IMPROVEMENT OF HYPERSONTICAL IMAGE CLASSIFICATION USING GENETIC ALGORITHM FOR FEATURE SELECTION AND SVMs PARAMETERS OPTIMIZATION

Nasehe JAMSHIDPOUR¹, Saeid HOMAYOUNI² and Farhad SAMADZADEGAN³

¹Graduate student, ²Assistant Professor, ³Associate Professor,
Geomatics Department, College of Engineering, University of Tehran
North Kargar St., Tehran, Iran
Tel: (98)-21-88008841 Fax: (98)-21-88008837
E-mail: {njamshidpour, homayounis, samadz}@ut.ac.ir

KEYWORDS: Hyperspectral Images, Classification, Genetic algorithm, Feature selection, Support Vector Machines.

ABSTRACT: In the recent decades hyperspectral remotely sensed data present the excellent potential for information extraction of the earth surfaces. These observations are normally sampled in several hundred narrow and continuous bands. Such high dimensionality of data provides more discrimination ability in classification task, but also imposes high computational cost and complexity in data modeling. In particular, because of inadequate number of training samples related to dimension of data, the proof of “curse of dimensionality” may occur. Therefore, as a necessary pre-processing step the data reduction techniques, are indispensable to perform by feature selection. Feature selection methods are divided into two categories: the wrapper methods and the filter methods. In the first methods, the “goodness” of selected feature is estimated based on classification accuracy, whereas the second methods use an independent criterion to find the most proper subset of features. This paper suggests a framework to combine filter and wrapper feature selection methods to find the optimal or the near optimal feature subset and optimize the Support Vector Machines (SVM) kernel parameters at the same time. The proposed approach includes two steps: in the first step the filter method selection is applied. Then, using four different filter feature selection methods, a feature pool is produced. In the second step, a single objective genetic algorithm, as a wrapper method, searches previous resulted feature pool to find the final feature subset and optimize SVM’s parameters. The Genetic Algorithm (GA) is a global optimizer which belongs to the randomized heuristic search techniques. Based on the Darwinian principle of “survival of the fittest”, the GA obtains the optimal solution after a series of iterative computations. The suggested algorithm is applied to the Hyperion images acquired over the Okavango Delta in Botswana. The results showed the great improvement in classification accuracy and the significant decrease in feature subset. The number of bands used for classification is reduced from 145 to 16, while the classification accuracy increased from 89.87% to 94.11%.

1. INTRODUCTION

The hyperspectral satellite/airborne sensors, like Hyperion, acquire images in several hundred narrow and continuous bands. This rich and detailed spectral information can be used to classifying and identifying a wide range of earth surface materials. This task normally do not lead to an acceptable accuracy using the multispectral images (Zhuo, Zheng et al. 2008). However, the processing of hyperspectral data with high dimensional feature vectors requires large amount of computational resources and increases the complexity of classification problem. Also, due to the insufficient training samples in comparison to high dimensionality of data, a problem arises, which is known as ‘curse of dimensionality’ (Raymer, Punch et al. 2000). But new generations of classifiers like Support Vector Machines can overcomes this problem. However, large dimension of hyperspectral data causes high computational time and complexity even in such classifiers. In addition, since the hyperspectral images are acquired in very close spectral bands, the resulting high dimensional feature sets contain the redundant information. Thus, without a considerable lose of information, the data dimensionality can be reduced (Fukunaga 1990). Data dimension reduction can be performed by feature selection methods, which reduce computational costs and improve the prediction ability of classification by removing the redundant and irrelevant features. Feature selection methods are divided into two groups: the filter methods and the wrapper methods. The first approaches require no feedback from the classifier and selects features using an intrinsic property of the data to select feature subsets (Vafaei and De Jong 1992). Contrary, the wrapper approaches utilize the classification accuracy to assess the ‘goodness’ of feature subset directly. Although wrapper methods can yield the better performance, but they are more time consumer and complicated than filter methods. Further, the selected features by filter methods cannot reach fine classification accuracy (Zhuo, Zheng et al. 2008).
Siedlecki and Sklansky (Jain and Zongker 1997) introduced the use of genetic algorithms (GA) for feature selection. Many other feature selection algorithms are proposed in literature. A comparative study of various feature selection methods is given in (Ferri, Pudil et al. 1994). Some feature selection methods introduced in pervious works, such as the branch and bound (Narendra and Fukunaga 1977), the max-min approach (Backer and Schipper 1977), the floating search method (Pudil, Novovicov et al. 1994) and the steepest ascent (Bruzzone and Serpico 2001).

