Fusion of Extreme Learning Machine and Graph-Based Optimization Methods for Active Classification of Remote Sensing Images

Mohamed A. Bencherif, *Member, IEEE*, Yakoub Bazi, *Senior Member, IEEE*, Abderrezak Guessoum, Naif Alajlan, *Senior Member, IEEE*, Farid Melgani, *Senior Member, IEEE*, and Haikel AlHichri, *Member, IEEE*

***Abstract*—In this letter, we propose an efficient multiclass active learning (AL) method for remote sensing image classification. We fuse the capabilities of an extreme learning machine (ELM) classifier and graph-based optimization methods to boost the classification accuracy while minimizing the user interaction. First, we use the ELM to generate an initial label estimation of the unlabeled image pixels. Then, we optimize a graph-based functional energy that integrates the ELM outputs as an initial estimation of the image structure. As for the ELM, the solution to this multiclass optimization problem leads to a system of linear equations. Due to the sparse Laplacian matrix built from the lattice graph defined on the image pixels, the optimization problem is solved in a linear time. In the experiments, we report and discuss the results of the proposed AL method on two very high resolution images acquired by IKONOS-2 and GoeEye-1, as well as the well-known Pavia University hyperspectral image.**

***Index Terms*—Active learning (AL), extreme learning machine (ELM), graph-based optimization, multiclass classification.**

I. Introduction

In THE last few years, active learning (AL) has become popular in the remote sensing community as a solution for improving the classification of images with limited training samples [1], [2]. In an AL setting, a user is asked to label a set of pixels or segments in an image through an iterative process. The aim of AL is to rank the learning set according to a criterion that allows us to select the most useful samples to improve the model, thus minimizing the number of training samples necessary to maintain discrimination capabilities as high as possible. From the works available in literature, it clearly appears that most of the research efforts were devoted to the definition of improved spectral–spatial criteria to select uncertain and diverse unlabeled pixels. See [1] and [2] for more details.

Another possible solution for reducing user interactions is to automatically enhance the classifier accuracy at each iteration of the AL process. To this end, the exploitation of semisupervised learning paradigms that aim to jointly learn labeled and unlabeled samples appears to be an interesting alternative [3]. Usually, this learning paradigm calls for graph theory to build the relationship between the labeled and unlabeled samples through what we call the combinatorial Laplacian matrix [4]. Then, the functional of the classifier is augmented with this regularizer, and the final solution is expressed in the kernel space. It is worth recalling that this concept has been commonly adopted for the support vector machine (SVM) classifiers [4]. However, in the AL context, this learning mode will: 1) overload the classifier; 2) only allow a partial exploitation of the spatial–contextual information; and 3) not permit a proper exploitation of the power of graph-based optimization methods.

To meet the requirements of AL, we view the problem in a different way. In particular, we formulate the optimization problem in the graph space by defining an appropriate graph-based energy functional that incorporates the classifier output as an initial estimation of the image structure. As a result, the solution to this multiclass optimization problem leads to a linear system of equations. Due to the sparse format of the combinatorial Laplacian matrix induced by the lattice graph defined on the image, this system of equations can be efficiently solved in a linear time. As a base classifier, we use in this letter an extreme learning machine (ELM) classifier for generating the initial estimations of the unlabeled samples [5]. Compared with the existing learning methods such as the SVM, the ELM classifier is characterized by several attractive proprieties: 1) it has a unified formulation for binary, multiclass, and regression problems, and the solution of these problems is also given in a unified analytical form; 2) the feature mapping could be done either in a known space similar to neural networks or in an infinite space similar to kernel methods; and 3) for multiclass classification, it uses a configuration of multioutput nodes where the number of nodes is equal to the number of classes. In a recent contribution, the ELM has shown interesting results for the classification of hyperspectral images [6]. It is worth noting that the new proposed classification method named as ELM random walker (ELMRW) can be viewed as a Markov random field (MRF) approach. However, compared with standard MRF methods [7], the image smoothness is introduced using a lattice graph, and the solution of the optimization problem is unique and given in a compact form.

