GA-SVM Algorithm for Improving Land-Cover Classification Using SAR and Optical Remote Sensing Data
Chanika Sukawattanavijit, Jie Chen, Member, IEEE, and Hongsheng Zhang, Member, IEEE

Abstract—Multisource remote sensing data have been widely used to improve land-cover classifications. The combination of synthetic aperture radar (SAR) and optical imagery can detect different land-cover types, and the use of genetic algorithms (GAs) and support vector machines (SVMs) can lead to improved classifications. Moreover, SVM kernel parameters and feature selection affect the classification accuracy. Thus, a GA was implemented for feature selection and parameter optimization. In this letter, a GA-SVM algorithm was proposed as a method of classifying multifrequency RADARSAT-2 (RS2) SAR images and Thaichote (THEOS) multispectral images. The results of the GA-SVM algorithm were compared with those of the grid search algorithm, a traditional method of parameter searching. The results showed that the GA-SVM algorithm outperformed the grid search approach and provided higher classification accuracy using fewer input features. The images obtained by fusing RS2 data and THEOS data provided high classification accuracy at over 95%. The results showed improved classification accuracy and demonstrated the advantages of using the GA-SVM algorithm, which provided the best accuracy using fewer features.

Index Terms—Genetic algorithms (GAs), image fusion, land-cover classification, multisource data, optical imagery, support vector machine (SVM), synthetic aperture radar (SAR).

I. INTRODUCTION

THE fusion of synthetic aperture radar (SAR) and optical images is one of the most important processes in land-cover classification. Currently, the availability of up-to-date, multisource remote sensing and crop type data are important for improving land-cover classifications. In Thailand, optical data are often limited by cloud cover. Thus, the main advantage of SAR is the all-weather capability of these systems. In optical systems, information obtained from the electromagnetic spectrum depends on the reflection and emission properties of the earth’s surface, whereas the SAR backscatter coefficient is measured using the structural and dielectric properties of the target surface. The combination of optical and SAR data can be applied to improve land-cover mapping, as has been shown in several studies [1], [2].

Support vector machines (SVMs) embody a number of theoretical machine-learning concepts. Initially, SVMs were developed with the investigation capabilities and capacity control of machine learning and used formalization to solve overfitting problems in high-dimensional feature spaces [3]. SVMs are able to minimize so-called structural risks when determining classification errors. SVMs use maximum likelihood techniques that empirically reduce the misclassification problem, which is directly defined by the distribution of training sets. Presently, SVMs are related to the nonparametric supervised classification method, which has been demonstrated to be a robust method and has been adopted in the field of pattern recognition and machine learning. Moreover, SVMs have been widely employed in many studies using remotely sensed imagery [4]–[7].

The optimal feature subset and parameter settings are important factors for improving SVM classification. In this letter, the SVM approach with radial basis function (RBF) kernel parameters was applied to classify land cover [8]. Two parameters were optimally identified to achieve the best accuracy: the penalty parameter C and the kernel function width γ. Grid search is the traditional method of finding the proper C and γ is then applied. However, this technique is time intensive, and is difficult to manage. Furthermore, the grid search approach cannot simultaneously process feature subset selection and SVM parameter optimization.

Genetic algorithms (GAs) can simultaneously identify the optimal feature subset and the SVM kernel parameters without decreasing the accuracy of the SVM classification. GAs were first proposed to optimize the parameter and feature selection for SVM classifiers in several studies [9], [10]; however, these letters did not address the combination of optical and SAR data. Moreover, the GA-SVM algorithm has been widely employed in a number of studies, including medical studies [11], financial data analyses [12], and biological studies [13]. Therefore, the objective of this letter is to determine the optimal
II. SUPPORT VECTOR MACHINES

The SVM classifier separates classes using a decision boundary, which maximizes the margin between two groups. This boundary is the so-called best separating hyperplane, and the support vectors are the points nearest the hyperplane. The support vectors are also the most important features of the training samples. The SVM method can be made into a nonlinear approach using nonlinear kernels. Because the SVM method is a binary classification method, it can function as a multiclass classifier by combining many binary SVM classifiers, and pairwise classification is used in this multiclass method. The results of the SVM approach are the decision values of each pixel, which are employed to estimate the probability values. The “true” probability values are represented based on probabilities between 0 and 1, and the sum of these values is equal to 1 in each pixel. The classification is then achieved through the selection of the maximum probability.

