Optik - International Journal for Light and Electron Optics 206 (2020) 163712

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.com/locate/ijleo

Original research article

Spatial-prior generalized fuzziness extreme learning machine autoencoder-based active learning for hyperspectral image classification

Muhammad Ahmad[a,b,*], Sidrah Shabbir[b], Diego Oliva[c], Manuel Mazzara[d], Salvatore Distefano[a]

[a] *Dipartimento di Matematica e Informatica—MIFT, University of Messina, Messina 98121, Italy*
[b] *Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan*
[c] *Department of Computer Sciences, Universidad de Guadalajara, Jalisco, Mexico*
[d] *Institute of Software Development and Engineering, Innopolis University, Innopolis 420500, Russia*

ARTICLE INFO

Keywords:
Extreme learning machine
Autoencoder
Active learning
Spatial spectral information
Fuzziness

ABSTRACT

Hyperspectral imaging has been extensively utilized in several fields, and it benefits from detailed spectral information contained in each pixel, generating a thematic map for classification to assign a unique label to each sample. However, the acquisition of labeled data for classification is expensive in terms of time and cost. Moreover, manual selection and labeling are often subjective and tend to induce redundancy into the classifier. In this paper, a spatial prior generalized fuzziness extreme learning machine autoencoder (GFELM-AE) based active learning is proposed, which contextualizes the manifold regularization to the objective of ELM-AE. Experiments on a benchmark dataset confirmed that the GFELM-AE presents competitive results compared to the state-of-the-art, leading to the improved statistical significance in terms of F1-score, precision, and recall.

1. Introduction

Hyperspectral imaging (HSI) is the field of science that includes all those activities necessary for the observation, acquisition, and interpretation of information related to objects, events, phenomena or any other item under investigation without making physical contact [1]. HSI is a technological tool that analyzes a wide spectrum of the electromagnetic wave instead of just assigning primary colors such as RGB to each pixel [2]. The light striking at each pixel is divided into many different spectral channels to provide detailed information on what is imaged.

HSI has attracted a formidable interest of the scientific community in recent years and has been extensively utilized for urban planning, agriculture land usage, natural phenomena, such as damage assessment due to earthquakes or floods, eruptions, climate change (e.g. glaciers), deforestation [3], meteorology, national security, natural resource management, change detection, man-made material identification, semantic annotation, unmixing [4], and classification [5].

Notably, the most important challenge for the HSI domain is connected to the characteristics of the data, which typically yields hundreds of contiguous and narrow spectral bands with very high spatial resolution throughout the electromagnetic waves to light waves [6]. Moreover, some challenges can arise from typical characteristics such as, the unavailability of training data which causes

[*] Corresponding author at: Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, 64200, Pakistan
E-mail addresses: mahmad00@gmail.com, m.ahmad@kfueit.edu.pk (M. Ahmad).

https://doi.org/10.1016/j.ijleo.2019.163712
Received 30 July 2019; Received in revised form 22 October 2019; Accepted 5 November 2019

inflated false discovery rate and low statistical performance [7]. This characteristic results in a relatively poor predictive performance [8,9] in terms of kappa κ and overall accuracy when addressing HSI classification (HSIC).

Some of the well-known machine learning methods deployed for HSIC include, multinomial logistic regression [10], random forests [11], ensemble learning [12], deep learning [13], support vector machine [14], and k-nearest neighbors [15]. However, these classification methods often underperform due to the "Hughes phenomenon" [16]. This phenomenon occurs in a case when the number of training samples is significantly inferior to the number of spectral bands required by the classification model [15].

Active learning (AL) poses as an alternative method to overcome the limitation of limited availability of training samples by iteratively selecting the useful samples as the training candidates [15]. The most commonly used samples selection methods utilized in AL are random selection [17], mutual information [18], breaking ties [19], modified breaking ties [20], uncertain sampling [21,22], query by committee (QBC) [23], Fisher information ratio [24], fuzziness [15], and spectral angle mapper and fuzziness (FSAM) based sample selection [1].

The aforementioned sample selection methods are categorized according to the information utilized for selecting the samples is either spectral or spatial-spectral. The later one entirely depends on the large degree on the spatial distribution of potential candidates. However, there are not many studies that have integrated spatial constraints into the AL methods [25–29]. Nevertheless, QBC has achieved remarkable results for many classification applications. QBC sample selection mechanism is based on the maximum disagreement of an ensemble of classifiers. However, such methods are highly computational complex due to the iterative training of the classifier for each new candidate [27].