Manuscript received May 13, 2014; revised July 4, 2014 and August 2, 2014; accepted August 13, 2014. This work was supported by the Deanship of Scientific Research of the King Saud University through the International Research Group under Project IRG14-20.

M. A. Bencherif and A. Guessoum are with the Department of Electronics, Faculty of Technology, Saad Dahlab University, 9000 Blida, Algeria (e-mail: mbencherif1@yahoo.com; abderguessoum@yahoo.com).

Y. Bazi, N. Alajlan, and H. AlHichri are with the ALISR Laboratory, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia (e-mail: ybazi@ksu.edu.sa; najlan@ksu.edu.sa; hhichri@ksu.edu.sa).

F. Melgani is with the Department of Information Engineering and Computer Science, University of Trento, 38213 Trento, Italy (e-mail: melgani@disi.unitn.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2014.2349538

II. Proposed AL Method

Let us consider a training set initially composed of n (assumed to be small) labeled samples $S = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x}_i$ represents the features of interest (e.g., the spectral bands and morphological profile (MP) features extracted from them), and y_i is a discrete label defined among P possible classes. The inclusion of MP features as an additional source of contextual information is justified by their success in enhancing the classification accuracy of very high resolution (VHR) images [1], [6]. We also consider a learning set composed of m unlabeled samples $U = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$, with $m \gg n$, which is usually the rest of image I (i.e., $U = I - S$). In the following, we detail the proposed ELMRW classification method, and then, we illustrate its application in the context of AL.

A. Classification With ELM

The ELM output $\mathbf{f}^0(\mathbf{x}_\ell) \in \Re^p$ for a test sample $\mathbf{x}_\ell$ can be expressed in the kernel space as follows [5], [6]:

$$\mathbf{f}^0(\mathbf{x}_\ell) = \begin{bmatrix} k(\mathbf{x}_\ell, \mathbf{x}_1) \\ \vdots \\ k(\mathbf{x}_\ell, \mathbf{x}_n) \end{bmatrix}^T (\mathbf{I}/C + \mathbf{K})^{-1}\mathbf{Y}. \quad (1)$$

The first term in (1) is a vector of length n and represents the kernel distances between test point $\mathbf{x}_\ell$ and the training samples. $\mathbf{K} \in \Re^{n\times n}$ is the kernel matrix computed from the training samples. Here, we use the common Gaussian kernel, i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, where γ represents a parameter that is inversely proportional to the width of the kernel. $\mathbf{I} \in \Re^{n\times n}$ is an identity matrix, and C is a regularization parameter. $\mathbf{Y} \in \Re^{n\times P}$ is the output matrix built from training labels y_i. Each row vector of $\mathbf{Y}$ (of dimension P) has all its values set to 0, except the entry matching class label y_i, which is set to 1. A test sample $\mathbf{x}_\ell$ will be assigned to the index of the output node that has the highest value. In other words, if we let $\mathbf{f}^0(\mathbf{x}_\ell) = [f_1^0(\mathbf{x}_\ell), \ldots, f_P^0(\mathbf{x}_\ell)]^T$, then the predicted class label for sample $\mathbf{x}_\ell$ is

$$y_\ell^* = \arg\max_{k\in\{1,\ldots,P\}} f_k^0(\mathbf{x}_\ell). \quad (2)$$

In the following, for computation convenience, we transform the ELM output into probabilities using the softmax function as follows:

$$f_k^*(\mathbf{x}_\ell) = \frac{\exp\left(f_k^0(\mathbf{x}_\ell)\right)}{\sum_{j=1}^{P}\exp\left(f_j^0(\mathbf{x}_\ell)\right)}, \quad k = 1, \ldots, P. \quad (3)$$