Kernel functions, such as linear functions, polynomial functions, RBFs, and sigmoid functions, have been widely used in many studies [14]. In this letter, we used the RBF kernel, because it is able to classify high-dimensional data, unlike the linear kernel function. In addition, the RBF has fewer parameters to define than does the polynomial kernel. The RBF kernel of two samples \( x \) and \( x' \), which are feature vectors in some input space, is defined as follows:

\[
K(x, x') = \exp(-\gamma \|x - x'\|^2)
\]

where \( \gamma \) is a parameter that sets the “spread” of the kernel. In this letter, the SVM classifier uses a well-known RBF kernel composed of the regularization parameter (\( C \)) and the width of the kernel (\( \gamma \)). These parameters must be determined for the RBF kernel because of the robustness of the kernel. In addition, a one-against-all (OAA) strategy was employed for land-cover classification. The OAA strategy involves building a total of \( k \) SVMs for the \( k \) classes. This boundary is the so-called best separating hyperplane, which maximizes the margin between two groups. The support vectors are the points nearest the hyperplane. The support vectors are also the most important features of the training samples. The SVM method can be made into a nonlinear approach using nonlinear kernels. Because the SVM method is a binary classification method, it can function as a multiclass classifier by combining many binary SVM classifiers, and pairwise classification is used in this multiclass method. The results of the SVM approach are the decision values of each pixel, which are employed to estimate the probability values. The “true” probability values are represented based on probabilities between 0 and 1, and the sum of these values is equal to 1 in each pixel. The classification is then achieved through the selection of the maximum probability.

Kernel functions, such as linear functions, polynomial functions, RBFs, and sigmoid functions, have been widely used in many studies [14]. In this letter, we used the RBF kernel, because it is able to classify high-dimensional data, unlike the linear kernel function. In addition, the RBF has fewer parameters to define than does the polynomial kernel. The RBF kernel of two samples \( x \) and \( x' \), which are feature vectors in some input space, is defined as follows:

\[
K(x, x') = \exp(-\gamma \|x - x'\|^2)
\]

where \( \gamma \) is a parameter that sets the “spread” of the kernel. In this letter, the SVM classifier uses a well-known RBF kernel composed of the regularization parameter (\( C \)) and the width of the kernel (\( \gamma \)). These parameters must be determined for the RBF kernel because of the robustness of the kernel. In addition, a one-against-all (OAA) strategy was employed for land-cover classification. The OAA strategy involves building a total of \( k \) SVMs for the \( k \) classes. Each SVM is trained to classify one class against all the other classes. Pixel \( x \) is then given the label of the class with the highest value of \( \Phi(x) + b_i \), where \( i = 1 \) to \( k \). The OAA strategy is among the most popular methods of land-cover classification and has been used in many studies [4], [15].

III. GENETIC ALGORITHM

GAs have been commonly applied in optimization analyses. GAs are based on the concept of Darwinian natural selection and are rooted in biological processes. GAs process a set of features called a population; choose individuals to be “parents” from the population, with the selection depending on the values of the fitness function; and then create “children” for the next generation. GAs create successive populations of optimal solutions that are represented by a chromosome. The fitness function is applied to evaluate the quality of the solutions in each step, including the crossover and mutation functions, which are the main operators in a GA. Moreover, GAs can address a large number of features, and can be used for feature selection.

