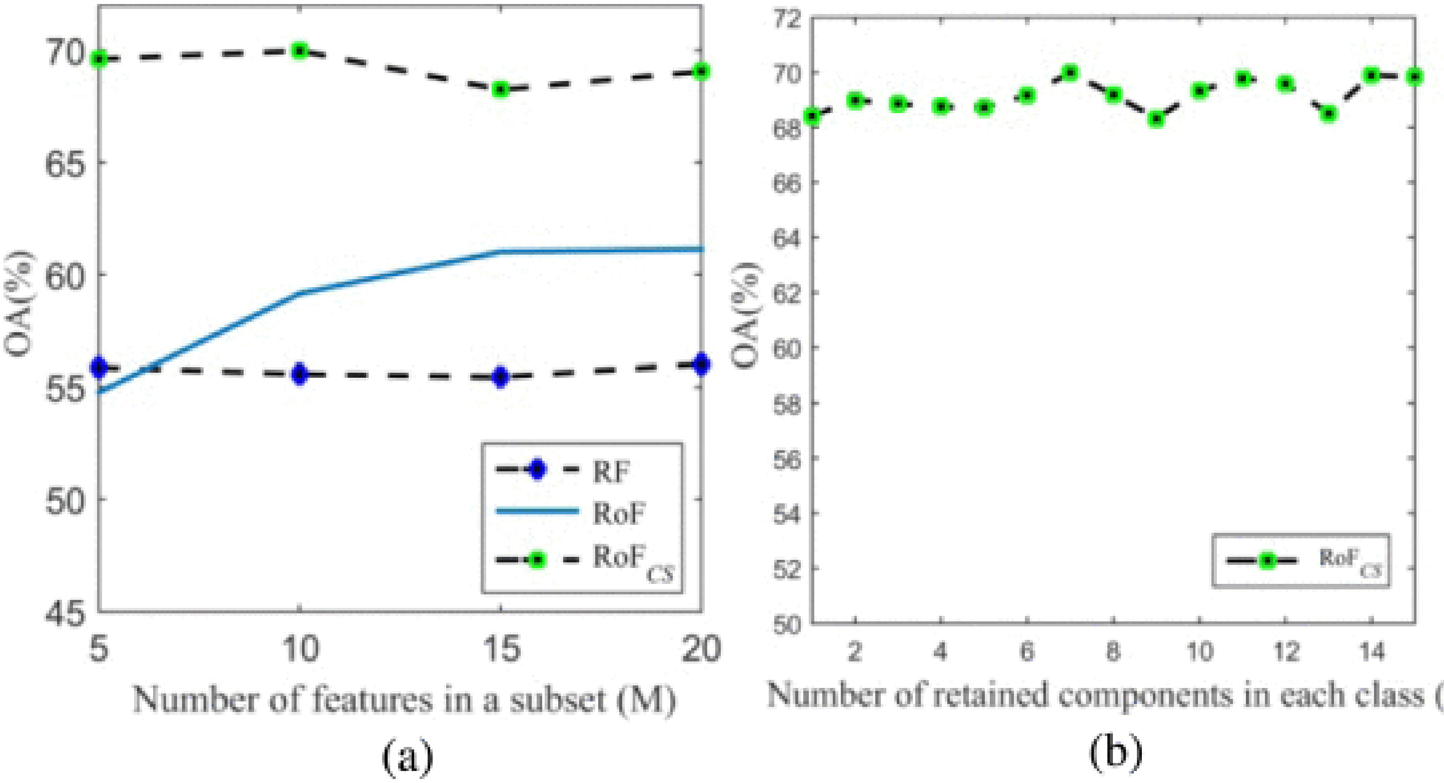
TABLE III Classification Results Using Different Training Sets. The Results Correspond to the Mean Values and Standard Deviations Over Ten Repetitions