B. Classification With ELMRW

Let us consider $G = (V, E)$ a graph with vertices $v \in V$ and edges $e \in E$. Each pixel in image I is associated with a vertex v_i, and the vertices are locally connected via an eight-connected lattice. An edge spanning two vertices v_i and v_j is denoted by e_{ij}, and the associated weight is denoted by $\omega(e_{ij})$, or simply ω_{ij}. A common choice for obtaining these weights is the Gaussian weighting function, i.e., $\omega_{ij} = \exp(-\beta\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, and β is a free parameter. The degree of a vertex v_i is $d_i = \Sigma\omega_{ij}$ for all edges e_{ij} incident on v_i. Then, the combinatorial Laplacian matrix indexed by vertices v_i and v_j is given by

$$L_{ij} = \begin{cases} d_{ij}, & \text{if } i = j \\ -\omega_{ij}, & \text{if } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

In the binary case, a possible definition of the energy minimization problem can be given in the following general form [11]:

$$\min_f \sum_{e_{ij}\in E} \omega_{ij}^p \|f_i - f_j\|^q + \lambda \sum_{v_i\in V} \mu_i^p \|f_i - f_i^*\|^q. \quad (5)$$

The first term is related to image smoothness. The second term (the data fidelity term) encodes an initial estimation of the image structure, and μ_i and λ are the local and global weights, respectively, enforcing that fidelity. In our case, the initial estimation is the ELM posterior probability f_i^*. Moreover, local weights μ_i could be set equal to d_i, but for simplicity, we suppose that $\mu_i = 1$ for all the pixels in the image.

Depending on the values of p and q, different energy models can be obtained [8]. Here, we focus on the case when $p = 1$ and $q = 2$, which leads to the so-called random walker algorithm [9]. Under this assumption, the corresponding energy is

$$\min_f \sum_{e_{ij}\in E} \omega_{ij} \|f_i - f_j\|^2 + \lambda \sum_{v_i\in V} \|f_i - f_i^*\|^2. \quad (6)$$

In the multiclass case, (6) can be written as follows:

$$\min_f \sum_{e_{ij}\in E} \omega_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 + \lambda \sum_{v_i\in V} \mu_i \|\mathbf{f}_i - \mathbf{f}_i^*\|^2 \quad (7)$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are now probability vectors (of dimension P) associated with vertices v_i and v_j, respectively. $\mathbf{f}_i^* \in \Re^P$ is the ELM posterior probability associated with a vertex v_i. It can be shown that (7) can be written in the following matrix form:

$$\mathcal{L} = \text{Trace}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \lambda\|\mathbf{F} - \mathbf{F}^*\|_F^2 \quad (8)$$

where $\|\cdot\|_F$ is the Frobenius norm. $\mathbf{F} \in \Re^{(n+m)\times P}$ and $\mathbf{F}^* \in \Re^{(n+m)\times P}$ are built from probability vectors $\mathbf{f}_k \in \Re^P$ and $\mathbf{f}_k^* \in \Re^P$, respectively, with $k = 1, \ldots, n+m$. $\mathbf{L} \in \Re^{(n+m)\times(n+m)}$ is the Laplacian matrix generated from all possible v_i and v_j pairs. We note that $\mathbf{L}$ is a sparse and positive semidefinite matrix.

Since $V = V_S \cup V_U$, we can decompose (8) to obtain

$$\mathcal{L} = \text{Trace}\left(\begin{bmatrix}\mathbf{F}_S \\ \mathbf{F}_U\end{bmatrix}^T \begin{bmatrix}\mathbf{L}_S & \mathbf{B} \\ \mathbf{B}^T & \mathbf{L}_U\end{bmatrix} \begin{bmatrix}\mathbf{F}_S \\ \mathbf{F}_U\end{bmatrix}\right) + \lambda \left\| \begin{bmatrix}\mathbf{F}_S \\ \mathbf{F}_U\end{bmatrix} - \begin{bmatrix}\mathbf{F}_S^* \\ \mathbf{F}_U^*\end{bmatrix} \right\|_F^2 \quad (9)$$