To address the above-mentioned issue, batch mode AL methods have been proposed which concomitantly considered the uncertainty and diversity of the newly selected samples [30]. The work [31] has highlighted the benefits of integrating spatio-contextual information into AL, however, it did not consider the spatial distribution of newly selected samples. Nevertheless, this method was later extended in [18], which considers the spatial position of newly selected samples in the feature space. One of the outcomes from such transformation is the point-wise dispersed distribution in the spatial domain, which incurs the risk of revisiting the same geographical locations[1] several times, especially in the HSI domain [32]. Therefore, the combination of spatial-spectral AL achieves better performance than its pixel-wise counterparts and represents a novel and promising research contribution yet to be fully explored in the HSIC domain [32].

Therefore, this work introduces a spatial prior generalized fuzziness extreme learning machine autoencoder (GFELM-AE) AL method for HSIC to reduce the sample selection bias whilst maintaining the data stability in the spatial domain. GFELM-AE distinguishes from standard AL methods in several aspects including but not limited to, instead of using uncertainty measure, GFELM-AE utilizes the fuzziness associated with the confidence of the training model in classifying unseen samples [1]. Later, GFELM-AE couples the samples fuzziness with their diversity to select the new training samples that have lesser similarity with the existing training sample. Finally, GFELM-AE keeps the pool of selected samples balance, giving equal representation to all classes to some extent which is achieved via softening the thresholds at run time. Experiments on benchmark HSI dataset demonstrate that GFELM-AE leads to increased predictive power in terms of statistical tests including precision, recall, and F1-score, and classification matrices such as kappa (κ) coefficient and overall accuracies.

2. Extreme learning machine autoencoder

Suppose $\mathbf{X} = [x_1, x_2, x_3,...,x_B]^T \in \mathcal{R}^{B\times(N\times M)}$ is an HI cube which is composed of B spectral bands and $(N \times M)$ samples per band belonging to Y classes where $x_i = [x_{1,i}, x_{2,i}, x_{3,i}, \ldots, x_{B,i}]^T$ is the ith sample in the cube. Let us assume $(x_i, y_i) \in (\mathcal{R}^{B\times(N\times M)}, \mathcal{R}^Y)$, where y_i is the class label of the ith sample. Let us further assume that n number of training samples are selected from $\mathbf{X}$ to create the training set $\mathbf{X}_T = \{(x_i, y_i)\}_{i=1}^{n}$. The rest of the samples form the validation set $\mathbf{X}_V = \{(x_i, y_i)\}_{i=1}^{m}$. Please note that $n \ll m$, and $(\mathbf{X}_T \cap \mathbf{X}_V) = \emptyset$.

ELM-AE is a relatively new expansion which consists of the two-stage building block with extremely high generalization performance. The first stage consists of mapping of hidden neurons n_B to the original data into the n_B dimensional feature space. In this regards, two different structures can be used to extract the features based on the size of the input number of samples $|X_T|$[2] and hidden neurons n_B.

2.1. *Compressed or loose, i.e., $|X_T| > n_B$.*

Compressed architecture [33] considers the Euclidean distance among each input sample and the corresponding Euclidean distances in lower dimensional space is equal. Therefore, the mapping of x_i to the n_B dimensional feature space can be calculated as $h(x_i) = g(a^T x_i + b)$ such that $a^T a = I$, $b^T b = 1$, where a_i is the input weights between the input and hidden layer which is a $|X_T| \times n_B$ matrix and b is the bias of hidden units, $h(x_i) \in R^{n_B}$ is the output vector of hidden layer with respect to x_i and $g(\circ)$ is an activation function with I is the identity matrix of order n_B [34].

2.2. *Sparse, i.e., $|X_T| < n_B$*

Sparse ELM-AE architecture considers the orthogonal hidden random parameters as $h(x_i) = g(a^T x_i + b)$ such that $aa^T = I$, $b^T b = 1$,

[1] Region of interests.

[2] $|\circ|$ is the cardinality of X_T.

where the output of the network is computed as $f(x_i) = h(x_i)^T\beta$. β is updated by adopting the squared loss of the prediction error. Thus, the training model is updated as;

$$\min_{\beta\in R^{n_B\times|X_T|}} L = \min_{\beta\in R^{n_B\times|X_T|}} \frac{1}{2}||\beta||^2 + \frac{C}{2}||\mathbf{X} - \mathbf{H}\beta||^2 \tag{1}$$

where C is a penalty coefficient and $\frac{1}{2}||\beta||^2$ controls the complexity of the model. Furthermore, by setting the gradient of L with respect to β zero, one can have;

$$\triangledown = \beta - C\mathbf{H}^T(\mathbf{X} - \mathbf{H}\beta) = 0 \tag{2}$$

Eq. (2) can easily be solved in two different cases. The first case assumes that the number of samples is greater than the hidden neurons;

$$\beta* = \left(\mathbf{H}^T\mathbf{H} + \frac{\mathbf{I}_{n_B}}{C}\right)^{-1}\mathbf{H}^T\mathbf{X} \tag{3}$$

