Analysis of Posterior Probability Uncertainty for Classification of Hyperspectral Images by Support Vector Machines

Xiaoxia SUN1, Liwei LI*, Bing Zhang1, Ling YANG2
1. Lab of Digital Earth, Chinese Academy of Sciences, Beijing, China
2. College of Environment and Planning, Henan University, Kaifeng, China
* Correspondent Author Email: lw1@ceode.ac.cn

Abstract—This paper analyzes the uncertainty of classification posterior probability of support vector machine (SVM) using urban hyperspectral images. The hyperspectral images in Zhangye are selected as the study zone, and the sample parameter data were acquired based on the high resolution images and the ground survey information, the images were classified with parameter-optimized SVM to obtain the posterior probability graph for each class, and the posterior probability graphs were truncated using the threshold values of 0.2, 0.4, 0.6, 0.8 and 0.9 for analysis of the accuracy change of ground object classification at different probabilities. The results show that with the increase of truncation probability, the user accuracy in the classification results increases continuously, while the producer accuracy shows a declining tendency, and the overall classification accuracy also shows a declining tendency. The analysis of the posterior probability distribution of various types of ground objects shows that it is difficult to distinguish the posterior probability of some mixed ground objects. The untrained water body targets can be easily distinguished by the truncation probability, but the posterior probabilities of untrained red materials and white materials are mixed together. This shows that there exist some conditions in which the posterior probability of optimized SVM can not directly and effectively indicate the distinction of ground objects. The posterior probability should be used optionally, and at the same time, it is necessary to construct a more robust calculation method for the posterior probability.

Index Terms—SVM, hyperspectral, posterior probability, classification, uncertainty.

I. INTRODUCTION

Supervised classification algorithm requiresthat training sample data can represent the statistical characteristics of the data to be classified [3], [4]. In practical applications, as some non-interest category is deliberately excluded in the training phase or some unknown category is ignored unconsciously, the training samples can not fully represent all ground objects [2]. In such case, the classification results and accuracy can not fully represent the classification results and accuracy of the whole image [1]. Classification posterior probability can indicate the uncertainty of classification results [8]. The analysis of classification posterior probability is able to guide the acquisition of the information on the unclassified ground objects to improve the accuracy of classification results. It has practical significance in remote sensing image classification applications [5].

In recent years, support vector machine (SVM), as a supervised classification algorithm, shows very good performances in remote sensing data classification and is especially widely used in hyperspectral data classification [6]. The statistic analysis based on two-two classification results can produce the posterior probability of classification by SVMs[7]. Presently, the research interests are mainly focusing on high-precision classification of ground objects by SVM when there are sufficient samples, while fewer efforts are made in analysis of the posterior probability of classification by SVMs. In this paper, we analyze the posterior probability uncertainty of classification by SVM based on urban hyperspectral images to guide the application of classification algorithms using SVMs on condition that there are incomplete samples.

II. DATA AND EXPERIMENTAL DESIGN

A. Data

The airborne hyperspectral data on the areas surrounding Hexi University, Zhangye City, Gansu province, are selected as the study zone. It has a size of 1000*1000 pixels, a band number of 48, a spectrum range of 380-1055nm and a spatial resolution of 1m (as shown in Fig. 1), in which R, G and B are displayed with 19 bands, 13 bands and 6 bands respectively.

Eight classes were selected based on the high resolution images and ground survey information respectively, i.e., artificial grasses, red materials and water bodies, and a certain number of samples are selected for each class as the test samples. 10% of the five classes of gray materials, white materials, blue roofs, bare lands and vegetation were randomly selected as the samples for supervising the training phase in classification, and artificial grasses, red materials and water bodies are combined into unclass for testing the stability of probability graph. Therefore, the training samples for the test include 5 classes, and the test samples include 6 classes, covering the category of unclass. The spatial distribution of the test and training samples are identified in Fig. 1.

B. Test design

Classify the study area by SVM using the training samples and optimize the parameters using the grid search method. The minimum and maximum values of the Kernel parameter are 0.1 and 1000 respectively, the search multiple was 10, the minimum and maximum values of the
Regularization parameter are 0.1 and 1000 respectively, the search index multiple is 10, the final search results are that the

<table>
<thead>
<tr>
<th>Class</th>
<th>Gray materials</th>
<th>White materials</th>
<th>Blue roofs</th>
<th>Bare lands</th>
<th>Vegetations</th>
<th>Unclass</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td>384</td>
<td>235</td>
<td>129</td>
<td>152</td>
<td>293</td>
<td>0</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>3835</td>
<td>2345</td>
<td>1291</td>
<td>1519</td>
<td>2926</td>
<td>4499</td>
</tr>
</tbody>
</table>

value of the Kernel parameter is 10, and the value of Regularization parameter is 100, and the Cross Validation method is used to assess the training results.