Samples per class		RF	RoF	RoF$_{CS}$
10 samples	OA (%)	47.14±2.57	51.11±2.35	**62.17±2.1**
	AA (%)	53.77±0.99	60.03±1.14	**68.21±2.0**
	κ	41.04±2.58	45.52±2.33	**57.50±2.1**
20 samples	OA (%)	55.58±2.31	59.15±2.61	**69.98±1.1**
	AA (%)	61.87±1.33	67.19±1.80	**76.10±0.9**
	κ	50.15±2.46	55.95±3.65	**66.26±1.1**
30 samples	OA (%)	59.44±1.31	62.47±1.89	**73.13±1.2**
	AA (%)	65.93±1.18	70.99±1.07	**79.00±0.8**
	κ	54.53±1.41	58.04±1.82	**69.78±1.3**
40 samples	OA (%)	61.47±1.42	65.64±1.81	**75.64±1.0**
	AA (%)	68.19±0.90	73.63±1.17	**81.36±0.7**
	κ	56.77±1.53	61.50±1.95	**72.57±1.2**
50 samples	OA (%)	64.94±1.10	68.05±1.35	**77.31±0.8**
	AA (%)	70.86±0.67	76.10±1.12	**83.11±0.6**
	κ	60.54±1.17	64.20±1.47	**74.46±0.8**

The effectiveness of the proposed $\text{RoF}_{\mathrm{CS}}$is also evaluated considering different numbers of training samples (i.e., 10, 20, 30, 40, and 50 samples per class). Here, the experimental analysis is performed on the full data set, since it constitutes a challenging problem due to high feature-to-instance ratio and noise. The obtained results are shown in Table III, which reports the mean values and standard deviations of “OA (%),” “AA (%),” and κ values. The obtained results show that the proposed RoF_{CS} provides the best classification accuracy in all the cases, outperforming both RF and RoF. In particular, when 40 samples per class are considered, RoF_{CS} obtains an OA of 75.64%, an AA of 81.36%, and a κ of 72.57%. This is significantly better than those obtained by the RF (with +14.17% of OA, +13.17% of AA, and +15.80% of κ) and the RoF (with +10.00% of OA, +7.73 of AA, and +11.07 of κ). Another observation that can be derived from Table III is that, having a lower standard deviation, the performance of RoF_{CS} is more stable than the ones of RF and RoF. More recent pixelwise classification results obtained for Indian Pines scene can be found in [21]–[22][23]. A direct comparison with their results, which are not reported due to the space limit, shows that the RoF_{CS} can be considered a competitive classification method.

The last part of our experimental analysis focuses on the analysis of the sensitivity of two parameters M, which represents the number of features in a subspace, and l, which represents the number of retained components in each class. The results of such analysis are depicted in Fig. 2. As the number of features in a subspace (M) increases, RoF tends to have better performance. RF and $\text{RoF}_{\mathrm{CS}}$are robust to this parameter. Moreover, the proposed RoF_{CS} is not sensitive to the number of retained components (l) in each class. Hence, we can conclude that the proposed RoF_{CS} is not sensitive to the two parameters, which is an essential additional advantage. This should be considered during the parameter tuning, since the computational time of the RoF_{CS} could be reduced by choosing larger values of M and smaller values of l.