Second case assumes that the number of samples is less than the hidden neurons i.e., $\mathbf{H}$ will have more columns than rows. This will also bring an additional constraint to $\beta = \mathbf{H}^T\alpha$, where $\alpha\in R^{|X_T|}$ [35];

$$\beta* = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T + \frac{\mathbf{I}_{n_B}}{C}\right)^{-1}\mathbf{X} \tag{4}$$

Given the training samples $\mathbf{X}_T$, a new representation of training space can be obtained as $\mathbf{X}_{New} = \mathbf{X}_T\beta^T$. The manifold learning exploits the marginal probability distribution for semi-supervised problems [36] with the following assumptions; (1): The both $\mathbf{X}_T$ and $\mathbf{X}_V$ come from the same marginal probability distribution of $p(\mathbf{X})$. (2): The conditional probability of $p(y_i|x_1)$ and $p(y_i|x_2)$ should be similar if x_1 and x_2 are close enough. Thus, the manifold learning is formulated as;

$$L = \frac{1}{2}\sum_{\mathrm{ij}} W_{\mathrm{ij}}||p(y_i|x_i) - p(y_j|x_j)||^2 \tag{5}$$

where W_{ij} is the pairwise similarity among x_i and x_j. The conditional probabilities referred in Eq. (5) are computed as follows $L = \frac{1}{2}\sum_{\mathrm{ij}} W_{\mathrm{ij}}||\hat{y}_i - \hat{y}_j||^2$, where $\hat{y}_i$ and $\hat{y}_j$ are the predicted class labels for x_i and x_j [34]. This can be expressed in the form of matrix as $\hat{h} = \mathrm{Tr}(\hat{Y}^T L\hat{Y})$. $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian, in which $\mathbf{W} = \nu_{ij}$ is sparse similarity matrix, $\mathbf{D}$ is a diagonal matrix as $D_{ii} = \nu_{ij}$, and $\nu_{ij} = (exp(-\parallel x_i - x_j \parallel^2)/2\sigma^2)$.

3. Fuzziness-based generalized extreme learning machine autoencoder

Let us assume $X_T = \{x_i\}_{i=1}^n$ be the set of training samples. Thus, the optimization of GELM-AE can be reformulated as;

$$\min_{\beta\in R^{n_B\times n_o}} \frac{1}{2}||\beta||^2 + \frac{k}{2}||\mathbf{X} - \mathbf{H}\beta|^2 + \frac{\lambda}{2}\mathrm{Tr}(\beta^2\mathbf{H}^T\mathbf{L}\mathbf{H}\beta) \tag{6}$$

$$\triangledown\mathbf{L} = \beta - \Psi\mathbf{H}^T(\mathbf{X} - \mathbf{H}\beta) + \lambda\mathbf{H}^T\mathbf{L}\mathbf{H}\beta \tag{7}$$

where Ψ is a $(n \times n)$ diagonal matrix with diagonal elements $[\Psi]_{ii} = k$. Further details on how to solve Eq. (7) can be found in [37]. Similar to Eq. (2), here we discussed the two possible cases individually.

1. If $|X_T| > n_B$.

$$\beta* = \left(\mathbf{I}_{n_B} + \mathbf{H}^T\Psi\mathbf{H} + \lambda\mathbf{H}^T\mathbf{L}\mathbf{H}\right)^{-1}\mathbf{H}^T\Psi\mathbf{X} \tag{8}$$

2. If $|X_T| < n_B$. This will induced with an additional constraints to $\beta = \mathbf{H}^T\alpha$ [37].

$$\beta* = \mathbf{H}^T(\mathbf{I}_n + \Psi\mathbf{H}\mathbf{H}^T + \lambda\mathbf{L}\mathbf{H}\mathbf{H}^T)^{-1}\Psi\mathbf{X} \tag{9}$$

This will introduce a new representation of training samples $\mathbf{X}_{New}$ in n_B dimensional space as $\mathbf{X}_{New} = \mathbf{X}_T\beta^T$. The main steps of GELM-AE are summarized in Algorithm 1.