where $\mathbf{F}_U \in \Re^{m\times P}$ is unknown. $\mathbf{B} \in \Re^{n\times m}$, $\mathbf{L}_S \in \Re^{n\times n}$, and $\mathbf{L}_U \in \Re^{m\times m}$ are the submatrices of $\mathbf{L}$ issued from the decomposition of V into labeled and unlabeled sets V_S and V_U, respectively. Then, after some mathematical manipulations, we obtain the following functional energy:

$$\mathcal{L} = \text{Trace}\left(\mathbf{F}_S^T\mathbf{L}_S\mathbf{F}_S + \mathbf{F}_U^T\mathbf{L}_U\mathbf{F}_U + 2\mathbf{F}_U^T\mathbf{B}^T\mathbf{F}_S\right) + \lambda\|\mathbf{F}_S - \mathbf{F}_S^*\|_F^2 + \lambda\|\mathbf{F}_U - \mathbf{F}_U^*\|_F^2. \quad (10)$$

To minimize (10) with respect to $\mathbf{F}_U$, we compute

$$\frac{d\mathcal{L}}{d\mathbf{F}_U} = 2\lambda(\mathbf{F}_U - \mathbf{F}_U^*) + 2\mathbf{L}_U\mathbf{F}_U + 2\mathbf{B}^T\mathbf{F}_S = 0. \quad (11)$$

Then, the probabilities associated with the unlabeled pixels are obtained by solving the following linear system of equations:

$$\mathbf{F}_U = (\mathbf{L}_U + \lambda \mathbf{I})^{-1} \left(-\mathbf{B}^T \mathbf{F}_S + \lambda \mathbf{F}_U^*\right) \tag{12}$$

where $\mathbf{L}_U + \lambda \mathbf{I}$ is a positive-semidefinite matrix. Know that $\mathbf{L}$ is a sparse matrix as it is generated from an eight-connected lattice graph; then, this linear system of equations can be solved in a linear time.

C. AL Algorithm

In an AL setting, we first start by training the ELMRW on training set S. The resulting classification model is used to sort the unlabeled samples of learning set U. In this letter, we evaluate each sample using the breaking ties (BT) criterion [10], which is based on the posterior probabilities of associating a sample to a given class. In a multiclass context, the difference between the two highest probabilities is indicative of the way that a sample is handled by the classifier. When the two highest probabilities are close, the classifier confidence is low. From the sorted samples, the most uncertain N_s samples are labeled by the human expert and added to training set S. The entire process is iterated until a predefined convergence condition is satisfied. The following algorithm summarizes the proposed AL strategy.

Algorithm: AL with ELMRW

Input: - Image I
- Initial training set $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$
- Number of active iterations $ITER$
- Number of samples to label at every iteration N_s
- Regularization parameter λ
- Gaussian weighting function parameter β

Output: - Classification result

Start

1: Construct an eight-connected lattice graph using the image pixels as vertices.
2: Compute Laplacian matrix $\mathbf{L}$ with (4).
3: Set the number of iterations $iter = 0$.
4: While $iter \leq ITER$ do
  4.1: Estimate the best parameters (C, γ) of the ELM using the differential evolution algorithm proposed in [6].
  4.2: Train the ELM on training set S using these parameters, and then compute prediction $\mathbf{F}^*$ for the unlabeled set using (1) and (3).
  4.3: Use (12) to compute the new prediction $\mathbf{F}_U$ for the unlabeled set.
  4.4: Apply the BT active criterion to rank the unlabeled set, and select the most N_s informative pixels (or segments).
  4.5: Ask an expert to label them, and augment S with these pixels.
  4.6: $iter = iter + 1$
5: Generate the final classification map.