In this paper we suggests a framework to combine wrapper and filter methods based on genetic algorithm for classification of hyperspectral data using Support Vector Machines (SVMs). Since the kernel parameter setting for SVM, has a great influence on accuracy of classification, we decide to optimize both feature subset and classifier parameters simultaneously. The paper is organized as follows: after the introduction, the most important feature selection methods are introduced in section. Section 2 describes our approach. Section 4 presents experimental results and discussion and finally a general conclusion is given in section 5.

2. FEATURE SELECTION METHODS

The feature selection techniques study how to select the most informative spectral channels to achieve the better classification of hyperspectral images. Feature selection is different from feature extraction methods which creates new features by mapping the feature space to a lower dimensional space (Tan, Fu et al. 2008). Feature selection methods involve both criterion function and algorithm search (Jain and Zongker 1997). The search algorithm creates and compares possible feature subset by applying criterion function (Bruzzone and Serpico 2001).

In the following, four feature selection criterions are introduced. These criterions are used in filter method in the experimental implementations. Because these criterions have no dependency to classification accuracy, all of four methods can be classified as the filter methods.

2.1 Information Entropy

The entropy measure evaluates each band separately, and defined as below (Dash and Liu 1999):

\[
E(\lambda) = - \sum_{i} p_i \ln p_i
\]

(1)

\(E\) is the entropy measure; \(p\) is the probability density function of reflectance values in hyperspectral feature and \(m\) is the number of reflectance values. The amount of information in data is related to the entropy value. As much as the entropy is higher, the amount of information is larger (Tan, Fu et al. 2008).

2.2 First Spectral Derivative

The bandwidth, or wavelength range of each band is a variable in a hyperspectral sensor design (Price 1994). This method explores the bandwidth variable as a function of added information. If the difference of reflectance values of two adjacent bands is small, then one of bands is redundant and can be excluded(Bajcsy and Groves 2004). The mathematical definition is:

\[
D(\lambda_j) = \sum_{x} ||I(x, \lambda_j) - I(x, \lambda_{j+1})||
\]

(2)

Where \(I\) is the hyperspectral value, \(x\) is the spatial location and \(\lambda\) is the central wavelength of each band.

2.3 Correlation Measure

One of the standard measures of band similarity is normalized correlation (Duda, Hart et al. 2001). This measure performs well when the signal to noise ratio is large enough, and calculates the normalized correlation for adjacent pairs of bands (Bajcsy and Groves 2004). The mathematical description is given below:

\[
\text{CorM}(\lambda_j) = \frac{E[I(x, \lambda_j) + I(x, \lambda_{j+1})]}{\sigma(I(x, \lambda_j) + I(x, \lambda_{j+1}))}
\]

(3)

Where \(I\) is the hyperspectral value, \(x\) is the spatial location, \(\lambda\) is the central wavelength. \(E\) represents the expected value and \(\sigma\) is the standard deviation.

2.4 Principal Component Analysis Ranking

Principal component analysis (PCA) is one of the most well- known feature extraction method that has been used very frequently for data reduction. PCA is a feature extraction process that first transforms the original image into a
principal component (PC) image through principal component transformation (PCT) (Richards and Jia 2006). The first several bands in the PC images contain most of the variance in the original image (Yao and Tina 2003). PCT has two sets of output: the eigenvalues and the eigenvectors or eigenimages. The eigenvalue indicates the amount of information in a new dimension and the eigenvectors indicate the influence of the original dimensions on the new dimension (Bajcsy and Groves 2004). The PCA ranking (PCAr) is defined as follows:

$$PCAr(\lambda_i) = \sum_j |c_{ij}|$$

where $\lambda_i$ is the eigenvalue for the $i^{th}$ principal component, and $c_{ij}$ is the mapping coefficient of the $i^{th}$ central wavelength in the $j^{th}$ principal component.