Algorithm 1. GELM-AE algorithm.

```
Data: $\mathbf{X}_T = \{x_i\}_{i=1}^n$, $n_B$, penalty coefficient $k$ and $\lambda$.
Initialize $n_B$ with random weights and biases;
if $n_B \leq n$ then
    Compute $\beta$ using Equation 8;
else
    Compute $\beta$ using Equation 9;
end
$\mathbf{X}_{New} = \mathbf{X}_T\beta^T$;
```

Here we further check the deep structure in terms of multi-layer GFELM-AE based AL framework to preserve the high-level underlying structure of HSI cube. The multi-layer GELM is a stacked feed-forward neural network that employs unsupervised GELM-AE as a base building block. Each layer as shown in Algorithm 1 is trained in feed-forward sense and uses the outputs of the first layer as the input of the second layer and by this analogy, one can train a multi-layer GFELM-AE. If multi-layer GELM-AE is trained well, it will capture the excellent potential structure and then can provide a better class marginal probability distribution μ_{ij} matrix of unseen samples from which one can compute the fuzziness of unseen samples. The marginal probability matrix should satisfy the properties defined in [15]. For the true class, the marginal probability would be approximated as close to 1, whereas, if the output is small, the marginal probability would be approximated as close to 0. However, AL methods do not require accurate probabilities, but only need a ranking of the samples scores according to their marginal probabilities which would help to estimate the fuzziness and the output of the sample.

Finally, the fuzziness ($E(\mu)$) upon ($N \times M$) samples for Y classes from the membership matrix (μ_{ij}) can then be defined as expressed in Eq. (10) which must satisfy the properties defined in [15].

$$E(\mu) = \frac{-1}{N.\,Y}\sum_{i=1}^{N}\sum_{j=1}^{Y}[\mu_{\mathrm{ij}}\log(\mu_{\mathrm{ij}}) + (1-\mu_{\mathrm{ij}})\log(1-\mu_{\mathrm{ij}})] \tag{10}$$

Then, we first associate $E(\mu)$, predicted class labels, and actual class labels with $\mathbf{X}_V$ and then sort the $\mathbf{X}_V$ in descending order based on the fuzziness values. We then heuristically select the $\hat{m}$ number of misclassified samples which have higher fuzziness to include into the training set, where $\hat{m} \ll m$. The proposed strategy keeps the pool of $\hat{m}$ new samples balanced, giving equal representation to all classes, which is achieved via softening the thresholds at run time.

4. Experimental dataset

The performance of our proposed pipeline is validated on a benchmark Salinas dataset acquired by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. Salinas dataset was acquired with 224 bands through the AVIRIS sensor over the Salinas Valley, California. Salinas scene is characterized by a high spatial resolution of 3.7 m pixels. The total area covered by the Salinas dataset comprises 512 × 217 samples per band. However, to make the system computationally less complex and for quick processing, we intentionally selected a subset of the Salinas dataset. This subset comprises 86 × 83 pixels per band located within Salinas at [*lines*, *samples*] = [158–240, 591–676] for total of six classes. There are 20 water absorption bands in the Salinas dataset, which were removed before the experiments. The removed bands are [108–112, 154–167, 224]. Salinas dataset is available only at-sensor radiance which includes vegetables, bare soils, and vineyard fields. Further information about the Salinas dataset can be found in [38,39].

5. Experimental results

Herein, a set of experiments are done to evaluate the performance of GFELM-AE using AVIRIS dataset. Evaluating such a dataset is a challenging task dominated by nested regions and complex classes. We further evaluate the influence of the number of training samples on classification performance. The performance of GFELM-AE based AL is measured using two well know metrics, namely, overall accuracy (OA) and kappa (κ) coefficient. Furthermore, all the important tuning parameters are carefully tuned and optimized before the experimental setup. All these experiments are carried out using MATLAB (2014b) on an Intel Core i5, 3.20 GHz CPU with 12 GB of RAM. To validate the performance of GFELM-AE, several statistical tests namely, recall, precision, and F1-Score are also conducted.

In all the experiments listed in Tables 1–7, the initial number of training samples size is set as 50 number of samples from the entire population. Later in each iteration, the size of the training set increases with 1% selected samples by GFELM-AE. From Tables 1–7, the observations confirm that the GFELM-AE can greatly improve the results based on a small portion from the entire population, i.e., the classifier trained using a limited number of selected samples can produce better generalization performance rather than randomly selecting the bulk amount of training samples.

The experiments shown in Tables 1–7 provide the number of training samples and the test samples which indicate the number of true versus estimated labels. It can also be observed from listed results, that fuzziness diversity-based AL is quite robust as it achieved

Table 1
Confusion metric for 1% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.47%, Weeds = 95.99%, 4wk = 92.36%, 5wk = 97.54%, 6wk = 99.39%, and 7wk = 98.59%.

Class	Training performance						Testing performance with kappa (κ) 96.34					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	15	0	0	0	0	0	374	0	0	**2**	0	0
Weeds	0	21	0	0	0	0	0	1269	**19**	**2**	0	**32**
4wk	0	0	12	**2**	0	0	0	0	556	**46**	0	0
5wk	0	0	0	20	**1**	0	0	0	**3**	1467	**34**	0
6wk	0	0	0	0	15	0	0	0	0	0	655	**4**
7wk	0	0	0	0	0	14	0	0	0	0	**11**	774

Table 2
Confusion metric for 2% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.61%, 4wk = 99.99%, 5wk = 93.62%, 6wk = 99.24%, and 7wk = 98.98%.