As shown in Fig. 1, the main GA operations are selection, crossover, and mutation. In the first step, the selection operator selects the parents. Then, the crossover operator combines two parents to create new individuals for the new generation using a single-point crossover, two-point crossover, or homologous crossover, as defined by the user. Finally, the mutation operator randomly changes a parent to create new children in a binary code chromosome.

IV. GA-SVM ALGORITHM FOR FEATURE SUBSET SELECTION AND PARAMETER OPTIMIZATION

A GA for efficient feature subset selection and parameter optimization related to the design of chromosomes, calculation of fitness functions, and system processing is presented in the following.

A. Chromosome Design

An SVM classifier with an RBF kernel was employed to classify the land-cover types in the study area; however, the parameters \( C \) and \( \gamma \) must first be determined. The GA-based algorithm can be used to find the best input feature and the SVM parameter values for the land-cover classification. Thus, the chromosome contained three parts: the selected features, \( C \), and \( \gamma \). A binary coding method was applied to determine the chromosome. In Fig. 2, the features are designated as \( F_{bf} \sim F_{bf} \), which represent input features. \( F_{bi} = 1 \), when a corresponding feature is selected, and \( F_{bi} = 0 \), when a feature is not selected. \( C_{bic} \sim C_{bic} \) is the value of \( C \), and \( \gamma_{by} \sim \gamma_{by} \) is the value of \( \gamma \). The term \( n_f \) is the number of bits representing the features, \( n_c \) is the number of bits representing the parameter \( C \), and \( n_\gamma \) is the number of bits representing the parameter \( \gamma \).

B. Fitness Function Design

The fitness function is an important part of evaluating whether an individual is “fit” to survive for reproduction purposes. In this letter, we employed two criteria to define the fitness function, namely, the classification accuracy and the number of features in the selected subset. If an individual has a high classification accuracy and a small number of features,
then the function has a high fitness value and a high probability of passing into the next generation, as shown in (2)

\[
\text{Fitness} = W_{OA} \times OASVM + \left( W_f \cdot \frac{\sum_{i=1}^{n_f} f_i}{n_f} \right)
\]

(2)

where

- \( W_{OA} \) classification accuracy weight;
- \( OASVM \) overall classification accuracy;
- \( W_f \) weight of the number of features;
- \( f_i \) mask value of the \( i \)th feature.

In this letter, we determined the fitness function for a classification accuracy weight (\( W_{OA} \)) of 0.8 and a weight of the number of features (\( W_f \)) of 0.2 for all data sets.

C. Proposed GA-Based Algorithm for Determining the Optimal Feature Subset and SVM Parameters

The main processes performed in the GA-based feature selection and parameter optimization (Fig. 3) are summarized as follows.

1) Chromosomes in the initial population, including the feature subset and the SVM kernel parameters (\( C, \gamma \)), are generated. The initial size of the population should be chosen by the user.
2) The fitness values for each chromosome, namely, \( C, \gamma \), and the selected feature subset, are calculated and evaluated.
3) In SVM classification, the training and testing samples for each class derived from the region of interest (ROI) are used based on professional interpretation. For the training set, a fused image is applied to train the SVM classifier, and the testing samples are employed to assess the classification accuracy.
4) Fitness values of the individuals are calculated using the fitness function in (2), and they are dependent on the classification accuracy and selected features.
5) The number of individuals with a high fitness value will be chosen and maintained for the next generation in the genetic operation (reproduction) step.
6) If the termination criteria are satisfied, the process stops; otherwise, the process will be continued with the next generation using the genetic operation, i.e., selection, crossover, and mutation.

V. STUDY AREA AND DATA SETS

The study area is located in Lopburi, Nakhon Ratchasima, and Saraburi provinces, central Thailand. The study area contains different land-cover types, such as residential, forest, and agricultural areas, as well as bodies of water. The study area is approximately 715 km². In this letter, we used two groups of multisource images. One group includes Thaichote (THEOS) and RADARSAT-2 (RS2) data, and the other group includes LANDSAT-8 and RS2 data (Fig. 4).