End

III. Experimental Results

A. Data Set Description

Jeddah: The first data set represents a multispectral VHR image of size 700 pixels × 650 pixels acquired by the IKONOS-2 sensor in July 2004. The image has three spectral bands with a spatial resolution of 1 m and refers to a portion of the city of Jeddah (Saudi Arabia), in which seven land-cover types are dominant, i.e., two types of *Asphalt*, *Bare soil*, *Grass*, *Roofs*, *Trees*, and *Water*.

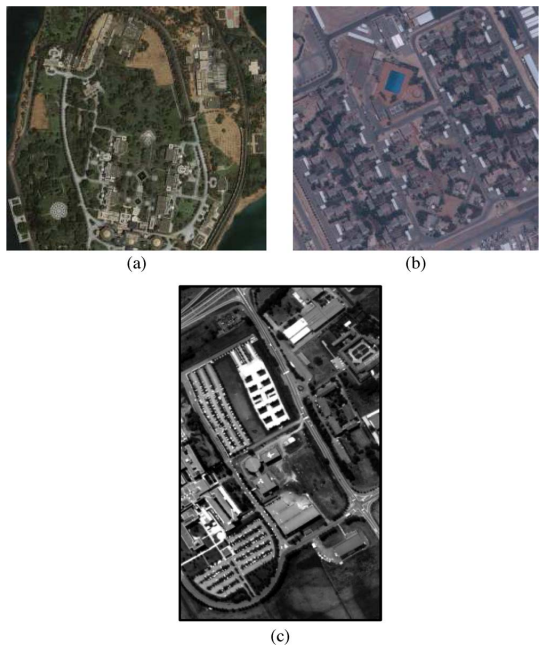

Fig. 1. Images used in the experiments. (a) Jeddah. (b) Riyadh. (c) Pavia University.

TABLE I
Data Sets Used in the Experiments

Dataset	Sensor	#Bands	#Classes	#Samples
Jeddah	IKONOS-2	3	7	70339
Riyadh	GeoEye-1	3	8	151347
Pavia University	ROSIS	102	9	42776

Riyadh: The second data set represents a multispectral VHR image of size 800 pixels × 800 pixels acquired by the GeoEye-1 sensor in August 2010. The image has three spectral bands with a spatial resolution of 0.5 m and refers to a portion of the city of Riyadh (Saudi Arabia). Eight classes are considered, i.e., two types of *Roofs*, *two types of Asphalt*, *Bare soil*, *Grass*, *Trees*, and *Water*.

Pavia University: The third data set is the well-known Pavia University hyperspectral image of size 610 pixels × 340 pixels. It is acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over a part of Pavia University (Italy) in July 2002. The image has 103 bands and is characterized by a spatial resolution of 1.3 m. Nine classes are considered, i.e., *Asphalt*, *Bare soil*, *Bitumen*, *Bricks*, *Meadows*, *Shadow*, *Tiles*, *Trees*, and *Water*.

Fig. 1 shows the corresponding images, whereas Table I summarizes the main information related to each data set.

B. Experiment Setup

For each data set, we split the available samples in two sets of equal size, which correspond to learning set U and the test set. The initial training set L (i.e., five samples per class) is randomly generated from learning set U. Then, we run the AL