3. THE PROPOSED APPROACH

3.1 SVM based classification

The Support Vector Machines is a new generation of classification system which was first introduced in (Vapnik 2000). It has been applied successfully in a range of problems including pattern recognition, bioinformatics, text classification, and image classification. SVMs classified samples into two classes by finding out a set of support vectors that outline a decision boundary, which is often called the optimal hyperplane (Huang and Wang 2006). If the data can be separated linearly, the construction of hyperplane is possible. Otherwise by using the kernel functions, the data map into a higher dimensional feature space, which are linearly separable, and the hyperplane can be found (Tan, Fu et al. 2008). In this research, we choose the Radial Base Function (RBF) as SVM’s kernel, because it works sufficiently well often (Hsu, Chang et al. 2003) and it has only two parameters, ($\gamma$, C), to set and control. Two critical problems arise when using SVMs: i) how to find the optimal feature subset for SVMs and ii) how to set the kernel’s parameters (Huang and Wang 2006). These two problems are important, because the feature subset choice influences the proper kernel parameters and conversely (Fröhlich, Chapelle et al. 2003).

3.2 Genetic Algorithm for optimization problem

Genetic algorithms is a general adaptive optimization search method based on a direct analogy to Darwinian natural selection and genetics in biological systems (Huang and Wang 2006). GA starts from an initial population which is composed of a set of possible solutions called individuals (chromosomes), and then evaluates the quality of each individual based on a fitness function. The fitter solutions have a better chance to survive or reproduce in the next generations. The population during consecutive generations evolves to be fitter in the problem’s conditions. Selection, crossover, and mutation are the main GA’s operators to reproduce future generations. The evolutionary process will not stop until termination condition satisfy (Zhuo, Zheng et al. 2008).

3.3 Feature Pool

The feature subsets produced by the four feature selection methods are fed into a feature pool. Instead of GA searches all the original features, the feature pool will be searched. Some of feature selection methods introduce a subset of features, whereas the rest of methods rank the feature based on a criterion. In the latter methods, determination of a cut-off point is necessary to find a feature subset in ranked list. One common practice is to select simply top-ranked features (Tan, Fu et al. 2008), for example 25 top-ranked.

3.4 Chromosome Design

GA is able to optimize the RBF’s parameters and the feature subset simultaneously. Therefore the chromosome is comprised of three different part; C, $\gamma$, and feature mask. There are different chromosome encoding system, here the binary coding system has been used (Engelbrecht 2007). As shown in Figure 1, the first part of chromosome represent the feature mask, and the middle and last parts show the values of parameters C and $\gamma$ respectively.

```

| g^1_1 | ... | g^1_n | g^2_1 | ... | g^2_n | g^3_1 | ... | g^3_k | ... | g^h_n |

Figure 1. The chromosome comprises three parts, C, $\gamma$, and the features mask (Huang and Wang 2006)
```

The number of bits, representing each parameters ($n_i$, $n_h$), are calculated according to the precision requirement (Huang and Wang 2006). The number of bits, representing the feature mask ($n_i$), is equivalent to the number of top-ranked features which has been fed into feature pool. For chromosome representing feature mask, ‘1’ shows the feature is selected and ‘0’ means feature is not selected. If the range of parameter is $[x_{min}, x_{max}]$ the genotype of chromosome representing parameters should be transformed as (Engelbrecht 2007):
3.5 Fitness Function