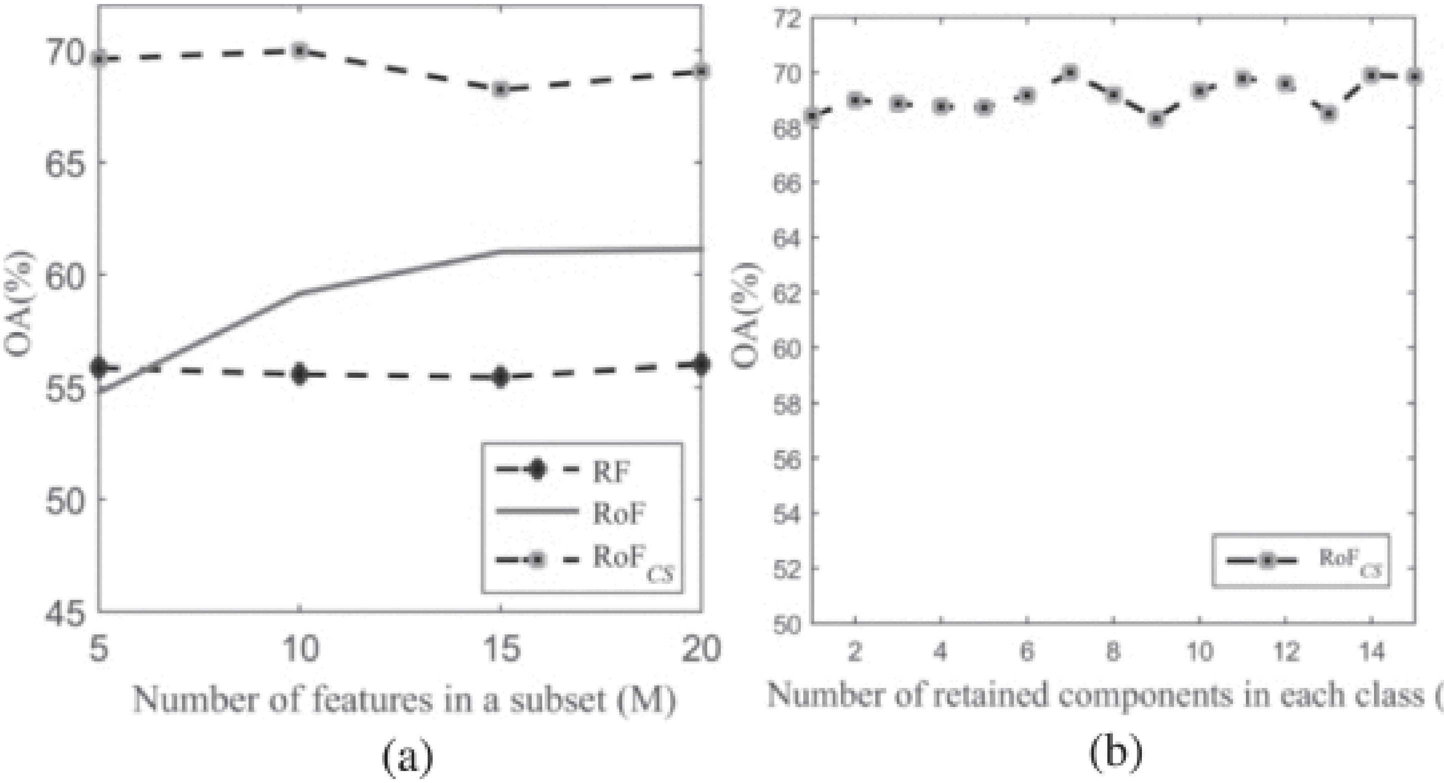


Fig. 2.

Sensitivity to the change of (a) number of features in a subset $(M)$and (b) number of retained components (l) in each class.

View All

SECTION IV.

Conclusion

In this letter, we have developed a new version of the RoF classifier. The proposed RoF_{CS} method integrates diverse DT classifiers that are trained on different feature sets. Each set is defined by the concatenation of class-specific features extracted by applying PCA. The experimental analysis was performed on Indian Pines test site. Two scenarios (i.e., the full spectral bands and the noisy bands removed) were considered, in order to evaluated the proposed approach in noisy conditions. Experimental results demonstrated the superiority of the proposed RoF_{CS}compared to SVM and RoF, indicating that the RoF_{CS} can better cope with the high ratio between high dimensionality of the feature space and the limited number of training samples, as well as the presence of mixed pixels and noise in the scene. The main reason for the powerful capability of RoF_{CS} is that more diversity and higher accuracy of member classifiers are introduced in the ensemble. An additional advantage of the RoF_{CS} is the noncritical parameter tuning. In practice, the users might select a high value of M and a low value of l to reduce the computational time. In future work, we will test the performances of other data transformation techniques in our proposed RoF_{CS} method.

ACKNOWLEDGMENT

The authors would like to thank Prof D. Landgrebe from Purdue University, USA, for providing the hyperspectral remote sensing data set.

$$e_c = \|\mathbf{X}^c - \mathbf{A}^c\mathbf{Z}^c\|_F^2 = \left\|\mathbf{X}^c - \sum_{i=1}^{n} \mathbf{a}_i \mathbf{z}_i^T\right\|_F^2 \tag{1}$$

$$idx = \arg\min_i \mathrm{err}(i) = \left\{ i \,|\, \min_i \|\mathbf{X}_l - \mathbf{a}_i \mathbf{z}_i^T\|_F^2 \right\} \tag{2}$$

$$\mathbf{X}_{l+1} \leftarrow \mathbf{X}_l - \mathbf{a}_{idx}\mathbf{z}_{idx}^T \tag{3}$$