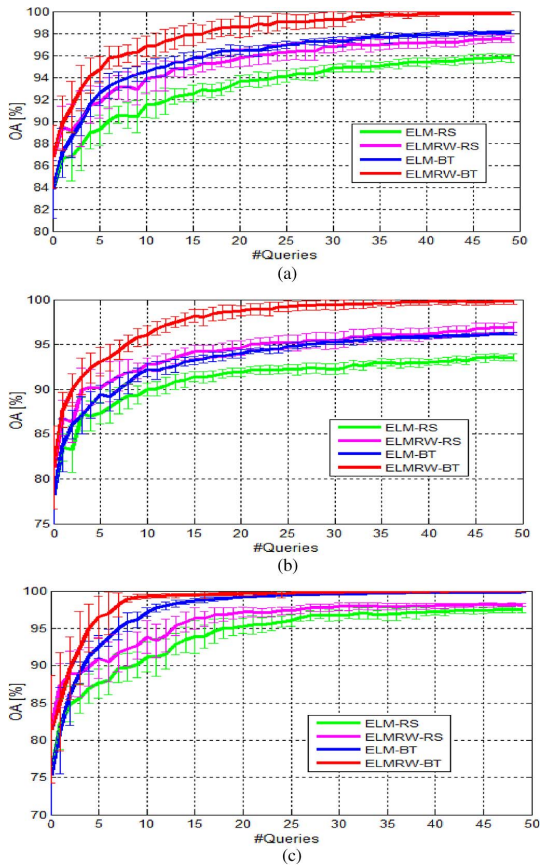

Fig. 2. OA versus the number of queries obtained by the ELM-RS, the ELM-BT, the ELMRW-RS, and the ELMRW-BT for the (a) Jeddah, (b) Riyadh, and (c) Pavia University data sets.

algorithm for 50 iterations by adding ten samples at each iteration. To statistically obtain reliable results, we run the entire AL process ten times, each time with different learning and test sets. The classification performance is presented in terms of the overall accuracy (OA), the average accuracy (AA), and the Kappa statistic [11], in addition to the standard deviations (σ) of the OA, the AA, and the Kappa. This last measure is useful for evaluating the stability of the AL method.

In the experiments, to generate the MP features for both Jeddah and Riyadh, we apply a set of opening and closing operations with reconstruction to the original spectral features using the disk-shaped structural elements of sizes 3, 6, 9, and 12, respectively. This leads to a feature vector of dimension 27. For Pavia University, we use the solution proposed in [6]. We first apply the principal component analysis (PCA) to the original spectral features, and then we generate the extra MP features, as done for Jeddah and Riyadh from the first five PCA components. This leads to a feature vector of dimension 45. For training the ELM, we estimate the best values of C and estimate γ according to a threefold cross-validation procedure. The search boundary for these values is set in the ranges [10^{-3} 1000] and [10^{-3} 10], respectively. For a better search ability, we use the model selection method based on differential evolution [6]. To compute the edge weights of the graph, we set parameter β to 1. However, we experimentally found that the choice of this parameter is not critical. We also set regularization parameter λ to 0.01 in all the experiments. A justification of this choice will be discussed in the experiments. For comparison purposes, we present the results of the proposed ELMRW versus the baseline ELM classifier with random sampling (RS) and the state-of-the-art BT strategy. However, any other active criteria could be used as well. In the rest of this letter, we term the investigated scenarios as ELM-RS, ELMRW-RS, ELM-BT, and ELMRW-BT. In the following, the experiments are carried out on an HP-Ultrabook with process Core i7 with 1.80 GHz and 8 GB of random access memory.

TABLE II
CLASSIFICATION RESULTS: (a) JEDDAH, (b) RIYADH, AND (c) PAVIA UNIVERSITY

(a)

	ELM	ELMRW	ELM		ELMRW	
Method	Initial		RS	BT	RS	BT
#Training samples	35		535			
OA	83.81	86.75	95.75	98.12	97.45	**99.87**
σ_{OA}	2.61	2.84	0.36	0.17	0.30	**0.15**
AA	81.12	83.92	91.36	96.73	93.92	**99.57**
σ_{OA}	3.01	3.59	1.88	1.06	2.10	**0.66**
Kappa	0.798	0.834	0.946	0.976	0.967	**0.998**
σ_{Kappa}	0.030	0.034	0.004	0.002	0.003	**0.001**

(b)

	ELM	ELMRW	ELM		ELMRW	
Method	Initial		RS	BT	RS	BT
#Training samples	40		540			
OA	78.23	81.26	93.55	96.22	96.88	**99.78**
σ_{OA}	3.96	4.62	0.38	0.13	0.56	**0.31**
AA	82.84	86.04	93.23	96.49	96.60	**99.64**
σ_{OA}	2.14	2.42	0.63	0.28	0.67	**0.54**
Kappa	0.733	0.770	0.919	0.953	0.961	**0.997**
σ_{Kappa}	0.046	0.053	0.004	0.001	0.007	**0.003**