VI. METHODOLOGY

A. Preimage Processing

The RS2 and multispectral (MS) data sets, including THEOS and LANDSAT-8 images, were georectified to map coordinates. The georectified images were applied using a MS sensor with a spatial resolution of 15 or 30 m. In addition, the RS2 fine mode had a spatial resolution of 8 m; thus, both the images were the same size (1250 × 1250 pixels).

Subsequently, the image-to-map registration algorithm was used to transform the RS2 and MS images to a WGS84 UTM zone 47 projection. There were 60 total tie points to the ground control points, and the root-mean-square error was less than 1 pixel. Then, Lee filtering of the 3 × 3 kernel size was applied to suppress the noise in the SAR data. Finally, the RS2 backscattered values were converted to decibels using (3)

\[
dB = 10 \cdot \log_{10}(DN^2).
\]

(3)

B. Image Fusion

Image fusion is the combination of two or more different data sets to form new data using a confidence-based algorithm. In this letter, the HH-polarized C-band RS2 images were fused
with the THEOS and LANDSAT-8 MS images using a principle component analysis (PCA) technique. The data fusion was performed at the pixel level. The PCA converts intercorrelated MS bands into a new set of uncorrelated components. The first component is similar to panchromatic (Pan) image and is replaced by a high-resolution Pan during the fusion process. The Pan image is fused in the low-resolution MS bands by performing a reverse PCA [16].

C. Feature Selection and Parameter Optimization Using the GA-SVM Algorithm

Our methodology was executed in the MATLAB 3.2 development environment by extending the LIBSVM library [17]. The proposed GA-SVM algorithm was used to select the feature subset and find the optimal SVM parameters. The main steps in this method for multisource remotely sensed data are as follows.

1) Create the Training Sets and Testing Sets: The training and testing data sets were created using the ENvironment for Visualizing Images (ENVI) program. The land cover in this letter was classified into four types: forest, agriculture, residential area, and water body.

2) Parameter Settings for the GA-SVM Algorithm: The parameters in the GA-SVM algorithm included the chromosome, the fitness function, and the other parameters in the GA-based approach. The other parameters were as follows: initial population number = 100; number of generations = 10; crossover rate = 0.8; mutation rate = 0.05; number in the remaining population = 0.025*number in the population; one-point crossover; roulette wheel selection; and uniform mutation. These parameters were based on our observations and were supported by the results.

3) Run the GA-SVM Algorithm: The GA-SVM algorithm was executed with training and testing data sets that used an ROI and that were created using ENVI software.

The results of the grid search algorithm, the traditional method for determining optimal SVM parameters, were compared with the results obtained using the GA-SVM algorithm. For the grid search, C and γ pairs were trained, and the pair that provided the optimal cross-validation accuracy was chosen. After the selection of the optimal (C, γ) pair, all training samples were used again to train the last classifier.

To evaluate the quality of the land-cover classification, the overall accuracy and kappa coefficient (k) were calculated for the classification results based on reference data using confusion metrics. In this letter, the training samples included 1000 pixels in each class based on an expert field survey, which used the optical and SAR data. The testing samples included 5000 pixels in each class. They were collected from the field survey and generated to evaluate the classification accuracy.

VII. RESULTS AND DISCUSSION

The performance of the SVM classifier using the GA-SVM algorithm and that of the grid search method using a supervised classification to form four categories (water body, forest, residential area, and agriculture) were compared for the RS2, THEOS, and LANDSAT-8 fused images. The classification results of the GA-based method are shown in Fig. 5.