Class	Training performance						Testing performance with kappa (κ) 97.29					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	17	0	0	0	0	0	374	0	0	0	0	0
Weeds	0	38	**2**	0	0	0	0	1298	**3**	**2**	0	0
4wk	0	0	28	0	0	0	0	0	588	0	0	0
5wk	0	0	**1**	33	0	0	0	0	**95**	1396	0	0
6wk	0	0	0	0	15	0	0	0	0	0	654	**5**
7wk	0	**1**	0	0	**1**	14	0	**2**	0	**2**	**4**	775

Table 3
Confusion metric for 3% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.77%, 4wk = 95.58%, 5wk = 99.99%, 6wk = 98.63%, and 7wk = 99.87%.

Class	Training performance						Testing performance with kappa (κ) 99.05					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	17	0	0	0	0	0	374	0	0	0	0	0
Weeds	0	39	**1**	2	0	0	0	1298	**3**	0	0	0
4wk	0	0	22	**6**	0	0	0	**11**	562	**15**	0	0
5wk	0	0	0	76	0	0	0	0	0	1449	0	0
6wk	0	0	0	0	15	**2**	0	0	0	6	648	**3**
7wk	0	**2**	0	0	**2**	16	0	**1**	0	0	0	778

Table 4
Confusion metric for 4% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.99%, 4wk = 94.84%, 5wk = 99.10%, 6wk = 94.58%, and 7wk = 99.61%.

Class	Training performance						Testing performance with kappa (κ) 98.03					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	19	0	0	0	0	0	372	0	0	0	0	0
Weeds	0	40	**5**	**2**	0	0	0	1296	0	0	0	0
4wk	0	**1**	45	**8**	0	0	0	0	533	**29**	0	0
5wk	0	0	**5**	75	0	0	0	0	**13**	1432	0	0
6wk	0	0	0	**1**	23	**4**	0	0	0	**35**	611	0
7wk	0	**3**	0	0	**3**	16	0	0	0	0	**3**	774

Table 5
Confusion metric for 5% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.85%, 4wk = 98.71%, 5wk = 939.39%, 6wk = 9.36%, and 7wk = 99.99%.

Class	Training performance						Testing performance with kappa (κ) 97.33					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	19	0	0	0	0	0	372	0	0	0	0	0
Weeds	0	40	**7**	0	0	0	0	1294	**2**	0	0	0
4wk	0	0	59	**14**	0	0	0	0	536	**7**	0	0
5wk	0	0	**5**	81	**2**	0	0	0	**75**	1342	**20**	0
6wk	0	0	0	**5**	42	**2**	0	0	0	**4**	621	0
7wk	0	**2**	0	0	**4**	18	0	0	0	0	0	775

higher classification results which are comparable or way better than several state-of-the-art techniques.

To better analyze the performance of GFELM-AE, Table 8 shows the statistical significance in terms of recall, precision, and F1-score tests. The experiments shown in Table 8 are performed with different numbers of actively selected samples from each class. Table 8 is produced to support the results shown in Tables 1–7. The global recall, precision, and F1-score for each iteration of these results are obtained using 5 Monte Carlo runs. Furthermore, Table 8 shows the statistical significance of GFELM-AE in terms of recall, precision, and F1-score with the 99% confidence interval. For any good model, precision, recall, and F1-score values should be greater than 80% on average, and in our case, these values are almost above 97%.

Table 6
Confusion metric for 6% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.99%, 4wk = 99.99%, 5wk = 99.99%, 6wk = 99.99%, and 7wk = 99.48%.

Class	Training performance						Testing performance with kappa (κ) 99.89					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	19	0	0	0	0	0	372	0	0	0	0	0
Weeds	0	41	**5**	**2**	0	0	0	1295	0	0	0	0
4wk	0	0	70	**6**	0	0	0	0	540	0	0	0
5wk	0	0	0	132	0	0	0	0	0	1393	0	0
6wk	0	0	0	0	50	**1**	0	0	0	0	623	0
7wk	0	0	0	0	**4**	20	0	0	0	0	**4**	771

Table 7
Confusion metric for 7% training samples selected through GFELM-AE. The overall accuracy per class is as follows: Weeds1 = 99.99%, Weeds = 99.99%, 4wk = 99.99%, 5wk = 99.99%, 6wk = 99.99%, and 7wk = 99.89%.