(c)

	ELM	ELMRW	ELM		ELMRW	
Method	Initial		RS	BT	RS	BT
#Training samples	45		545			
OA	75.30	81.44	97.47	99.77	98.17	**99.85**
σ_{OA}	6.55	7.20	0.43	0.032	0.19	**0.032**
AA	80.79	84.15	96.94	99.68	97.18	**99.76**
σ_{OA}	4.97	5.50	0.58	0.068	0.26	**0.060**
Kappa	0.683	0.759	0.966	0.997	0.975	**0.998**
σ_{Kappa}	0.080	0.091	0.005	0	0.002	**0**

C. Results

Active Classification With ELM and ELMRW: Fig. 2 shows the OA versus the number of queries obtained by the ELM-RS, the ELMRW-RS, the ELM-BT, and the ELMRW-BT for the three data sets, respectively. Based on this figure, we first notice the advantage of using the BT criterion compared with the RS. In all cases, the accuracy of the ELMRW is clearly better than that yielded by the ELM for both the RS and BT selection schemes. Among the investigated scenarios, the ELMRW-BT is the most accurate and stable, in addition to its fast convergence with less iterations compared with the other scenarios. Table II provides the classification results obtained for the initial training set and after 50 queries. For Jeddah, the ELM and the ELMRW yield an (OA, AA, Kappa) equal to (83.81%, 81.12%, 0.798) and (86.75%, 83.92%, 0.834), respectively. For Riyadh, they yield (78.23%, 82.84%,

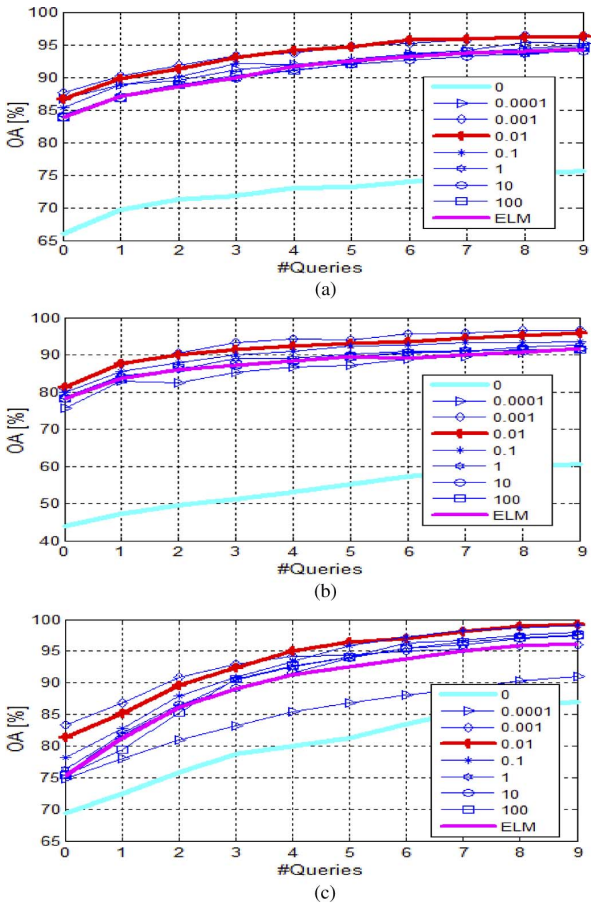

Fig. 3. OA versus the number of queries obtained by the ELMRW-BT for different values of λ for the (a) Jeddah, (b) Riyadh, and (c) Pavia University data sets.