The genetic algorithm should succeed two goals: maximizing the classification accuracy, and minimizing the number of selected features. These criteria used to create a single objective function as follows:

\[ F = w \cdot C(x) + (1 - w) \cdot \frac{1}{N(x)} \]  

where \( x \) is the feature subset, \( C(x) \) represents the classification accuracy, \( N(x) \) is the size of selected feature subset, and \( w \) is a parameter between 0 and 1 which adjusts the influence of each criterion. As value of \( w \) is higher the weight of classification accuracy in fitness function is greater. On the other hand, reducing the value of \( w \) will give more penalties on the size of \( x \) (Tan, Fu et al. 2008). By adjusting \( w \), we can achieve a tradeoff between the accuracy and the size of the feature subset obtained. We defined \( w = 0.8 \) for our experiment to avoiding large decrease in classification accuracy.

3.6 Architecture of the Proposed Approach

This paper suggests a hybrid framework, presented in Figure 2. This approach combines the filter and wrapper methods to provide sufficient information and optimize the SVM kernel’s parameters simultaneously. The goal is to compromise different feature selection methods to find the optimal feature subset with smaller dimension and higher/equivalent classification performance. This algorithm includes two steps; in the first step multiple of filter feature selection methods performed and the outcomes of these methods compose a feature pool which is searched in the next step. In the second step, using the genetic algorithms (GA) final feature subset are determined and two parameters of RBF kernel (\( \gamma \), C) are optimized.

Figure 2. System architectures of Top-ranked GA-based feature selection and SVM parameters optimization
4. EXPERIMENTS AND DISCUSSION

4.1 Data Description

The hyperspectral data used in this study is Hyperion data acquired over the Okavango Delta in Botswana. The Hyperion system is a hyperspectral imager capable of sampling 220 spectral bands (from 0.4 to 2.5 μm) with a 30-meter resolution (Zhuo, Zheng et al. 2008). This dataset consists of observations from 14 identified classes representing the land cover types in the study area. The uncalibrated and noisy bands concerning to the water absorption phenomena are removed and the remaining 145 bands are included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220] (Korycinski, Crawford et al. 2003).

4.2 Experimental Results

Our implementation was carried out on the Matlab 7.8.0 environment by extending the LIBSVM which is originally developed by (Chang and Lin 2001). The empirical evaluation was performed on Intel® Core ™ 2 Duo CPU running at 2.53 GHz and 2 GB RAM.

Applying the k-fold cross-validation is a common way to guarantee the valid results for predicting on new data. In this experiment 10-fold cross-validation was used to evaluate the classification performance. Data scaling is an important pre-processing step to avoid numerical difficulty during the SVM calculation (Hsu, Chang et al. 2003), and prevents the domineering of variables in greater value range. Each feature should be mapped to [0 1] range.

As we can see from the Table 1, The top 25 features ranked by these feature selection algorithms are represented in Table 1. Next, the 25 top ranked features from each method are picked up, after removing multiplied features. Finally the 70 remaining features build the feature pool for using in GA optimizations. As we can see from the Table 1, the different feature selection criteria ranked the features dissimilarly. So performing various selection methods provides a diverse feature pool.

<table>
<thead>
<tr>
<th>Feature selection method</th>
<th>Top 25-features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy Information</td>
<td>33, 30, 34, 24, 31, 23, 25, 26, 21, 26, 40, 29, 43, 27, 44, 42, 41, 20, 17, 22, 25, 18, 19, 39, 16</td>
</tr>
<tr>
<td>First Spectral Derivative</td>
<td>32, 31, 46, 61, 62, 47, 28, 29, 60, 36, 45, 26, 44, 4, 48, 78, 38, 80, 39, 29, 76, 2, 59, 72, 77</td>
</tr>
<tr>
<td>Correlation Measure</td>
<td>56, 58, 54, 70, 52, 100, 74, 96, 86, 94, 102, 88, 80, 98, 90, 104, 66, 84, 68, 50, 72, 64, 78, 106, 92</td>
</tr>
<tr>
<td>PCA Ranking</td>
<td>34, 33, 30, 35, 44, 36, 31, 43, 42, 40, 41, 39, 29, 45, 37, 38, 50, 52, 51, 53, 27, 54, 56, 46, 55</td>
</tr>
<tr>
<td>Feature Pool</td>
<td>2, 4, 6, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 56, 58, 59, 60, 61, 62, 64, 66, 68, 70, 72, 74, 76, 77, 78, 79, 80, 84, 86, 88, 90, 92, 94, 96, 98, 100, 102, 104, 106</td>
</tr>
</tbody>
</table>