Class	Training performance						Testing performance with kappa (κ) 99.98					
	Weeds1	Weeds	4wk	5wk	6wk	7wk	Weeds1	Weeds	4wk	5wk	6wk	7wk
Weeds1	22	0	0	0	0	0	369	0	0	0	0	0
Weeds	0	53	**6**	**2**	0	0	0	1282	0	0	0	0
4wk	0	0	58	**20**	0	0	0	0	538	0	0	0
5wk	0	0	0	150	0	0	0	0	0	1375	0	0
6wk	0	0	0	0	52	**3**	0	0	0	0	619	0
7wk	0	**1**	0	0	**3**	30	0	0	0	0	**1**	764

Table 8
Statistical confidence of GFELM-AE with 99% confidence interval.

Iteration	Recall	Precision	F1-score
1	0.9722 ± 0.0089	0.9700 ± 0.0083	0.9708 ± 0.0059
2	0.9858 ± 0.0080	0.9734 ± 0.0186	0.9785 ± 0.0097
3	0.9897 ± 0.0057	0.9946 ± 0.0018	0.9921 ± 0.0029
4	0.9802 ± 0.0085	0.9881 ± 0.0058	0.9840 ± 0.0056
5	0.9855 ± 0.0084	0.9725 ± 0.0162	0.9783 ± 0.0095
6	0.9991 ± 0.0007	0.9989 ± 0.0009	0.9990 ± 0.0005
7	0.9998 ± 0.0001	0.9995 ± 0.0004	0.9999 ± 0.0001

6. Comparisons and discussions

The most recent advancements in AL methods consist of hybrid and single-pass context learning. These methods jointly investigate adaptive and incremental learning from the field of traditional and online machine learning. Such investigations have resulted in several nontrivial AL methods, such as [40,41] considered online learning frameworks in a way to handle online single pass settings in which the data stream samples arrive without interruption i.e., continuously, thus does not allow classifier retraining. Moreover, these works only focused on close concepts of ignorance[3] and conflict.[4]

The GFELM-AE-AL has been rigorously compared with some significant works published in the HIC area adopting different sample selection methods in recent years. These samples' selection methods include random sampling (RS),[5] breaking ties (BT),[6] modified breaking ties (MBT),[7] and mutual information (MI).[8]

The comparative results are based on 5 Monte Carlo runs with 50 number of randomly selected initial training samples. However,

[3] Ignorance represents the distance between already seen training samples and new sample.

[4] Conflict models how close a query point is to the actual class boundary.

[5] RS [27] relies on the random selection of the samples without considering any specific criteria.

[6] BT [42] relies on the smallest difference of the posterior probabilities for each sample. In multiclass settings, BT can be applied by calculating the difference between the two highest probabilities. As a result, BT finds the samples minimizing the distance between the first two most probable classes. The BT method generally focuses on the boundaries comprising many samples, possibly disregarding boundaries with fewer samples.

[7] MBT [19] includes more diversity in the sampling process as compared to BT. The samples are selected by maximizing the probability of the largest class for each class. MBT takes into account all the class boundaries by conducting the sampling in cyclic fashion, making sure that the MBT does not get trapped in any class whereas BT could be trapped in a single boundary.

[8] MI [18] of two samples is a measure of the mutual dependence between the two samples.

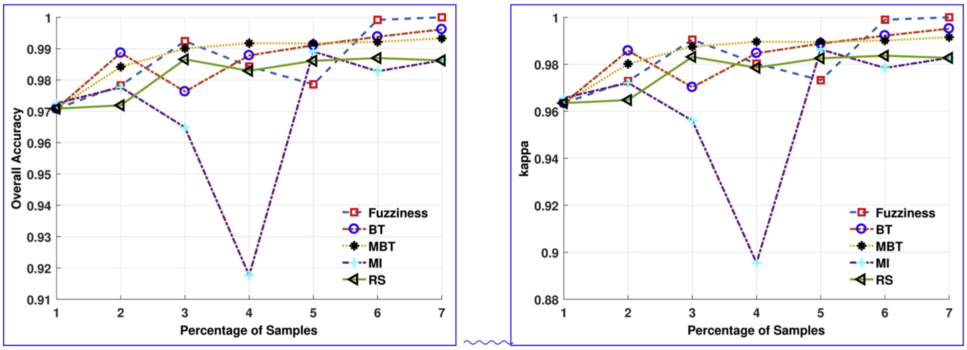

Fig. 1. Overall and Kappa accuracy with different percentage (%) of training samples selected in each iteration using different sample selection methods. It is perceived from the above figure that by including the samples back to the training set, the classification results in terms of accuracy are significantly improved.

in each iteration, 50 newly selected samples (with each sample selection method, individually) are selected and added back to the training set to retrain the classifier. The obtained results are presented in Fig. 1, and based on the results, one can argue that the GFELM-AE outperforms with almost all kinds of sample selection methods. By this comparison, we experimentally observed that fuzziness-based samples selection slightly outperforms than other sample selection methods because all other methods are more often subjective and tend to include redundancy into the classifier. In such situations, there is a risk the model may get overwhelmed due to the spatially miscellaneous and uninformative samples. Of course, ELM-AE can be easily integrated with any type of sample selection method based on the requirements.