0.733) and (81.26%, 86.04%, 0.770), respectively. For Pavia University, they yield (75.30%, 80.79%, 0.863) and (81.44%, 84.15%, 0.759), respectively. Here, we see that the improvements obtained by the ELMRW with respect to the ELM can reach up to 4% of the difference in terms of accuracy.

After 50 queries, the ELMRW-BT provides an (OA, AA, Kappa) of (99.87%, 99.57%, 0.998), (99.78%, 99.64%, 0.997), and (99.85%, 99.76%, 0.998) for Jeddah, Riyadh, and Pavia University, respectively. On the other side, the ELM-BT provides an (OA, AA, Kappa) equal to (98.12%, 96.73%, 0.976), (96.22%, 96.49%, 0.953), and (99.77%, 99.68%, 0.997) for these three data sets, respectively. From these results, we notice that the ELMRW-BT yields better classification accuracy for both the Jeddah and Riyadh data sets. For Pavia University, the classifiers provide the same result, but the ELMRW-BT converges earlier than the ELM-BT (around 10 queries versus 25 queries). Regarding the computation time, the ELMRW adds an extra time of (4, 6, and 2) s for the {Jeddah, Riyadh, and Pavia University} data sets at each iteration of the AL process compared with the ELM.

Sensitivity Analysis With Respect to λ: Fig. 3 depicts the classification results obtained by the ELMRW-BT for different values of λ. We observe in this figure that, by discarding the prior estimation brought by the ELM (i.e., $\lambda = 0$), the ELM-RW provides poor results. In this case, the ELM-RW acts here as the standard RW algorithm. In this case, only the available training samples are considered prior information for estimating the labels of the unlabeled pixels. By increasing the values of λ, the ELM-RW benefits from the initial conditions provided by the ELM to yield better classification accuracy compared with the ELM. We observe for all three data sets that setting $\lambda = 0.01$ results in a stable behavior. Finally, for increased values of λ, the ELM-RW tends to provide results close to those given by the ELM.

IV. Conclusion

In this letter, we have proposed an efficient AL method for the classification of remote sensing images. We defined a graph-based energy functional that integrates the ELM outputs and the available training samples as prior knowledge. This formulation exhibits several attractive properties as follows: 1) it is intrinsically formulated as a multiclass classification problem; and 2) the solution of this optimization problem leads to a system of sparse linear equations that can be efficiently solved in a linear time using sparse matrix solvers. The results obtained on three different data sets show that the proposed solution is stable and can significantly reduce the expert interaction while maximizing the gain in accuracy with respect to the results provided by the base classifier. Future developments can be defined in many directions as follows: 1) investigate other possible graph models in formulating the energy functional, as defined in (5); and 2) define new active selection criteria that are particularly suitable for this configuration.

Acknowledgment

The authors would like to thank Prof. P. Gamba of the University of Pavia, Pavia, Italy, for providing the hyperspectral data set used in the experiments.

References

[1] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2217–2233, Apr. 2014.

[2] M. M. Crawford, D. Tuia, and H. L. Yang, "Active learning: Any value for classification of remotely sensed," *Proc. IEEE*, vol. 101, no. 3, pp. 593–608, Jul. 2013.

[3] O. Chappele, B. Scholkopf, and A. Zienn, *Semisupervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[4] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[5] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[6] Y. Bazi *et al.*, "Differential evolution extreme learning machine for the classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 6, pp. 1066–1070, Jun. 2014.

[7] N. Alajlan, Y. Bazi, F. Melgani, and R. R. Yager, "Fusion of supervised and unsupervised learning for improved classification of hyperspectral images," *Inf. Sci.*, vol. 217, pp. 39–55, Dec. 2012.

[8] C. Couprie, L. Grady, L. Najman, and H. Talbot, "Power watershed: A unifying graph-based optimization framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1384–1398, Jul. 2011.

[9] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[10] T. Luo *et al.*, "Active learning to recognize multiple types of plankton," *J. Mach. Learn. Res.*, vol. 6, pp. 589–613, 2005.

[11] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.