Table 1. Top-25 Features from Entropy Information, First Spectral Derivative, Feature Pool on Hyperion data

The parameter setting for genetic algorithms is done as the followings: population size 30, number of generations 50, probability of cross over 0.8, probability of mutation 0.05. The obtained results include the optimized kernel parameter number and classification accuracy on each fold of the data set is shown in Table 2.

<table>
<thead>
<tr>
<th>Fold #</th>
<th>Overall Accuracy%</th>
<th>Optimized C</th>
<th>Optimized γ</th>
<th>Selected features</th>
<th>Overall Accuracy%</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96.01</td>
<td>67212</td>
<td>0.4720</td>
<td>18</td>
<td>86.81</td>
<td>145</td>
</tr>
<tr>
<td>2</td>
<td>95.08</td>
<td>39185</td>
<td>0.5669</td>
<td>18</td>
<td>91.38</td>
<td>145</td>
</tr>
<tr>
<td>3</td>
<td>95.11</td>
<td>74599</td>
<td>0.5512</td>
<td>18</td>
<td>92.05</td>
<td>145</td>
</tr>
<tr>
<td>4</td>
<td>95.71</td>
<td>68364</td>
<td>0.2520</td>
<td>18</td>
<td>90.44</td>
<td>145</td>
</tr>
<tr>
<td>5</td>
<td>90.18</td>
<td>82131</td>
<td>0.8189</td>
<td>12</td>
<td>86.81</td>
<td>145</td>
</tr>
<tr>
<td>6</td>
<td>93.83</td>
<td>62130</td>
<td>0.3150</td>
<td>13</td>
<td>88.89</td>
<td>145</td>
</tr>
<tr>
<td>7</td>
<td>94.74</td>
<td>33083</td>
<td>0.4404</td>
<td>17</td>
<td>93.50</td>
<td>145</td>
</tr>
<tr>
<td>8</td>
<td>94.43</td>
<td>90410</td>
<td>0.1260</td>
<td>17</td>
<td>87.00</td>
<td>145</td>
</tr>
<tr>
<td>9</td>
<td>95.69</td>
<td>30343</td>
<td>0.4330</td>
<td>17</td>
<td>91.69</td>
<td>145</td>
</tr>
<tr>
<td>10</td>
<td>90.40</td>
<td>51838</td>
<td>0.8346</td>
<td>13</td>
<td>90.10</td>
<td>145</td>
</tr>
<tr>
<td>Average</td>
<td>94.11</td>
<td>-</td>
<td>-</td>
<td>16.1</td>
<td>89.87</td>
<td>145</td>
</tr>
</tbody>
</table>

Table 2. Experimental results for Hyperion data using top-ranked GA-based feature
For Top-ranked GA approach average the classification accuracy measured by Overall Accuracy is 94.11%, and average number of features is 16.1 features. While the average accuracy of SVM classification without any feature selection method and with the default values for kernel parameter (\( \gamma \text{=1/145, C=1} \)) is 89.87%.

5. COCLUSION

The results showed that various feature selection methods provide different outcomes when they are applied to a same dataset. It has been proved that no single criterion is the best for all applications. This paper proposed a framework based on the genetic algorithm to accommodate multiple feature selection criteria and optimized the feature subset and the kernel parameters, simultaneously. In the experimental tests, we evaluated the classification accuracy of the proposed top-ranked features (bands) of Hyperion imagery data, using the genetic algorithm. The average number of bands used for classification was reduced from 145 to 16.1, while the classification accuracy increased from 89.87% to 94.11%. The number of selected feature subset in comparison to dimension of original feature space decreased significantly.

REFERENCES