As earlier explained, we start evaluating the model based on randomly selected 50 number of samples, and it is proven the fact that randomly adding samples back to the training set slightly increases the accuracies however the model becomes computationally complex. Therefore, we first identify the set of misclassified samples and then select a specific percentage of these samples to add them back to the original training samples. Classifier trained on such samples provides better generalization performance on samples which were initially misclassified.

7. Conclusion

The spatial-spectral classification of HSI with a limited number of training samples is a challenging task. To effectively tackle such issues, this paper introduces a GFELM-AE-AL method to reduce the sample selection bias while maintaining the data stability in the spatial domain. Experimental results on a benchmark dataset show that GFELM-AE leads to an increased predictive power regarding the statistical significance i.e., precision, F1-score, and recall rates. A comparative study has also been carried out confirming that GFELM-AE is an effective classification method with a fewer number of training samples.

Declaration of interest statement

The authors declare no financial interest exists.

References

[1] M. Ahmad, A. Khan, A. Khan, M. Mazzara, S. Distefano, A. Sohaib, O. Nibouche, Spatial prior fuzziness pool-based interactive classification of hyperspectral images, Remote Sens. 11 (May (5)) (2019).
[2] A. Schneider, H. Feussner, Chapter 5 – Diagnostic Procedures, Institute of Minimally Invasive Interdisciplinary Therapeutic Interventions (MITI), Technische Universität München (TUM), Biomedical Engineering in Gastrointestinal Surgery, London, 2017.
[3] Y. Qu, H. Qi, C. Kwan, Unsupervised sparse dirichlet-net for hyperspectral image super-resolution, CVPR'18, CoRR abs/1804.05042 (2018).
[4] M. Ahmad, A.M. Khan, R. Hussain, S. Protasov, F. Chow, A.M. Khattak, Unsupervised geometrical feature learning from hyperspectral data, 2016 IEEE Symposium Series on Computational Intelligence (SSCI) (2016) 1–6.
[5] M. Ahmad, A.M. Khan, R. Hussain, Graph-based spatial-spectral feature learning for hyperspectral image classification, IET Image Process. 11 (December (12)) (2017) 1310–1316.
[6] M. Ahmad, A.K. Bashir, A.M. Khan, Metric similarity regularizer to enhance pixel similarity performance for hyperspectral unmixing, Optik 140C (2017) 86–95.
[7] L. Yang, A.M. MacEachren, P. Mitra, T. Onorati, Visually-enabled active deep learning for (geo) text and image classification: a review, ISPRS Int. J. Geo-Inf. 7 (2) (2018) 65.
[8] C. Liu, L. He, Z. Li, J. Li, Feature-driven active learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 56 (January) (2018) 341–354.
[9] G. Mountrakis, J. Im, C. Ogole, Support vector machines in remote sensing: a review, ISPRS J. Photogramm. Remote Sens. 66 (3) (2011) 247–259.
[10] J. Li, J.M. Bioucas-Dias, A. Plaza, Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields, IEEE Trans. Geosci. Remote Sens. 50 (3) (2012) 809–823.
[11] J. Xia, D. Peijun, H. Xiyan, J. Chanussot, Hyperspectral remote sensing image classification based on rotation forest, IEEE Geosci. Remote Sens. Lett. 11 (1) (2014) 239–243.
[12] B. Pan, Z. Shi, X. Xu, Hierarchical guidance filtering-based ensemble classification for hyperspectral images, IEEE Trans. Geosci. Remote Sens. 55 (7) (2017) 4177–4189.
[13] B. Pan, Z. Shi, X. Xia, R-vcanet: a new deep-learning-based hyperspectral image classification method, IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. 10 (5) (2017) 1975–1986.

[14] K. Tan, P. Du, Hyperspectral remote sensing image classification based on support vector machine, J. Infrared Millim. Waves 27 (2) (2013) 123–128.

[15] M. Ahmad, S. Protasov, A.M. Khan, R. Hussian, A.M. Khattak, W.A. Khan, Fuzziness-based active learning framework to enhance hyperspectral image classification performance for discriminative and generative classifiers, PLOS ONE 13 (1) (2018) e0188996.

[16] G. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Inf. Theory 14 (January) (1968) 55–63.

[17] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, W.J. Emery, Svm active learning approach for image classification using spatial information, IEEE Trans. Geosci. Remote Sens. 52 (April) (2014) 2217–2233.