To illustrate the guidance role of probability graph in classification results, select the six threshold values of 0, 0.2, 0.4, 0.6, 0.8 and 0.9 to carry out threshold value truncation for the probability graph. Calculate the confusion matrix of the graph for each class after different threshold values are truncated using the same test samples, and then calculate the producer and user accuracy of the graph for each class and the overall accuracy of the classification results.

III. RESULT ANALYSIS

The classification results with threshold values of 0, 0.4, 0.8 and 0.9 are showed in Fig. 2-5. The comparison of the graphs of the classification results for each threshold value show that there are more and more unclassified black zones that are distributed increasingly wide ranges. When the threshold value is 0, it is the graph of the classification result of the five classes; when the threshold value is 0.2, it is impossible to distinguish the unclass; when the threshold value is 0.4, the water bodies are clear distinguished; when the threshold values is 0.6 and 0.8, the results are basically the same (the water bodies and a small amount of red materials are clearly distinguished and the shadow of part of the roofs and the boundary area of vegetations and bare lands are distinguished); when the threshold value is 0.9, the water bodies, the roads in red materials, a small part of bright red roofs and part of artificial grasses are distinguished. But the red roofs and artificial grass are still can not be distinguished.
Table 2 Statistic results of accuracy

<table>
<thead>
<tr>
<th>Class</th>
<th>Threshold 0</th>
<th>Threshold 0.4</th>
<th>Threshold 0.6</th>
<th>Threshold 0.8</th>
<th>Threshold 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User</td>
<td>Prod</td>
<td>User</td>
<td>Prod</td>
<td>User</td>
</tr>
<tr>
<td>Gray materials</td>
<td>95.74</td>
<td>97.84</td>
<td>95.89</td>
<td>97.84</td>
<td>97.16</td>
</tr>
<tr>
<td>White materials</td>
<td>58.08</td>
<td>97.19</td>
<td>57.98</td>
<td>96.43</td>
<td>58.2</td>
</tr>
<tr>
<td>Blue roofs</td>
<td>100</td>
<td>99.54</td>
<td>100</td>
<td>99.54</td>
<td>100</td>
</tr>
<tr>
<td>Bare lands</td>
<td>75.01</td>
<td>98.42</td>
<td>75.09</td>
<td>98.42</td>
<td>76.31</td>
</tr>
<tr>
<td>Vegetations</td>
<td>55.21</td>
<td>99.9</td>
<td>64.34</td>
<td>99.9</td>
<td>65.71</td>
</tr>
<tr>
<td>Unclassified</td>
<td>0</td>
<td>0</td>
<td>98.83</td>
<td>16.87</td>
<td>88.35</td>
</tr>
</tbody>
</table>

The statistics for the accuracy of the classification results when different threshold values are truncated is shown Table 2.

From the producer perspective, with the increase of the truncated threshold values, the accuracy of almost all interest classes decreases, but the uncertain zones obtained are increasingly larger, the number of unclassified zones is increasing, and the calculation accuracy of unclassified zones is also increasingly higher. From the user perspective, with the increase of the truncated threshold values, the accuracy of almost all interest classes is improved. Among them, the accuracy of blue roofs and gray materials always maintains the highest, followed by the accuracy of bare lands and vegetations, and the accuracy of white materials is lowest.

As shown by the classification results and the analysis of the spectral data of the image samples, some bare soil among crops are mixed with a small amount of weeds, making it difficult to distinguish spectrally the boundary area of bare lands and vegetations, some bare soil among buildings are mixed with impurities and stone, causing the confusion between it and other materials, and it is difficult to distinguish the posterior probability of the confusing ground objects. The untrained water body targets can be easily distinguished by the truncation probability. However, although there are noticeable spectral difference between white materials and red materials, their posterior probabilities are not easily distinguished. This shows that, the posterior probability of optimized SVM cannot directly and effectively indicate the distinguishability between red materials and white materials, so it is necessary to further study the causes for the abnormalities of the posterior probability of combine with probability of classification by SVMs in combination with the spectral analysis of ground objects and the posterior probability generating mechanism so as to construct the more robust calculation method for posterior probability.

IV. CONCLUSION

This paper analyzes the uncertainty of classification posterior probability of support vector machine (SVM) using urban hyperspectral images. The results show that the posterior probability can be used to detect the untrained classes in the classification results of SVM and with the increase of the threshold values, more classes are detected, but there are some circumstances in which the posterior probability of optimized SVM cannot directly and effectively indicate the distinguishability of ground objects. Therefore, the posterior probability should be used optionally, and at the same time, it
is necessary to construct a more robust calculation method for the posterior probability.

ACKNOWLEDGMENT

This research is funded by National 863 programs (No. 2009AA12Z102). Special thanks to colleagues in the Heihe experiment for providing the data.

REFERENCES