![Fig. 5. Land-cover classifications using the GA-based approach. (a) THEOS and RS2 image (4, 3, 2). (b) Classification results for the THEOS and RS2 fused image. (c) LANDSAT-8 and RS2 fused image. (d) Classification results for the LANDSAT-8 and RS2 fused image.](image)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Overall classification accuracy (%)</th>
<th>GA-based</th>
<th>Grid Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>THEOS</td>
<td>85.02</td>
<td>0.74</td>
<td>82.00</td>
</tr>
<tr>
<td>THEOS +RS2</td>
<td>96.30</td>
<td>0.94</td>
<td>95.70</td>
</tr>
<tr>
<td>LANDSAT-8 (LS8)</td>
<td>81.80</td>
<td>0.67</td>
<td>79.58</td>
</tr>
<tr>
<td>LS8 + RS2</td>
<td>89.20</td>
<td>0.83</td>
<td>87.60</td>
</tr>
</tbody>
</table>

The overall accuracies using the GA-based and grid algorithm methods were determined for different data sets, as shown in Table I. The overall GA-based accuracy of the original THEOS image was approximately 85.02%, and the PCA-fused image that combined RS2 and THEOS using the GA-based algorithm provided the best classification accuracy of 96.30%.

The GA-based approach only used two bands of the four total bands (Fig. 6), and the results differed from the default values of ENVI and the parameter optimization of the grid algorithm, which used all the bands. Moreover, the optimized SVM kernel parameters (C and γ) of RS2 with THEOS were 346.24 and 284.24, respectively, for the GA-based approach, and 3.83 and 76.11, respectively, for the grid search algorithm, as presented in Table II. For the fused image of RS2 with LS8, the values were 2048.00 and 32.00, respectively, for the GA-based approach, and 591.33 and 880.80, respectively, for the grid search algorithm.

The relationship between the fitness value and the number of generations was determined using the THEOS with RS2 and LS8 with RS2 fused images and the GA-based
fitness level was reached in the sixth generation for the THEOS (Sukawattanavijit and Chen, 2012). The average fitness values with RS2 image. The results of the GA-based approach were compared with the results of the grid search algorithm to determine the ability of each method to optimize SVM kernel parameters and feature selection. The results showed that the classification accuracy of the GA-SVM approach is greater than that of the grid search algorithm. Although the grid search method is a commonly used method for optimizing SVM parameters, this method only identifies the best cross-validation accuracies for two parameters of an RBF kernel (C and γ), and uses all the features for the SVM classification. By contrast, the GA-based approach provides adequate accuracy using fewer features.

VIII. Conclusion

The parameter optimization and feature selection processes of the SVM classifier are important techniques for improving land-cover classifications. The HH-polarized C-band RS2 image and the THEOS MS image were fused via the PCA. The results of the GA-based approach were compared with the results of the grid search algorithm to determine the ability of each method to optimize SVM kernel parameters and feature selection. The results showed that the classification accuracy of the GA-SVM algorithm is greater than that of the grid search algorithm. Although the grid search method is a commonly used method for optimizing SVM parameters, this method only identifies the best cross-validation accuracies for two parameters of an RBF kernel (C and γ), and uses all the features for the SVM classification. By contrast, the GA-based approach provides adequate accuracy using fewer features.

APPENDIX A

<table>
<thead>
<tr>
<th>Methods</th>
<th>C</th>
<th>γ</th>
<th>Selected bands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid algorithm</td>
<td>100</td>
<td>0.25</td>
<td>All 4 bands</td>
</tr>
<tr>
<td>GA-SVM</td>
<td>7.83</td>
<td>76.11</td>
<td>All 4 bands</td>
</tr>
<tr>
<td>RS2 + THEOS</td>
<td>2048.00</td>
<td>32.00</td>
<td>All 4 bands</td>
</tr>
<tr>
<td>RS2 + LS8</td>
<td>346.24</td>
<td>284.24</td>
<td>B1 (green), B2 (red)</td>
</tr>
<tr>
<td>RS2 + LS8</td>
<td>591.33</td>
<td>880.80</td>
<td>B2 (green), B5 (NIR)</td>
</tr>
</tbody>
</table>

REFERENCES