[18] J. Li, J.M. Bioucas-Dias, A. Plaza, Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning, IEEE Trans. Geosci. Remote Sens. 51 (February) (2013) 844–856.

[19] Q. Shi, B. Du, L. Zhang, Spatial coherence-based batch-mode active learning for remote sensing image classification, IEEE Trans. Image Process. 24 (July) (2015) 2037–2050.

[20] J. Li, J.M. Bioucas-Dias, A. Plaza, Hyperspectral image segmentation using a new bayesian approach with active learning, IEEE Trans. Geosci. Remote Sens. 49 (October) (2011) 3947–3960.

[21] D.D. Lewis, W.A. Gale, A sequential algorithm for training text classifiers, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'94, Springer-Verlag New York, Inc., New York, NY, USA, 1994, pp. 3–12.

[22] W. Di, M.M. Crawford, Active learning via multi-view and local proximity co-regularization for hyperspectral image classification, IEEE J. Sel. Topics Signal Process. 5 (June) (2011) 618–628.

[23] H.S. Seung, M. Opper, H. Sompolinsky, Query by committee, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT'92, ACM, New York, NY, USA, 1992, pp. 287–294.

[24] S. Rajan, J. Ghosh, M.M. Crawford, An active learning approach to hyperspectral data classification, IEEE Trans. Geosci. Remote Sens. 46 (April) (2008) 1231–1242.

[25] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, W.J. Emery, Improving active learning methods using spatial information, 2011 IEEE International Geoscience and Remote Sensing Symposium (2011 July) 3923–3926.

[26] A. Liu, G. Jun, J. Ghosh, Active learning of hyperspectral data with spatially dependent label acquisition costs, 2009 IEEE International Geoscience and Remote Sensing Symposium, vol. 5, July, 2009 pp. V-256–V-259.

[27] D. Tuia, M. Volpi, L. Copa, M. Kanevski, J. Munoz-Mari, A survey of active learning algorithms for supervised remote sensing image classification, IEEE J. Sel. Topics Signal Process. 5 (June) (2011) 606–617.

[28] T. Haines, T. Xiang, Active learning using dirichlet processes for rare class discovery and classification, Proceedings of the British Machine Vision Conference (2011) pp. 9.1–9.11.

[29] J. Michel, J. Malik, J. Inglada, Lazy yet efficient land-cover map generation for hr optical images, 2010 IEEE International Geoscience and Remote Sensing Symposium, July, 2010, pp. 1863–1866.

[30] B. Demir, C. Persello, L. Bruzzone, Batch-mode active-learning methods for the interactive classification of remote sensing images, IEEE Trans. Geosci. Remote Sens. 49 (March) (2011) 1014–1031.

[31] J. Munoz-Mari, D. Tuia, G. Camps-Valls, Semisupervised classification of remote sensing images with active queries, IEEE Trans. Geosci. Remote Sens. 50 (October) (2012) 3751–3763.

[32] L. He, J. Li, C. Liu, S. Li, Recent advances on spectral-spatial hyperspectral image classification: an overview and new guidelines, IEEE Trans. Geosci. Remote Sens. 56 (March) (2018) 1579–1597.

[33] W.B. Johnson, J. Lindenstrauss, G. Schechtman, Extensions of lipschitz maps into banach spaces, Israel J. Math. 54 (June) (1986) 129–138.

[34] K. Sun, J. Zhang, C. Zhang, J. Hu, Generalized extreme learning machine autoencoder and a new deep neural network, Neuro Comput. 230 (2016).

[35] C. Pan, D.S. Park, Y. Yang, H.M. Yoo, Leukocyte image segmentation by visual attention and extreme learning machine, Neural Comput. Appl. 21 (September) (2012) 1217–1227.

[36] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (December) (2006) 2399–2434.

[37] M.D. Tissera, M. Mcdonnell, Deep extreme learning machines: supervised autoencoding architecture for classification, Neurocomputing 174 (2015).

[38] M. Ahmad, M.A. Alqarni, A.M. Khan, R. Hussain, M. Mazzara, S. Distefano, Segmented and non-segmented stacked denoising autoencoder for hyperspectral band reduction, Optik 180 (2019) 370–378.

[39] Hyperspectral Datasets Description, (2018) (accessed 30.06.18), http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

[40] E. Lughofer, Single-pass active learning with conflict and ignorance, Evol. Syst. 3 (December) (2012) 251–271.

[41] M. Woodward, C. Finn, Active one-shot learning, CoRR abs/1702.06559 (2017).

[42] T. Luo, K. Kramer, S. Samson, A. Remsen, D.B. Goldgof, L.O. Hall, T. Hopkins, Active learning to recognize multiple types of plankton, Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 3 (2004 Aug) 478–481.